Ph.D Thesis

Thematic Knowledge Extraction

Shaomin Zhang

March 2003

Computing Department
The Nottingham Trent University
Burton Street, Nottingham NG1 4BU, UK

Email: shaomin.zhang@ntu.ac.uk
Tel: +44 115 8408516

Supervisors

Dr Heather Powell
Dr Dominic Palmer-Brown
Dr Lindsay Evett

Abstract

This thesis describes research into automatic knowledge extraction (AKE) from text. In particular, the automatic knowledge extraction is to produce or help to produce knowledge in the format of an existing hypermedia knowledge-based system: HyperTutor. The whole AKE process is divided into two main stages: concept extraction and relation extraction.

Automatic thematic concept (keyword) extraction (TCE) is described in detail. Two approaches for TCE are presented and evaluated. One of them is a machine learning approach based on artificial neural networks (ANNs). New measures of evaluation of this novel approach are introduced, based on the concept of generalisation. These include natural generalisation (NG) and pure generalisation (PG). Measures commonly used in knowledge extraction research, i.e. recall and precision, are applied in their normal binary form, but analogue versions are developed to assess the performance of the ANN-based approach. A comparison with chance (CWC) measure is also applied to the results. A stemming analysis method has also been attempted at sense-level and word-level. The results show that thematic concepts can be automatically extracted from text using an ANN plus a lexical semantic resource. The ANN alone produces best result for non-keywords and overall. Word level stemming analysis alone is the best for identifying keywords, while sense level analysis provides the most balanced results between keywords and non-keywords. The baseline comparison for the ANN method shows that the ANN method adds value to the external lexicon. The CWC measure shows that both the ANN and stemming methods work much better than chance.

Domain portability of the keyword extraction techniques developed is addressed. Although the ANN itself in not transferable, the ANN method is transferable with consistent performance between domains. The stemming analysis approach also transfers well between domains, although not as well as the ANN method.

An attempt to answer the question of how the ANN learns the problem of thematic concept extraction is also presented, based on the analysis of the weights in the trained ANN. Analysis is carried out on three different aspects: relation level analysis tries to find out if some kind of relations are more important than others, or if they are equally important to the ANN; path level analysis aims to identify what kinds of paths are more likely to lead to a noun being classified as a keyword; and analysis of category information helps to explain how category affects keyword recognition. This analysis has confirmed the hypothesis of close relationships that are distinct and characteristic of the seed word-KW relationship.

An important type of relation, named verb-noun relation, is targeted in the attempt of relation extraction. This is novel. Parse tree based and tagger based approaches have been investigated. The parse tree based method produces high precision and low recall. The main reason for the low recall is parser failure. This reflects the current limitation of the parser techniques is not advanced enough to process texts in the real world. The tagger-based approach produces high recall and low precision. The style of writing may have great impact on the results of relation extraction. The experiments have shown that both of the approaches perform better for one of the documents compared with the other.

This thesis also proposes issues for future work in AKE research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis describes research into the use of artificial neural networks (ANNs) for automatic knowledge extraction from text. In particular, to produce or help to produce knowledge in the format required for an existing hypermedia knowledge-based system: HyperTutor [167].

1.1 Background

As the amount of hypermedia information available is dramatically increasing, driven largely by the explosion in the use of the Internet, the ability to access information rapidly and effectively is becoming critical. On the one hand, what information to retrieve and how to present it are not only dependent on the information content itself, but also on the user's needs. On the other hand, it is very difficult for an average user to find information, or even to know whether there *is* any relevant information. To find information, users use keyword searches and point and click mechanisms to jump between different nodes or documents. It is easy to get lost in the information jungle and to lose track of the current position as well as the information that has been found.

Work has been done at the NTU [167] to tackle this problem of disorientation and to organise information into accessible knowledge. By combining knowledge-based systems, hypermedia, and natural language processing, a novel and generic formalism for structuring and interrogating hypermedia-based knowledge via a natural language interface has been developed. The system engages users in a dialogue with knowledge as well as allowing them to browse. It also has pedagogic features for tutoring. The system employs an augmented hyperlink network to represent knowledge.

In addition to the tutoring system (called Hypertutor), an authoring environment called HyperLab is used by an author to organise their knowledge into the knowledge representation structure. The authoring system is a kind of knowledge acquisition tool: it can acquire knowledge via interaction with human experts.

An example of a concept network in Hypertutor is shown in figure 1.1. The concepts and relations between concepts in this example were extracted manually from the text in figure 1.2.

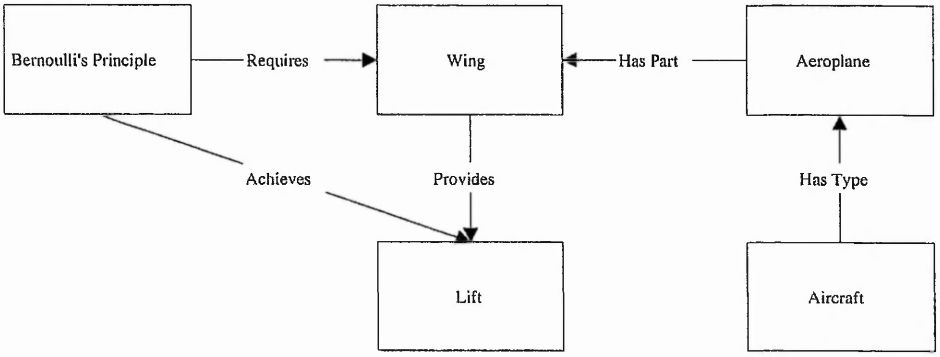
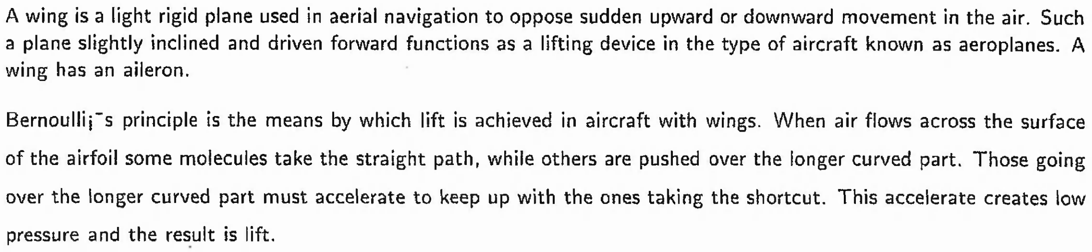


Figure 1.1: An example Hypertutor concept network



A wing is a light rigid plane used in aerial navigation to oppose sudden upward or downward movement in the air. Such a plane slightly inclined and driven forward functions as a lifting device in the type of aircraft known as aeroplanes. A wing has an aileron.

Bernoulli¡¯s principle is the means by which lift is achieved in aircraft with wings. When air flows across the surface of the airfoil some molecules take the straight path, while others are pushed over the longer curved part. Those going over the longer curved part must accelerate to keep up with the ones taking the shortcut. This accelerate creates low pressure and the result is lift.

Figure 1.2: An example text from Hypertutor

HyperTutor is a generic environment, therefore generic knowledge acquisition (KA) techniques are required. Although HyperLab is expressive enough to represent knowledge in the relevant domain, the knowledge has to be defined by human experts. KA is generally a difficult and time-consuming process[185]. It will therefore be of great benefit to automate the knowledge acquisition process so that knowledge can be automatically extracted from text with minimum human involvement.

This thesis presents research into enabling the important concepts (keywords) and their relationships in a domain to be automatically identified.

1.2 Research Objectives

1.3 Thematic knowledge extraction

Automatic thematic concepts extraction (TCE) is one of the key challenges for constructing knowledge bases directly from text. TCE involves extracting only the concepts that are relevant to a given theme. Very little work has focused directly on this issue.

This thesis presents research into approaches for constructing knowledge bases (KBs) by automatic knowledge extraction from text. It investigates the nature of semantic relations between concepts in text, specifically to determine the level and type of lexical information that is required to define the scope of the set of relations within a given topic. The core concepts of a given subject are referred to as Thematic Concepts (TCs) or keywords of that subject. This research aims at extracting only TCs from documents and building a KB from these TCs.

Existing techniques are successful in generating indexes for information retrieval. However, they are difficult to use in thematic concept extraction because they tend to produce a long list of concepts that cover nearly all the details of the information in the text, and these are usually not suitable to be incorporated into a knowledge base. Using these methods to perform thematic concept extraction results in high recall and very low precision[128]. Because the extracted knowledge is to be put into knowledge bases, techniques that produce high precision are needed to avoid introducing false knowledge into the knowledge bases.

Constructing a knowledge base is a time-consuming task, especially if the knowledge is extracted directly from natural language text. Traditional knowledge acquisition (KA) techniques, focusing on transferring knowledge from humans to computers, are not adequate to solve this problem. Researchers have tried to use machine learning methods to induce knowledge from structured data sources [11], [151] and unrestricted text [26]. However, KA to automatically construct KBs remains a difficult and unsolved task. Therefore, their construction is mainly carried out by experts.

Automatic knowledge acquisition direct from natural language sources has emerged as a new research field[182, 185, 207]. Research efforts have varied widely in terms of the level of knowledge that researchers have tried to extract from text. These levels range from simple keyword extraction [223, 225, 75], concept extraction [163], information extraction [168, 169, 170, 171, 172], ontology maintenance [88], question answering [31, 104, 7], and knowledge base construction [162, 54], through to full text understanding. The complexity and difficulty of the tasks increases steadily, from keyword extraction to full text understanding. This implies that there are different levels of knowledge contained in the text sources.

A common problem with most reported techniques for extracting knowledge from text is domain-portability. Extraction systems using these techniques work well in the domain for which the techniques were developed. However, they have to be adjusted when being transferred to new domains to incorporate the new features. The adaptations are usually time-consuming. To make the situation worse, with some systems only the technique developers have the required expertise to make the adaptation. Therefore, apart from the extraction itself, a major issue in developing successful automatic knowledge extraction systems is the feasibility and ease of domain portability. It is desirable to minimise the difficulty of extracting information on new domains by reducing the input required from the extraction technique expert and also the domain expert.

The main overall objective of the Ph.D investigation is to develop techniques for automatic knowledge extraction directly from electronic information. The investigation will focus on methods for extracting knowledge that can be organised into the existing formalism used in Hypertutor. This can be achieved in three stages:

1. Extract concepts and their definitions automatically from text.
2. Identify and extract semantic relationships between extracted concepts.
3. Convert concepts and relationships into the existing HyperTutor formalism.

The scope of this research covers stages 1 and 2.

1.4 Structure of the Thesis

The thesis is structured as follows. Chapter 2 presents a literature review of research relevant to thematic concept extraction research. The main areas are natural language processing, information extraction, information retrieval, machine learning and artificial neural networks. Other areas of research contributing, to a lesser extent, to TCE are not included in this thesis, e.g. machine translation. (Because of the difference in nature between thematic concept extraction and relation extraction, and because the two processes are applied at different stages, the literature review for relation extraction is presented separately in chapter 8.)

In chapter 3, lexical semantic resources and corpora commonly used in TCE research are described. This chapter does not provide an exhaustive list, but presents those that are relevant and potentially useful to this TCE research. When choosing lexical resources, it is important to consider the type of information they provide. This research requires diversity in the kinds of linguistic and semantic information represented. Where semantic lexicons provide the same kind of information, the preference is for the most commonly used. For corpora, coverage and availability are the important factors.

Chapter 4 is devoted to automatic keyword extraction, the first stage of this research. A machine learning approach for automatic keyword extraction based on artificial neural networks (ANNs) is presented and evaluated. To evaluate this novel approach, relatively new measures based on the concept of generalisation are used. These are natural generalisation (NG) and pure generalisation (PG) developed by Tepper et al [216]. Standard binary recall and precision measures [128] commonly used in knowledge extraction research are also applied. Further, more sophisticated measures are developed and introduced to give a more detailed picture of performance. These are analogue measures of recall and precision developed for the ANN approach. Comparison with chance measure is also used provide an overall assessment of effectiveness.

Chapter 5 describes the results of the investigation into the use of sense level information in keyword extraction. In the previous experiments in chapter 4, a noun was treated as a whole, i.e. the different meanings (or senses) are not distinguished. This means the information presented to the network in the previous experiments may be not sufficiently detailed for the network to learn the problem. The investigation aims to exploit the semantic lexicon to draw sense level information and to use this kind of information to improve the performance of keyword extraction. Two approaches of using the sense level information, sense level path and stemming analysis, were investigated.

Chapter 6 assesses the domain portability of the keyword extraction techniques developed in chapter 4 and 5. Domain portability experiments were carried out on two levels. One is to test the generality of the ANN itself, i.e. if an ANN trained on one domain is portable to another domain. The other is to test the generality of the methods, i.e. if the whole of the processes (training the ANN etc) are portable to other domains. Approaches are applied to a document from a new domain.

Chapter 7 attempts to answer the question of how the ANN learns the problem of thematic concept extraction is also presented, based on the analysis of the weights of the trained ANN. Analysis is carried out on three different aspects: relation level analysis tries to find out if some kinds of relations are more important than others or if they are equally important to the ANN; path level aims to identify what kinds of paths are more likely to lead to a noun being classified as keyword; and analysis on category information to explain how it affects keyword recognition.

Chapter 8 reviews previous work in relation extraction research. Techniques reviewed include pattern recognition, machine learning, and parser-based approaches. This is separate from the main literature review because the relevant approaches are quite separate.

Chapter 9 focuses on the relation extraction methods investigated in this research. It explains the stages for information re-construction from sentences in order to extract relations and the methodology used in relation extraction. Two approaches to relation extraction are taken: parse-tree based and tagger-based. Results from these two approaches are presented. Pros and cons of them are also discussed.

Finally, chapter 10 gives conclusions, summarises the achievements and proposes issues for future work in AKE research.

Chapter 2

Literature Review of Thematic Concept Extraction

2.1 Introduction

In the past ten years, NLP has been continuing to progress and develop. One reason for the growth of NLP is the adoption of empirical (corpus-based) approaches involving learning. These approaches are thought to be the most promising way for solving the problem of the knowledge acquisition bottleneck. These empirical approaches include statistical, machine learning and artificial neural network methods. These methods have been used in many NLP research directions such as machine translation, information retrieval and information extraction.

In this chapter, a literature review of research relevant to thematic concept extraction will be presented. These research areas include natural language processing, information extraction and information retrieval. The techniques applied include machine learning and artificial neural networks.

Because of the difference in nature between thematic concept extraction and relation extraction, such as information needed and techniques employed, the literature review for relation extraction is presented separately in chapter 8.

2.2 Natural Language Processing

2.2.1 Automatic Terminology Acquisition

Automatic terminology acquisition (ATA) is aimed at extracting technical terms (TTs) from text automatically. TTs tend to be domain-dependent. However, because the aim is for generic TCE, domain-independent techniques for ATA are the main concern in this review. Term dictionaries are not used in ATA because two features of TTs, quick emergence and constant change of meaning, make it very difficult to keep the dictionaries up to date. Alternative approaches have been proposed by Daille [63], Nakagawa [156] and Lauriston [123]. They are based on shallow text processing, such as pattern-matching.

Justeson and Katz [110] propose a method for ATA, based on their observations that TTs are mainly multi-word noun groups and usually appear more than once in a document, whatever the domain. These features are domain-independent and form the basis of their domain-independent techniques for ATA. Justeson and Katz use a regular expression to find noun phrases (NPs). Those NPs that matched the regular expression and appear more than once in the document are deemed to be TTs.

However, there are some shortcomings of this method. Firstly, it cannot find one-word TTs. Secondly, it fails to detect TTs that appeared only once in a document. Finally, it requires a long processing time. (The method cuts a sentence into various possible fragments with different lengths according to the regular expression and then compares all the fragments with the regular expression.)

Bowden [26] attempts to develop domain-independent techniques to produce glossaries automatically for technical papers. His system called Knowledge Extraction Program (KEP), can identify TTs, acronyms, concepts and their definitions, and relationships (including hypernym, partition and exemplification) between concepts. It uses an adapted version of Justeson and Katz's algorithm, extending the algorithm to be able to find one-word TTs. The processing time is greatly reduced by utilizing the special features of the CLAWS [127] tagset (the document being processed is tagged by the CLAWS tagger in a pre-processing stage). In finding relationships between concepts, a list of triggers is used to pick up sentences which potentially contain particular relationships. Patterns are then used to extract the relationships from the selected sentences.

Bowden uses text from the British National Corpus (BNC), 'B1G' [39], as the test document, and recall and precision [128] evaluation metrics. The results of the TT acquisition are 88% for precision and 19% for recall. The low recall is due to failing to identify TTs that appear only once in the document.

Krulwich [120] uses structural heuristics to attempt to extract concepts from text. The heuristics include font style, section headers, acronyms, etc.

Techniques for ATA have shown that performing automatic knowledge acquisition from text does not necessarily require complicated approaches. Shallow text processing methods, pattern-matching being the dominant method, can perform the task and produce reasonable results. A key issue is the quantity and quality of the patterns used for the pattern-matching. The patterns used have great influence on the performance of the acquisition. Automatic pattern generation via machine learning may be helpful in the collection of patterns. Another finding of this review is that the main generic techniques do not find single word technical terms.

2.2.2 Symbolic (Rule-based) Natural Language Processing

Ever since the first computer was built, understanding natural language (NL) has been the hope of AI researchers. Although early attempts for machine translation (MT) [18] failed, they revealed that MT is impossible without understanding language first.

Several NL programs developed during the 1970s are worth mentioning. SHRDLU [234] used the concepts of a block world, mapping sentences to programs and words to program steps. In theory, sentences could be understood by transforming them into programs. LUNAR [235] was a NL front end to a database containing information on moon rock samples. SHRDLU and LUNAR were significant advances on preceding systems. However, both of them worked in narrow closed worlds and required complete knowledge of their world, and they are thus non-portable and non-extensible. PLANES [229] was a NL front end to a database of aircraft flight and maintenance data. PLANES received NL sentences as input, turned the input sentence into query-language to the database and returned the answer in English. All these programs required a kind of rule-based hand-coded knowledge to perform the tasks. Other important work during this period included speech-act theory [86], frames [146], scripts [195], and knowledge representation [30].

During the 1980s, mainstream NL research focused on developing natural language systems using hand-coded symbolic grammars and knowledge bases. Allen [8] can be viewed as a summarization of the attempts of full understanding of NL from the early 1970s to early 1990s.

Some questions about language and the world must be answered if computers are to understand NL [228]. These are summarised below:

First, what is the function and purpose of language? Waltz [228] answered that "Language is, in general, used by a speaker to achieve goals". Thus, to understand language, a listener must understand the speaker's goal first. However, goals are complex, including to inform, correct, mislead a listener, cause the listener to perform an action, answer questions, undergo experiences, etc. Some sentences can be used for different goals in different contexts or even may serve more than one goal at the same time. These are sources of different levels of ambiguity which are very hard for a computer to process.

According to Wittgenstein's theory [208], language consists of a finite number of words that may be used and reused in an unlimited number of language games (contexts or domains); the same words used in different games to express different kinds of things. The mapping from words to entities or concepts in the world is multiplicitous. Therefore ambiguity in NL is unavoidable, even with careful use of language.

Second, how can it be shown that a computer understands NL? Understanding individual sentences is not enough: contexts such as dialogues, scene descriptions, stories, and so on must be taken into account. This is the scope of pragmatics research.

Third, how can novel structures in language be dealt with? Although there have been major advances in dealing with novel syntactic structures, a significant obstacle remains in dealing with novel semantic structure and novel concepts in the world. These are developing all the time.

Fourth, how should sentences that do not make semantic sense be interpreted. For example: "The bird flies quickly" is meaningful whereas "The pen swallowed the goldfish" is meaningless. To tackle these problems, common sense or world knowledge is necessary.

Many of these problems remain intractable after twenty years of research. Sowa [208] argued that unrestricted language understanding is impossible or at least impractical with a traditional symbolic rule-based approach to natural language. Brill [33] also pointed out that developing natural language processing (NLP) systems using hand-coded and symbolic rule-based methods remains difficult because of the extreme complexity of NL and because knowledge engineering is time-consuming, error-prone and domain-dependent. Furthermore, systems built this way are brittle and cannot work properly outside the scope for which they are built.

F. Gomez [84] uses a rule-based symbolic NLP method to acquire concepts and conceptual relations about the dietary of animals from encyclopedic texts. This system employs a skimmer to scan the document to find sentences relevant to this topic and these sentences are fully parsed to extract concepts and relations. However, how it is decided whether a sentence is relevant or not is not reported in his paper.

2.2.3 Empirical (Corpus-based) Natural Language Processing

One of the biggest difficulties in developing NLP systems is the knowledge acquisition bottleneck, i.e. furnishing systems with the information necessary to perform robust, large scale natural language processing. It is difficult to address this problem in traditional rule-based NLP systems.

Recently, there has been a great research interest in empirical NLP (ENLP), i.e. corpus-based approaches to NLP. A feature of these approaches is that they are more data-driven than rule-driven. Another feature is that they are automated (at least partially) with the inclusion of statistical and/or machine learning methods.

As Armstrong-Warwick [14] pointed out, there are four main reasons for the flourishing of ENLP. Firstly, it is possible to acquire knowledge automatically. All or the majority of the necessary knowledge is identified and coded automatically by the system itself, instead of being hand-coded by human experts. Thus the knowledge engineering and human involvement is greatly reduced. Secondly, ENLP methods can accommodate all linguistic phenomena in a given domain and application. Thirdly, large sets of real world data that contain noise and aspects not noticed by the developers can inform the processing. Finally, ENLP approaches can be extended and ported to new data and domains, i.e. they are potentially scaleable and domain-independent.

These reasons give hope of solving the tricky, long-standing and interrelated problems faced by the researchers into symbolic rule-based NLP .

Church [52] attributed the great increase in ENLP to ever-growing computing power and the availability of large corpora and lexical sources necessary for it. Another reason is that both industry and government (particularly the United States) have encouraged the development of practical NLP systems rather than "toy" programs in "toy" domains.

One of the successful tasks performed by ENLP is part-of-speech tagging [127, 142, 137, 29, 79, 161]. A problem for English annotation is that there is no standard agreed among researchers. Atwell et al [16] developed a system that can tag English texts using virtually all state-of-the-art English corpus annotation schemes. The system is also able to map different annotation formalisms for English language corpus. This is a very important work in English corpus annotation because it allows researchers to compare and to take advantages of different annotation schemes.

Demetriou and Atwell [64] propose domain-independent semantic tagger for English. They suggest that a semantic tagger should annotate a word with a set of semantic primitive attributes or features, instead of an atomic semantic tag. The features may include the root, subject categories, selectional restrictions and meaning of the word. They also show that such a domain-independent semantic tagger can be derived from MRDs.

ENLP is also employed in current research into word sense disambiguation [158], semantic analysis [159], anaphora resolution [12], text segmentation [133], machine translation [117], parsing [46, 216], speech recognition [211], and word clustering [232, 198].

ENLP can include statistical methods based on probabilistic and statistical models and information theory; machine learning (ML); and artificial neural networks (ANNs).

Before the 1980s most NLP systems were knowledge based with hand-coded knowledge of the lexicon, syntax and semantics. Since the early 1990s, empirical NLP emerged and has gained popularity because of the increase of computing power, enabling large corpora, and statistical methods to be processed. However, as Haas [87] pointed out, knowledge-based NLP techniques are still important approaches in NLP research. McMahon also noticed that non-statistical methods are still dominant in natural language understanding applications [141]. Many researchers have realized that knowledge based methods and empirical methods are not mutually exclusive: both of them can have a place in a full NLP system. There is considerable research on how to find out the right roles for both kinds of approach and how to fit the outputs from them together [101].

Brill [33] provides an excellent review of ENLP. McMahon in [141] is also good although it focuses on statistical methods of ENLP.

2.2.4 Relevance to the project

The recent development of NLP, especially the empirical NLP techniques, affect the choices of research methods for this project. Although rule-based NLP approaches remain important, empirical NLP is more promising in producing more robust systems, processing more realistic problems and getting over the limitations of "toy" systems in "toy" domains.

2.3 Information Retrieval

2.3.1 Introduction

Information Retrieval (IR) is concerned with techniques for a user to find information in large collections of documents. This involves retrieving a subset of documents which are considered relevant to the needs of the user. A user's information needs are presented to an IR system in the form of a query which may be a single word, a handful of words or even a paragraph of text expressing the user's interests.

IR systems usually create a representation composed of a finite set of features for documents based on the content of the documents and representation for queries about their meanings. By computing some similarities between representations of documents and a query, IR systems can decide which documents are relevant to the query and return them to the users.

2.3.2 Retrieval Models

A retrieval model specifies the details of how to generate representations for documents and queries, and how to calculate the similarities. Traditionally, there are four prominent IR models in use [61]. They are: Boolean model; vector space model; probabilistic model; and cluster-based model.

In the Boolean model [76], the presentation of documents and queries is determined by a set of features or concepts, $R = \{r_1, r_2, ..., r_n\}$ on the basis of the existence of these features in documents and queries. Thus, documents and queries are presented as a binary-valued vector of length n, i.e. r_k, where r_k $(1 \leq k \leq n)$ takes either 1 or 0 as its value according to the existence of feature r_k in documents or queries. A query can be considered as a boolean expression in which operands are features or concepts. Any documents whose presentation satisfies the query expression are judged to have met the user's need and all other documents are judged to be not relevant to the query. This model does not rank the documents that do not fully satisfy the query.

The vector space[76] model is similar to the Boolean model in that both of them use a set of features or concepts to describe the documents and queries and use a vector to represent the descriptions. However in the Vector Space model, each element of a vector takes a real number called a weight as its value. Several methods can be used to calculate the weights for a vector, the most commonly used is *tf·idf* (term frequency times inverse document frequency) [194]. The Cosine of the angle between the vectors of a document and a query is the most common function used to decide on the relevance of a document to a query.

The probabilistic model [186] is based on the belief that assigning ranks of probability of relevance to documents will contribute to the effectiveness of retrieval. Different methods of estimating the probability of relevance result in different probabilistic formulations. Suppose the representation scheme is the same as in the Boolean model, i.e. a set of features and a vector are used to represent documents and queries. Furthermore, let F and D be the set representation of queries and documents respectively, then an event space can be defined as $F \times D$. Determining which request-document pairs are relevant to a query will fulfil the IR task. The method is to estimate the probability of relevance, $P(R|(f,d))$, where f and d represent a query and a document respectively. The probability can be used in ranking the relevant documents.

In the cluster-based model [181], clusters of documents are generated using a clustering algorithm and a set of similarity measures. For each cluster a representative document is selected as the representation of the whole cluster. It is not the documents in the cluster but this representative document that is compared to the representation of the query to determine which clusters match. The documents in the matched clusters are then retrieved. Relevance of a document to a query is determined by three factors: comparing the cluster representative to the query, similarity measures and the clustering algorithm.

The probabilistic model is similar to the vector space model and both of them can be considered as a generalisation of the Boolean model.

In this research we are interested in extracting concepts from documents rather than the documents themselves.

2.3.2.1 Automatic Index Generation

Other related work is automatic index generation (AIG) [130]. The generated indexes may be for human readers or for use by information retrieval software. The latter kind is not suitable for humans, since it tends to include every occurrence of every word in the document. Indexes for humans tend to be much smaller, since they index only important occurrences of interesting words and phrases. Recent work on automatically creating indexes are mostly orientated towards human readers [156, 130].

A major difference between concept extraction and index generation is the length of the lists produced. Concept lists are usually short, only containing the most important, topical phrases for a document. Indexes, for both humans and retrieval systems, are usually much longer than concept lists. They can contain many less important, more unrelated phrases. A concept list is involved in characterising the central idea of a document or topic concisely, rather than providing detailed information of everything in a document. Thus, AIG is not suitable for TCE.

2.3.2.2 Answer Extraction

Answer extraction (AE) is another area of research similar to concept extraction. This tries to extract sentences directly related to the answer to a question, instead of returning the whole document to answer the question [31, 104]. However, when the extracted sentences are taken outside their context, anaphors are left unresolved. This often means that the extracted sentences lack cohesion so that they are very difficult to understand [31]. The difficulty of using AE in TCE is that AE extracts whole sentences rather than concepts.

2.3.3 Similarity and Relatedness

Current IR research is no longer only concerned with document retrieval. It is also involved in passage/paragraph retrieval, document classification and information extraction. NLP and AI have already been used in IR to provide intelligent IR systems. New techniques aimed at improving the effectiveness of IR have been developed, including automatic identification of concepts in documents, automatic generation of index terms, the recognition of the same concept described using different words, and the recognition of similar concepts [131, 102, 163].

A very basic and important underlying concept in these techniques is **Similarity** between concepts. There are two kinds of similarity [67]: the first is **Paradigmatic Similarity** between two words which is defined as *one word can substitute for the other in a context.* This is sometimes just called 'semantic similarity'. The second is **Syntagmatic Similarity** between two words which is defined as *two words appear in the same context.* This kind is usually called 'semantic relatedness'.

Agirre [5] further distinguishes relatedness into 'local' (relatedness exists in the same sentence) and 'global' (relatedness exists in a context larger than a sentence).

There are two ways, taxonomy-based and distribution-based, to compute semantic similarity and relatedness for words.

2.3.3.1 Similarity: Taxonomy-based Methods

Rada [176] and Lee [126] compute similarity with a taxonomy-based network based on the shortest distance between the nodes which correspond to the words being evaluated. The similarity is inversely proportional to the shortest distance. Having realised that other semantic relationships other than hypernym (is-a) also help compute similarity, Nagao [155] uses both hypernyms and synonyms with a distance evaluation. Nirenburg [160] also uses antonyms.

Distance-based measures have a problem of non-unified distance, i.e. the distances of links in a taxonomy are different, because some parts in a taxonomy are denser than other parts. Taking this problem into account, Resnik [178] proposes a method that gets rid of path distance but is based on a notion of **Information Content**. In this method, similarity between two words is evaluated using the entropy value of the most informative concept subsuming the two in a semantic taxonomy. Agirre [6] proposes a measure called **Conceptual Density** which still uses the shortest path distance, but takes the depth and density of the taxonomy into account.

2.3.3.2 Similarity: Distribution-based Methods

Distribution-based methods are based on the belief that the semantic content of a word can be described by the cooccurrence patterns with other words in a large corpus. Two words are deemed similar if they are accompanied by a significant set of common words.

Brown [36] uses a vector to characterise the context of a given word. The vector holds the number of occurrences following the word of each of a set of possible context words. The similarity between two words is based on the similarity of their respective vectors. Other work considers the cooccurrence patterns in a larger context[78].

Linguistic knowledge has also been used to characterise the cooccurrence patterns of words. Pereira [164] uses verb-object relationships between words. Grefenstette considers subject, object, indirect-object and modifier. Sekine [198] gives a similar model. Federici [71, 70] uses both verb-subject and verb-object relationships.

2.3.3.3 Relatedness

According to the definition: "related words are those that occur together in a context", all the distribution-based methods mentioned in the last section are suitable for computing semantic relatedness. In addition, there is much work on obtaining cooccurrence data from large text corpora [53, 42, 95, 204, 180, 224].

Some work has gone beyond word similarity and relatedness. Salton's [192] experiment aims at providing IR systems with the ability to retrieve parts of a document relevant to a user's needs by revealing relationships between pieces of text, accessing particular text portions or skipping from one section to other related parts, all based on similarity and relatedness between parts of text.

2.3.4 Relevance to the project

The basic techniques in IR research, no matter in whatever retrieval model, are representing the content of documents and queries, and matching the representations of documents to those of queries. The choice of features used to represent the content is important.

The reason that IR techniques, especially those for evaluating similarity and relatedness, are of interest for the project is that they can be potentially used to identify concepts (keywords) related to a theme, to identify similar or the same concepts appearing in different words and to allocate text fragments to the related concept. These function are crucial to this research.

2.4 Information Extraction

2.4.1 Introduction

Information Extraction (IE) aims to identify instances of a particular class of event or relationship in unrestricted natural language text. Relevant arguments concerning events and relationships are extracted and encoded in a format suitable for incorporation into a database [128]. Compared with full text understanding which attempts to extract and represents all information in the text explicitly, IE is only concerned with the facts related to a specific domain that has been decided before the extraction starts. Although IE is less comprehensive than full text understanding and puts more emphasis on the facts themselves than on the relationships between the facts, it is more feasible in practice than full text understanding.

2.4.2 Evaluation

The American Advanced Research Projects Agency (ARPA) sponsored the first message understanding conference (MUC) performance evaluation [168, 169, 170, 171, 172] of IE systems. This is an important event in the development of IE technology. The evaluation consists of two phases. Before the evaluation, a corpus of text along with their answer keys from a predefined domain were issued to the participating research groups. The answer keys to a text are the information that should be extracted from the text. The texts and answer keys can be used for tuning the system under development. This is the development phase. In the evaluation phase, each participating system is presented with the same new set of text to process. The extracted information is then compared with the answers for the new text. In both phases, the answer keys are hand-coded by human experts.

The measures used to evaluate the system are recall and precision [128]. Recall, measures the ratio of correct information ($N_{correct}$) extracted from the text against all the information (N_{all}) available in the text. Precision, measures the ratio of correct information that was extracted against all the information extracted ($N_{extracted}$). These metrics can be represented in the formulae below.

$$recall = \frac{N_{correct}}{N_{all}} \tag{2.1}$$

$$precision = \frac{N_{correct}}{N_{extracted}} \tag{2.2}$$

A series of MUC evaluations have been performed, with MUC3 and MUC4 for the domain: "Latin America terrorism activity", MUC5 for both a business domain (joint ventures) and a technical domain (microelectronics), and MUC6 for aircraft orders and labour negotiations.

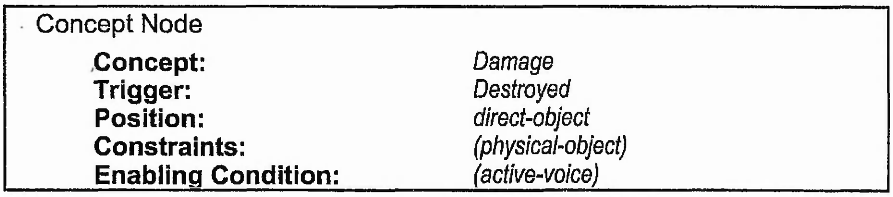
MUC3 to MUC5 had focused on a single task of 'information extraction'. As the MUC series has progressed, IE systems have become more complicated and sophisticated. MUC6 is different from MUC3 to MUC5 in that it aims at not only information extraction, but other basic language analysis as well, including: named entity recognition; coreference; and template elements. The aim is to make IE more portable between domains.

As a result of the sponsorship of DARPA as well as the demand for IE systems in industry, education, government and many other users, information extraction has become viable for special-purpose text processing tasks. Currently, there are IE systems used in analysis of life insurance applications [82], electronic patient records [207], news wire and transcripts, legal documents [97] and the Internet [59, 220, 206].

2.4.3 Automatic Extraction Pattern Acquisition

The methods used in the extraction can be very simple, for example, extraction of just keywords without any understanding about the language and the event to be extracted. They can also be very complicated, e.g. in-depth natural language understanding by full syntactic, semantic and discourse analysis. However, most IE systems currently use pattern-matching methods to extract relevant information from text. These methods represent a middle way between the two extremes (simple keyword match and in-depth natural language understanding). The first task when developing an IE system of this kind is to construct patterns which will be used to extract information. The quality and quantity of patterns strongly influences the resulting performance. Patterns construction is usually performed manually by human experts. It is a time-consuming, knowledge-intensive and tedious task. Recently, there has been a trend in this field to attempt to construct the patterns for extraction automatically [184, 207]. Automatic pattern construction usually uses corpus-based and/or machine learning methods.

Attempts to automatically construct extraction patterns for extracting events or entities from source text by corpus-based, machine learning methods have been made by a number of researchers. The methods differ in that they require different kinds of training corpus, preprocessing, background knowledge and learning strategy.



Concept Node

Concept:	*Damage*
Trigger:	*Destroyed*
Position:	*direct-object*
Constraints:	*(physical-object)*
Enabling Condition:	*(active-voice)*

Figure 2.1: An example extraction pattern

AUTOSLOG [185] is one of the systems that can create extraction patterns (called "Concept Nodes" in AUTOSLOG) by learning. Figure 2.1 shows an example of a concept node, specifying the concept extracted (damage), trigger phrase (destroyed), syntactic position (direct-object), constraints (physical object) and enabling conditions (active voice). For example, using this pattern with the following sentence, "two office buildings" will be extracted.

The hurricane destroyed two office buildings.

To learn concept nodes, AUTOSLOG requires training texts and answer keys, a partial parser, a small lexicon with semantic information and a small set of general linguistic patterns. Answer keys are noun phrases which are used to pick up relevant sentences from the training text. The picked sentences are passed to the partial parser which identifies the subject, verb group, object, prepositional phrase and clause. Then the linguistic patterns, which contain pragmatic role information about the answer keys based on their syntactic roles, are used to generate a 'Concept Node' from the matched constituents and their context. The semantic lexicon is used by the parser to assign semantic classes to noun groups and other constituents.

Most linguistic patterns are domain-independent, thus there is no or little modification when transferring to other domains. However, the answer keys are domain-dependent. The extraction patterns learned by AUTOSLOG require some human post processing because some patterns are not correct due to the shallow parsing.

AUTOSLOG has been used to derive extraction patterns automatically in three domains, including terrorism, business and joint ventures and microelectronics [184]. Experiments have shown that AUTOSLOG can construct in five hours extraction patterns that would need 1200 hours of manual work and that the patterns learned by AUTOSLOG achieve 98% of the performance of the hand-constructed patterns.

Another important feature of AUTOSLOG is that the constructed patterns can be examined by domain experts instead of the computational linguist who must be familiar with the IE system. Thus the deployment of an IE system in a new domain can be carried out by the end user.

A new version of AUTOSLOG called AUTOSLOG-TS [183], takes semantic preferences into account when extracting pattern and merges the syntactically compatible patterns to produce multi-slot case frames which generate more cohesive output and produce fewer false hits than the original extraction patterns.

CRYSTAL [207] uses a more detailed triggering pattern to specify linguistic context and an inductive learning method to construct extraction patterns. The potential extraction patterns are generated first and then combined to produce more general patterns. These are then tested against the training corpus to decide their validity. If a pattern is valid, it is put into the extraction pattern set to replace the patterns that were combined to produce it. This process continues until no new patterns are created. Experiments in a medical diagnosis domain showed that CRYSTAL can achieve 50% to 80% precision and 45% to 75% recall, depending on the value of the threshold.

CRYSTAL needs no human input concerning the proposed patterns but requires more background knowledge than AUTOSLOG for learning. In contrast, AuotSlog needs human intervention to process the proposed patterns and less background knowledge.

Other work on automatic extraction pattern construction includes: Palka [116] which is similar to AUTOSLOG, requiring a concept hierarchy, a set of keywords to pick up relevant sentences from training text and a semantic lexicon; Liep [98] which can identify semantic relationships between noun phrases; Cardie's [44] work, using a symbolic machine learning method to construct extraction patterns and Califf's [41] work, using a relational learning method to perform the task.

2.4.4 Relevance to the project

The suitability of the methods described above for the extraction of syntactic types other than noun phrase is not clear. However, these methods perform well in the extraction of noun phrases or concepts. This is the main concern of the first stage of this research. Some of the ideas and methods can be used for reference. Also Huffman [98] can recognise semantic relationships between noun phrases which is relevant to the relation extraction stage of the research.

However, the aim of constructing complete knowledge bases, cannot be achieved by just using these methods. The main reason is that techniques are needed that extract a much wider range of information about a subject or theme contained in the text. These techniques must identify thematic information and extract it. By contrast the methods mentioned in this section extract specific information or facts which are only a part of a theme. This means that these methods are not comprehensive enough for the project aims.

2.5 Machine Learning

2.5.1 Introduction

Machine Learning (ML) research aims to provide computer systems that can improve their performance for solving a task by learning from experience. Researchers using ML for NLP hope that through learning, computer systems can automatically acquire enormous quantities of knowledge regarding language and the world for use in NLP systems.

ML has produced some successful applications, including speech recognition, learning to classify astronomical structures and learning to play chess. However, in the area of NLP, ML research has progressed slowly [55].

Collier [55] classifies the ML strategies which had been tried by researchers according to the partition of effort of learning between the tutor and the learner. The classifications are listed in ascending order of prerequisite knowledge required as: rote learning, learning by instruction, learning by deduction, learning by analogy and learning by induction which includes learning from example and learning by observation and discovery. From rote learning, the most naive learning, to inductive learning, the most knowledge-rich one, there is a transfer of effort from tutor side to learner side. In rote learning, the effort is nearly all done by the tutor who must find out how to present examples in an appropriate format to the learner and the learner simply remembers what is taught. The quality of this kind of system depends on how the tutor organises the material and on the training process. By contrast, in inductive learning, there is a greater reliance on the learner, who finds out what should be learned without human intervention. The quality of this kind of system is determined by the quality of observation and the inductive ability the system.

ML for NLP primarily takes three paradigms: knowledge-based, statistical/corpus-based and artificial neural networks (ANN) based. ANN-based methods are covered in a separate section (2.6).

2.5.2 Phrase Identification

Turney [225] and Frank [75] use machine learning methods to identify noun phrases from documents. They both use supervised learning techniques to perform concept extraction. [225] uses decision trees and genetic algorithms in the learning strategy in two different experiments. The decision tree chooses keyphrases based on eleven features of phrases. In the genetic algorithm method, they extract phrases first and then use genetic algorithms to select key-phrases. All the features used in selecting phrases are from the document itself, not from an electronic lexicon. [75] uses a Bayesian approach. [225] and [75] both aim to produce a short list of keyphrases for a document, journal articles in particular. The number of keyphrases they produce is usually between 1 and 10. Turney reported the results (in the form of recall and precision) from his methods. He used the key-phrases provided by the authors of the documents as the answer keys. The recall from three documents (from different domains) was 0%, 50% and 50%, while precision was 0%, 33% and 44% respectively. However, it is difficult to cover the topical concepts of a given subject of a document with reasonable length using less than 10 concepts. Our experience of building knowledge bases indicate that for a KB to be of practical value at least 20 nodes is necessary [167].

2.5.3 Knowledge-based ML Approaches

Knowledge-based ML approaches are dominant in ML research. They share many features with knowledge-based NLP approaches and can be applied at different stages of NLP, e.g. syntactic, semantic and pragmatic processing.

2.5.3.1 Syntactic Knowledge-based ML

Syntactic processing tries to learn syntactic aspects of natural language from sentences which are positive examples. There are two main ways to perform the task. One uses **theoretical models** based on linguistic theories, trying to learn the structure of formal language. The other uses **cognitive models** based on psychological theories concerning how children acquire linguistic knowledge.

Theoretical models were more popular during the early language acquisition research and are more concerned with the complexity of language learning. Chomsky [47, 48] forms the basis of this research. Gold [83] demonstrates that *"a set of rules for any non-trivial language cannot be learned from positive examples only"*. Pinker [165] gave an excellent review of theoretical models and some important systems in language learning. Berwick [23] also included some work on using ML methods for syntactic processing.

Kelly [115] reports a cognitive language acquisition system which simulated the early stage of syntactic development, Hill [94] tries to understand and generate language at the level of two year old children and MacWhinney [134] describes a system starting from single words and then proceeding to more complex language.

NLP research has tended to become more empirical so that shallow parsing has become more important [2]. Some NLP tasks, e.g. information extraction, do not require complete parsing. Thus a partial parse (shallow parse) is enough. A partial parser only identifies the main constituents of a sentence, such as the noun phrase (subject and object), verb phrase, preposition phrase, etc. Much of the work on partial parsing employs a machine learning approach (Ramshau [177], Argamon [13] and Munoz [153]).

2.5.3.2 Semantic Knowledge-based ML

Semantic processing uses the meanings of words in phrases in sentences and represents the phrase and sentence meanings in formal formats (conceptual representation). It maps sentences that have the same or similar meaning but using different words or structures onto the same conceptual representation. Wilks's Preference Semantics [233] is an example of a conceptual representation.

RINA, developed by Zernik [237] can generate semantic representations for words and whole sentences. RINA requires an initial lexical hierarchy and takes phrases to be learned and their answers as training examples. It can extend the initial lexicon. The shortcoming of RINA is that the planning knowledge is primitive, so it is limited to simple domains. MAIMRA [202] takes scenes and sentences about spatial movement of objects as input and produces as output a lexicon about word categories and meanings. It requires no background knowledge, but it only works in very restricted domains. It represents semantic information in the form of Lexical Conceptual Structures [100]. The system consists of three modules: a parser generating parse trees for input sentences, a linker relating the meaning of words to the parse tree and creating semantic structure, and an inference module checking the validity of the semantic structures created.

Many IE researchers use ML approaches to tackle the problem of transferring IE systems across domains. Most work has focused on the automatic acquisition of extraction patterns. Supervised learning approaches have been widely used, including Cardie [44, 43], Riloff [185, 183], Soderland [207], Huffman [98] and Freitag [77]. Detailed information can be found in Section 2.3.

ML has also been applied to Word Sense Disambiguation (WSD), including, among many others, Duda [66], Quinlan [174] and Mooney [149].

2.5.3.3 Pragmatic Knowledge-based ML

Pragmatic processing, following semantic processing, takes the context in which events happen into account when performing NLP and other learning tasks with the hope that the knowledge and conventions of the context will help to solve problems, such as ambiguities of language or incomplete information.

GENSIS [150] aims at providing techniques which avoid hand-coding of schemas of stories by dynamically augmenting the existing world knowledge. This can analyse short stories and create schemas. In this system, a parser is used to process input sentences and to create conceptual representations. An understander builds a model which contains causal/support links. A schema library stores background knowledge required when the system infers causal links. A question answering module and a natural language generator inspects the model and generates answers in English. It may create new schema if the analysis achieves an important goal. Collier [55] points out the pros and cons of GENSIS. the integrated partial parser (IPP) in [125] focuses on developing generalisation abilities and robust parsing techniques. the IPP can parse newspaper stories by modelling the way people read text.

Reference resolution is important in discourse analysis. Several researchers have tried using machine learning techniques to solve the problem [57, 12, 139, 114, 81] . In these studies, machine learning methods such as Bayesian induction, decision trees, and maximum entropy modelling are used to learn from corpora annotated with coreference relations.

2.5.3.4 ML on Shallow Parsing

Parsing is the basis for most NLP problems. However, not all NLP applications require a complete syntactic analysis. For example, in IR, it is enough to find simple Noun Phrases (NPs) and Verb Phrases (VPs). In IE, Summary Generation and Question Answering, information about specific syntactico-semantic relations such as agent, object, location, time, etc (basically, who did what to whom, when, where and why) is more useful than detailed syntactic analyses.

Partial or shallow parsing, which recovers only a limited amount of syntactic information from sentences, has proven to be a useful technology for written and spoken language domains. For instance, shallow parsers were used to add robustness to a large speech-to-speech translation system [226]. Shallow parsers are also used to reduce the search space for full parsers [56]. Other applications of shallow parsing include question answering on the World Wide Web [37, 209] and text mining applications [196].

Because shallow parsers have to deal with natural languages, building a shallow parser is a time consuming task. To relieve the difficulties of building shallow parsers, machine learning techniques have already been used.

Ramshaw and Marcus [177] is an important work on using ML to shallow parse, treating the task of NP-chunking as a tagging task. Buchholz [38] uses a similar technique on other types of chunks and for finding relations. Skut and Brants [203] adapts a HMM approach for tagging and chunking. The most recent work on using ML for shallow parsing is in a Special Issue of the Journal on Machine Learning Research on Shallow Parsing [89]. This contains six research papers on shallow parsing using various ML techniques.

2.5.4 Statistical/Corpus-based ML Approaches

Probabilistic approaches, assume that information content is governed by probability distributions and that decisions can be made on the basis of reasoning about these probabilities and observed data.

Probability theory is important in capturing linguistic knowledge in that nearly all other models such as state machines, logic systems and formal rule systems can be augmented with probabilities. It can also be used to solve various kinds of ambiguity problems. Statistical-based models are important in ML, especially in corpus-based empirical NLP.

Baker [17] proposes an algorithm for learning a Probabilistic Context-Free Grammar for parsing. Brent [32] describes a cue-based method for learning subcategorization from verbs. Manning [136] performs the same task using a finite-state automaton. Abney [1] uses ML to attempt to solve the prepositional-phrase attachment problem.

Mitchell [148] presents a general learning algorithm using a naive Bayesian classifier to classify text. Similar work can be found in [132, 122, 107]. Some non-Bayesian statistical text learning algorithms are given by Rocchio [187] and Salton [193].

Recently, research has involved extracting knowledge from world wide web (WWW); Soderland [206] and Craven [59].

In ML research, a new technique of using both labelled and unlabelled data to improve training has emerged. This technique uses a small amount of labelled (e.g. annotated or classified) data and a large quantity of unlabelled data in training. It trains the system using the labelled data, uses the trained system to classify the unlabelled data and then adds the most confidently classified samples into the labelled data set. This technique is very useful in reducing the tasks of labelling or annotating training data or in situations in which labelled data is hard to obtain. Jones [109], Riloff [182], Mitchell [147] and Blum [25] all use similar techniques.

2.5.5 Relevance to the project

The acquisition of domain-specific knowledge is one of the main problems in developing techniques for domain-independent knowledge acquisition systems. ML is a feasible way to address this problem.

The choice of learning strategy is critical to the design of a learning system because this choice determines the partition of the learning effort between the tutor and learner, the representation of knowledge and the format of the input (training examples) and output (what is to be learned).

Learning linguistic knowledge is a difficult task for a computer. Although there is some progress, the knowledge learned so far is much less than is needed for a full natural language understanding system. Furthermore, learning systems usually require prerequisite knowledge to start to learn. Learning from scratch or step by step like a child is so far proving impossible for a computer. Humans can learn new knowledge based on what they have learned, i.e. they can use learned knowledge in learning new knowledge. This process can proceed continually. At present, no work has been reported on successfully simulating this kind of ability of human beings in a computer learning system.

Possibly the most difficult task for machine learning systems is how to fit together the techniques that learn different aspects of language into an integrated learning system.

2.6 Artificial Neural Networks for Natural Language Processing

2.6.1 Introduction

Traditionally, NLP has been considered to be a strongly symbolic task. Symbol processing has been applied to the study of natural language processing, aimed at modelling the human mind as a symbol processor. This approach has achieved a lot, but it cannot give a reasonable explanation or model of some phenomena, such as the human ability to understand ungrammatical and incomplete sentences.

ANN (also called subsymbolic or connectionist) approaches are promising in modelling the cognitive foundations of natural language. ANN models derive many cognitive effects from their distributed representations that represent statistical regularities of language [140]. Compared with symbolic NLP, connectionist NLP (CNLP) has only a short history, beginning in the early 1980s. The reason for this late start is that the multi-layer network learning algorithm, which overcomes the limitations of the simple perceptron of early ANN research and which made ANNs much more powerful and suitable for complex tasks, was not published until 1986 [190].

2.6.2 History and development of CNLP

Early application of ANNs to NLP, used connectionist methods in parsing [205, 227]. Other early work involved representing natural language knowledge [96]. Pollack [166] discussed representation strategies in ANNs.

Although the history of CNLP is short, nearly all areas of NLP research have been investigated from a connectionist perspective. These include phrase generation [80], anaphora resolution [9], goals and plans [201], learning syntax [199], semantic interpretation [222], prepositional phrase attachment [154], morphology [58], and discourse topic identification [111] .

Learning to capture grammatical structure has been one of the active CNLP areas and much work has been done. For example [199] shows that a simple recurrent network (SRN) can learn a finite-state grammar. Elman [68] uses a SRN to predict the next word in a context-free language with embedded clauses. Although the network can not learn the language completely, its performance is good and behaves like human beings in understanding language. Christiansen [50] shows that SRNs can learn a complex grammar and that they have similar behaviour to humans on the same grammar structures. Christiansen and Charter [51] also report that this type of network has good generalisation abilities.

Most connectionist models for syntax processing have used toy grammars and small vocabularies. Therefore the problem is the scaleability and suitability for realistic data. Some work has been done to answer these questions. Tabor [215] presents an SRN-based dynamic parser which performed well on time data. Christiansen [49] also reports an SRN, trained on recursive sentence structure, which can process empirical data. Tepper et al [217, 216] use a hybrid model, combining connectionist and symbolic approaches. A mixture of MlP (multi-layer perceptron) and SRN architecture are trained and tested using the Lancaster Parsed Corpus. The system *"is able to process large and varied samples of naturally occurring English text and sentences of arbitrary length and complexity. The system exhibits high levels of syntactic generalisation..."*. This system represents good progress in connectionist parsing. A thorough review of ANNs for parsing is in [216].

McClelland and Kawamoto [140] show how to use the capabilities of ANNs for language understanding. A network takes the syntactic role assignment of sentences as the input and assigns the correct case roles for each constituent. It also performs semantic enrichment on the word representations and performs word sense disambiguation as well. St.John and McClelland [214] use an ANN to interpret sentences. They suggest that various constraints, syntactic, semantic and thematic etc. play their roles in sentence comprehension and also how this kind of knowledge can be put into a network by training. The model consists of two parts. The first is an SRN to receive the word sequence in the sentence and build a structure during the process. The second part is a three-layer backpropagation network trained to answer questions about the sentence structure. In the process of training, syntactic, semantic and thematic knowledge is put into the network. St.John [213] uses a similar idea to process script-based stories.

Modular and structured architectures have also been proposed and used to build larger CNLP systems. DISCERN [145] is a system of this kind. Parsing, reasoning, lexical processing and generation are all implemented using ANNs and integrated into a system which learns to read and answer questions. Jain [103] describes a parser composed of modules, which can understand sentences with complicated structure and ungrammatical and incomplete sentences.

Recursive Auto-Associative Memory (RAAM) networks have also been investigated [24, 121]. RAAM networks are trained to map sentences in the active voice to the corresponding sentence in the passive voice [45] and to translate sentences between different languages. RAAM networks have also been used with SRNs to perform parsing [21, 200].

[153] uses unsupervised learning methods for the keyword extraction task. Fuzzy Adaptive Resonance Theory (ART) neural networks are used to cluster words from documents to semantic classes and then consider these together with the co-occurrence information to generate meaningful compound keywords.

An interesting work by Karen [111] uses ANNs to identify topic entities in written narrative discourse. The system uses only surface features and minimal semantic information to perform the task. This model develops an internal representation incorporating syntactic, semantic, and contextual information of past events in the preceding narrative as well as that of the current event. This work shows that "*It is reasonable to assume that connectionist implementations offer a viable alternative for modelling discourse processes*".

The results in CNLP show that ANNs can be used to represent and process language. CNLP models have the potential to be more sensitive to context than symbolic NLP. They also show that complex structure can be processed and large systems can be built from ANN components.

2.6.3 Issues of CNLP

Although CNLP has shown promise, there are some criticisms of it. One is that an ANN may appear to be a "black box" making the learning process and internal representations difficult for humans to understand. Some researchers believe that CNLP will be able to replace the symbolic model of language processing [144]. They give an excellent argument for how connectionist models match human behaviours better than symbolic models. Others think it is too early to claim that connectionist models can take over areas where symbolic methods dominate. There are still others who doubt that connectionism alone can deal with the complexity of natural language. The latter two groups have suggested that connectionist and symbolic models should be viewed as complementary and be combined to take advantage of the strength of both kinds. Another concern about the CNLP model is that although Tepper, Powell and Palmer-Brown's work [216] is on realistic language, most of them work only on "toy problems", such as simplified grammar or reduced vocabulary.

Generally speaking, CNLP models are at an early stage of development. Although they have shown very interesting results and features, more evidence is required to judge whether the results and features will still exist when dealing with the complexities of real language. Before CNLP will challenge the symbolic NLP models, some issues must be resolved: (1) representation of complex linguistic structures, (2) construction of training examples for realistic language processing (3) combining existing processing ability into an integrated system.

2.6.4 Relevance to the project

CNLP is of interest in that: firstly, ANNs learn from experience, rather than being fully hand-coded by designers; secondly, ANNs have the ability to generalise to novel cases. As most aspects of language cannot be learned just by rote, the ability to generalise is crucial to learning language; and thirdly, ANNs have the so-called "single route" feature. In many aspects of language, there exist "quasi-regularities", i.e. regularities that usually hold but which have exceptions. To address this kind of problem, symbolic models employ separate sets of rules for regular and exceptional cases by providing rules and a list of rules for exceptions. By contrast CNLP is able to provide a single mechanism incorporating both regular and exceptional cases.

2.7 Summary and Conclusions

Although shallow text processing methods, e.g. pattern-matching, can perform some tasks and produce reasonable results, they are limited to the domain of technical papers. Apart form the issue of the quantity and quality of the patterns used for pattern-matching, it processes texts regardless of semantics which is inevitable in extracting thematic concepts.

The use of rules alone has been unable to fully process natural language. It is difficult to incorporate the necessary level of abstraction required to process levels of natural language using rule-based systems. They are also difficult to adapt as language develops and new rules are required. Although the incorporation of semantic and pragmatic processing in symbolic systems has achieved some success, this success is within specific tasks or domains and thus their performance for other tasks or domains is unclear. Scaling up is also a problem with these systems.

ENLP methods are also used in extracting terms from text. F. Xu [236] uses statistical method (if.idf) to acquire domain relevant single-terms from text. [210] extracts two-word keywords using statistical methods, i.e. by calculating mutual information between word pairs. In contrast to symbolic systems, ENLP, which is data-driven and corpus-based, enables systems to learn knowledge from real-world linguistic resources. This allows NLP systems to learn automatically form linguistic rules and add the learnt rules to the systems, thus processing a wide range and realistic coverage of natural language.

It is acknowledged in the IR community that no current representation techniques completely capture the meanings of a document and it is doubted that adequate representations will be developed in the near future due to the richness of meanings in texts. Compared with representation used in natural language understanding, traditional IR representations are usually based on simple, very general features of documents (such as words and citations) and simple relationships between features. Because they are simple but general, they can be applied to most texts. However, because of this simplicity, they are insufficient to express thematic concepts in a document. Different documents tends to have their own themes which can not be captured by simple and general representations.

Automatically generated indexes are mainly used for information retrieval systems, including nearly all the nouns or nouns phrases in a document. This may be what is required, for example, for information retrieval. However, this is not focused enough for the TCE, since it is not constrained by a given theme. It is not a thematic technique thus it is not suitable for thematic concept extraction. A question about this technique is whether it is possible to extract keywords from the generated indexes instead of from the original document directly. This technique itself introduces errors. Unless the indexes are 100% correct in terms of recall and precision, errors will be passed on to the thematic extraction stage.

Answer Extraction tries to find sentences most relevant to some concepts or keywords. It is a very useful technique in finding definition of concepts, the second stage of this research. However, this technique is not viable for thematic concept extraction because it reverses the procedure: finding relevant texts for given keywords rather than finding keywords from text.

IE techniques are very successful in extracting specific predefined information. Thus a major concern about these techniques for TCE is its generality and domain portability. Although automatic pattern construction is a possible solution to this issue, there is still a long way to go to adapt this technique to be domain-portable and able to capture a much wider range of information about a subject or theme contained in the text. TCE must identify thematic information and extract it. By contrast the methods mentioned in this section extract specific information or facts which are only a part of a theme. This means that these methods are not comprehensive enough for the project aims. However, the methodology and evaluation used in IE is useful in this research.

Most of the traditional ML systems are rule-based symbolic systems. They have the same drawbacks as symbolic NLP systems. Rules are hand crafted into the systems, which is time consuming, error-prone and difficult to amend. Some of these systems with domain specific knowledge are powerful and flexible enough in the target domain, but difficult to port to other domains. Other systems are designed with no need of domain specific knowledge, i.e. they take a general approach and can be applied across many domains. However, due to the huge problem in NLP, these systems are generally less reliable.

In contrast, statistical/corpus based and ANN ML systems are more promising in regards to the domain portability. They are robust to changes of context and domain, and can process regularity and exceptions without additional human effort.

The aim of this research is to investigate the implementation of a robust, realistic, domain-independent and automatic knowledge acquisition system to populate a knowledge base from text information. The robustness and domain-independence can be achieved by employing a machine learning methods. However, the symbolic rule-based ML methods are not suitable. The statistical/corpus-based methods are suitable with respect to the domain robustness and domain portability issues. However, to date they have not been used to take semantic information into account. Semantic information is necessary for thematic techniques. ANNs offer a convenient means for investigating the integration of this type of information into the process. Another major reason for choosing ANN method was that ANN expertise was available within the supervisory team.

Semantic similarity is an important notion when finding related concepts from text. There are two ways that have been used to calculate semantic similarity between two concepts: on the basis of taxonomic relationships as hypernym/hyponym or though distributional evidence. The former method takes the semantic information implied in the taxonomic relationships, and is more reliable than the latter method. Most current methods for calculating similarity based on taxonomy only use hypernym/hyponym relationships. Thus, concepts in the same hierarchy will have high similarity measures. Concepts in different hierarchies will have low or zero similarity measures if just hypernym and hyponym relationships are used. This is a problem for TCE. For example, in most ontologies, object and person are in different hierarchies. Thus, textbook and teacher would be in different hypernym/hyponym hierarchies. The former is an object and the latter is a person. Textbook and teacher are thematic concepts in the domain education. It is not reasonable that two thematic concepts have no similarity at all. Methods using distributional evidence can overcome this problem, but lose the semantic information in the taxonomy-based method. Therefore, the usual semantic similarity is not sufficient to the thematic concept extraction. More information that connects thematic concepts must be take into account in calculating similarity measures for thematic concepts.

Thus, the approach for concept extraction will be machine learning, using ANNs. Semantic similarity will be calculated between words using a broad range of semantic relationships and supported by a semantic resource.

Chapter 3

Lexical Semantic Resources and Corpora

3.1 Introduction

Statistical methods for empirical NLP are flourishing. This is made possible by the availability of large corpora and electronic lexical sources. In this chapter, lexical semantic resources and corpora commonly used in AKE research are presented. The aim is to review those that are relevant and potentially useful to this research.

Semantic lexicons represented in this chapter include The Longman Dictionary of Contemporary English and The Longman Language Activator, WordNet, Cyc and EDR. They need to be evaluated according to the type of information provided. The corpora reviewed are the British National Corpus (BNC) and the Brown Corpus.

To be of use these must have high coverage and they must be readily available.

3.2 Longman Dictionary

3.2.1 The Longman Dictionary of Contemporary English

The Longman Dictionary of Contemporary English (LDOCE) was the first machine-readable dictionary available and widely used in language research. An important feature of it is that all the definitions in the dictionary are written in a controlled set of about 2000 basic words; the Longman Defining Vocabulary. The newest version of LDOCE is the LDOCE3-NLP Database, a machine-readable NLP version of LDOCE specially designed for people working in Linguistics, Information Technology, Artificial Intelligence, Natural Language Processing and related areas. It is issued in CD-Rom format.

The entries of LDOCE3 are in Standard Generalized Markup Language (SGML) format. LDOCE3 is still written using the Longman Defining Vocabulary, but recompiled using the resources of the 100 million word British National Corpus of written and spoken English, the Longman Lancaster Corpus, and the Longman Corpus of Learners' English.

LDOCE3 has the following features which provide valuable information for language research: homographs and senses are arranged by corpus frequency order; explicit frequency information; coding for each of 80,000 senses which includes Subject Field codes and Semantic codes; Grammatical Patterns and collocational information; and mutual information statistics derived from the BNC.

3.2.2 The Longman Language Activator

The Longman Language Activator (LLA) is a dictionary of core English, which organises words with meanings within the same semantic set into concept groups such as ANGRY, WALK, FINISH, HAPPY. For example, the concept ANGRY contains 68 meanings divided into 11 groups, such as, among many others, the groups of meanings for feeling angry-{angry, mad, annoyed, be in a temper, etc}, the group with words of feeling extremely angry-{irate, furious, seething, etc} and the group of feeling angry because of wrong or unfair treatment-{indignant, be up in arms, etc}. Altogether, the dictionary contains 1052 concepts accommodating 20,000 meanings.

This dictionary is valuable for word clustering (grouping words into subsets according to their similarity and relatedness), semantic networks, machine translation and many other areas of NLP research.

There are links between the LLA and the LDOCE3-NLP database. Every sense in the LDOCE3 is either labelled with a subject (or domain) or an LLA code (semantic code). The former indicates that the sense is most likely to appear in the given subject, while the latter gives information about the semantic group to which the word belongs.

3.2.3 Relevance to the project

A lot of research [208] has focused on the extraction of useful information from an existing dictionary and much of the research used LDOCE as the information source. Three features of LDOCE attract the attention of researchers. First, it is a real world dictionary. Its usability and adequacy of linguistic knowledge are proven. Second, as all word sense definitions are written using a limited vocabulary, it is possible to understand and process a core set of English words and extend this to words outside this set [208].

The LLA, which explicitly organises words with similar meaning into concept groups, is a valuable feature for word clustering, a technique for grouping words into subsets according to their similarity and relatedness. This is very relevant to the first stage of the research, i.e. keyword extraction.

The electronic form of LDOCE is issued on CD-ROMs. It is not free.

3.3 WordNet

3.3.1 Introduction

WordNet [72] is an on-line lexical reference system. In WordNet, nouns, verbs, adjectives and adverbs are all organized into the smallest semantic unit: Synonym Set (called Synset in WordNet) which represent a single concept in English. The Synsets are interconnected by semantic relationships.

3.3.2 Concepts, Nouns, and Categories

There are more than 57000 nouns in WordNet (as WordNet is updated the exact number increases). Most of them are compound nouns and a few are proper nouns. They are organised into about 48800 Synsets (concepts) and are represented as a kind of semantic inheritance network.

WordNet divides almost all concepts, except those that denote the most general concepts in English, into 25 separate lexicographer files. Each set of concepts forms an inheritance hierarchy headed with a concept called a unique beginner, and represents relatively distinct semantic fields or semantic primitives. Those concepts that are not included in the 25 hierarchies form the levels above these 25 unique beginners in the single inheritance hierarchy of English. The root of the single hierarchy is an empty concept. There are nine concepts directly beneath the root. These are shown in Table 3.1. Figure 3.1 shows part of the top of the hierarchy down to the unique beginner level.

Table 3.1: Second Level Concepts in WordNet

Concept	**Meaning**
entity, something	anything having existence (living or nonliving)
psychological feature	a feature of the mental life of a living organism
abstraction	a general concept formed by extracting common features from specific examples
state	the way something is with respect to its main attributes
event	something that happens at a given place and time
act, human-action, human-activity	something that people do or cause to happen
group, grouping	any number of entities (members) considered as a unit
possession	anything owned or possessed
phenomenon	any state or process known through the senses rather than by intuition or reasoning

We can note that the concepts above the 25 unique beginners are too general for most tasks that require nouns or concept classification knowledge. This is one reason that the creators of WordNet use the 25 files. The practical benefit of using 25 separate files is that it divides the information into more manageable units, both for maintenance and use.

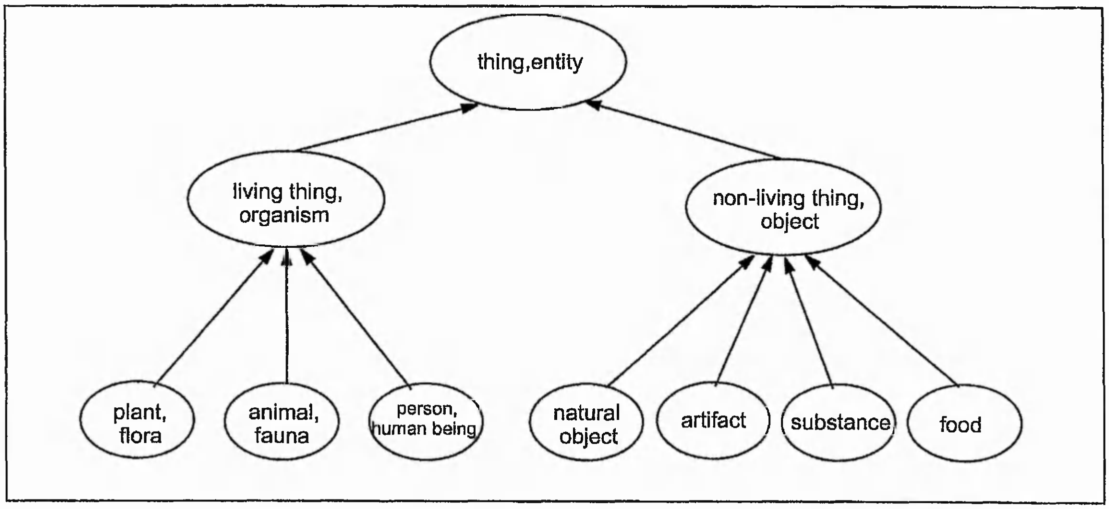


Figure 3.1: Part of the top of the hierarchy down to the unique beginner level.

Nouns can express several concepts. All concepts, except those above the 25 unique beginners, belong to one and only one of the 25 hierarchies. These hierarchies will therefore be referred to from now as "high-level categories". An example of this hierarchy is shown in Figure 3.2, from 'student', the lowest level, to 'person', one of the unique beginners. Each level in the hierarchy represents a category. The 25 high-level categories in WordNet are shown in Table 3.2.

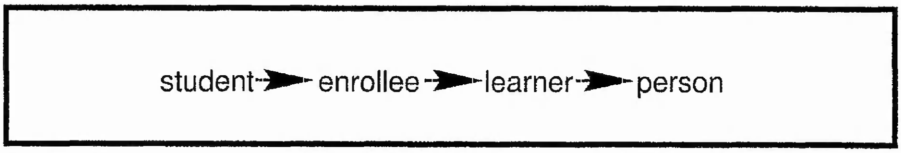


Figure 3.2: An example from the inheritance hierarchy

3.3.3 Relationships between Nouns

There are seven semantic relationships that interconnect the noun Synsets in WordNet. They are synonym, antonym, hypernym, hyponym, meronym, holonym and coordinate. If X is a kind of Y then Y is a hypernym of X and X is a hyponym of Y. If X is a part of Y then X is a meronym of Y and Y is a holonym of X. Coordinate is the relationship between concepts that have the same hypernym.

Table 3.2: High-level Categories in WordNet

Name	**Category**
Act	Nouns denoting acts or actions
Animal	Nouns denoting animals
Artifact	Nouns denoting man-made objects
Attribute	Nouns denoting attributes of people and objects
Body	Nouns denoting body parts
Cognition	Nouns denoting cognitive processes and contents
Communication	Nouns denoting communicative processes and contents
Event	Nouns denoting natural events
Feeling	Nouns denoting feelings and emotions
Food	Nouns denoting foods and drinks
Group	Nouns denoting groupings of people or objects
Location	Nouns denoting spatial position
Motive	Nouns denoting goals
Object	Nouns denoting natural objects (not man-made)
Person	Nouns denoting people
Phenomenon	Nouns denoting natural phenomena
Plant	Nouns denoting plants
Possession	Nouns denoting possession and transfer of possession
Process	Nouns denoting natural processes
Quantity	Nouns denoting quantities and units of measure
Relation	Nouns denoting relations between people or things or ideas
Shape	Nouns denoting two and three dimensional shapes
State	Nouns denoting stable states of affairs
Substance	Nouns denoting substances
Time	Nouns denoting time and temporal relations

3.3.4 Relevance to the project

WordNet has been widely used in various tasks, e.g. word sense disambiguation [6], word clustering [73]. It was created on psycholinguistic principles rather than on a computational model. The difference between WordNet and machine readable dictionaries (MRDs) is the organization of words or word senses. WordNet associates word senses according to semantic relations such as synonym, antonym, hypernym/hyponym, etc, between word senses and forms a semantic inheritance hierarchy. This feature which is very useful and important to automatic knowledge extraction can not be found in MRDs.

WordNet is free and downloadable from the WordNet home page [65].

3.4 CYC

3.4.1 An Overview of Cyc

Cyc [124] is the largest knowledge base project in the world. The intention of the Cyc project is to solve the problem of brittleness and the knowledge acquisition bottleneck of NLP and other knowledge-based systems. The main assumption underlying Cyc is that a huge set of basic 'common sense' knowledge about the world such as terms, rules and relations, etc will facilitate the solution to these problems and that the construction of this kind of knowledge base will make a variety of knowledge-intensive systems practically possible.

The project consists of a huge multi-contextual knowledge base (KB), an inference engine, interface tools (CYC APIs), and special-purpose application modules. The knowledge base and inference engine together enable CYC to understand and reason about its application domains.

3.4.2 The Cyc Knowledge Base

Most kinds of fundamental human knowledge, such as facts, rules of thumb, and heuristics for reasoning about the objects and events of everyday life, etc, are included in the Cyc knowledge base and are represented in a formal way using a formal language called CycL.

The Cyc KB currently contains over 1,000,000 hand-coded assertions or "rules" which are considered to represent common sense knowledge about the world. The size of the Cyc KB is ever-growing. New knowledge hand-coded by human experts is continually added to it. Unlike other KB systems, Cyc can add knowledge to its KB automatically by itself as the result of the inferencing process. There is no obvious end to this process.

To enable the effective inferencing on such a large knowledge base, the whole Cyc KB is divided into many smaller constituents called 'microtheories'. The division is done according to rules so that within the same microtheory the assertions share some kind of common features, such as a particular domain of knowledge or a certain level of detail. Another advantage of the microtheory mechanism is that assertions that are contradictory within the whole Cyc KB can be maintained. Currently, there are several hundred microtheories.

3.4.3 The CycL Language

CycL is the knowledge representation language used in Cyc. It is an augmented first-order predicate calculus, extended to deal with equality, default reasoning, and skolemization. CycL also includes some second-order features which are necessary for representing common sense knowledge.

A KB using CycL as its knowledge representation language is composed of a set of CycL sentences. The core part of a CycL KB is a set of abstract and very generic terms and fundamental axioms constructed from these generic terms.

3.4.4 The Upper Cyc Ontology

While most of the functionality of Cyc is not publicly available, a small part of Cyc KB called Upper Cyc Ontology is publicly available. Upper Cyc Ontology contains about 3,000 terms thought to represent 'the most general concepts of human consensus reality'.

3.4.5 Relevance to the project

Although empirical NLP is gaining in popularity, knowledge-based NLP is still an important area of research [87]. Non-statistical methods are still dominant in natural language understanding applications [141]. Cyc is an attempt to capture common sense knowledge that can be used in any domain and represents a very important research area. In addition, it has been successfully used in several domains and applications [62]. Cyc's huge knowledge-base and inference power can be invaluable to an NLP or an automatic knowledge extraction system. However, Cyc is currently not available because it has not been publicly released.

3.5 EDR

3.5.1 An Overview of EDR

The intention of the development of EDR [105] is to provide dictionaries for advanced NLP. The EDR Electronic Dictionary is the result of a nine-year project from 1986 to 1994, funded by the Japan Key Technology Center, together with commercial sponsors from the computing industry. The EDR Electronic Dictionary is composed of five types of dictionary, i.e. Word, Bilingual, Concept, Co-occurrence, and Technical Terminology, plus the EDR Corpus.

The Concept Classification Dictionary in EDR is an IS-A network of concepts. It contains a list of super-sub relations between concepts. The concept of a 'school', for example, has ' organization', 'building' and 'function' as its super-concepts, and has 'elementary school', 'university' and so forth, as its sub-concepts.

The Word Dictionary provides the relationships between words and concepts and the grammatical attributes regarding these relationships. The Bilingual Dictionary lists the correspondences between words in English and Japanese. The purpose of the Concept Dictionary is to provide information about concepts occurring in the Word Dictionaries, the Bilingual Dictionaries and the Co-occurrence Dictionaries. The Co-occurrence Dictionary gives information about co-occurrences including their frequency, and the situations in which they occur, part-of-speech classification and syntactic structure of collocated words. The Technical Terminology Dictionary covers the field of information processing, and consists of 8 sub-dictionaries in Japanese and English.

The EDR Corpus contains large amount of linguistic data and the analysis of that data. This dictionary provides: sample sentences and the contexts of the sentences; information about each constituent of the sample sentences; a syntactic tree of the sentences and concepts' relationships in the sentence.

3.5.2 Relevance to the project

The EDR is one of the most complete large-scale lexical resources available. It contains rich information on syntactic and semantic categorization and organises word senses into a hierarchy of concepts. It also provides details of semantic relations between concepts. As [67] points out, "The combination of syntactic and semantic information given in the EDR is particularly well suited for addressing questions concerning the syntax-semantics interface. Another strong relation to lexical semantic notions concerns the use of semantic relations, although the classification adopted goes well beyond the characterization of event participant roles."

The EDR has been used in a variety of applications including machine translation, lexical acquisition and word sense disambiguation. However, This dictionary is not free. It can be purchased in the form of a CD-ROM.

3.6 British National Corpus (BNC)

3.6.1 Introduction

The British National Corpus (BNC) [39] is a very large corpus of modern English used in the United Kingdom, including both written (90%) and spoken (10%) (100 million words). Samples in the written part of the BNC were selected from a wide variety of sources, such as newspapers (both regional and national), periodicals and journals on various subjects, books, fiction, letters, memoranda, essays, etc. For the spoken part, samples include many different conversations in different contexts.

The BNC has been used in a variety of different types of research related to the English Language, such as NLP, AI, speech recognition and synthesis, lexicography and computational linguistics.

3.6.2 Format

The BNC comprises more than 100 million words divided into 4124 texts. Out of these, 863 are transcribed from spoken conversations and monologues. CLAWS (the Constituent Likelihood Automatic Word-tagging System) developed by UCREL (University Center for Computer Corpus Research on Language) at Lancaster University was used to automatically segment texts into sentences and assign a word class (part of speech) to each word in the sentences. The segmentation produced over 6 million sentences from the corpus. The output from CLAWS and other structural information about the sentences, texts, etc, are encoded using SGML (Standard Generalised Markup Language).

Single-user editions of the BNC cost 50 pounds whilst network editions cost 250 pounds.

3.7 Brown Corpus

3.7.1 Introduction

The Brown corpus [93] is one of the earliest corpora widely used in NLP and other language-related research. It contains more than 1 million words which are divided into 500 samples, each of about 2000 words. Each sample starts at the beginning of a sentence and ends at the end of a sentence exceeding the 2000 word boundary. All the samples are from ordinary American written English, and are chosen from various printed materials in the USA published during 1961. The creator hoped that the corpus would be useful for various studies in language-related research.

The Brown corpus is composed of texts from four different categories which are divided into thirteen subcategories. For example, “press reportage” is subdivided into political, sports, society, cultural etc while “religion” is subdivided into books, periodicals and tracts.

Each text in the corpus is assigned an ‘identification code’ giving the name of its category and its position in the category.

3.7.2 Versions of the Corpus

There are six versions of the Brown corpus, all of which have the same basic text but in different formats. They include the basic form, and tagged (with part of speech) version and a version where sentences are represented as variable length records.

The Brown corpus is freely available.

3.8 Summary

Lexical semantic resources and corpora, which are widely used in NLP research, are reviewed in this chapter in terms of their appropriateness for the chosen project objectives of integrating semantic information into a TCE system using ANNs and an is-a type of relationship.

The approach chosen for this research (see 2.7) is to assess semantic similarity based on a broad range of semantic relationships rather than just hypernym/hyponym relationships. LDOCE is a dictionary, providing useful information, such as concept groups. However, it provides no relations between the concept groups and does not form a concept network. It does not satisfy the minimum requirement of the lexicon resource. CYC is a traditional rule-based knowledge base. It was not available outside the company when the lexicon for this research was chosen.

The Concept Classification Dictionary in EDR is an IS-A network of concepts. Thus it is suitable. WordNet is also suitable. It is a fully implemented semantic lexicon. Apart from hypernym/hyponym relation, it also contains synonym/antonym, meronym/holonym and coordinate relations which provide the possibility to utilise more relationships in calculating similarity between nouns.

The final choice is WordNet in that firstly, Wordnet is well researched and used in various NLP research [6, 73]. Secondly, it is freely available. In contrast there has been much less work using EDR in NLP. Thus, it would take much more time to use it in this research because of the lack of references. Also EDR is not freely available.

WordNet has 7 relationships (see 3.3.3). For symmetry, a new relationship called coordinatee is introduced here. This relates "nouns that have the same hyponym".

Chapter 4

Thematic Concept Extraction: Word Level

4.1 Introduction

As mentioned, the main purpose of this research is to develop a knowledge acquisition front end for HyperTutor. HyperTutor uses a kind of knowledge representation formalism similar to a semantic network. The ultimate aim of this research is to organise knowledge extracted into the same formalism. It represents knowledge as a network of nodes interconnected by links where the nodes denote concepts and the links denote relationships between concepts. In each node, there is text relating to the node including some derived from the link relationships. In this chapter, we refer to the names of nodes as keywords and are concerned with identifying them automatically as the first stage in a complete KA process.

4.2 Approach to Keyword Extraction

4.2.1 Proposed Solution

4.2.1.1 Introduction

The work presented here differs from most of that reviewed in that the task of TCE requires relatively focused extraction results. This work is also unique among reported concept extraction work in utilizing ANNs in the task and in combining ANNs with linguistic methods.

Two different methods in the extraction of concepts are used. The first is a machine learning approach based on ANNs. ANNs can learn from experience and examples, rather than being fully hand-coded by designers. The aim is to harness the ability to generalise from a wide range of examples to novel cases. Because most aspects of language cannot be learned by rote, the ability to generalise is crucial to learning language. Another feature of ANNs is the so-called "single route" feature. In many aspects of language, there are "quasi-regularities", i.e. regularities that usually hold but which have exceptions. To address this kind of problem, symbolic models employ separate sets of rules for regular and exceptional cases by providing rules and a list of rules for exceptions. By contrast ANNs are able to provide a single mechanism incorporating both regular and exceptional cases, and importantly, they are compatible with statistical and corpus-based NLP approaches. This method does not involve full NLU and so is potentially more tractable. However it also avoids the very domain-specific pattern-matching techniques of IE. The approach is novel in that although ANNs have been used in parsing [216], there have been no similar applications of ANNs reported for KA.

The other approach utilised for keyword extraction is stemming analysis (SA) [189]. SA is commonly used to extract concepts in IE and information retrieval (IR). The motivation for this approach and its combination with the ANN method are explained in chapter 5.

4.2.1.2 Overview of the ANN Approach

A supervised learning method is used in extracting keywords from a document. The basic idea is to choose a word called a "seed word" which defines the domain. For example, for the domain education, the word "education" would be a suitable seed word. Then similarity between the nouns from the document and the seed word can be calculated. However, instead of calculating the similarity of nouns as in [131, 102, 163] and presenting it to an ANN, this method presents information to calculate similarity to the ANN. This information includes the category of the noun and the path of the relations from the noun to the seed word. In the training process, the ANN should be able to learn to use this information to judge a noun to be a keyword or non-keyword.

Thus approach is to train an ANN to differentiate between keywords and non-keywords based on an input representation of their relationships to a seed word. The relationships between each potential keyword and the seed word are obtained by searching an electronic semantic lexicon. The central idea is to find relationships between nouns and the seed word in a semantic lexicon and to use these relationships to differentiate between keywords (KWs) and non-keywords (NKWs). ANNs are trained with patterns representing these relationships. The idea is based on the hypothesis that in a semantic lexicon, nouns that are strongly relevant to the seed word in semantic terms will have different relationships to it than words that are not relevant. Thus, nouns that are closely related to the seed word (i.e. KWs) should be differentiable from other nouns (NKWs). Once trained, the network should be able to recognise input patterns/relationships that correspond to keywords of the domain represented by the original seed word. This is a typical pattern classification problem.

Training data consists of input patterns for KW and NKW examples where the KW/NKW distinction has been judged by humans. The input patterns of a noun are composed of representations of the category the noun belongs to and the shortest paths from the noun to the seed word. A category is a set of nouns with the same feature(s). For example, LivingThing is a category containing all animals and plants. A path between two nouns is defined as the relationship types between the adjacent nouns on the route from one noun to the other in the lexicon. Both kinds of information can be obtained from an is-a hierarchy such as is provided by most semantic lexicons.

The only requirement of the semantic lexicon is that it provides an is-a hierarchy. WordNet is suitable, but other lexicons (e.g. EDR [105]) could also be used. This means that the approach is not limited to any particular lexicon.

Domain independence is feasible because the category and path information can be obtained for any seed word and noun. To take advantage of this, a domain independent lexicon must be used. The approach can then be implemented for any domain.

The approach is outlined in the following steps:

1. The nouns in documents relevant to the seed word domain are divided into two groups for training and testing respectively, see section 4.4.1. These are each judged as being KWs or NKWs by humans. The nouns in the training set form the basis for the training data.
2. All training nouns and their relationships to seed words are identified automatically according to a universal (domain-independent) semantic lexicon. All the information for a noun (mainly derived from paths in the lexicon between the noun and the seed word) is organised into a pattern that will be input to an ANN for training. The output target is 1 or 0 depending on whether the noun is a KW of the seed word or not.
3. The ANN is trained. The ANN architecture is optimised on the training data by finding the minimum number of hidden neurons capable of learning all of the training patterns.
4. The trained ANN is tested to see how well it can extract keywords from the test nouns. A threshold, 0.5 is applied to the output to decide between KWs and NKWs.

4.3 Evaluation Approaches

4.3.1 Introduction to Methodologies

Basic neural network theory tells us that if a problem is linear, it can be solved without the use of hidden neurons, i.e. with a single layer of connections between input neurons and output neurons. In the case of this study, hidden neurons are required to solve the problem to any reasonable level of accuracy, therefore the problem is non-linear and non-trivial.

To evaluate this novel approach, measures are introduced based on the concepts of generalisation in ANN research, and on recall and precision widely accepted in KA research. The most basic measure (generalisation known here as natural generalisation) states what proportion of nouns are correctly classified (as KWs or NKWs) in the test text. Standard binary recall and precision measures are also applied [128]. Further, more sophisticated measures are developed and introduced to give a more detailed picture of performance. These are pure generalisation and analogue measures of recall and precision. Finally, a single "comparison with chance (CWC)" measure is used in the results evaluation. These measures are described in overview below and in detail in the following sections.

As stated, both binary and analogue recall and precision metrics are used. In traditional information retrieval, recall and precision are binary metrics. However, the analogue nature of the ANN output and the desire to have a single overall performance measure that is unbiased according to the ratio of KWs to NKWs, has led to the development of novel analogue measures of recall and precision. These incorporate confidence into the measurement of performance.

Generalisation is appropriate to evaluate ANN results. However, the linguistic problem domain suggests two types of generalisation: pure and natural [218, 219]. Pure generalisation evaluates the effectiveness of the ANN learning of the problem in terms of its ability to classify unseen patterns. Natural generalisation evaluates utility in terms of the effectiveness of the classification on unseen text. This is more appropriate for evaluating the overall ability of the trained network in performing the text processing task.

Natural generalisation and pure generalisation, in many applications (e.g. image processing) are the same because, for example, no two photographs are identical. In applications where the source of the data gives continuous values, natural and pure generalisation tend to be the same, because patterns are unlikely to be repeated. However, in domains where the source data is symbolic, natural and pure generalisation are different because patterns tend to be repeated. This is particularly so with language within a given domain where some words are bound to be repeated.

The comparison with chance measure gives a single numerical output between 1 and 2. A result of 1 means that the method is performing no better than chance. A result of 2 is perfect performance.

Table 4.1 shows the definitions used in the following section. Here, "unique" means words or patterns that only exist in the test set i.e. not in the training set.

Table 4.1: Definitions of symbols-all relate to the test set

Definition	**Symbol**
Number of keywords patterns	N_{kw}
Number of non-keywords patterns	N_{nkw}
Number of unique keywords patterns	N_{ukw}
Number of unique non-keywords patterns	N_{unkw}
Number of patterns identified as keyword patterns	N_{ikw}
Number of patterns identified as non-keyword patterns	N_{inkw}
Number of patterns correctly identified as keyword patterns	N_{ickw}
Number of patterns correctly identified as non-keyword patterns	N_{icnkw}
Number of unique patterns correctly identified as keyword patterns	N_{icukw}
Number of unique patterns correctly identified as non-keyword patterns	N_{icunkw}

4.3.2 Natural Generalisation

Generalisation refers to how well a network performs with new data sets after training. The ability to generalise is the main reason that ANNs attract researchers. Generalisation refers to the ability to learn not only by memory but more importantly, by induction. Therefore generalisation forms the basis of the evaluation of ANNs.

As previously mentioned, two types, natural and pure, were defined. Natural generalisation (NG) is the percentage of nouns in the test data that are correctly categorised as KWs or NKWs. This can be evaluated for the total test set (NG_{total}) or evaluated separately for keywords (NG_{kw}) and non-keywords (NG_{nkw}). Therefore:

$$NG_{total} = \frac{N_{ickw} + N_{icnkw}}{N_{kw} + N_{nkw}} \tag{4.1}$$

$$NG_{kw} = \frac{N_{ickw}}{N_{kw}} \tag{4.2}$$

$$NG_{nkw} = \frac{N_{icnkw}}{N_{nkw}} \tag{4.3}$$

In TCE, the thematic concepts (KWs) are only a very small part of the overall concepts in a document. Thus the KWs and NKWs in this situation are extremely unbalanced. It is therefore necessary to apply performance measures separately for KWs and NKWs. It is obvious that if we only calculate the overall performance, this would be swamped by the performance for the NKWs, leaving the performance on KWs nearly unknown. The results (see later) support this hypothesis. Thus, measuring them separately is absolutely necessary. This is one of the main differences of this research from other concept extraction research.

For completeness, overall performance is also measured although these values are usually very close to those for NKWs.

4.3.3 Pure Generalisation

NG is indicative of the overall performance on unseen text, but in terms of neural network learning may include data that is repeated from the training set. This means that a component of the natural generalisation measure may involve memorisation. NG alone is therefore not sufficient to fully evaluate the learning of the ANN. Let us consider an extreme situation: suppose all words in the test set had also occurred in the training set. Because the network can memorise all the patterns in the training data set (provided there are enough hidden neurons in the network), then all the patterns in the test set will be identified correctly but it may be that no general learning has taken place. Using NG to evaluate the performance of the network could result in a very high score (1.0), but this says nothing about how much knowledge about the nature of relationships has been derived from the training examples. Pure generalisation (PG) has been introduced to measure the amount of induced knowledge. PG is the percentage of nouns with previously unseen input patterns in the test data that are correctly classified. Again it can be applied to the total result and to KWs and NKWs separately. PG measures are defined as:

$$PG_{total} = \frac{N_{icukw} + N_{icunkw}}{N_{ukw} + N_{unkw}} \quad (4.4)$$

$$PG_{kw} = \frac{N_{icukw}}{N_{ukw}} \quad (4.5)$$

$$PG_{nkw} = \frac{N_{icunkw}}{N_{unkw}} \quad (4.6)$$

4.3.4 Recall and Precision: Binary

As we have seen, NG is the percentage of the target knowledge that has been extracted and PG is the percentage of the unique target knowledge that has been extracted. However, not all of the extracted knowledge is correct. To know how accurate the extraction is, other measures must be introduced because neither NG nor PG can be used to measure the accuracy of extraction. Precision [128] commonly used in KA research is the right metric for accuracy. A metric usually used together with precision is recall.

Recall [128] measures the ratio of correct information ($N_{correct}$) extracted from the text against all the information (N_{all}) available in the text. Precision measures the ratio of correct information that was extracted against all the information extracted ($N_{extracted}$). Thus,

$$recall = \frac{N_{correct}}{N_{all}} \quad (4.7)$$

$$precision = \frac{N_{correct}}{N_{extracted}} \quad (4.8)$$

These are applicable to keywords and non-keywords separately and are defined for KWs as

$$recall_{kw} = \frac{N_{ickw}}{N_{kw}} \quad (4.9)$$

$$precision_{kw} = \frac{N_{ickw}}{N_{ikw}} \quad (4.10)$$

and for NKWs as

$$recall_{nkw} = \frac{N_{icnkw}}{N_{nkw}} \quad (4.11)$$

$$precision_{nkw} = \frac{N_{icnkw}}{N_{inkw}} \tag{4.12}$$

From the definition, we know that recall is the same as NG. It is included here for convention and completeness. It does not provide any new information about the performance.

4.3.5 Precision and Recall: Analogue

Precision and recall measures have two limitations. Firstly, they do not immediately provide an overall performance measure because they take no account of the ratio of KWs to NKWs. Secondly, they do not accommodate the analogue nature of the ANN response which provides extra information about the level of confidence of the decisions. Therefore, the basic formulae 4.7 and 4.8 for recall and precision have been adapted.

Suppose the target and actual output of a pattern P in the test data set, TS, are T_p and A_p respectively, where T_p is either 1 or 0 and $0 <= A_p <= 1$.

Correctness is defined as decreasing in proportion to the output error, but also increasing in proportion to the deviation from 0.5, since that is the point of zero correctness. This gives a correctness scale of {0,1}, as follows:

$$N_{correct_p} = 2|A_p - 0.5| * (1 - |T_p - A_p|) \tag{4.13}$$

Therefore $N_{correct}$ for all patterns is:

$$N_{correct} = 2 \sum_{p \in TS} |A_p - 0.5| * (1 - |T_p - A_p|) \tag{4.14}$$

Extraction is defined in terms of the decisiveness or responsiveness of the network i.e. its deviation from a natural response. Since A_p is in the range {0,1}, an output of 0.5 means the network makes no response to P. So

$$N_{extracted_p} = 2 * |A_p - 0.5| \tag{4.15}$$

A coeffcient of 2 puts $N_{extracted_p}$ in the range of 0 and 1. Therefore, for all patterns $N_{extracted}$ is

$$N_{extracted} = 2 \sum_{p \in TS} (|A_p - 0.5|) \tag{4.16}$$

The number of patterns is the sum of the number of KW patterns and the number of NKW patterns, thus

$$N_{all} = N_{ikw} + N_{inkw} \tag{4.17}$$

Thus, the formulae of recall and precision suitable for an ANN-based approach are:

$$\begin{aligned} R &= \frac{2\sum_{p\in TS}|A_p - 0.5| * (1 - |T_p - A_p|)}{N_{ikw} + N_{inkw}} && (4.18) \\ P &= \frac{\sum_{p\in TS}|A_p - 0.5| * (1 - |T_p - A_p|)}{\sum_{p\in TS}(|A_p - 0.5|)} && (4.19) \end{aligned}$$

From the formulae, it can be seen that analogue recall is the average level of correctness and certainty. Analogue precision is the average correctness.

4.3.6 Comparison with Chance

In section 4.3.2 the overall NG is swamped by the performance for the non-keywords. This leads to the NG measures being calculated separately for keywords and non-keywords. However, a single measure that can provide a single numerical output giving equal weight to keywords and non-keywords is still desirable. Such a measure would also enable comparison of different keyword extraction methods.

If a keyword/non-keyword recogniser is run that works totally by chance, it will divide words into KWs and NKWs in a ratio that adds up to 1 (0.7:0.3, 0.5:0.5 etc). The natural generalisation measures for keywords and non-keywords would be based on the same ratios (70% and 30%, 50% and 50% etc). Therefore summing the KW and NKW natural generalisation measures gives a comparison with chance. A figure of 100 (or 1) is no better than chance, a figure of 200 (or 2) is perfect classification. The same approach can be used for PG.

4.3.7 Summary

Several evaluation measures have been defined. This is justified by the fact that this is quite a new research area and each of the measures (except binary recall and natural generalisation) evaluate different aspects of the approach. No measures can be removed without losing information which may be useful to understand and evaluate the approach. Each measure tells us something different (see section 4.4.7).

4.4 Experiments with Category and Path Information

4.4.1 Training Data Preparation

Initial experiments were performed with the domain defined by the seed word "education". The document chosen is a research paper entitled "Mediated Learning: A New Model of Networked Instruction and Learning" [15]. The document can be found in appendix A. See appendix B for the nouns in the document. It is comprised of 19672 words with 707 unique nouns occurring 8334 times. 54 words were identified as keywords. Three human judges based their keyword classification on considering education in the sense of "education in a formal setting" and decided which nouns were strongly associated with this. Where there was a discrepancy between the independent judgements, the humans conferred and a joint decision was made.

The whole text was divided into two sections so as to distribute unique keywords approximately evenly between the three parts, i.e. half in each part. The result of the division is shown in Table 4.2. In this table, unique KWs in the test set are those that do not occur in the training set.

Sample documents are used to mimic the situation where an author is converting a document concerning a given domain into the HyperTutor knowledge representation scheme. The nouns in the sample document and sets of training and test are given in Appendix B, together with the human-generated classification between keywords and non-keywords.

Table 4.2: Result of document division

Part	Nouns	Keywords	Unique Keywords
Whole document	707	54	54
Training set	111	19	19
Test set	344	39	20

4.4.2 Input Pattern Design

The ANN architecture used was a Feed-Forward Multi-Layer Percepton (FF-MLP) with Back-propagation. As stated in section 4.2.1.2, the identification of KW/NKW is treated as a pattern classification problem. FF-MLP is widely used in solving this kind of problem. Thus, a three-layer (one hidden layer) FF-MLP architecture is used. Because there is mathematical proof [118] which states that theoretically, the computational power of more than one hidden layer is equivalent to that of one hidden layer, one hidden layer was considered to be sufficient.

For each noun in the training document, there is an input-output pattern pair in the training data set. Each input pattern is composed of two parts. The first is the category information of the noun. This information should be useful because it seems that a KW is likely to have the same category as the seed word or be restricted to a small set of categories if the granularity of categories is fine enough. The twenty-five high-level categories in Table 3.2 in the inheritance hierarchy are used in this part, so there are twenty-five bits to represent category information. The bit corresponding to the top-level category that the noun belongs to is set to 1, the remaining bits being set to 0.

The second part of the input pattern is more complicated. It represents aspects of paths in WordNet between the noun and the seed word in terms of the length of the paths as well as the relationships between the words on the linking paths. A path from one noun to another is composed of a list of the relationship types between all the adjacent nouns. For example, a path from "university" to "education" is shown in figure 4.1. (The intermediate words on the path are not represented on the input as the structural information about them in WordNet is confined to their relationships to other words.) The distance from university to education is 2, consisting of two of the eight types of possible relationship. The second part of the input pattern consists of **the relationship information of the shortest N paths up to a maximum length of M.** The criteria for choosing M and N are described later.

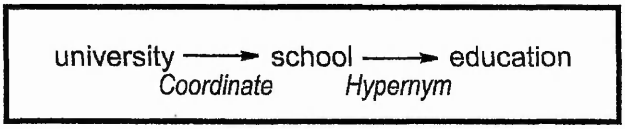


Figure 4.1: An example path in WordNet between 'university' and 'education'

It is noted that two kinds of information are encoded into an input pattern: category and paths. Are they both necessary? It is obvious that the category itself is not sufficient to differentiate keywords from non-keywords. If only the category information is used, all nouns from the same hierarchy must be KWs or all NKWs, otherwise, contradictions would arise. So, because the category is too general, KWs and NKWs must exist in the same category. However, investigation has shown that without using category information, there are contradictions in the training patterns for the ANN. This indicates two things. Firstly, paths tend to repeat. This means that the **path information is in some way fundamental to the relationship between nouns.** This supports the use of path information in training the ANN. Secondly, it indicates that both category and path information are needed to differentiate KWs from NKWs. Analysis of the trained ANN has shown that both category and path information are required. The ANN uses both kinds of information in determining if a noun is a keyword or not. A detailed analysis can be found in chapter 7.

4.4.3 The Selection of Paths

Nearly all nouns have more than one path to a seed word, so: how many paths are enough for training; what maximum length should be used; and which paths should be selected? The criterion set to find a solution to these questions is to present enough information for the network to learn the problem. This was used to choose M and N. M should be large enough for all KWs in the training data to have at least one path with a length equal to or shorter than M. If M is too small, some KWs will be presented to the ANN with no path information, which would give the network no information on which to base its selection.

In order to learn the problem, enough information needs to be presented for there to be no contradictions in the training data. A contradiction occurs when two patterns have the same inputs and different target outputs. If there are contradictions in the training data, the ANN will not be able to acquire any sensible knowledge about them.

If two nouns belong to the same WordNet category, and have the same paths to the seed word but one is classified as a keyword and the other a non-keyword, there will be a contradiction. See Figure 4.2, where "week" and "semester" both belong to the same category and have the same path to education. Identical path information can also be generated when the intermediate words are different between the two paths but the relationships are the same.

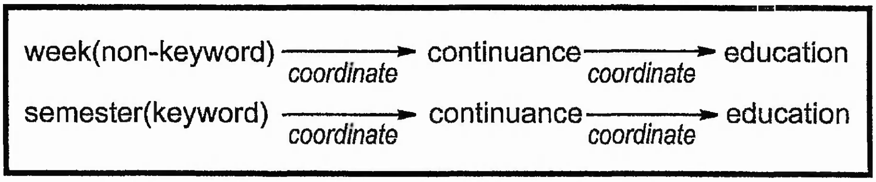


Figure 4.2: Example of contradictory Paths

Therefore, one path for each noun is not enough to distinguish between them. A combination of M and N is required such that there are no contradictions in the training data set. However, the M-N combination should also minimize the amount of training data in order to make the training as easy as possible. For the nouns that have more than N shortest paths to choose from, the first N paths found are chosen. For those that have less than N shortest paths, the path length is increased until N paths are found or the path length exceeds M.

A further complication is that there is no systematic way of ordering paths on the input. If only one ordering of path is used in training, the same word tested with the same path but with a different ordering of path will not be recognised correctly. Therefore, training data is generated with input patterns for all the possible ways of ordering the inputs. This aims to allow the network to recognise path features regardless of the order in which they were found when WordNet was searched, e.g. for N=3 paths, 6 training patterns containing the 6 path order permutations (together with the category information) are generated.

4.4.4 Path Representation on ANN Inputs

A major part of an input pattern represents the relationship pattern on the N shortest paths of max length M between the noun and the seed word. There are eight relationships, seven from WordNet and one added for symmetry.

Suppose the maximum path length is 4. A path will therefore have a length in the range 1 to 4. In the scheme used this means that there would be 4 fields, A to D, each representing one of the 4 possible path lengths. Each field contains sub-fields that allow the relationship type for each link on the path to be represented. The sub-field (relationship type) is represented using 8 bits. Each bit corresponds to one relationship so that, as with the classification coding, only 1 bit is high at a time. The bit distribution for 1 path with M=4 is shown in figure 4.3.

A	B	C	D
8 bits	16 bits	24 bits	32 bits

Figure 4.3: Bit pattern for a path with maximum length M=4 (80 bits in total).

A denotes a path of length 1, B denotes a path of length 2, C a path of 3 and D, 4. If a path is of length 1, then all bits in the B, C and D fields are set to 0. The A field is set according to the relationship i.e. the bit corresponding to the relevant relationship is set high. If the path is of length 2, then the bits in the A, C and D fields are all set to 0 and the B field is set according to the relationships in the path: the first 8 bits is used to represent the first relationship and the second 8 bits is used to represent the second relationship. The same principle applies to paths of length 3 and 4.

For the example in figure 4.1, the length of the path is 2. The pattern of this path is shown in figure 4.4.

A	B		C	D
0...0,	00000100	00010000	0...0,	0...0
8 bits	*coordinate*	*hypernym*	24 bits	32 bits

Figure 4.4: Bit pattern of the path of length 2 in Figure 4.1

Up to N paths can be represented; thus the total input pattern to represent the shortest N paths with maximum length M between a noun and seed word and with, for example N=3 and M=4 is shown in figure 4.5.

category	**path1**	**path2**	**path3**
	ABCD	ABCD	ABCD
25 bits	80 bits	80 bits	80 bits

Figure 4.5: Total input pattern of 3 paths with maximum length 4

The output pattern is one bit which has a target of 1 for a keyword or 0 for a non-keyword.

4.4.5 Pattern Coding Scheme

Neural networks are adaptive systems that have automatic learning properties. They adapt their internal parameters (weights) in order to satisfy constraints imposed by training data and the training algorithms. Jackson [74] has pointed out that traditional pattern recognition failed to solve many classification problems partially because the task of identifying and extracting appropriate features from raw data is too complex. However, the main purpose of ANNs in this context is to discover the relevant feature in the raw data as suggested by Wasserman [230]. Consequently, it may be advantageous to present large, unprocessed data vectors to networks and let the training procedure identify the useful information during training.

To express the type of the relation, eight bits are used because there are eight types of relations from WordNet. Two possible coding schemes are shown in table 4.3.

Table 4.3: Coding Scheme for a Relation

Number	Relation	Bit Pattern	Binary Code
1	Synonym	00000001	000
2	Antenym	00000010	001
3	Coordinate	00000100	010
4	Coordinatee	00001000	011
5	Hypernym	00010000	100
6	Hyponnym	00100000	101
7	Meronym	01000000	110
8	Holonym	10000000	111

It is possible to use binary coding to express the eight relation types to reduce the input vector space. The Binary column shows a possible binary coding scheme. However, binary coding is biased in that some classes are more similar in terms of shared bits than others. In contrast, 00000001 and 00000010 have equal similarity to 0000001 and 10000000, etc. So the fully redundant 8 bit pattern is used in order to maintain neutrality. The input vectors are all mutually or orthogonal.

4.4.6 Training the ANN

The initial weight range of the FF-MLP was set to between {-0.5,0.5}, and the error threshold was set to 0.2. The momentum used was 0.3.

Pattern-orientated adaptive learning, a modified form of backpropagation was used[218, 219]. Suppose the current learning rate and the learning error for pattern P are α and E respectively, then the new learning rate for P, α', will be:

$$\alpha' = \alpha + (1 - \alpha) * |E| \qquad 0 < \alpha < 1 \tag{4.20}$$

This method requires E to be in the range {-1,1}. The Sigmoid output satisfies this requirement. Experiments show that this is a very efficient learning method. The network using this method converged within 44 iterations while it needed more than 4110 iterations using a constant learning rate.

A series of tests was carried out to establish the optimum combination for M=4 and N=5 that would remove all contradictions.

According to the representation scheme, for the 111 training nouns there should be 13320 (111*N! = 111*5!) training patterns. Patterns representing keywords were repeated in the training set to balance the number of keyword patterns and non-keyword patterns because without balancing the distribution of patterns in the training set is very biased. The ratio of non-keywords to keywords is about 5 : 1. By duplicating all keyword patterns 4 times, the training data was balanced. The total of extra patterns is 19*5!*(5-1)=9120. Thus, altogether, 22440 training patterns were presented to the network and the network learnt all the patterns in 44 iterations.

Experiments were also performed to minimise the number of hidden-neurons. A binary search method was used for this. First, the number of hidden-neurons was set large enough (e.g. 40) so that the problem can be learned by the network. Then, the number was halved (20) and the network was trained again (many times with different initial weights). If the network could not learn the problem with this number of hidden-neurons, the number was set to midway between the two numbers (30) otherwise the number was halved again (10). Using this binary search method, the minimum number of hidden-neurons was found to be 2.

Thus the architecture of the ANN used is shown in figure 4.6. There are 425 nodes in the input layer, and 2 and 1 in the hidden-layer and output layer respectively. As 5 paths with maximum length 4 were used for each input pattern and each path needs 80 bits, 5 paths need 400 bits. 400 plus 25 bits of category information makes 425 bits for input.

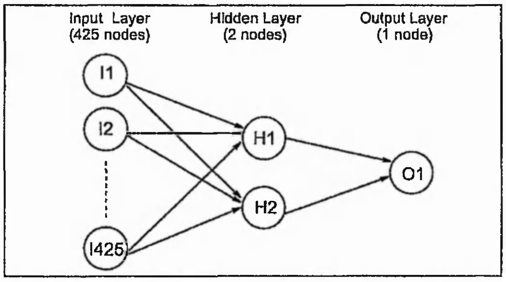


Figure 4.6: The architecture of the ANN used

Training was deemed to be complete when all the training data had been acquired so that the outputs were all within 0.2 of the targets.

4.4.7 Results of Experiments

After training, the network was presented with 41280 (344*5!) patterns in the test data set to see how well the network had learnt the problem. A threshold of 0.5 was used to classify a tested pattern, i.e. if the output of a tested pattern is larger than 0.5, the network had classified it as a KW. If the output is less than 0.5, it was considered to have been classified as a NKW. The results are shown in Table 4.4.

Table 4.4: Test set results with ANN trained on Category and Path information (M=4 and N=5)

Word Type	Total Number of Words	Number of Patterns	Percentage
Total Nouns	344	41280	84%
Keywords	39	4680	62%
Non-Keywords	305	36600	87%
Unique Nouns	252	30240	82%
Unique Keywords	20	2400	47%
Unique Non-Keywords	232	27840	83%

Applying the evaluation measures to these experimental results, gives the results in table 4.5.

Table 4.5: Experiment Results

			Analogue	Analogue	Binary	Binary	Comparison with Chance	
Data Set	NG	PG	Recall	Precision	Recall	Precision	Natural	Pure
Total	0.84	0.82	0.81	0.86	n/a	n/a	1.49	1.30
Keywords	0.62	0.47	0.59	0.63	0.62	0.38		
Non Keywords	0.87	0.83	0.84	0.88	0.87	0.95		

Table 4.5 shows that the NG and PG for all the nouns in the test set are 0.84 and 0.82 respectively. This means the network can achieve 84% accuracy in classifying all the nouns (both keyword and non-keyword) and 82% accuracy in classifying unseen nouns. Considering the complexity of the problem, these are good results. The natural CWC is 1.49, significantly above chance (1.0). Pure CWC, 1.30, is also much better than chance. The generalisation measures of NKWs is slightly higher than the generalisation measures of the overall test nouns. However NG and PG for KWs is lower. The results for NKWs are much better than those for KWs. From table 4.5, the same trends can be observed for recall and precision measures, both analogue and binary, yet there were also some significant differences between the various measures.

Precision is a difficult measure to optimise since it requires detection of all KWs and rejection of all NKWs, whereas recall is only dependent on the level of detection. High KW precision is more difficult to achieve than high NKW precision, because the ratio of KWs to NKWs is much lower than 1/2. Thus, it is not surprising that the worst statistic of performance in table 4.5 is binary precision. However, analogue precision for KWs is much higher, at 0.63. This is because of the range of responses between 0 and 1: in general KWs give higher responses than NKWs even when the response is below the 0.5 threshold, so that they are classified as NKWs. This means that when the network output results in a wrong classification, the level of confidence tends to be lower than when the output results in a correct classification.

Is is noticeable that the difference between the binary recall and binary precision of KWs is much bigger than that of NKWs. The reason is again the low ratio of KWs to NKWs (39 to 344). The binary recall for KWs is reasonable good. However, 13% of NKWs were incorrectly identified as KWs. The number is nearly the same as the number of KWs, making the binary precision quite low.

It was thought that the reason for the low performance on KWs in initial experiments was the imbalance between the numbers of KWs and NKWs, i.e. as the number of NKWs is much higher than the number of KWs in the training set, the unbalanced training data made the network biased towards classifying a noun as a NKW. The results in tables 4.4 and 4.5 prove this hypothesis wrong because the training data has been balanced (see section 4.4.6).

Another possible reason for the low performance on KWs is that the information presented to the network is not sufficient for the network to learn to differentiate the KWs from the NKWs. It is based on this hypothesis that the next experiment described in chapter 5, using category and sense-level path information, was proposed.

4.4.8 Baseline Comparison

Table 4.6: Baseline and ANN Results

	Decision Rule with Steps(N)				
Measure	**1**	**2**	**3**	**4**	**ANN**
$BinaryRecall_{kw}$	0.31	0.39	0.49	0.64	0.62
$BinaryPrecision_{kw}$	0.27	0.17	0.13	0.11	0.38
$BinaryRecall_{nkw}$	0.89	0.76	0.58	0.35	0.87
$BinaryPrecision_{nkw}$	0.91	0.91	0.90	0.88	0.95
$NaturalGeneralisation_{total}$	0.83	0.72	0.57	0.38	0.84
$NaturalGeneralisation_{kw}$	0.31	0.39	0.49	0.64	0.62
$NaturalGeneralisation_{nkw}$	0.89	0.76	0.58	0.35	0.87
$ComparisonWithChance_{Natural}$	1.20	1.15	1.07	0.99	1.49

A simple method involving just WordNet is used to give baseline results to evaluate the contribution of the ANN to the solution which involved WordNet information and ANN processing. Instead of evaluating the relationships along the paths between a word and the seed word, a simple decision rule is applied: any word within N steps of the seed word is deemed to be closely related to it and is therefore classified as a KW. This gives the results in Table 4.6 for comparison with the ANN-based method.

The results in Table 4.6 show that the task of extracting KWs is complex. The simple 'distance from seed word' rule is inadequate: it fails to extract most KWs until the step size is so large that a high proportion of NKWs are mistaken for KWs. The ANN approach is a significant improvement on this situation. To a precision of one decimal point, the results in Table 4.6 show the ANN to equal or improve on every metric for all step sizes. Considered overall, across the various metrics, the ANN is a significant improvement, irrespective of step size.

Chapter 5

Investigation into the Use of Sense Level Information

5.1 Experiments with Category and Sense-Level Path Information

It is observed from table 4.5 that the performance on keywords is low. One possible reason is that in the previous experiments, a noun was treated as a whole, i.e. the different meanings (or senses) are not distinguished. WordNet can provide path information based on the more accurate word senses. This means the information presented to the network in the previous experiments may not be sufficient for the network to learn the problem. If the differentiation of word sense were taken into account, the performance on the keywords might be improved. Thus the following experiments were undertaken.

5.1.1 Word-Level and Sense-Level Paths

In WordNet, the relationships are actually the relationships between senses of nouns rather than between nouns. Examples are shown in arrows in figure 5.1 and 5.2 where W1 to W6 are words and S1 to SN below each are the N separate senses of the word. The arrows show where there is a relationship between two senses. In the previous experiments relationships between senses were taken as relationships between the words concerned, so word level paths were used.

In figure 5.1, W1 and W3 are connected at the word level and the sense level by virtue of the connections shown. In contrast, W4 and W6 in figure 5.2 would only be connected at the word level. There is no sense level path between them.

At word-level, different senses of the same word are not distinguished, so all links between senses of two words are considered as the links between words. Thus, word-level paths are much more common than sense-level paths in WordNet.

From figure 5.1 and 5.2, it is clear that sense-level paths are more accurate than word-level paths

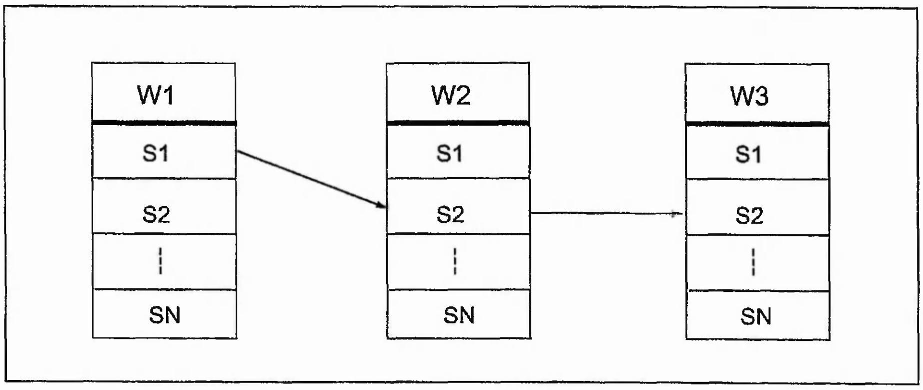


Figure 5.1: Sense-level path in WordNet

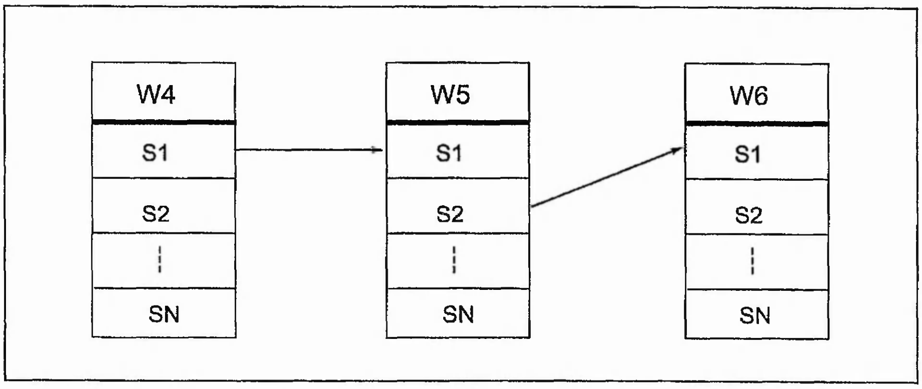


Figure 5.2: Word-level path in WordNet

and should be better in differentiating between keywords and non-keywords. In an extreme case, a word level path including the word "match" could be made via the sense meaning "an implement to start a fire" and the sense meaning "a sporting event". Clearly this path would not be semantically valid. In the series of experiments described in this chapter, nouns are treated as separate senses. Those senses for which a noun was judged as a keyword were identified as **key senses**. Two senses of the seed word "education" were identified as seed senses; they are sense 1, " activities that impart knowledge" and sense 4, "the profession of teaching (especially at a school or college or university)". As before for word level, the hypothesis is that the paths from key senses to a seed sense have some features that other paths do not have, e.g. the lengths of paths from key senses to seed senses should be shorter than the lengths of other paths. This and other features of these paths should be useful for the network to differentiate between key senses and other senses.

5.1.2 Word Sense Information

The problem of distinguishing between different meanings or senses of a word is generally known as word sense disambiguation (WSD). WSD has long been considered to be able to increase the accuracy of natural language processing, e.g. IR [212] and machine translation [99]. This section presents experiments carried out to investigate the possibility of using WSD to increase the rate of correct KW extraction. The aim is not to decide which sense a noun should take in a given context (which is the usual aim of WSD), but is to use word sense information to increase the accuracy of keyword identification.

In WordNet, the relationships between nouns are actually the relationships between senses of nouns. For example, in WordNet, although it is thought that "course" is represented as a hyponym of "education", the full representation is actually that sense one of "course" (defined as: Education imparted in a series of lessons or class meetings) is a hyponym of sense one of "education" (Activities that impart knowledge). Another example is sense three of "experience" (The accumulation of knowledge or skill that results from direct participation in events or activities) which is a hyponym of sense three of "education" (The gradual process of acquiring knowledge) which is not a key-sense according to the definition of the domain education used here. In the second example, the path from sense one of experience to sense three of education is not of concern because only senses one and four of education are seed senses. Therefore a noun should be identified as a keyword only when at least one of its senses is a key sense. If the identification is carried out via paths through WordNet then at least one sense of a keyword should have a qualifying path via intermediate word senses to one or more key senses of the seed word. In WordNet, "education" has six meanings, but only two of them are relevant to the domain definition used in these experiments. Some nouns may have paths to "education" but not to the seed senses. These paths are therefore spurious and should be excluded.

In order for experiments like those presented in section 4.4 to be carried out at the sense level, there must be patterns in the training data for each sense of a noun in the training set. This also applies to the test data.

Before the generation of sense-level paths, all senses of the nouns in the document were independently examined as before. The three human judges repeated the assessment of whether nouns were key or not by looking at all of the senses of the nouns and judging whether they related to the seed senses of "education". A final decision was agreed between the three judges for senses on which there were disagreements. The nouns are again taken from the document used in previous experiments. The resulting classifications are given in Appendix C.

5.1.3 The Limits of WordNet

Finding paths for a sense needs to be strictly based on the sense information. This leads to the discovery that WordNet has a lack of links between concepts. For example, sense two of "college" is defined as "An institution of higher education created to educate and grant degrees; often a part of a university" and yet has no direct link to either key sense of the seed word "education". It has no path to "higher education" either although "higher education" appears in its definition. However it has paths to "educational institution" but this is also not connected to "education" and "higher education". This is an aspect of the implementation of WordNet.

It is therefore concluded that there is no close relationship between how the noun synsets are connected together in the WordNet hierarchy and the definitions of the noun synsets. This leads to a lack of paths and hence the conclusion that keyword identification can not be carried out by just using the sense-level path information as had been hoped. There is a clear need for an extended form of WordNet.

5.2 Experiments with Stemming Information

By examining the sense definitions of the synsets it was found that words in the synset definitions of KWs usually have a strong verbal relevance to the seed word, as was seen for "college" in 5.1.3. The possibility of using the definition information for KW identification has therefore been investigated. Rose and Evett [189] used similar techniques.

5.2.1 Stemming analysis

Two possible ways of utilising the sense definitions suggest themselves. The first way is to syntactically analyse the structure of the sense definition and extract constituents which are semantically linked to one or more seed senses. The other way is to perform stemming analysis on the words in the sense definitions, which can produce numeric values of similarity between the word senses and the seed senses. The latter is much easier to carry out than syntactic analysis and its numeric result is also easier to combine with the results from an ANN or to be presented to an ANN (as ANNs process numeric values). As well as the difficulty of performing syntactic analysis, some kind of semantic analysis would also need to be performed to extract the relevant constituents. This is even more difficult. Still another problem is how to combine the results of the syntactic and semantic analyses which are not numeric, with the results from an ANN. Thus, in utilising the sense definition, stemming analysis instead of syntactic structure analysis was chosen. The purpose of stemming analysis is to identify when two words have the same root.

In order to test the applicability of stemming analysis of the word sense definitions in WordNet, the following steps were carried out:

1. Construct a relation word set (RWS) for each key sense of the seed word by combining all the synonyms, hypernyms, hyponyms, holonyms, meronyms and coordinatees of the sense. Merge the RWSs for different key senses of the seed word into a single RWS for the seed word by unioning them. Antonyms are not included in the RWS because they represent the opposite meaning of the seed sense. Also not included are coordinates, because this relationship is too loose, introducing many irrelevant nouns to the list. A RWS for the seed word was created for finding stemming information instead of just using the seed word itself because the seed

word itself is normally too narrow to provide enough stemming information.

2. For a sense of a noun in the data set, extract all nouns in the sense definition to form a noun set (NS).

3. Generate stemming information of all noun senses in the NS against all the words in the RWS of the seed word. The highest value of common stem is taken as the result for a given noun sense.

4. When the stemming information is to be combined with the ANN method, information at word level rather than sense level must be used. The final stage converts the sense-level stemming information into word-level stemming information by selecting the highest similarity from all the senses of the word and taking this as the similarity of the word.

5.2.2 Stemming Algorithms

There are four popular automatic approaches to stemming, namely affix removal, n-gram, table lookup and successor variety[76]. Table lookup is not an option because of the time necessary to build a comprehensive stem dictionary. Experiments using n-gram stemming produced high associations for nearly all keywords, but it also leads to a high level of false associations for non-key senses, for example, "profession" and "impression" were associated with similarity of 70%, even though they are not from the same stem. The main reason that nouns are associated spuriously is that they have the same suffix which contributes much to the similarity measure used in the method. Therefore suffix removal has been combined with the n-gram method to keep the associations for key senses while decreasing the spurious associations of non-key senses.

The similarity of all nouns in the NS to all nouns in the RWS are calculated and the highest similarly value is chosen as the relationship between the noun sense and the seed word sense. The result is a number between 0 and 1, which can be considered as the confidence of the sense being a key sense.

5.2.3 Results

5.2.3.1 Sense Level Stemming Analysis Results

The results of sense-level stemming analysis (steps 1 to 3 in 5.2.1) is shown in table 5.1. There is no training as such in obtaining the stemming information; neither set has been used to develop the algorithm. In the table, the column "Correct Number" means the number of items that are correctly classified by the stemming analysis. That is to say, for a key sense, there is a high similarity; for a non-key sense, there is a low similarity.

The table shows good performance on keywords, but the performances on non-key senses are not as good as the previous experiments. The motivation of combining the good performance on non-key senses of the previous experiments and the good performance for key senses of stemming analysis leads to the proposition of the experiment described in next section which combines both approaches.

Table 5.1: Result of Sense-Level Stemming Analysis

Date Set 1				
Data Set	**Number**	**Correct Number**	**Percentage**	**Comparison with Chance**
Total	489	376	0.77	1.65
Keyword	27	24	0.89	
Non-keyword	462	352	0.76	
Data Set 2				
Data Set	**Number**	**Correct Number**	**Percentage**	**Comparison with Chance**
Total	1353	1069	0.79	1.62
Keyword	52	43	0.83	
Non-keyword	1301	1026	0.79	

Data set 1 and 2 are respectively the training and test sets used in the experiment in chapter 4. They are effectively both test sets for the stemming method but are shown separately in case the reader wishes to make a direct comparison with the results for the same data in the other experiments.

5.2.3.2 Word Level Stemming Analysis Results

After the stemming analysis at the sense level has been done, the result of the analysis can be converted to word-level by choosing the highest stemming similarity of all the senses of a word (step 4 in 5.2.1). The results are shown in table 5.2

Table 5.2: Word-Level Stemming Information Converted from Sense-Level Stemming Results

Data Set 1				
Data Set	**Number**	**Correct Number**	**Percentage**	**Comparison with Chance**
Total	111	67	0.60	1.43
Keyword	18	17	0.94	
Non-keyword	93	46	0.49	
Data Set 2				
Data Set	**Number**	**Correct Number**	**Percentage**	**Comparison with Chance**
Total	344	215	0.63	1.51
Keyword	37	34	0.92	
Non-keyword	307	181	0.59	

In this table it is noticed that the word-level stemming information is not as good as that of the sense-level in table 5.1. The reason is that selecting the highest sense-level similarity as the similarity of a word makes many non-keywords have a high similarity. These words are then mistaken as keywords and result in the word-level performance going down. These results imply that sense level data is more reliable than word level data. They support the experiments proposed in section 5.1 that were not possible because of the sparsity of connections in WordNet.

5.2.3.3 Results of Combining Stemming Analysis with the ANN Output

The formula $(A_p + S_p)/2$ is used to combine the ANN output A_p for a test pattern P with word level stemming analysis result S_p for the pattern. A threshold of 0.5 is then applied to obtain the decision regarding whether a word is considered a keyword. This simple calculation produces the result in Table 5.3.

Table 5.3: Comparison of combined stemming and ANN analogue with just ANN analogue

Average of stemming and ANN decisions						
Data Set	**NG**	**PG**	**Recall**	**Precision**	**CWC Natural**	**CWC Pure**
Total	0.71	0.69	0.66	0.70	1.48	1.26
Keywords	0.77	0.56	0.71	0.77		
Non Keywords	0.71	0.70	0.66	0.70		
Previous Results (ANN alone)						
Data Set	**NG**	**PG**	**Recall**	**Precision**	**CWC Natural**	**CWC Pure**
Total	0.84	0.82	0.81	0.86	1.49	1.30
Keywords	0.62	0.47	0.59	0.63		
Non Keywords	0.87	0.83	0.84	0.88		

Although the total PG of 0.82 was high for the previous results, the PG for keywords was poor at 0.47. As identification of keywords is the main purpose of this KA stage, the incorporation of stemming information represents an improvement (0.56) whilst maintaining a high rejection rate (0.70) for non-keywords. The same pattern is seen for all of the other measures (natural generalisation, recall and precision).

From tables 5.1, 5.2 and 5.3, it can be concluded that the ANN alone is best overall and for rejecting NKWs. Word level stemming analysis alone is the best for identifying KWs, while sense level analysis provides the most balanced results between KWs and NKWs. However, it should be noted that the ANN and stemming methods use different kinds of information. ANN uses word level path information while stemming analysis (both sense level and word level) use word definition information. Although both kinds of information are from the WordNet, there is no direct relation between them, as stated in section 5.1.3. Thus, the ANN results and the stemming analysis results are not directly comparable. Because word definition information works well in stemming analysis, it would be interesting to find out if it can be used in the ANN to improve the performance of ANNs.

Chapter 6

Domain Portability

A major concern for knowledge acquisition researchers is domain portability of the techniques developed. Although the techniques performed well on the education domain, their portability must be tested.

To test whether the approaches are portable across different domains, experiments were carried out on another domain: Law. The document is chosen from the encyclopaedia Britannica titled as "The Profession and Practice of Law"[35]. The seed word used is "law". The document can be found in appendix A. See appendix B for the nouns in the document.

6.1 Levels of Portability

Portability experiments were carried out on two levels. One is to test the generality of the ANN, i.e. is an ANN trained on one domain portable to another domain? The other is to test the generality of the methods, i.e. are the ANN and stemming analysis methods portable to other domains?

6.2 Generic ANN for All Domains

To test whether the ANN is generic, test data from a new domain is input into a trained ANN. Therefore, test data from the Law domain was input to the ANN trained on the Education domain. The results are shown in table 6.1. When compared with table 5.3, the results show measures for NKWs are only slightly lower than those from using the same domain in training and test. However, measures for KWs are not good: only 60% (0.28/0.47) of the training results are portable.

It is clear that an ANN trained on a specific domain is not in general sufficient to be used on another domain without further processing, i.e. the ANN trained on one domain is not portable to another domain.

Table 6.1: Results of running Law domain test data on an ANN trained using Education domain data

Data Set	NG	PG	Recall	Precision	CWC Natural	CWC Pure
Total	0.74	0.74	0.71	0.86	1.04	1.05
Keywords	0.27	0.28	0.24	0.26		
Non Keywords	0.77	0.77	0.74	0.79		

6.3 Portability of the ANN method

Data from the Law domain was divided into training and test sets as before. An ANN was trained in the same way. The maximum length of path M and number of shortest paths N are the same as with the education domain (4 and 5 respectively). However, the hidden neurons of the ANN for the new domain was 4. This was arrived at using the binary search method described previously in section 4.4.6. The results are in table 6.2.

Table 6.2: Results for the ANN approach in domain: Law

Data Set	NG	PG	Recall	Precision	CWC Natural	CWC Pure
Total	0.80	0.75	0.77	0.82	1.54	1.30
Keywords	0.74	0.54	0.66	0.71		
Non Keywords	0.80	0.76	0.78	0.83		

Comparing the results of the ANN from education and Law domains shown in table 4.5 and 6.2, it is clear that the performance of the ANN approach on the law domain is as good as on the education domain. Although the measures on NKWs are slightly worse than from the education domain, the measures on keywords are all better. For example, pure generalisation in the Law domain is 7% higher than in the education domain. The comparison with chance figures are virtually the same implying that the work done by the network is consistent and repeatable.

Thus, we can conclude that the technique using ANN to identify keywords is domain portable.

6.4 Portability of the Stemming Method

Word-level stemming analysis was carried out for the Law domain data, see table 6.3 for the results.

Table 6.3: Results of Word-level Stemming Analysis in the Domain: Law

Data Set	Number	Identified	Percentage	Comparison with Chance
Total	521	370	0.72	1.27
Keywords	41	19	0.54	
Non Keywords	351	351	0.73	

Comparing results of stemming analysis from the two domains (table 6.3 and table 5.2 respectively), it can be seen that the stemming analysis results on the law domain are not as good as in the education domain in terms of keyword identification, but it produced a reasonably good result for NKWs. The percentage of identified keywords is only 54%, much worse than that of the education domain, of 92%. The low rate of keyword identification does not necessarily mean a failure of stemming analysis. Higher rates of non-keyword identification can help to reject spurious keywords recognised by the ANN.

By comparing the keywords, shown in table 6.4, from both domains, it can be seen that legal terms are more specialised and therefore possibly not as well represented in WordNet which is intended as a general purpose lexicon. This may be the cause of the low keyword recognition rate. Conversely, educational terms are in more commonly use and are also applied outside the domain itself. Their representation and interconnectivity in WordNet is therefore richer.

These results appear to show that unlike the ANN method which shows a consistent pattern and level of performance in different domains, the stemming analysis method is less robust to domain change.

Table 6.4: Keywords from education and law domain

Keywords from education domain
Academic, academic-year, assessment, baccalaureate, campus, class, classroom, coaching, college, college-level, competence, comprehension, course, curriculum, degree, education, enrollment, exam, examination, grade, grading, graduate, graduation, higher-education, homework, instruction, instructor, knowledge, learner, learning, lecture, lecturer, lecturing, lesson, polytechnic, remediation, school, semester, student, study, studying, subject, syllabus, teach, teacher, teaching, term, textbook, training, tuition, tutor, tutoring, undergraduate, university
Keywords from law domain
accused, act, action, advocacy, advocate, appeal, appearance, assessor, attorney, bar, barrister, case, case-law, casebook, chancery, civil-law, client, commission, common-law, compensation, contract, conviction, counsel, counselor, court, crime, criminal, defendant, defense, disbarment, enactment, evidence, guilt, imprisonment, judge, judgeship, judgement, judiciary, jurisdiction, jurisprudence, jurist, jury, justice, law, lawyer, legality, legislation, legislature, litigant, litigation, magistracy, magistrate, notary, offense, precedent, pretrial, proceeding, procurator, prosecution, prosecutor, punishment, regulating, regulation, right, solicitation, solicitor, statute, testimony, trial, tribunal, witness

6.5 Portability of ANN and Stemming Analysis Combined

The results from the combination of ANN and stemming analysis methods in the new domain are shown in table 6.5. It can be seen that all the measures follow the same trends as in the Education domain, e.g. the pure generalisation for keywords increases by 11% compared with using the ANN alone.

Table 6.5: Results for ANN approach alone and when combined with Stemming Analysis for the domain: "law"

Results of the Combination						
Data Set	**Natural Generalisation**	**Pure Generalisation**	**Recall**	**Precision**	**Comparison with Chance**	
					Natural	**Pure**
Total	0.74	0.69	0.60	0.78	1.53	1.34
Keywords	0.80	0.65	0.68	0.84		
Non Keywords	0.73	0.69	0.59	0.78		
Results of Using ANN Only						
Data Set	**Natural Generalisation**	**Pure Generalisation**	**Recall**	**Precision**	**Comparison with Chance**	
					Natural	**Pure**
Total	0.80	0.75	0.77	0.82	1.54	1.3
Keywords	0.74	0.54	0.66	0.71		
Non Keywords	0.80	0.76	0.78	0.83		

Results from stemming analysis in the law domain are worse than in the education domain, but the results from the ANN are better. The overall performance (by combining the two approaches) in the new domain is better than in the education domain.

Chapter 7

Trained Weight Analysis

The ANN used for extracting keywords from text has been evaluated in chapters 4 and 5. Its portability to other domains has also been analysed in chapter 6. However, how the ANN learns to solve the keyword identification problem is still not clear. This chapter attempts to give the answer to this question based on the analysis of the weights of the trained ANN. The aim is to find out some rules understandable to humans, rather than leaving the trained ANN as a black box.

The analysis is based on how the ANN learns the problem, how it forms rules and how it uses the rules in differentiating KWs and NKWs. The ANN generates rules based on the category and path information in the training set. Rules contain key features and are triggered by these key features when applied to the test set. There should be two kinds of rules: one for identifying KWs and the other for NKWs. A rule that is mainly for identifying KW may draw false examples i.e. it identifies NKWs as KWs. From a human perspective, this rule does not seem helpful. However, for the ANN, there is no fault with this because the KW features are in these NKW examples. This rule needs to be taken in the context of all the other rules the ANN is applying. Each rule can be learnt as a rule of thumb and the network's decision is the combined effect. All evaluation on rules is made from the perspective of their use to the ANN.

In the trained ANN, there are many rules with different importance in solving the problem of KW extraction. In this analysis, only the important ones will be explained.

7.1 Introduction

There are two hidden nodes in the trained ANN. However, it is noticed that the second of the two is much more important than the first. The weights from the input layer to the first hidden node are always small, unlike from those to the second hidden node. Thus, investigation was done to find out what are their exact roles in the ANN. By calculating the final output of the ANN without hidden neuron 1, it is found that contribution from this hidden node to the final output is small. However, this hidden node is clearly necessary in some small aspect of the problem as training without it consistently failed. The trained weight analysis will focus on the weights from inputs to the second hidden node.

It is hoped that by separately analysing the weights from the input nodes representing relations, paths and categories, it is possible to know how the ANN uses these different types of information in differentiating between KWs and NKWs.

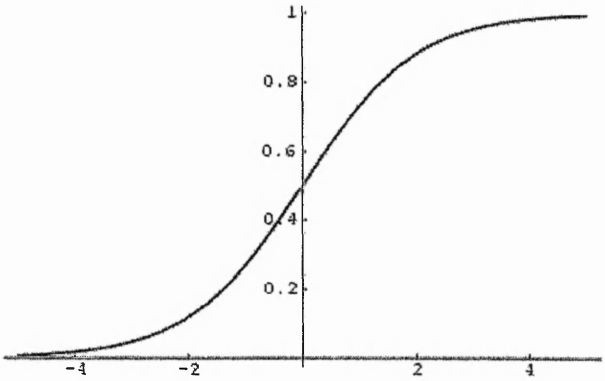


Figure 7.1: Curve of Sigmoid Function

The Sigmoid function is used in calculating outputs from hidden nodes and output node of the ANN. From the curve of the function shown in figure 7.1, it is clear that the value of the function is always positive and the larger the input to it, the larger the output value. 0.5 is used as the decision point on the output neuron, i.e. a pattern will be recognised as a key if the final output from this is greater than or equal to 0.5.

In the trained ANN, the weights from hidden layer to the output layer are -4.76 (from bias node), 11.2 (from hidden node 1) and 10.6 (from hidden node 2). Because the weights from the two hidden nodes to the output node are both positive, larger outputs from the hidden nodes will produce a larger final output which tends to make the input pattern be identified as a key pattern. A larger output from a hidden node depends on the inputs, either 1 or 0, to the input nodes of the ANN. Thus a positive weight between an input node and a hidden node will make the input to the hidden node bigger which in turn tends to produce a bigger output from the hidden nodes. This bigger value tends to produce a key, as mentioned. Thus, a positive weight encourages a pattern to be classified as a key, while a negative one tends to encourage classification as a NKW.

7.2 Path Information

7.2.1 Introduction

In order to allow the network to recognise path features regardless of the order in which they were found when WordNet was searched, 120 (N=5) path order permutations are generated for each noun in the training set (see section 4.4.3). Figure 7.2 shows the curves of the trained weights corresponding to the five paths. In this figure, weight position 0 to 79 denote path 1, 80 to 159 to path 2 and so forth. The curves for the five paths reveal the same pattern, with very small absolute value changes. This shows that the multiple permutation training worked so that as expected, paths found in any order have the same effect on the output. Because of the similar distribution of the weights for the five paths, the weights for one path can represent all paths for the purpose of analysing their effects.

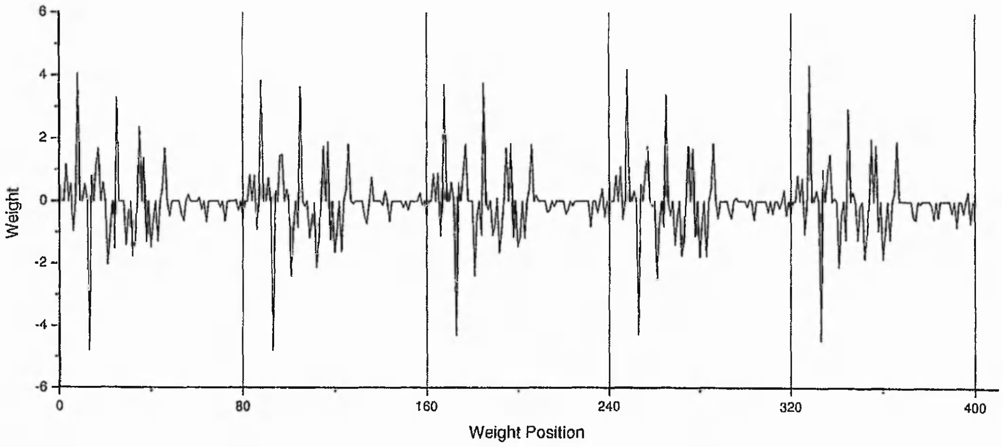


Figure 7.2: Distributions of Trained Weights from Input to Hidden Nodes for All Five Paths

In this chapter, weights corresponding to the first path are used to demonstrate how rules can be extracted at path level from trained weights. However, the final judgement of whether a noun is a KW or not must be based on all paths plus the category information.

Table 7.1 list the weights from input nodes of the first path (80 bits, see section 4.4.4) to the second hidden node.

A Key Path (KP) is defined as a path whose contribution to a hidden node is positive. It is defined so because a positive input value to a Sigmoid function will produce an output value larger than 0.5 and if the contributions from the other 4 paths are zero, the noun will be classified as a KW. A KP cannot guarantee a noun to be identified as a KW because negative contributions from other paths may be great enough to lead to a final output being below 0.5. If all the paths of a noun are KPs, this noun is very likely to be a keyword although responses from the category input nodes must be considered as well. However, it is not necessarily for all five paths of a keyword to be KPs. More detailed discussion about the relation between KP and KWs will be presented in the next section.

An Important Path (IP) is defined as a path which very strongly supports the overall ANN decision such that it could be considered isolation, i.e. it concurs with the ANN decision for greater than 50% (for KPs, 75% for non-KPs) of the cases where this path occurs. An IP may be positive or negative. A positive IP (PIP) is an indicator of a KW while a negative IP (NIP) is an indicator of NKW. In identifying rules from the trained ANN, PIPs and NIPs are treated differently in terms of the number of nouns containing the PIP/NIP. This is because of the ratio of KWs to NKWs. As there are many more NKWs in both training set and test set, a NIP supported by a small number

Table 7.1: Weights from Input Nodes to Hidden Node 2

Weight Position	Relation Name	Weight	Number of Occurrence in Training Set	Length Path
0	Holonym	0.042039	0	1
1	Meronym	0.416257	0	
2	Hyponym	0.861905	2	
3	Hypernym	-0.106342	0	
4	Coordinatee	0.586079	1	
5	Coordinate	-0.913129	11	
6	Antenym	-0.419279	0	
7	Synonym	3.798928	7	
8	Holonym	0.376858	0	2
9	Meronym	0.222587	0	
10	Hyponym	0.288554	6	
11	Hypernym	0.407498	0	
12	Coordinatee	-4.490516	40	
13	Coordinate	0.472735	160	
14	Antenym	0.410825	0	
15	Synonym	0.891865	77	
16	Holonym	1.902968	2	
17	Meronym	0.214499	0	
18	Hyponym	0.183683	33	
19	Hypernym	-0.129597	0	
20	Coordinatee	-2.296845	18	
21	Coordinate	-0.685661	175	
22	Antenym	0.267998	0	
23	Synonym	-1.054257	55	
24	Holonym	3.401821	1	3
25	Meronym	-0.145222	0	
26	Hyponym	-0.107417	5	
27	Hypernym	-0.105914	0	
28	Coordinatee	-1.019322	26	
29	Coordinate	-0.630883	72	
30	Antenym	-0.229679	1	
31	Synonym	-2.087179	121	
32	Holonym	-0.756108	1	
33	Meronym	0.186361	0	
34	Hyponym	2.06014	9	
35	Hypernym	0.076617	0	
36	Coordinatee	1.622537	13	
37	Coordinate	-0.835712	157	
38	Antenym	-0.36346	0	
39	Synonym	-1.962879	46	

Table 7.1: Weights from Input Nodes to Hidden Node 2 (Continued)

Weight Position	Relation Name	Weight	Number of Occurrence in Training Set	Length Path
40	Holonym	-0.953957	4	
41	Meronym	0.457183	0	
42	Hyponym	-1.581043	34	
43	Hypernym	-0.372677	0	
44	Coordinatee	0.14153	22	
45	Coordinate	1.312237	125	
46	Antenym	0.142384	0	
47	Synonym	-0.551402	41	
48	Holonym	0.243126	0	4
49	Meronym	-0.254784	0	
50	Hyponym	0.268639	0	
51	Hypernym	-0.007981	0	
52	Coordinatee	-0.043361	2	
53	Coordinate	-0.260528	5	
54	Antenym	-0.085803	0	
55	Synonym	0.687955	18	
56	Holonym	-0.345393	0	
57	Meronym	-0.416379	0	
58	Hyponym	-0.315149	0	
59	Hypernym	0.47766	0	
60	Coordinatee	-0.333422	9	
61	Coordinate	-0.512647	8	
62	Antenym	0.469756	0	
63	Synonym	-0.81137	8	
64	Holonym	-0.020432	0	
65	Meronym	-0.414365	0	
66	Hyponym	0.054552	0	
67	Hypernym	-0.122394	0	
68	Coordinatee	-0.035783	0	
69	Coordinate	-0.050881	22	
70	Antenym	-0.140065	0	
71	Synonym	-1.002031	3	
72	Holonym	0.468657	0	
73	Meronym	0.069445	0	
74	Hyponym	-0.18144	3	
75	Hypernym	0.096606	0	
76	Coordinatee	0.351181	1	
77	Coordinate	-0.361985	19	
78	Antenym	0.293939	0	
79	Synonym	-0.858105	2	

of examples is considered as not reliable. This is different for PIP. KWs are distinguished from the mass of NKWs with their special features that may be exclusive to a single keyword. So KPs producing good test performance, even though with a small number of supporting examples, are still treated as a PIP.

7.2.2 Relation level

The objective of this section is to find out relative importance of the eight relations. A Major Weight (MW) is a weight which is the most different from its neighbour weights. In identifying important relations, two aspects need to be considered. The first is to identify the MWs. The second is the location of the MWs. This is important because the position of a MW determines its meaning. For example, a MW in the place of the synonym bit in the first relation and another MW on the coordinate bit in the second relation of a path of length of 2 means this path is a KP (supposing both of the MWs are positive). The combination of the two relations means that a path composed of a coordinate of a noun which is a synonym of the seed word is a KP. However, if the two locations of MWs are swapped, i.e. the first representing a coordinate and the second a synonym, it means that a synonym of a noun which is a coordinate of the seed word is a KP. Obviously, the two results are quite different.

Apart from MWs (≥ 1.5 for positive weights and ≤ -1.5 for negative weights), there are also two different levels of weight- reasonably strong weights (RSWs) (≥ 0.8 for positive weights and ≤ -0.8 for negative weights) and weak weights (WWs). In the analysis below, MWs and RSWs are labeled with **, * and referred to as "strong indicators" and "indicators" respectively.

For each noun, 5 paths of length up to 4 are used. The length of a path is the same as the number of relations in the path. According to the input data scheme (see section 4.4.4), each path is represented by 80 bits. Within the 80 bits, there are 10 bits associated with every relation type. In this section, weights for each relation will be draw together in a table for the sake of convenience of discussion. The 10 weights form 4 groups for the four possible path length: row 1 is for a path of length 1; rows 2 and 3 for path of length 2; rows 4 to 6 for length 3 and rows 7 to 10 for length 4. The location column indicates the places of the weight in the 80 bits. Weights with ** are identified as MWs. Weights with * is not as important as a MW, but are still quite important and cannot be ignored. Combined with other relations, these weights may have a critical effect.

7.2.2.1 Holonym relations

Table 7.2 shows the weights connecting all the holonym bits in an input pattern to the second hidden node.

Weights holonym22 and holonym31 are identified as MWs. Holonym33 is identified as RSWs.

Holonym22 locates in the second relation of a path of length 2. Its positive value indicates that a path of length 2 with holonym appearing as the second relation of the path is a KP if the weight from the first relation of this path is positive or the weight is negative but its absolute value is

Table 7.2: Weights from Holonym Relations to Hidden Node 2

Name	Location	Path1	Length
holonym11	0	0.04	1
holonym21	8	0.38	2
holonym22**	16	1.90	
holonym31**	24	3.40	3
holonym32	32	-0.76	
holonym33*	40	-0.95	
holonym41	48	0.24	4
holonym42	56	-0.35	
holonym43	64	-0.02	
holonym44	72	0.47	

smaller than Holonym22.

Patterns drawn from Holonym relations are shown in table 7.3, where * means any of the eight relations and the plus sign (+) means this is a positive indicator.

Table 7.3: Holonym Relation Patterns

Patterns			Meaning
*	+Holonym		a path of length 2 with a holonym as the second relation of the path is a strong indicator of a KP
+Holonym	*	*	A path of length 3 with a holonym as the first relation of the path is a strong indicator of a KP
*	*	-Holonym	A path of length 3 with a holonym as the third relation of the path is an indicator of a NKP

Although MWs in one relation and KPs are strong indicators, analysis of single relations is not enough to reveal the whole mystery of the ANN. Various relations must be considered together. Similar analysis has been done for the other relations. This is shown in appendix E. The conclusions about relations are in section 7.2.2.2.

7.2.2.2 Conclusion

Because there is no relation data in the training set for meronym, antonym and hypernym relations, the weights for these relations do not change in the training process of the ANN. This means that no rules can be extracted for these relations. Meronym is a sparse relation because it is not fully implemented in WordNet. Antonym is mainly implemented for adjectives and adverbs. It is very rare for nouns. Thus, it is reasonable that there are no meronym and antonym relations in the training set.

The reason for the lack of hypernym relation in the training set relates to the order of relations in which the algorithm searches WordNet. Hypernym is the last relation in the order. Before it is used, enough paths have already been found because synonym, hyponym, coordinate and coordinatee are all very rich in WordNet. Thus, the reason that there are no hypernym relations in the training set is not because of sparsity of the relation. To verify this explanation, an experiment was done. In the experiment, the hypernym relation is put first in the search list. It is found that nearly all of the paths contain hypernym relations. No hypernym relations in the training data does not necessary mean the total loss of hypernym information provided by WordNet, because the coordinate relation has the hypernym relation as its constituent.

From the analysis of the relation weights, it can be concluded that synonym, hyponym, holonym, and coordinatee relations are more important than other relations. Coordinate, from which no MW can be identified, is considered not as important as synonym, hyponym, holonym, and coordinatee relations. Theoretically, meronym (a word that names a part of a given word) would be an important relation. Unfortunately, there are no relations of this type in the training set because of the implementation of WordNet. It is therefore impossible to reach any conclusion regarding this relation.

7.2.3 Path Level

The order of the relations in a path is important when considering path level patterns. Table 7.4 shows all the paths in the training data. With five paths for each noun, there are 555 paths altogether for the 111 nouns in the training set. However, there are only 72 exclusive paths. In table 7.4, Path Relations are the relations comprising a path, with the exact order shown in the table. Number is the number of times a path occurs in the training set. No KP is how many of the paths are for a KW, and No NKP is the number of the paths for NKWs. Path Weight is the sum of all the weights on this path. Take the 12th row in the table as an example. This is a path of length 2, comprised of the relations "coordinate" and "hyponym". There are 28 occurrences of this path in the training set, 8 of them for keywords and 20 for non-keywords. The input from this path to the hidden node is 0.656418.

An interesting learning rule of the ANN can be found from table 7.4. In the training data, assume there are k occurrences of a path, there m of them are for KWs and n for NKWs. The rule is that if $m > n$, this path is learned as a KP; otherwise, it is learned as a NKP. This learning rule seems reasonable intuitively. Thus it is called "learn by intuition" (LBI) rule here. The ANN also learns categories by this rule. The ANN learned most of the paths in the training set by this rule, with several exceptions.

The following sections compare the path level decision with the overall ANN i.e. word level decisions in order to determine the most important rules.

Table 7.4: Constituents and frequency all paths in the training set

Row No	Length	Path Relations	Total Number	No KP	No NKP	Path Weight
1	1	(Synonym)	7	7	0	3.79893
2	1	(Coordinate)	11	0	11	-0.913129
3	1	(Coordinatee)	1	1	0	0.586079
4	1	(Hyponym)	2	2	0	0.861905
5	2	(Synonym, Synonym)	17	7	10	-0.162392
6	2	(Synonym, Coordinate)	50	1	49	0.206204
7	2	(Synonym, Coordinatee)	5	1	4	-1.40498
8	2	(Synonym, Hyponym)	5	4	1	1.07555
9	2	(Coordinate, Synonym)	31	5	26	-0.581522
10	2	(Coordinate, Coordinate)	96	22	74	-0.212926
11	2	(Coordinate, Coordinatee)	4	1	3	-1.82411
12	2	(Coordinate, Hyponym)	28	8	20	0.656418
13	2	(Coordinate, Holonym)	1	1	0	2.3757
14	2	(Coordinatee, Synonym)	7	0	7	-5.54477
15	2	(Coordinatee, Coordinate)	24	0	24	-5.17618
16	2	(Coordinatee, Coordinatee)	9	0	9	-6.78736
17	2	(Hyponym, Coordinate)	5	0	5	-0.397107
18	2	(Hyponym, Holonym)	1	1	0	2.19152
19	3	(Antonym, Coordinate, Coordinate)	1	0	1	0.246846
20	3	(Synonym, Synonym, Coordinate)	17	2	15	-2.73782
21	3	(Synonym, Synonym, Synonym)	8	0	8	-4.60146
22	3	(Synonym, Hyponym, Coordinate)	2	0	2	1.2852
23	3	(Synonym, Coordinatee, Coordinate)	4	1	3	0.847595
24	3	(Hyponym, Coordinate, Coordinate)	2	0	2	0.369108
25	3	(Synonym, Coordinatee, Synonym)	1	1	0	-1.01604
26	3	(Holonym, Coordinate, Coordinate)	1	1	0	3.87835
27	3	(Coordinatee, Coordinate, Coordinatee)	5	2	3	-1.7135
28	3	(Coordinate, Synonym, Synonym)	3	0	3	-3.14516
29	3	(Coordinate, Coordinatee, Coordinatee)	1	0	1	1.13318
30	3	(Synonym, Coordinate, Synonym)	14	2	12	-3.47429
31	3	(Synonym, Coordinate, Coordinate)	43	2	41	-1.61065
32	3	(Coordinate, Synonym, Hyponym)	1	0	1	-4.1748
33	3	(Hyponym, Coordinate, Hyponym)	1	0	1	-2.52417
34	3	(Coordinatee, Coordinate, Coordinate)	5	0	5	-0.542797
35	3	(Coordinatee, Coordinate, Synonym)	7	0	7	-2.40644
36	3	(Synonym, Synonym, Hyponym)	1	0	1	-5.6311

Table 7.4: Constituents and frequency all paths in the training set (Continued)

Row No	Length	Path Relations	Total Number	No KP	No NKP	Path Weight
37	3	(Synonym, Coordinate, Holonym)	1	0	1	-3.87685
38	3	(Coordinatee, Coordinatee, Coordinate)	2	0	2	1.91545
39	3	(Synonym, Synonym, Coordinatee)	4	0	4	-3.90853
40	3	(Coordinate, Coordinate, Synonym)	4	0	4	-2.018
41	3	(Coordinate, Hyponym, Hyponym)	1	1	0	-0.151786
42	3	(Coordinatee, Hyponym, Coordinate)	1	1	0	2.35306
43	3	(Coordinate, Coordinate, Coordinate)	30	7	23	-0.154358
44	3	(Synonym, Coordinate, Coordinatee)	10	2	8	-2.78136
45	3	(Synonym, Hyponym, Synonym)	1	0	1	-0.578441
46	3	(Hyponym, Hyponym, Holonym)	1	1	0	0.998766
47	3	(Hyponym, Coordinate, Synonym)	1	0	1	-1.49453
48	3	(Coordinate, Coordinate, Hyponym)	16	0	16	-3.04764
49	3	(Synonym, Coordinate, Hyponym)	13	1	12	-4.50393
50	3	(Coordinate, Synonym, Coordinate)	7	1	6	-1.28152
51	3	(Coordinatee, Synonym, Coordinate)	5	0	5	-1.66996
52	3	(Coordinate, Coordinatee, Coordinate)	3	0	3	2.30389
53	3	(Coordinate, Coordinate, Holonym)	2	0	2	-2.42055
54	3	(Synonym, Coordinatee, Coordinatee)	1	0	1	-0.323112
55	3	(Coordinate, Hyponym, Synonym)	1	0	1	0.877855
56	3	(Synonym, Hyponym, Coordinatee)	1	0	1	0.114491
57	3	(Coordinate, Coordinatee, Synonym)	1	0	1	0.440252
58	3	(Coordinate, Hyponym, Coordinate)	1	0	1	2.74149
59	3	(Coordinate, Holonym, Coordinate)	1	0	1	-0.074754
60	3	(Coordinatee, Coordinate, Hyponym)	1	0	1	-3.43608
61	4	(Synonym, Synonym, Coordinate, Coordinate)	4	0	4	-0.536281
62	4	(Coordinate, Coordinatee, Coordinate, Synonym)	2	0	2	-1.50294
63	4	(Coordinate, Coordinate, Coordinate, Hyponym)	1	0	1	-1.0055
64	4	(Synonym, Synonym, Coordinate, Coordinatee)	1	1	0	0.176885
65	4	(Synonym, Coordinate, Synonym, Coordinate)	1	1	0	-1.18871
66	4	(Synonym, Coordinatee, Coordinate, Hyponym)	2	1	1	0.122212
67	4	(Synonym, Coordinatee, Coordinate, Coordinate)	3	3	0	-0.058333
68	4	(Synonym, Coordinate, Coordinate, Coordinate)	5	2	3	-0.237558
69	4	(Coordinatee, Coordinate, Coordinate, Coordinate)	1	0	1	-0.968874
70	4	(Coordinatee, Synonym, Coordinate, Coordinate)	1	0	1	-1.2676
71	4	(Coordinate, Synonym, Coordinate, Coordinate)	2	1	1	-1.48476
72	4	(Synonym, Coordinatee, Synonym, Coordinate)	2	0	2	-1.00948

7.2.3.1 Paths of Length 1

The training and test results involving length 1 paths are shown in table 7.5.

Table 7.5: Training and Test Results for Length 1 Paths

Path Relations	Training Set			Path	NP	Test Set			Concurrence
	Total No	No KPs	No NKPs	Weight	or NKP	Total No	No KPs	No NKPs	with overall ANN decision
(Synonym)	7	7	0	3.79893	KP	12	12	0	100%
(Hyponym)	2	2	0	0.861905	KP	8	7	1	88%
(Coordinatee)	1	1	0	0.586079	KP	4	3	1	75%
(Coordinate)	11	0	11	-0.913129	NKP	29	7	22	76%

In this table, "Total No" is the occurrences of a path in the training or test set. "No KPs" and "No NKPs" are the number of words that contain this path that are identified as KWs and NKWs by the ANN respectively. "Concurrence with overall ANN decision" is the percentage of the words that contain this path where the path level response gives the same KW/NKW classification as is given by the ANN. Note that the percentage is based on how the ANN learns, not on the basis of how humans classify the nouns. This column is calculated just on the path level not word level which is considered later. For example, "lecture" is a KW as judged by humans, which contains a path of "coordinatee". The ANN recognised it correctly. "Cognition" also contains this path, but it is not a keyword judged by human. However, as the path level decision is the same as the overall ANN decision, it contributes to the concurrence percentage. The concurrence percentage therefore indicates how strongly the path level result correlates with or contributes to the overall ANN performance.

Table 7.6: Patterns Extracted for Path of Length 1

Important Path	Type	Meaning
(Synonym)	PIP	a path composed of a synonym of the seed word is an IP
(Hyponym)	PIP	a path composed of a hyponym of the seed word is an IP
(Coordinatee)	PIP	a path composed of a coordinatee of the seed word is an IP
(Coordinate)	NIP	a path composed of a coordinate of the seed word is an IP

LBI is true for all the 4 paths of length 1. For the positive patterns, all nouns which are synonyms of the seed word in the test set are identified as keywords. For hyponym, there are 8 nouns in the test set that have a path composed of hyponym, 7 of them are identified as KWs. For coordinatee, 3 out of 4 nouns involved in this path are classified as KWs. As for coordinate, 29 nouns in test set have this kind of path, 22 of them are considered as NKWs by the ANN. From the results, it can be concluded that these rules work very well towards the ANN differentiation between KWs and NKWs.

Thus, 4 rules concerning path of length 1 can be extracted from the ANN. They are shown in table 7.6.

7.2.3.2 Paths of Length 2

Training and test results for length 2 paths are shown in table 7.7. Rows with an asterisk in the Path Relations column mean that the path is not learned by LBI. 12 out of 14 are learned under the LBI rule. There were two exceptions. The ANN classifies 5 of two-leg paths as KPs and 9 as NKPs. 12 rules concerning paths of length 2 can be extracted from the ANN. They are shown in table 7.8.

Table 7.7: Training and Test Results for Length 2 Paths

Path Relations	Training Set			Path	NP	Test Set			Con
	Total No	No KPs	No NKPs	Weight	or NKP	Total No	No KPs	No NKPs	
(Coordinate, Holonym)	1	1	0	2.3757	KP	6	3	3	50%
(Hyponym, Holonym)	1	1	0	2.19152	KP	2	2	0	100%
(Synonym, Hyponym)	5	4	1	1.07555	KP	16	12	4	75%
*(Coordinate, Hyponym)	28	8	20	0.656418	KP	52	12	40	23%
*(Synonym, Coordinate)	50	1	49	0.206204	KP	101	15	86	15%
(Coordinatee, Coordinatee)	9	0	9	-6.78736	NKP	26	0	26	100%
(Coordinatee, Synonym)	7	0	7	-5.54477	NKP	22	1	21	95%
(Coordinatee, Coordinate)	24	0	24	-5.17618	NKP	54	1	53	98%
(Coordinate, Coordinatee)	4	1	3	-1.82411	NKP	16	2	14	88%
(Synonym Coordinatee)	5	1	4	-1.40498	NKP	21	4	19	81%
(Coordinate, Synonym)	31	5	26	-0.581522	NKP	54	6	48	89%
(Hyponym, Coordinate)	5	0	5	-0.397107	NKP	15	0	15	100%
(Coordinate, Coordinate)	96	22	74	-0.212926	NKP	152	28	124	83%
(Synonym, Synonym)	17	7	10	-0.162392	NKP	42	11	31	74%

Con: Concurrence with overall ANN decision

These rules work well towards the ANN differentiation between KWs and NKWs, with a minimum concurrence percentage of 50%. Most of them produce more than 74% concurrence.

The two paths that are not learned by the LBI rule can not be treated as a rule because of their performance in testing, with only 23% and 15% concurrence. These two paths are not critical to the nouns containing them. Other paths or combination must be used to determine whether these nouns are keywords or not.

What is the reason for the two exceptions to LBI? Taking (synonym, coordinate) as an example, most of the nouns that have this type of path have a strong negative category response, or strong negative path responses from other paths. Therefore the incorrect classification of this path is not very important for these nouns to be identified as NKWs, because the response from this path i.e. (synonym, soordinate) is small (only 0.21). However, the only keyword ("class") having this path has a strong negative response. The responses from the paths are not strong enough to counteract the negative category response. Thus, this path is quite important for this word to be classified as a KW. The result is the ANN assigning it a positive weight, so that it can classify "class" correctly.

Table 7.8: Patterns Extracted for Path of Length 2

Important Path	Type	Meaning
(Coordinate, Holonym)	PIP	This path is a strong indicator of a keyword
(Hyponym, Holonym)	PIP	This path is a strong indicator of a keyword
(Synonym, Hyponym)	PIP	This path is a strong indicator of a keyword
(Coordinatee, Coordinatee)	NIP	This path is a strong indicator of a non-keyword
(Coordinatee, Synonym)	NIP	This path is a strong indicator of a non-keyword
(Coordinatee, Coordinate)	NIP	This path is a strong indicator of a non-keyword
(Coordinate, Coordinatee)	NIP	This path is a strong indicator of a non-keyword
(Synonym, Coordinatee)	NIP	This path is a strong indicator of a non-keyword
(Coordinate, Synonym)	NIP	This path is a strong indicator of a non-keyword
(Hyponym, Coordinate)	NIP	This path is a strong indicator of a non-keyword
(Coordinate, Coordinate)	NIP	This path is a strong indicator of a non-keyword
(Synonym, Synonym)	NIP	This path is a strong indicator of a non-keyword

Note that path (synonym, synonym) is not a PIP. This is important because it is against intuition. First, that is how the ANN is trained. For 17 of (synonym, synonym) paths, 10 are NKWs. That is to say, some words containing this path in the training set are not identified as KWs by humans. The reason that these words are not identified as keywords by human is that the human analysis did not take phrase into account whereas WordNet does model some phrases. For example, the noun “action” is considered by humans as a NKW in the education domain under the definition in chapter 4. This is correct. However, “educational activity” is a synonym of education. The path search algorithm does not take phrases like “educational activity” into account. It takes the nouns in the WordNet synset of the seed word as synonyms. Thus, “activity” is considered by the search algorithm as a synonym. As “action” is a synonym of “activity”, leading to “action” the resulting path is (synonym, synonym). This results in the above phenomenon: paths of (synonym, synonym) which are not KPs. Analysis of phrases is suggested as further research.

7.2.3.3 Paths of Length 3

Training and test results for length 3 paths are shown in table 7.9. For length 3 paths, 29 out of 42 are learned by LBI rule.

Rules extracted regarding paths of length 3 are shown in table 7.10.

7.2.3.4 Paths of Length 4

Training and test results for length 4 paths are shown in table 7.11. Rules extracted regarding paths of length 4 are shown in table 7.12.

For the 12 length 4 paths, 2 of them are not learned by the LBI rule. The two KPs in this group are

Table 7.9: Training and Test Results for Length 3 Paths

Path Relations	Training Set			Path	NP	Test Set			Con
	T No	No KPs	No NKPs	Weight	or NKP	T No	No KPs	No NKPs	
(Holonym, Coordinate, Coordinate)	1	1	0	3.87835	KP	2	1	1	50%
*(Coordinate, Hyponym, Coordinate)	1	0	1	2.74149	KP	6	3	3	50%
(Coordinatee, Hyponym, Coordinate)	1	1	0	2.35306	KP	1	1	0	100%
*(Coordinate, Coordinatee, Coordinate)	3	0	3	2.30389	KP	8	1	8	13%
*(Coordinatee, Coordinatee, Coordinate)	2	0	2	1.91545	KP	7	1	6	14%
*(Synonym, Hyponym, Coordinate)	2	0	2	1.2852	KP	5	1	4	20%
*(Coordinate, Coordinatee, Coordinatee)	1	0	1	1.13318	KP	0	0	0	N/A
(Hyponym, Hyponym, Holonym)	1	1	0	0.998766	KP	1	1	0	100%
*(Coordinate, Hyponym, Synonym)	1	0	1	0.877855	KP	1	0	1	0%
*(Synonym, Coordinatee, Coordinate)	4	1	3	0.847595	KP	9	1	8	11%
*(Coordinate, Coordinatee, Synonym)	1	0	1	0.440252	KP	2	0	2	0%
*(Hyponym, Coordinate, Coordinate)	2	0	2	0.369108	KP	4	1	3	25%
*(Antonym, Coordinate, Coordinate)	1	0	1	0.246846	KP	5	0	5	0%
*(Synonym, Hyponym, Coordinatee)	1	0	1	0.114491	KP	5	3	2	60%
(Synonym, Synonym, Hyponym)	1	0	1	-5.6311	NKP	9	0	9	100%
(Synonym, Synonym, Synonym)	8	0	8	-4.60146	NKP	14	0	14	100%
(Synonym, Coordinate, Hyponym)	13	1	12	-4.50393	NKP	39	3	36	92%
(Coordinate, Synonym, Hyponym)	1	0	1	-4.1748	NKP	7	0	7	100%
(Synonym, Synonym, Coordinatee)	4	0	4	-3.90853	NKP	10	0	10	100%
(Synonym, Coordinate, Holonym)	1	0	1	-3.87685	NKP	1	0	1	100%
(Synonym, Coordinate, Synonym)	14	2	12	-3.47429	NKP	47	5	42	89%
(Coordinatee, Coordinate, Hyponym)	1	0	1	-3.43608	NKP	7	0	7	100%
(Coordinate, Synonym, Synonym)	3	0	3	-3.14516	NKP	3	0	3	100%
(Coordinate, Coordinate, Hyponym)	16	0	16	-3.04764	NKP	40	2	38	95%
(Synonym, Coordinate, Coordinatee)	10	2	8	-2.78136	NKP	23	2	21	91%
(Synonym, Synonym, Coordinate)	17	2	15	-2.73782	NKP	67	5	62	93%
(Hyponym, Coordinate, Hyponym)	1	0	1	-2.52417	NKP	2	0	2	100%
(Coordinate, Coordinate, Holonym)	2	0	2	-2.42055	NKP	2	2	0	0%
(Coordinatee, Coordinate, Synonym)	7	0	7	-2.40644	NKP	11	1	10	90%
(Coordinate, Coordinate, Synonym)	4	0	4	-2.018	NKP	15	0	15	100%
(Coordinatee, Coordinate, Coordinatee)	5	2	3	-1.7135	NKP	7	4	3	43%
(Coordinatee, Synonym, Coordinate)	5	0	5	-1.66996	NKP	10	0	10	100%
(Synonym, Coordinate, Coordinate)	43	2	41	-1.61065	NKP	112	15	97	87%
(Hyponym, Coordinate, Synonym)	1	0	1	-1.49453	NKP	2	0	2	100%
(Coordinate, Synonym, Coordinate)	7	1	6	-1.28152	NKP	26	4	22	85%

T No: Total Number in training/test set

Con: Concurrence with overall ANN decision

Table 7.9: Training and Test Results for Length 3 Paths (Continued)

Path Relations	Training Set			Path	NP	Test Set			Con
	T No	No KPs	No NKPs	Weight	or NKP	T No	No KPs	No NKPs	
*(Synonym, Coordinatee, Synonym)	1	1	0	-1.01604	NKP	2	1	1	50%
(Synonym, Hyponym, Synonym)	1	0	1	-0.578441	NKP	1	1	0	0%
(Coordinatee, Coordinate, Coordinate)	5	0	5	-0.542797	NKP	20	6	14	70%
(Synonym, Coordinatee, Coordinatee)	1	0	1	-0.323112	NIP	6	1	5	83%
(Coordinate, Coordinate, Coordinate)	30	7	23	-0.154358	NIP	71	17	54	76%
*(Coordinate, Hyponym, Hyponym)	1	1	0	-0.151786	NIP	4	3	1	25%
(Coordinate, Holonym, Coordinate)	1	0	1	-0.074754	NIP	1	0	1	100%

T No: Total Number in training/test set

Con: Concurrence with overall ANN decision

Table 7.10: Patterns Extracted for Path of Length 3

Important Path	Type	Meaning
(Holonym, Coordinate, Coordinate)	PIP	This path is a strong indicator of a keyword
(Coordinatee, Hyponym, Coordinate)	PIP	This path is a strong indicator of a keyword
(Coordinate, Hyponym, Coordinate)	PIP	This path is a strong indicator of a keyword
(Hyponym, Hyponym, Holonym)	NIP	This path is a strong indicator of a non-keyword
(Synonym, Hyponym, Coordinatee)	PIP	This path is a strong indicator of a keyword
(Synonym, Synonym, Hyponym)	NIP	This path is a strong indicator of a non-keyword
(Synonym, Synonym, Synonym)	NIP	This path is a strong indicator of a non-keyword
(Synonym, Coordinate, Hyponym)	NIP	This path is a strong indicator of a non-keyword
(Coordinate, Synonym, Hyponym)	NIP	This path is a strong indicator of a non-keyword
(Synonym, Synonym, Coordinatee)	NIP	This path is a strong indicator of a non-keyword
(Synonym, Coordinate, Synonym)	NIP	This path is a strong indicator of a non-keyword
(Coordinatee, Coordinate, Hyponym)	NIP	This path is a strong indicator of a non-keyword
(Coordinate, Coordinate, Hyponym)	NIP	This path is a strong indicator of a non-keyword
(Synonym, Coordinate, Coordinatee)	NIP	This path is a strong indicator of a non-keyword
(Synonym, Synonym, Coordinate)	NIP	This path is a strong indicator of a non-keyword
(Coordinatee, Coordinate, Synonym)	NIP	This path is a strong indicator of a non-keyword
(Coordinate, Coordinate, Synonym)	NIP	This path is a strong indicator of a non-keyword
(Coordinatee, Synonym, Coordinate)	NIP	This path is a strong indicator of a non-keyword
(Synonym, Coordinate, Coordinate)	NIP	This path is a strong indicator of a non-keyword
(Coordinate, Synonym, Coordinate)	NIP	This path is a strong indicator of a non-keyword
(Coordinatee, Coordinate, Coordinate)	NIP	This path is a strong indicator of a non-keyword
(Synonym, Coordinatee, Coordinatee)	NIP	This path is a strong indicator of a non-keyword
(Coordinate, Coordinate, Coordinate)	NIP	This path is a strong indicator of a non-keyword

Table 7.11: Training and Test Results for Length 4 Paths

Path Relations	Training Set			Path	NP	Test Set			Con
	T No	No KPs	No NKPs	Weight	or NKP	T No	No KPs	No NKPs	
(Synonym, Synonym, Coordinate, Coordinatee)	1	1	0	0.176885	KP	1	1	0	100%
(Synonym, Coordinatee, Coordinate, Hyponym)	2	1	1	0.122212	KP	2	1	1	50%
(Coordinate, Coordinatee, Coordinate, Synonym)	2	0	2	-1.50294	NKP	1	0	1	100%
(Coordinate, Synonym, Coordinate, Coordinate)	2	1	1	-1.48476	NKP	3	1	2	67%
(Coordinatee, Synonym, Coordinate, Coordinate)	1	0	1	-1.2676	NKP	3	0	3	100%
*(Synonym, Coordinate, Synonym, Coordinate)	1	1	0	-1.18871	NKP	2	2	0	0%
(Synonym, Coordinatee, Synonym, Coordinate)	2	0	2	-1.00948	NKP	1	0	1	100%
(Coordinate, Coordinate, Coordinate, Hyponym)	1	0	1	-1.0055	NKP	3	1	2	67%
(Coordinatee, Coordinate, Coordinate, Coordinate)	1	0	1	-0.968874	NKP	1	0	1	100%
(Synonym, Synonym, Coordinate, Coordinate)	4	0	4	-0.536281	NKP	1	1	0	0%
(Synonym, Coordinate, Coordinate, Coordinate)	5	2	3	-0.237558	NKP	10	6	4	40%
*(Synonym, Coordinatee, Coordinate, Coordinate)	3	3	0	-0.058333	NKP	3	2	1	33%

T No: Total Number in training/test set

Con: Concurrence with overall ANN decision

Table 7.12: Patterns Extracted for Path of Length 4

Important Path	Type	Meaning
(Synonym, Synonym, Coordinate, Coordinatee)	PIP	This path is a indicator of a keyword
(Synonym, Coordinatee, Coordinate, Hyponym)	PIP	This path is a indicator of a keyword

quite weak. This suggests that length of 4 paths are not very important for keyword identification. Only two weak rules are extracted from length 4 paths which are rare in both training and test sets compared with lengths 2 and 3. From table 7.11, it can be seen that only a few words have length 4 paths in the training set. It is in questionable whether any rule drawn from this number of examples is reliable. For all the length 4 paths, only one is associated with more than 3 words in the test set. Thus, it is also difficult to evaluate the rules properly.

7.2.3.5 Conclusion

Coordinate and coordinatee are necessary relations in terms of paths. However, all of the paths that have a different sign from the balance of KWs and NKWs that activate that path, involve coordinate and/or coordiantee. 88% of paths contain coordinate and/or coordinatee. However, 100% of wrong paths involved them.

It is also the case that a significant proportion of paths with the correct sign contain at least one of those relations. However, this proportion is only 84%. The coordinate and coordinatee relations are necessary for training the ANN. Without them the paths are too sparse, but it is clear from the analysis that they often provide misleading information.

It can be seen that there are more rules for NIP than for PIP. This is because of the high ratio of NKWs to KWs. The longer a path, the less likely it is to be a PIP.

The percentage of positive paths in length of 1, 2, 3 and 4 are 75%, 35%, 33% and 16% respectively. The longer a path, the less likely it is to be a KP. This indicates that KWs tend to have short paths to the seed word. This is intuitively reasonable as if a word has long path to the seed word, it is probably not a KW, because there is no direct relation between them.

7.3 Category Information

In the trained weights, there are 25 bit for the 25 categories in WordNet. Because word level paths are used and a noun usually has more than one sense, a noun may fall into more than one category. Table 7.13 shows which keyword are in which categories for all the keywords in the training data.

It is clear from table 7.14 which shows the trained weights for the 25 categories that some categories are strong positive and some are strong negative, i.e. nouns in some categories are likely to be KWs and nouns from other categories are likely to NKWs.

How does the ANN learn the categories for positive and negative responses? By analysing the training data and the weights on the categories, the number of KWs/NKWs in a category in the training is found. In general, if the number of KWs is greater than the number of the NKWs, the ANN makes this category produce a positive response; if the number of KWs is smaller than the number of the NKWs, the ANN makes this category produce a negative response. If the number of KWs and the number of NKWs are both zeros, the ANN makes this category produce a small response, either positive or negative. This is the LBI rule introduced in last section.

Table 7.13: Categories of Training Keywords

	Academic	Campus	Class	Classroom	College	Course	Education	Enrollment	Grade	Instruction	Instructor	Learner	Learning	Lecture	Semester	Student	Teaching	Term	University
Act			1			1	1	1		1				1			1		
Animal																			
*Artifact				1	1	1													1
Attribute			1				1		1										
Body					1														1
Cognition							1		1				1					1	
Communication										1				1			1	1	
Event																			
*Feeling																			
Food						1													
Group			1		1	1	1		1										1
Location		1				1													
Motive																			
*Object		1		1	1	1													1
*Person	1										1	1				1			
Phenomenon																			
Plant																			
Possession																			
*Process							1		1				1						
*Quantity																		1	
*Relation									1	1				1				1	
Shape																			
*State									1										
Substance						1													
*Time									1						1			1	

Table 7.14: Trained Weights for Categories

Category	Weight	Number of KWs	Number of NKWs
Act	-2.70795	7	42
Animal	0.007797	0	0
*Artifact	0.998207	4	20
Attribute	-0.2075	3	28
Body	-1.4837	2	4
Cognition	-1.77333	5	37
Communication	-4.60152	3	33
Event	-5.73979	0	15
*Feeling	0.151903	0	2
Food	-1.72955	1	1
Group	-2.08959	6	20
Location	-0.10671	2	13
Motive	0.366939	0	0
*Object	2.229406	5	30
*Person	5.107476	4	10
Phenomenon	-4.67399	0	8
Plant	-0.12178	0	0
Possession	-4.93744	0	11
*Process	4.791475	3	8
*Quantity	3.787848	1	3
*Relation	1.89677	4	37
Shape	-0.13024	0	1
*State	0.498404	1	16
Substance	-3.69179	1	3
*Time	1.897714	3	17

However, there are some exceptions to the LBI rule. In the training data, there are some KWs whose path information is negative and so need positive category responses to learn them.

Take the "process" category as an example. In the training data, there are 3 KWs and 8 NKWs in this category (see table 7.15). The KW "grade" has very strong negative path information. It needs a positive category for it to be learned as a KW, thus the ANN adjusts the weights on some of the categories in which "grade" falls. "Grade" falls into 7 categories (see table 7.13). Which weights of these categories should be adjusted? Assigning a positive weight to the "process" category will only help KWs to be classified as KWs. What may be affected are the NKWs in this category, because a positive category weight would reduce the chance for NKWs to be learned correctly. However, most of the NKWs in this category have quite strong negative path and category information (some are very strong) as shown in table 7.15. Assigning this category a positive value of 4.79 which is the final weight of this bit will not prevent any of the NKWs from being learned correctly. Thus, even though there are more NKWs in this category in the training data, the ANN still assigns this category a positive weight. Table 7.15 shows the final path and category information of all the nouns in this category.

Table 7.15: Path and Category Responses for Training Nouns in the "Process" Category

Noun	Path Response	Category Response	Final Input	Final Output	KW Target
education:	4.24	-0.54	3.71	1.00	y
grade:	-5.13	6.47	1.33	0.97	y
learning:	8.96	4.47	13.43	1.00	y
action:	-1.67	-0.33	-1.99	0.03	n
development:	-4.64	-4.26	-8.90	0.01	n
feedback:	-3.12	-1.14	-4.25	0.01	n
identification:	-14.49	-1.00	-15.49	0.01	n
increase:	-2.67	-5.19	-7.86	0.01	n
passing:	-15.11	-8.65	-23.77	0.01	n
research:	-6.49	1.76	-4.73	0.01	n
review:	-4.25	2.28	-1.97	0.03	n

As already mentioned, the noun "grade" falls into seven categories (see table 7.13). Positive weights are assigned to the categories "relation", "state", and "time" against the LBI rule because the noun "grade" is in these categories.

Why does the ANN fail to learn "grade" by changing the path information to produce positive path responses? A possible reason is that the paths contained in "grade" are also used by other NKWs which must be satisfied, leading these paths to producing a negative response.

7.4 Word Level

The decision function for a noun to be a KW or not can be expressed as: $Sigomoid(category+path)$, where category is the sum of the weights of the ANN from the category bits and path is the sum of the weights of the ANN from the path bits. If the input to the function is greater than zero, the noun is recognised as a KW. Thus the test nouns can be divided into four groups according to this function. The first group contains those nouns with positive category and path values. The second group contains the nouns with positive category but negative path, and the third with negative category and positive path and the fourth negative category and path. A noun falling into group 4 cannot be identified as a KW and a noun in group 1 cannot be identified as a NKW.

Table 7.16 tells us two things. Firstly, category information is important in determining if a noun is a KW or not. Secondly, however, by comparing group 2 with group 3, it can be seen that category information introduces more false KWs than path information does. In group 2, where category information is positive and path information is negative (this means that IKWs are due to positive category information), out of the 34 identified KWs, only 10 are true KWs (i.e. recognised by

Table 7.16: Test Results Grouped by Path and Category Information

Group	Category	Path	*IKWs**	Correct	Percent	*INKWs**	Correct	Percent
1	+	+	13	5	0.38	0	0	N/A
2	+	-	34	10	0.29	85	81	0.95
3	-	+	15	10	0.67	11	10	0.91
4	-	-	0	0	N/A	184	175	0.95

IKWs: Number of keywords identified by the ANN
INKWs: Number of non-keywords identified by the ANN

humans). While in group 3, where IKWs are due to positive path information, 10 out of 15 are correct. The latter is much more accurate than the former.

What is the role of the category information in identifying NKWs? Group 2, drawing more false KWs, is very accurate in recognising NKWs with 95% correctness. Does this mean that category is more useful in doing this than path information? Group 4 also produces 95% correctness in identifying NKWs. The common feature in these two groups is that they have negative path information. From this comparison, it can be concluded that path information is as accurate as category information in identifying NKWs. Group 3, which is good at identifying KW, is also good at spotting non-keywords.

7.5 Uneven Number of Paths with Different Lengths

It is can be seen from the table 7.4 that paths of length 2 and 3 (283 + 226) are much more then paths with length of 1 and 4 (21 + 25). Will the ANN bias to paths of length 2 and 3 because of the dominant number of them? To answer this question, the test set is separated according to the path lengths. An ideal separation would divide the test set into four groups, one for each of the lengths. For example, group 1 contains the nouns which only have paths of length 1, and so forth. In practice, this is impossible because very few nouns contain paths of the same length. By examining the test data, it is found out that the number of combinations of path lengths is small. Thus, the test set was divided into seven groups, listed in table 7.17. The groups are numbered by four digits. The positions filled with 0 in the Group column means that nouns in this group have no that length of path. For example, group 0230 means that all the nouns in this group have paths of length 2 and 3 only, no paths of length of 1 and 4. Then tests are done on each group separately. The NG and PG results of the tests are shown in table 7.17.

From the test results, there is no evidence that the ANN are biased to paths of length 2 and 3. For the two set involved in paths of length 1, i.e. group 1200 and 1230, no of the results are the lowest among all the groups. Although results from group 0034 are not as goods as from other groups, results from group 0004 are as good as other groups.

Table 7.17: Test Results on Separated Test Data

	Pattern	NG			PG		
Group	Number	Overall	KWs	NKWs	Overall	KWs	NKWs
all	41400	85%	64%	88%	81%	33%	85%
0004	720	83%	100%	75%	75%	100%	67%
0034	8760	64%	50%	66%	58%	0%	63%
0030	1800	85%	60%	87%	81%	0%	84%
0230	6840	89%	33%	94%	86%	0%	93%
0200	17640	84%	50%	87%	80%	0%	83%
1230	5160	75%	67%	100%	67%	50%	100%
1200	480	83%	100%	77%	77%	100%	72%

7.6 Summary and Conclusion

From the rules identified by analysing the trained weights, it can be seen that paths of less than 4 are more useful to the ANN. Length 4 paths are not as important. Category information is also useful in identifying keywords because some keywords can only be identified by using this information, as in group 2 in table 7.16.

Weight analysis shows that the LBI rule is nature in the learning of the ANN.

Of the 8 relations used in the training of the ANN, synonym, holonym and hypernym have been identified as important relations. Coordniate and coordinatee relations are necessary to get enough paths for training. However, they often give misleading information to the ANN.

At path level, some IPs are identified which provide an insight into how the ANN solves the problem after learning. Humans can predict the ANN's KW/NKW classification, by looking at the paths for the word. This means that the ANN is no longer a black box. It is understandable by humans.

In learning the problem of KW/NKW classification, the ANN draws most of its conclusions from lengths 1, 2 and 3 paths and nearly no definite conclusion from length 4. The ratios of PIP to NIP in lengths 1, 2 and 3 paths are in reverse proportion to the length of path. This fact indicates that the ANN did learn the length of path as a key feature in recognizing KWs. The ANN did not learn this feature from the dominance (in terms of occurrences) of length 2 and 3 paths. This is evident because the ANN also learns from the length 1 paths which are as sparse as those of length 4. The experiment of separating the training data according to length of paths and training ANNs separately also proved that the ANN is not biased to length 2 and 3 paths.

Category information is useful in identifying KWs. However, it also draws more false KWs than path information does. It is not as accurate as path information. A possible reason is that the granularity is too crude. If a category is learnt as a key category, all nouns in this category are strongly affected. The resolution to this problem is a finer category system. Further research on using category information should be done.

Chapter 8

Literature Review for Relation Extraction

8.1 Introduction

Automatically extracting non-taxonomic relations from text is an intricate task and is not well-researched. Because of the diversity of relations, it is not clear what kind of relationship is in a particular section of text, hence it is hard to clearly define targets for extraction.

Some methods have previously been proposed for automatically extracting relations from text, especially from domain-specific text [10, 90, 92, 60, 22, 4]. However, most of the approaches only consider taxonomic relations, or other specific kinds of relations, such as has-a-part, definition, etc. These approaches have contributed a lot to the automatic construction of knowledge. Nevertheless, non-taxonomic relations between concepts are even more important than taxonomic relations for knowledge bases that are of real use. In fact, these relations take most of the time in building knowledge bases.

Relation extraction approaches vary in terms of the techniques, data sources, target relations, and evaluation methodologies. Pattern recognition is widely used in the attempt to extract relations from text [92, 22, 10, 138, 106, 157, 26, 34, 4]. Approaches based on shallow parsing are also popular [135]. Naive Bayes machine learning is also used [60]. The data sources from which relations are to be extracted also vary. Examples include from text[3, 4, 40, 88, 92, 28, 135], large language corpora [92, 28, 22, 3, 4, 152], World Wide Web [60, 34] and semi-structured data sources [54]. As pointed out, most work is concentrated on taxonomic relations [92], or on small sets of pre-defined relations like part-whole [22] or organisation-location [3]. Byrd [40] and Maedche [135] have attempted the most comprehensive relation set extraction. A more detailed review of the important works mentioned will be presented in following sections of this chapter.

8.2 Pattern Recognition Approaches

Hearst [92] uses a pattern recognition method for extracting hyponym relations from unrestricted text. A set of lexicon-syntactic patterns that are easy to identify, occur frequently and indicate the hyponyms reliably is identified. It is noted that there are many ways that the structure of a language can indicate a relation (hyponym in this case) of interest, and that a pattern can only identify a subset of the possible instances of a relation. It is difficult to find structures that frequently and indisputably indicate a particular relation, so as many patterns as possible were used. Because of the difficulty of identifying linguistic patterns manually, the approach is to use an algorithm to discover patterns automatically and then to verify this process manually. However, the automatic version of this algorithm was not implemented because of the complexity of the context in which a pattern occurs. The method is evaluated by comparing the extractions with the hyponym relations in WordNet. However no direct measures such as recall and precision were reported.

Berland and Charniak [22] use a similar technique to extract meronym relations from large corpora. Hearst reports failing with this relation in his research. Five patterns that are usually used to express part-whole relations are used. To use these patterns, a seed word (e.g. car) which is the whole of the part-whole relations is decided. All of the occurrences of these patterns against the seed word in the corpus are recorded and a rank based on statistical models is assigned to the identified parts. The ranked list is then presented to humans to judge the correctness of the extractions. For the first 50 top extracted words, the accuracy is 55%.

Hearst [92] and Berland [22] both use similar techniques and both report sparsity of examples as a problem. If there are only a small number of positive examples in a large corpus, the large number of negative examples will tend to dominate the data.

Other researchers also tried pattern recognition to extract semantic relations from machine readable dictionaries (MRDs). Alshawi [10] uses patterns consisting of part-of-speech indicators and wildcard characters to interpret LDOCE definitions. Markowitz [138], Jensen and Binot [106] and Nakamura and Nagao [157] also use patterns to extract taxonomy from dictionaries.

The most recent work in extracting relations using pattern recognition is [26]. Pattern-matching methods in relation extraction are used. Three kinds of relation were targeted: definition, exemplification and partition. First, two sets of trigger phrases were defined for each relation of interest to identify sentences containing potential instances of relations and to de-trigger sentences containing false instances. Then the triggered sentences are segmented by applying a list of patterns to them. Beside the segmenting patterns, each relation also has a set of templates which are to match some tokens. Tokens that match one or more templates are passed on to the validation stage in which a list of part-of-speech patterns are used to check whether the text fragments are valid noun phrases which are concepts and elucidations to be extracted. The patterns used to trigger sentences are totally non-syntactic and the validation requires that the sentences are correctly part-of-speech tagged.

Recall and precision measures are used to evaluate his approach. Table 8.1 shows the results reported.

Table 8.1: Results from Bowden Relation Extraction

Relations	Precision	Recall
Exemplification	100%	24%
Definition	100%	33%
Partition	100%	66%

This method is claimed to be domain independent. However, the method is relation dependent. Although it is possible to extend it to new kinds of relations, triggering phrases, segmenting patterns and validation patterns have to be constructed for new relations. Building these patterns is not an easy job and may need special knowledge about language and some domain knowledge, thus it is difficult for an end user to add new relations.

Some other researchers also employ pattern matching techniques with success in various parsing and extraction systems ([10, 92]). An unanswered question with these techniques is to determine how complete a pattern set is. This is important in that the completeness of the used pattern set has a significant impact on the extraction. No results are reported on this issue. It is doubtful if a complete pattern set can ever be established.

Another problem is that this technique extracts knowledge without attempting to understand what is to be extracted. This is usually the case when using pattern matching in knowledge extraction. An example of this kind can be found in Bowden [28]. For the sentence "An example of a sorting criterion is alphabetical order", the system extracted "sorting criterion" as the concept and "alphabetical order" as the example, which is correct. However, for the sentence "An example of use of loop is in the bill program in Chapter 1", it extracts "use of loop" as the concept and "bill program in Chapter 1" as the example, which is wrong. This problem can be avoid by identifying the concepts first and then finding the relations between them. The first stage of this research, identifying concepts, will effectively remove this type of error.

Bowden[26] also reports another problem with pattern matching. Sometimes there is no suitable pattern that can correctly segment a sentence involving a complex sentence structure. An example of this kind of sentence is "Defined as a function which orders a list of items a sorting routine is a common function in a large program". This problem reveals the limits of using pattern-matching techniques in knowledge extracting. The suggested solution, using parsing techniques, also means that pattern matching alone is not enough for full KE.

This research is trying to extract named relations that will be the links between concepts. Apparently, the relations involved are much wider than several pre-defined relations. Due to the limitations of pattern-matching techniques, the relation dependent method [26] is not suitable for this purpose. However, it provides a good base line for comparison.

Brin [34] and Agichtein [3] use a quite different strategy in their attempt to extract pre-defined relations. The strategy is based on an assumption that a target of extraction occurs redundantly in the corpora. Thus it is not necessary to extract every occurrence of the target as long as the system extracts one of the instances of the target from the corpus.

In the Snowball-VS system [3] for extracting organisation-location relations, the system starts with a small number of example relations, such as (Microsoft, Redmond). It identifies the sentences in which the examples occur. From these sentences, patterns are generated and used to search more examples. For example, the system may learn that $< Location > -based < Organizatio >$ is a pattern. New examples are examined and the most reliable ones are then used to find more patterns. A pattern is expressed as a 5-tuple of

$$\{< left - context, weight1 >, Location, < middle - context, weight2 >, Organisation, < right - context, weight3 >\}$$

A 5-tuple is generated for each of the occurrence contexts of known examples. Then the generated tuples are clustered. The centroid of each cluster is treated as a new pattern. In finding new examples, a tuple is generated for each of the text segments in which a location and an organisation occur together. The new tuple is compared with all the existed patterns by calculating a matching function, to decide which pattern (or patterns) generates it. The information regards which pattern or patterns matches the candidate tuple and the degree of match is stored for this candidate tuple. The system generates a confidence measure for each pattern. Tuples generated by patterns with low confidence are discarded.

The system is trained and tested on a large collection of real newspapers from the North American News Text corpus[85]. Because of the large size of the test set, it is not feasible to validate the extraction manually. Thus, the evaluation involves using automatic construction of the answer set against which the extraction is tested. It achieved high recall and precision compared with other information extraction systems. There are several reasons for the good results. As mentioned above, this method is designed to extract relations from a large collection of documents, not for a small number of documents. It is doubtful whether this method will work well for a small number of documents. An instance needs only to be extracted once from the large set of documents and an instance of a relation is very likely to appear in one form of the patterns. For example, instance (O, L) can be expressed in many different ways each of which needs a special pattern for extraction. In a single document, the instance can only be expressed in a certain way, then a pattern, say P1, must be composed to extract this instance in this document. However, in a large document, P1 may not be necessary because it is very likely that this instance can be extracted by another pattern which is easy to construct. In this case, P1, which may be more difficult to construct, is not necessary. Thus, in this work, pattern construction is not as important as in other IE systems. A concern about the evaluation is the involvement of automatic construction of the answer set. This is not as rigorous as a manually constructed answer set. Thus, the conclusion is that it is unfair to compare this work to other IE systems.

Snowball-SMT [4] is a new version of Snowball, which takes into account the order of the words in context. Thus Snowball-VS focuses on the presence or absence of keywords that indicate the correct relation, while Snowball focuses on the order of words in the context. The combination of the two versions increases the precision of extractions.

8.3 Machine Learning Approaches

Craven [60] reports attempts to build knowledge bases from the World Wide Web, based on knowledge extraction using a supervised learning system. The aim is to extract instances of classes and relations between the instances. The instances are the nodes of the knowledge and relations the arcs connecting the nodes (as in Hypertutor). The learning system requires an ontology to define the classes to be extracted. The classes have names and relation slots that can be filled by the relations extracted. To train the system, a set of training examples specifying instances of classes and relations is also required.

Classifiers were trained to identify instances of classes. Both naive Bayes statistical models and first-order models were used. The learning algorithm used in training the system is similar to FOIL [173], a first-order system. The hypothesis on which the relation extraction system training is based is that relations between instances of classes are usually presented by hyper-link paths in the web. Thus the system focuses on learning rules that characterise prototypical paths of the relations.

The system was trained in the domain of "Computer Science Department". The experimental ontology included seven classes: Department, Faculty, Staff, Student, Research Project, Course and Other. Three relations were tried, including department-of-person, member-of-project and instructor-of-course. The system was tested in an on-line environment: a Web site of a computer department comprising 2722 Web pages. The system achieved high accuracy in the extraction of instances of both classes and relations (74% 78% and 85% for the three target relations). One reason for the high accuracy is that web page structures are much simpler than sentence structures in text. So it is difficult to compare this work to others that extract relations between words from text. Another reason is that the ontology is very limited, having only seven classes and only three target relations.

This technique requires an ontology which is domain-dependent, thus a new ontology has to be constructed for a new domain. New training examples also have to be built for new relations. Constructing new ontologies and building new training examples are both difficult and time-consuming. So the main problem with this method is domain transferability and limited scope. The pre-defined relation types of this work are too narrow to compare with our relation extraction objectives here. The difference between sentence structures and Web page structures, makes it is difficult to employ this technique to build Hypertutor KBs.

Brin [34] also tried to extract organisation-location relations from the WWW using pattern recognition.

8.4 Parser-Based Approach

Maedche [135] tries to extract non-taxonomic relations between concepts, using shallow parsing and domain-specific text analysis, based on a taxonomic ontology. To extract relations, shallow parsing is used to identify linguistically related pairs of words. Statistical analysis is then done on the identified word pairs to discover generalised association rules. Relations at appropriate levels of abstraction are proposed based on an ontology. Support and confidence metrics are used for evaluation. Support for a word pair is the percentage of occurrences of the pair in all of the word pairs. Confidence is the percentage of occurrences of the pair in all of the pairs involved with the first word of the pair. Statistical correlation data is used to calculate the support and confidence measures for each word.

The following is an example of the extractions from the tourism domain. The sentence is:

Costs at the youth hostel amount to $20 per night.

Suppose that the shallow parser finds out that the word "cost" frequently co-occurs with "hotel", "guest house" and "youth hostel", then three pairs of words can be identified: (cost, hotel), (cost, guest house), and (cost, youth hostel). In the ontology, a word that is at higher level than "hotel", "guest house" and "youth hostel" can be found. Suppose these three words are not at the highest level in the ontology. In this case, "accommodation" is such a word. Support and confidence measures are also calculated for word pair (cost, accommodation). Finally, a suitable level of abstraction will be determined to be the result of the extraction, based on the calculated support and confidence information of the pairs and the threshold support and confidence value.

Test results from this system vary according to the threshold value of support and confidence. The results reported in the test domain are shown in table 8.2.

Table 8.2: Results from Maedche Relation Extraction

Support	**Confidence**			
	0.01		**0.4**	
	Recall	**Precision**	**Recall**	**Precision**
0.0001	66%	2%	2%	1%
0.04	13%	11%	0%	0%
0.06	6%	9%	0%	0%

However, two aspects of this approach are not clear. Firstly, the paper did not address the issue of domain transferability. This is important because a domain-specific taxonomy was required to perform the domain dependent analysis to propose suitable levels of abstraction. Building a domain-specific taxonomy is not an easy task, although there is reported work on automatic taxonomy construction. It would be more useful if a general taxonomy e.g. WordNet could be used. Secondly, it is claimed that the system can identify the name of the extracted relations. However, how this was carried out was not stated.

This technique tries to extract non-taxonomic relationships and is not confined to any specific types of relations. This is very similar to the relation extraction objectives of this project. They both aim to extract relations based on sentence analysis. However, the techniques are not the same. One difference is that in this work, since keywords have already been extracted before the relation extraction stage, it is only necessary to concentrate on keyword pairs. Thus, statistical analysis, for example support and confidence information of word pairs is not necessary. The task of this research is to find out what the relation between the pair of keywords is and the name of this relation. It is not necessary to decide whether this pair of words should be extracted or not. In order to use this technique a suitable and reliable trunk parser would be necessary.

Chapter 9

Relation Extraction

9.1 Introduction

As defined in Chapter 1, this research is to extract knowledge from text and to organise it into the existing formalism used in Hypertutor. This formalism includes concepts and hyperlinks where hyperlinks denote the relationships between the concepts. Techniques for extracting concepts are presented in chapters 4 and 5. This chapter presents techniques for extracting relationships between concepts.

Relations between concepts is a major component in knowledge bases. This is also true in the target hyper-knowledge base. In fact, defining relations for a knowledge base is more time consuming than finding concepts when manually building a KB. Relation extraction (RE) is harder than concept extraction when automatically constructing a KB. This is not a well-researched area, thus there is plenty of scope for investigation.

Relations are important to KBs in that they connect concepts together, forming an interlinked network of concepts. An interlinked network of concepts about a domain or world is much more understandable than a set of isolated concepts.

9.1.1 Difficulty of RE

Human languages reflecting the world and human thoughts, express overwhelmingly complicated relations between objects and concepts in the world. Complexity in human languages exist in lexicons, syntax, semantics, pragmatics, context and in speaking and writing style, to name a few.

The complex structure of human language poses difficulties for attempts to extract relations automatically from text. There are so many different ways of expressing relations in a language that it is impossible even to extract all instances of a single kind of relation from text. Another problem is when there is a component missing or omitted, such as in infinitive, gerund structures. For example, in the sentence "In Anglo-American systems a lawyer investigates the facts and the evidence by conferring with his client, interviewing witnesses, and reviewing documents", the subject of verbs "confer", "interview" and "review" is "lawyer". However, it is omitted. It is not a problem for human readers, but it poses a problem for computer extraction. Relations may be within a sentence or across sentences. Anaphora resolution which relates to the identification of the correct relation between words has proved to be an extremely difficult task in natural language processing [81, 57, 12, 9].

9.1.2 Relation Extraction Objectives

All of the reported work in this area is trying to extract specific kinds of relations chosen by the researchers [92, 28, 60, 22, 3]. This is because there are so many different kinds of relations in the world and hence in a natural language that it could be very difficult to attempt to extract them all. Thus, a definition of the kind of relations is needed. Relation components for Hypertutor incorporate verbs as connections between nouns, i.e. hyperlinks, therefore only verb-noun relations in a sentence are targeted. Apart from the main verbs in a sentence, other verb forms such as infinitive and gerund structures are also under consideration. This kind of relation is termed here as named relations and the verb is the name of the relation. Some examples of target relations are shown in tables 9.1, 9.2, and 9.3 in section 9.2.2.

Other relations, such as those between nouns in a compound noun, are not considered.

There are several reasons that verb relations are chosen. Firstly, these are required by the target knowledge base, Hypertutor. Secondly, this is the most comprehensive and most widely used type of relation. According to linguistic theory, the most important relations in a sentence are expressed in this form, even if there are other relations in a sentence, such as relations expressed in prepositions. For example, in the sentence "He beat her with his hand", the basic fact conveyed is that he beat her. In answering the question of "what happened?" after reading this sentence, most readers would reply "He beat her", not "He used his hand". Thirdly, even though this is such an important kind of relation, very little work has been done to extract it because of the complexity of verbs.

9.2 New Approaches to Relation Extraction

Most of the work reviewed in chapter 8 is for knowledge base construction and is corpus-based. Some used a strategy that does not try to extract as many instances as possible from a single document, but tries to extract from all the documents [3]. This work does not employ this kind of strategy, because the task is to understand texts which demand extracting as much knowledge as possible from a single document.

Generally speaking, a sentence can be viewed as a set of relations and a relation can be expressed as $Relation = LinkWord(word, word, ...)$. Here the LinkWord can be a verb or a preposition. Obviously, a word can appear in more than one relation, thus link-words link words together and the linked words can connect relations together. A link-word may link 2, 3 or more than three words. The main job of RE is to find out which link-word connects which words together.

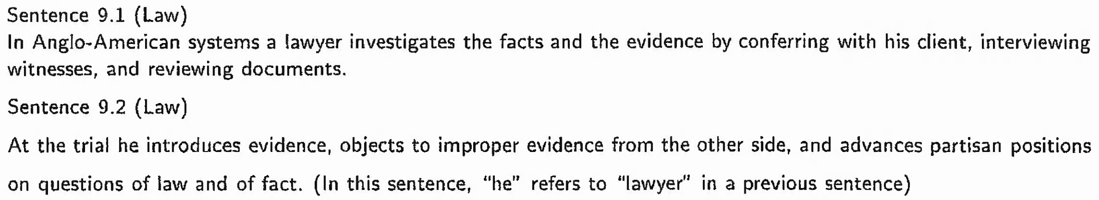
One important fact about natural language is that the relations in a sentence are not generally expressed in a direct way. There are omissions, anaphora etc. This increases efficiency and enriches the aesthetic qualities of natural language, but poses problems for automatic extraction. An important part of relation extraction is therefore to re-arrange sentences so that things are expressed directly and to restore all the nouns that participate in a relation.

As only verb relations are to be extracted, the extraction can be carried out in two steps. The first is to identify verbs and the second is to find out nouns linked by verbs. It is trivial to find verbs in a sentence, either by parsing or by part-of-speech tagging. Thus, the difficulties lie in finding which nouns are linked by a verb. It seems that the difficulty should be reduced by only finding the verbs that link two or more keywords which are already known. However, this is not true because of the complicated verb structures, sentence structures and anaphora.

Compared to infinitive, participles and gerund, it is easy to find nouns for main verbs. Because most of the omissions occur in infinitive, participles and gerund structures and most anaphora are in compound and complicated sentence, it is easier to extract relations in a simple sentence than in a compound sentence and complex sentence. (A clause is a group of words containing a subject and verb which forms part of a sentence. A simple sentence contains only one clause. A compound sentence is a sentence made up of two or more independent clauses but no subordinate clauses. The clauses in a compound sentence are usually joined by conjunctions and/or some kind of punctuation. A complex sentence is a sentence made up of one main clause and at least one subordinate clause[221].)

9.2.1 Information Re-construction

As stated above, not all the components of a relation to be extracted are in the parse tree or even in the sentences. Even if a parse tree of the sentence is available, some of them must be deduced, based on grammar rules, verb sub-categorisation or sentence structure. For example, in the sentence 9.1 below, relations $investigate(lawyer, facts)$ and $investigate(lawyer, evidence)$ can be extracted easily because all the elements of the relations are in the sentence and related together in a direct way. The other three relations of interest, $confer(lawyer, client)$, $interview(lawyer, witnesses)$ and $review(lawyer, documents)$, are not that easy to extract. As human readers of this sentence, we know that the actor of the "conferring" is the lawyer, i.e. the subject of the verb "confer" is the lawyer. However, this subject will not be explicitly represented in any parse tree for this sentence. To extract this relation, some kind of reasoning has to be done even though this subject is already in the sentence. The reasoning may be based on grammar rules, semantic analysis, etc. In sentence 9.2 shown in figure 9.1, if it is known that "he" refers to a "lawyer" in a previous sentence, then relations $introduce(lawyer, evidence)$ and $object - to(lawyer, evidence)$ can be extracted. It is impossible to find in any parse tree for this sentence that the subject of the verbs "introduce" and "object to", is "lawyer". The only "subject" in the parse tree is the word "he" which is definitely not a keyword. This problem is harder to solve than the last one in that the element ("lawyer") of the relation to extract is not in the sentence at all. Thus, reasoning in the scope of the sentence will not work. This is the well known problem of anaphoric resolution.



Sentence 9.1 (Law)
In Anglo-American systems a lawyer investigates the facts and the evidence by conferring with his client, interviewing witnesses, and reviewing documents.

Sentence 9.2 (Law)
At the trial he introduces evidence, objects to improper evidence from the other side, and advances partisan positions on questions of law and of fact. (In this sentence, "he" refers to "lawyer" in a previous sentence)

Figure 9.1: Two Sentences from the Law Domain

It can be concluded from the above that parsing itself cannot totally solve the relation extraction problem. Some kind of post-analysis and/or reasoning are necessary.

Information re-construction is done after sentences from the document are parsed and before the relations are extracted. Although there are components omitted, a parse still holds the information needed for the relation extraction.

The task of information re-construction is mainly concerned with finding subjects or objects for infinitives, gerunds and participles. There are many ways of using these forms of verbs. In this research, the rules for finding components for these verbs are based on analysis of simple sentences [221]. However, the sentence structures in the target documents are much more complicated. Another grammatical phenomenon that is not considered is the passive voice of infinitives and participles. The aim of the experiments is to investigate methods to identify subjects for non-finite verbs.

9.2.2 Methodology

Recall and precision metrics are used to quantify the comprehensiveness and accuracy of the extraction. Algorithms are developed and tested on the two documents used in the keyword extraction. These consisted of 54 sentences from the domain education and 57 sentences from the domain law. They were given to four human extractors to extract the relations in the sentences. Keywords are already known and underlined so that they are identified easily. The human extraction task is composed of the parts described below. Three of the extractors are native English speakers. The other human extractor is the author of this thesis. Extractions from the algorithms are then checked against the human extractions to calculate the recall and precision metrics.

The human judges were asked to identify relations and categorise them as one of the three relation types to the level of difficulty of the extraction. The first type is relationships whose keywords are connected in one clause, i.e. without involving indirect references. The second type is relationships between keywords which rely on one or more indirect references within a sentence. The third type is relationships between keywords which involve one or more indirect references between sentences in the whole document scope.

Tables 9.1, 9.2, and 9.3 show the three different types of relations with examples used to explain the differences to the human extractors.

Table 9.1: Type 1 Extraction Examples

Sentence	**Extractions of Type 1**
Cows produce milk.	produce(cows, milk)
The carrot is a very peculiar species of vegetable.	is a species of(carrot, vegetable)

In Anglo-American systems a lawyer investigates the facts and the evidence by conferring with his client, interviewing witnesses, and reviewing documents.	investigate(lawyer, facts)
	investigate(lawyer, evidence)
	No extraction for client, witnesses and documents as lawyer is only implicitly referred to

If a lawyer loses his client's case, he may seek a new trial or relief in an appellate court.	No extraction for lawyer and case because the relationship between them is not generally true. No extraction for "lawyer" and "trial" as anaphoric reference used to connect them.
At the trial he introduces evidence, objects to improper evidence from the other side, and advances partisan positions on questions of law and of fact.	Although he refers to "lawyer", no extract anything here–indirect reference to another sentence.

Table 9.2: Type 2 Extraction Examples

Sentence	**Extractions of Type 2**

In Anglo-American systems a lawyer investigates the facts and the evidence by conferring with his client, interviewing witnesses, and reviewing documents.	conferring(lawyer, client)
	interviewing(lawyer, witness)
	reviewing(lawyer, document)

If a lawyer loses his client's case, he may seek a new trial or relief in an appellate court.	seek(lawyer, trial)
At the trial he introduces evidence, objects to improper evidence from the other side, and advances partisan positions on questions of law and of fact.	Although he refers to "lawyer", no extract anything here–indirect reference to another sentence.

Table 9.3: Type 3 Extraction Examples

Sentence	**Extractions of Type 3**
At the trial he introduces evidence, objects to improper evidence from the other side, and advances partisan positions on questions of law and of fact.	introduces(lawyer, evidence) objects to(lawyer, evidence)

The results from the human judges and sentences used in the test are listed in Appendix G and F respectively. Extractor 1 is a native English speaking researcher, not specialized in language processing. Extractors 2 and 3 are native English speaker and working in natural language processing. Extractor 4 is not a native English speaker, but doing research in NLP. It is clear that any two of the human extractors are not totally agreed on what should be extracted. Sometimes, the difference is quite big. Extractor 1 produced some relations that are expressed in prepositions. Although this type of relation is not part of the aim here, it does indicate that preposition relations could be important. Extractor 1 also extracted more relations than the other three extractors in direct relations, but less in indirect relations.

The human extraction results for type 3 relations are not presented because the automatic extraction algorithm is designed for the first 2 types of relations only. The extraction of the third type, involving anaphoric and co-reference resolution, remains as a task for further research.

An important phenomenon is that in most of the extractions, the concepts composing the relation are not single nouns, but are phrases containing a keyword. If the keywords alone are included in the extraction without the modifier words in the phrase, the meaning will change. This fact tells us that using a single keyword to express a key concept is not enough. Using key phrases to express key concepts is an issue for further research.

After further analysis, target extractions were defined as those in tables 9.4, 9.5, 9.6 and 9.7.

Table 9.4: Direct Relation within Sentences for Document 1

Sentence No	**Keyword 1**	**Keyword 2**	**Linking Words**
8	Students	Their learning	Analyse
21	Mediated learning	Instruction, learning and assessment	Is a form of
47	Students	Courses	Take

Table 9.5: Direct Relation within Sentences for Document 2

Sentence No	**Keyword 1**	**Keyword 2**	**Linking Words**
3	Lawyer	The evidence	Investigates
6	Lawyers	Judge	Suggest lines of inquiry to
12	Lawyer	Rights	Helps to particularize
25	Procurators	Litigation	Attended to the formal steps in
26	Advocate	Client	Gave advice to
28	Procurator	Advocate	Swallowed up
59	Advocate	Rights	Fights for

Table 9.6: Indirect Relation within Sentences for Document 1

Sentence No	Keyword 1	Keyword 2	Linking Words
3	Mediated learning	college	Being implemented in
	Mediated learning	University	Being implemented in
5	Learning activities	Instructor + students	Provide assistance
24	Instructor	Teach	
	Students	Learn	
30	Students	Course	Complete
32	Instructor	Students	Take on
34	Students	College	Enter
35	Students	Courses	Are placed in
46	Students	Colleges	Attend
50	Students	College	Arrive on

Table 9.7: Indirect Relation within Sentences for Document 2

Sentence No	Keyword 1	Keyword 2	Linking Words
1	Law	Cases	Is applied to
2	Judge	Cases	Works in the process of trying and deciding
	Advocate	Cases	Works in the process of trying and deciding
3	Lawyer	Client	Confers with
	Lawyer	Witnesses	Interviews
6	Lawyer	The law	Argue
7	A lawyer	A trial	May seek
14	Lawyer	Client	Has loyalty to
21	Procurator	Litigation	Attended to formal aspects of
26	Advocate	Client	Gave advice to
	Advocate	Court	Presented oral arguments in
31	Judges	Advocates	Appointed from
37	Most mentor	Queen's counsel	Is
39	Barrister	A solicitor	Is employed by
40	Bar	Solicitors	Gives legal advice to
57	Advocate	Law	Has a duty to the

9.3 Parse-tree Based Approach

In Maedche [135], shallow parsing was used to identify linguistically related pairs of words. Statistical analysis was then done on the identified word pairs to discover generalised association rules. In order to use a similar technique, a suitable trunk parser is necessary. However, because keywords are already known and a pair of keywords is definitely the target of extraction, no statistical analysis is needed here.

9.3.1 Evaluation of Different Parsers

It was not possible to purchase a commercial parser, so free parsers from the Internet were tried.

9.3.1.1 Apple Pie Parser

Apple pie parser [197] is a bottom-up probabilistic chart parser which finds the parse tree with the best score using a best-first search algorithm. Its grammar (of English) is a semi-context sensitive grammar with two non-terminal symbols (changed to 5 non-terminal symbols in version 5.8). It was automatically extracted from the Penn Tree Bank (PTB), syntactically tagged corpus [137].

This is, as far as the author knows, the first publicly available parser based on the new strategy [197] for grammar generation, i.e. a fully automatic acquisition of grammar from a syntactically tagged corpus (as opposed to manual or statistically aided manual grammar generation used in conventional projects). Although there are some problems with this strategy, such as the availability of such a corpus, and domain restrictions, the performance of the grammar is fairly good. The author believes the idea shows one of the promising directions for the future of natural language research.

The parser generates a syntactic tree with bracketing as with the PTB. The parser did not take advantage of argument structure labels in the latest release (Version 2.0) of the PTB. It is trying to make a parse tree as accurate as possible for reasonable sentences. It was known before the experiment that the performance was not good compared with that for state of art parsers.

There is a serious problem in applying this parser to the documents in this project. For several sentences it failed to complete a parse. Thus, it cannot be used for this task.

9.3.1.2 Link Grammar Parser

The Link Grammar Parser [129] is a syntactic parser of English, based on link grammar, an original theory of English syntax. Given a sentence, the system assigns to it a syntactic structure, which consists of a set of labeled links connecting pairs of words.

The parser has a dictionary of about 60000 word forms. It has coverage of a wide variety of syntactic constructions, including many rare and idiomatic ones. The parser is robust; it is able to skip over portions of the sentence that it cannot understand, and assign some structure to the rest of the sentence. It is able to handle unknown vocabulary, and make intelligent guesses from context and spelling about the syntactic categories of unknown words. It has knowledge of capitalization, numerical expressions, and a variety of punctuation symbols.

The same problem arises in applying the link Grammar parser to our experiment documents. It kept running trying to parse the sentences and not stopping. Therefore, it cannot be used.

9.3.1.3 EngCG-2 Parser

The third is a parser called EngCG-2 based on English Functional Dependency Grammar [112]. It employs a grammar comprising of 27000 hand-written rules.

Although it is one of the best English parsers on the market, it can only produce a complete parse tree for 38% of the sentences of the document from the law domain and only 20% of the sentences of the document from the education domain. For the other sentences, it produces a forest, not a tree. This means it failed to parse the sentences.

The online demo page which allows a user to parse 5 sentences each time was used. Because the problem is not as serious as the first two, this parser is the final choice.

33 dependency functions are produced by this parser, such as subject, subject complement, object, object complement, etc. This parser also tags the sentence at word level. Based on these tags verbs can be identified in infinitive, gerund and participle forms.

A parse tree for sentence 9.1 in section 9.2.1 from this parser is shown in figure 9.2.

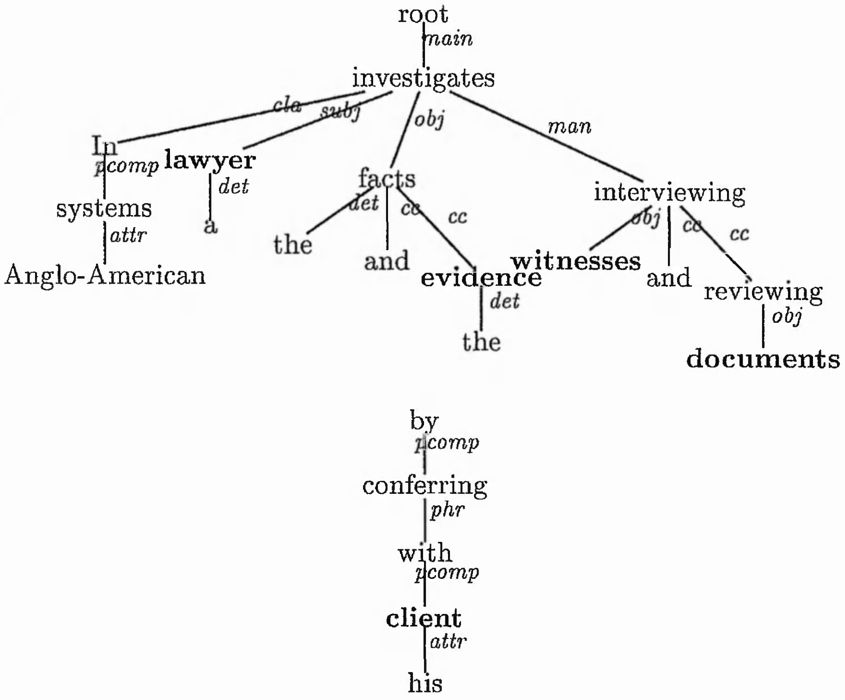


Figure 9.2: Parse Tree for Sentence 1 Produced by EngCG-2 Parser

Another parser tried was a hybrid parser developed by Tepper [217]. This uses a combination of ANN and symbolic techniques. However, this parser needs re-training to use it to its full advantage. This retraining would be time consuming. In contrast, EngCG-2 is easy to use being a web service..

9.3.2 Finding Elements for Verbs

It has already been noted that some kind of post-analysis and/or reasoning based on the parse tree must be done. The initial aim is to try to extract the first two kinds of relations, i.e. those that are located within sentences. Because the target is verb-noun relations, the main work of post analysis is to find subjects, objects, including direct and indirect, and complements for verbs. Verbs here means all the verb forms including the main verb of a sentence and infinitives, gerunds and particles appearing in the sentences.

Having obtained a parse tree for a sentence, it is necessary to find constituents, such as subjects, objects, and complements for any form of the verbs. Subjects for infinitives, gerunds and participles are usually missing. Thus, the main work for post-analysis is to find subjects for these forms of verbs. This is based on the usage of these forms of verbs in English [221]. Compared to finding subjects, it is easy to find objects and complements for verbs because they are already in the parse tree, assuming it is a correct parse for the sentence. However, finding subjects for the main verb of a sentence is also simple, as they are also in the parser tree.

Rules for finding subjects of infinitives, gerunds and participles are shown in tables 9.8, 9.9, and 9.10. In these tables, syntactic structure column means the structure in which an infinitive, gerund or participles appears. In row 3 in table 9.8, for instance, syntactic structure "Infinitives as complements of a verb" means if an infinitive follows the main verb of the sentence and functions as a complement, the subject of the infinitive is the subject of the main verb (main subject), as shown in the example column.

In deriving these rules, [221] was consulted. It listed the usage of these non-finite verbs, but does not say which constituent is the subject and/or object.

9.3.2.1 Finding Subjects for Infinitives

A general rule about infinitives is that if an infinitive has no formal subject as in "It is impossible for him to finish the work in one day", its subject is usually the subject of the main verb of the sentence in which the infinitive appears. As with other rules in English (actually in any natural language), there are exceptions to this rule. Special cases of the usage of infinitives considered in the extraction algorithm are listed in table 9.8.

For "The man is too weak to move", rule 7 identifies the actor of "move" to be "the man". However, the meaning is negative, i.e. "the man" does not move because he cannot move. This does not necessarily mean that the rule is wrong, but that a fuller understanding requires semantic analysis. For rule 3, the subject for the infinitive depends on the object of the main verb. If it can perform an action, it is the subject of the infinitive. If it cannot act, the subject of the main verb is the subject of the infinitive.

Table 9.8: Rules of Finding Subjects for Infinitives

No	Syntactic Structure	Subject	Example
1	Infinitives as Subjects	No	To save money seems impossible. It seems impossible to save money.
2	Infinitives as objects of a verb	Main subject	I managed to put the fire out.
3	Infinitives as complements of a verb	Main subject Object of the main verb	He made an effort to stand up I want him to go
4	Verb + how/what/when/where/ which/why/whether + infinitive	Main subject Main subject	I found out how to buy fruits cheaply.
5	Noun + Infinitive + preposition	Main subject	I have no pen to write with
6	Noun + infinitive	No for "to let"	A house to let is not easy to be found
7	Too + adjective + infinitive	Main subject	The man is too weak to move
8	Adjective + enough + infinitive	Main subject	He is old enough to go to school

9.3.2.2 Finding Subjects for Gerunds

Finding subject for gerunds is simpler than for infinitives because gerunds tend to be nouns rather than verbs and the usage of gerunds in English is simpler than that of infinitives.

If a gerund has its own formal subject e.g. "his" in sentence "His coming makes us very happy", it is the subject of the gerund. Other rules are listed in table 9.9.

Table 9.9: Rules of Finding Subjects for Gerunds

No	Syntactic Structure	Subject	Example
1	Gerunds as subjects	No	Resting is easier than working
2	Gerunds as subject complement	No	My favorite sport is swimming
3	Gerunds as object	Main subject	He admitted taking the money
4	Preposition + gerund	Main subject	I am against saying anything.

9.3.2.3 Finding Subjects for Participles

Participles serve as adjectives in sentences, so they can not be the subject or object of a sentence. The main function of present participles is to modify nouns and to be objective complements.

9.3.3 Verification of the Algorithm for Subject Extraction

The implementation of the aforementioned rule tables needs two kinds of information: the parse tree and the tags for words in sentences. It needs the tags to find out the verbs that are not the main verbs of a clause and to determine their type, i.e. infinitive, gerund or participle. Then, the

Table 9.10: Rules of Finding Subjects for Participles

No	Syntactic Structure	Subject	Example
1	Transitive verb + object + present participle	Object of the main verb	I heard someone knocking at the door
2	Present participle + noun	The noun	The setting sun looks beautiful
3	Present participle + noun	The noun	The man standing over there is my brother
4	Intransitive verb + present participle as complement	Main subject	We stood listening to him
5	Transitive verb + object + past participle	Object of the main verb, passive voice	You can hear English spoken all over the world
6	Past participle + noun	The noun, passive voice	The fallen leaves have covered the path
7	Noun + past participle	The noun, passive voice	The language spoken in Canada is English
8	Intransitive verb +	Main subject, passive voice	They returned utterly exhausted
8	past participle as complement	Main subject, passive voice	They returned utterly exhausted
9 9	Participial construction	Main subject	Walking along the street yesterday, I met a friend of mine.
			Born in better time, he would have become famous
10	Absolute Participial construction	Sense subject of the construction (usually the noun or pronoun before the participle)	The weather being fine, we went on a picnic
			The lesson ended, the boys rushed out of classroom to play balls

grammatical function of the verb must be found out. And usually, the context in which it appears is also needed.

Sentences from [221] were used to help design the algorithm for finding subjects for infinitive, gerund and participle. These sentences can be found in appendix H. The adjusted algorithm was then applied to a part of the document from the law domain to test it. This part is composed of 29 sentences. The algorithm found correct subjects for 26 of the 44 non-finite verbs in this part, i.e. 59%. For the other 18 where the algorithm failed to find correct subjects, 12 failed at the parse stage, i.e. there is not a complete parse for the these sentences. Because of the complicated sentence structures, it is difficult to say how many correct subjects could be found for the 12 non-finite verbs if the sentences in which they appear had complete parse trees. However, it seems that at least four more could have been found because other non-finite verbs in similar structures with a complete parse tree have had the correct subject identified.

9.3.4 Relation Extraction Algorithm

Because the subjects for non-finite verbs are already recorded, to extract the relation is just to find out if the subject and object of a verb are both keywords.

9.3.5 Relation Extraction Test Results

Tests are done by applying the relation extraction algorithm to the same sets of sentences that were identified in the human extraction. Results are shown in table 9.11. Although the precision measures are quite good, the recall measures are disappointing.

Table 9.11: Relation Extraction Results from Parse Tree-based Method on Both Domains

Domain	Targets	Extractions	Target Extractions	Recall	Precision
Education	14	2	1	4%	50%
Law	23	8	4	34%	50%

Table 9.12: Relation Extracted from Parse Tree-based Method on the Law Domain

No	Extraction	Target Extraction	Subject
1	deciding(judge, cases)	yes	yes
2	interviewing(lawyer, witnesses)	yes	yes
3	reviewing(lawyer, documents)	yes	yes
4	argue(lawyers, law)	no	no
5	loses(lawyer, case)	no	no
6	particularize(lawyer, rights)	no	yes
7	swallowed(procurators, advocates)	yes	no
8	secured(barristers, right)	no	no

Table 9.13: Relation Extracted from Parse Tree-based Method on the Education Domain

No	Extraction	Target Extraction	Subject
1	plan(students, courses)	no	yes
2	take(instructors, students)	yes	no

Tables 9.12 and 9.13 show all the extractions from the two domains. The Target Extraction column in these two tables indicates whether an extraction is in the human extraction set or not. The Subject column means that the subject of the verb in the extraction is found by the subject-finding algorithm.

Table 9.12 shows that half, 4 out of 8, of the subjects are found this way and three of the four are in correct extractions. This means that the algorithm does work and increases both of the recall and precision measures.

9.3.6 Discussion of the Test Results

Sentences involved in this discussion are listed below. All sentences used in the extraction documents can be found in appendix F.

Sentence 9.3 (Law)
Continental lawyers suggest lines of factual inquiry to the judge and, like their Anglo-American counterparts, advance legal theories and argue the law in accord with the interests of their clients.

Sentence 9.4 (Law)
When in the 16th century the Court of Chancery was established as the dispenser of "equity," the appropriate agent for litigation was called a solicitor, but the common-law serjeants and barristers secured the right of advocacy in that court.

Sentence 9.5 (Law)
In structuring these arrangements the lawyer is helping to particularize the legal rights of the parties.

Sentence 9.6 (Law)
In either system, if a lawyer loses his client's case, he may seek a new trial or relief in an appellate court.

Sentence 9.7 (Education)
This information on learner progress and achievement could be analysed by instructors and their students to plan possible future courses of action.

Sentence 9.8 (Law)
In structuring these arrangements the lawyer is helping to particularize the legal rights of the parties.

Sentence 9.9 (Law)
At trial he plays an active role in taking evidence, questioning witnesses, and framing the issues.

Figure 9.3: Sentences Used in the Discussion of Parse-tree Based Relation Extraction Test Result

Of the 4 false extractions in table 9.12, numbers 4 and 8 are from the sentences 9.3 and 9.4 in figure 9.3 respectively. From the meaning of the sentence, they should be viewed as correct extractions even though the human judges did not extract them.

Extraction number 6 in table 9.12 is from sentence 9.5. The standard extraction agreed by human extractors is $helps-to-particularize(lawyer, right)$. The extraction from the automatic extraction is not the same, but it is very close.

The algorithm extracts relation number 5 in table 9.12 from the conditional clause in sentence 9.6. However, a lawyer does not always lose cases, the extraction is not suitable to put into a knowledge base. This is the reason that the human extractors excluded it from their extractions. However it is a correct extraction of a relation between a subject and an object. This extraction indicates that extraction based only on syntactic analysis and parse tree is not enough for this task. Some semantic analysis is necessary. This problem also appears in the education domain. The false extraction from this domain is $plan(student, courses)$ from sentence 9.7. In this sentence, the sense of "courses" implied is not a key sense. It is also conditional. Both aspects mean it should not be in the KB but neither can be detected without semantic analysis.

Table 9.11 shows a big difference between the recall performances for the two domains. From section 9.3.1.3 it is already known that the parser used is far from ideal. This problem is worse for the document from the education domain than that from the law domain.

Table 9.14 shows the performances of the EngCG-2 on both domains. The number column is the number of trees in the parse forest produced by the parser for a sentence. Education and Law columns are the numbers of sentences that have a parse forest with the number of trees indicated in the number column. 12 sentences in the education domain have a complete parse tree and 4 sentences have a parse forest comprising of 2 trees. The percentage is cumulative, so 30% of sentences from the education domain have a parse forest comprising of less than two trees. It can be seen that the performance of the EngCG-2 parser on the document from the law domain is much better than that on the document from the education domain. 36% of the sentences from the law document have a complete parse, while for the education domain this figure is only 23%. 53% of the sentences from the law domain have a parse forest comprising of more than two separate parse trees, whereas this figure for sentences from the education domain is as high as 70%.

Table 9.14: Performances of the EngCG-2 Parser on the Documents from Two Domains

Tree Number	Education Domain	Percentage	Law Domain	Percentage
1	12	23%	21	36%
2	4	30%	6	47%
3	6	42%	10	65%
4	3	59%	7	77%
5	7	61%	4	84%
6	3	67%	2	88%
7	2	71%	1	89%
8	2	75%	5	98%
9	5	85%	0	98%
>9	8	100%	1	100%

This difference in performance of the parser on the two domains explains the difference of the extraction results in table 9.11. The different natures of the two documents may account for the difference in parser performance. The education document is a research paper which includes many very complicated sentence structures and usages of English, typical of a research paper. The law document is from an encyclopaedia and should therefore be easier to read. The sentence structures seem to be much simpler.

9.3.7 Extraction Relations Involving Only One Keyword

By relaxing the extraction restriction to extract relations containing only one keyword, more interesting relations can be extracted. For example, three extractions are suggested by the extraction algorithm from sentence 9.8 from the law domain. The human extraction from this sentence is $helps - to - particularize(lawyer, rights)$ which is actually the transform and combination of extraction 2 and 3 in table 9.15. It can be imagined that by post-processing, this human extraction can be generated. What is missing from the algorithm is that it did not attempt to consider extraction at phrase level. The algorithm is designed to work at word level. However, the humans take phrases into account when extracting relations. This also indicates that extraction at phrase level should be the subject of further research. Another interesting aspect shown by this experiment is that the algorithm does to some extent grasp the main relation chain in a sentence.

Table 9.15: Extractions from Sentence 9.8 and 9.9 by Loosing Extraction Condition

No	Extraction	Subject	From Sentence
1	structuring(lawyer, arrangements)	yes	8
2	helping(lawyer, particularize)	no	8
3	particularize(lawyer, rights)	yes	8
5	taking(, evidence)	yes	9
6	questioning(he, witnesses)	yes	9

Another example is from sentence 9.9 for which the two extractions are shown in Table 9.15. For extraction 6, the algorithm found the subject of the verb "questioning" to be "he" which refers to "lawyer" in a previous sentence. However, the final algorithm excludes it from the extraction because it contains only one keyword. If the algorithm could resolve the anaphoric reference, this would be a valid extraction. References between sentences results in the third kind of relation defined. This is another further research issue.

In sentence 9.9, the subject of participle "taking" is also "he". Unfortunately, the sentence is split into three trees shown in figure 9.4, with "the issues" in one tree, "in taking evidence" in another and the rest words of the sentence in the third. Because "he" and "taking" are not in the same tree, the algorithm failed to find the subject "he" for "taking".

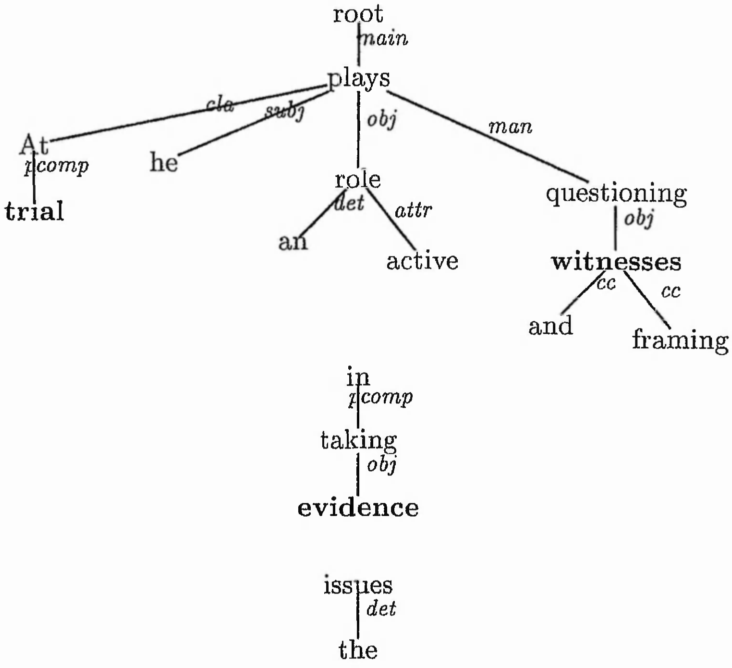


Figure 9.4: Parse Tree for Sentence 9.9

The algorithm also failed to extract *investigates*(*lawyer*, *evidence*). However, it suggests an extraction of $investigates(lawyer, facts)$. In this case, "facts" which is not a keyword and "evidence" which is a keyword, are both the object of the verb "investigate". The algorithm is successful in extracting the first but failed in the second. The reason is that the algorithm did not take into account the situation where a verb has two objects.

9.3.8 Problems with the Parse-tree Based Approach

The parse tree-based extraction algorithm produced high precision but low recall measures. The low recall measures are largely because of parser failure. It failed to produce a parse tree for most of the sentences. If a verb connects two keywords, it is a valid extraction. However, if for a sentence a complete parse is not available, but a parse forest instead and the keywords are located in different trees of the forest, the extraction will not produce a good result.

Another issue concerns how many components there should be for a verb. This depends on the sub-categorisation [188, 238] of the verb. This work extracts the subject and object of the verb in a relation. This is only a part of verb relations. If verb sub-categoration is taken into account, the extraction could be improved. For example, human extractors produce an extraction of *suggest lines of inquiry to(lawyer, judge)* from sentence 3 in section 9.3.6. The algorithm failed to extract this because "suggest" and "to the judge" are in separate parse trees, i.e. because of parser error (It suggest an extraction of $suggest(lawyer, lines)$). However, even if they were in the same parse tree, the algorithm would still fail. For the "lines of factual inquiry" is the indirect object of the "suggest" and the "judge" is the direct object, i.e. there are two objects in the sentence. The algorithm only extracts one of them because the other noun of a relation has to be the subject. If the verb sub-categorisation is taken into account, we would be able to extract $suggest(lawyer, linesoffactualinquiry, tojudge)$. The human extraction is actually a compromise for only extracting two components for a verb.

9.4 Tagger-based Approach

An alternative to using a parser as the basis of relation extraction is to use tags. Taggers are more reliable and robust than parsers. They produce more information for the words themselves but no information for the function of the words in a sentence.

9.4.1 Evaluation and Baseline

One of the main objectives of this approach is to increase the recall which is low in the parse tree based approach. The two main reasons leading to the low recall are failing to produce a parse tree for most of the sentences and not taking phrases into account. The latter not only reduces the number of extractions but also poses a danger of taking a single word out of the phrase context therefore changing its meaning and making the extraction not reflect its true meaning.

Thus, the tagger-based approach will work at phrase level, including verb groups and noun phrases.

As keyword extraction and human extraction against which the automatic extraction is evaluated are both at word level, the automatic relation extraction at phrase level introduces an evaluation problem, i.e. how to evaluate the automatic extractions. Keywords will be drawn from the noun phrases that serve as the subject and object of a verb and the extraction is composed the subject, the object and the verb. These composed extractions will be evaluated against human extractions.

Because a tagger does not provide the functions of words in a sentence, the extraction cannot be based on the words' functions in the sentence. Language rules have to be utilised directly without the help of parse trees.

An observation is that the subject of a finite verb is usually the nearest left noun and the object of a verb is usually the nearest right noun. This is the basis for a baseline evaluation of this method.

The tagger used is CLAWS [127] tagger version 4, one of the best taggers available. It consistently achieves 96-97% accuracy. The tagset used in CLAWS4 is C7 which can be found in appendix I.

9.4.2 An Overview of the Approach

From section 9.3 it can be seen that finding the subject for finite verbs is different from finding the subject for non-finite verbs. In the parse tree based approach, the parser determines which verb is finite and which is non-finite. This information is not available with the tagger approach. An alternative way of finding out if a verb is finite or not is therefore needed. Also finding subjects for a non-finite verb with the parse tree based method is heavily dependent on the grammatical functions that are from the parser. In the tagger-based approach, the method for finding subjects also needs changing. The basic unit to process in the tagger-based approach is the clause which contains a single finite verb.

The extraction algorithm is composed of the following tasks: 1. find noun phrase; 2. find verb groups; 3. judge verb type and voice (infinitive, participle, gerund and main verb) 4. find clause marks and marking clauses; 5. find subjects for verbs; 6. find Object for verbs; and 7. extract relations.

9.4.3 Algorithm for Relation Extraction

9.4.3.1 Identify Noun Phrases

Finding noun phrases is based on the generation rules shown in table 9.16. These rules were augmented from the rules in [40].

Rules 1 to 3 are common in works for identifying noun phrases. Rule 4 means that prepositions are considered as a part of noun phrases. Also two noun phrases connected by a coordinate conjunction are treated as one noun phrase. By doing this, it is easy to process more than one object of the same verb as required in sentence 9.9 in figure 9.3 . Cardinal, ordinal and quantity modifiers of noun phrases are not processed (Rule 2).

Table 9.16: Rules for Identifying Noun Phrases

No	Rules
1	$NP \rightarrow pronoun \vert propernoun \vert noun$
2	$NP \rightarrow (article)(cardinal)(ordinal)(quantity)(possessive)(ADJ)NP$
3	$ADJ \rightarrow adjective \vert ADJ$
4	$NP \rightarrow NP \bullet PP$
5	$PP \rightarrow preposition \bullet NP$
6	$NP \rightarrow NP \bullet coordinate - conjunction \bullet NP$

9.4.3.2 Identify Verb Groups and Their Type and Voice

A verb group includes the verb, auxiliary verbs, modal auxiliary, adverb and preposition adverbs. For example, in the sentence "the attorney is not permitted to concurrently represent two or more clients", the two identified verb groups will be "is not permitted" and "represent". For the verb "represent", the "concurrently" is not included in the group because adverbs are only processed after a verb.

The methods used in finding subjects and objects for a finite verb, an infinitive and a participle are different. It is necessary to determine the type of a verb i.e. whether it is a finite or a infinitive or a participle. All these verb forms have different tenses and voices.

9.4.3.3 Marking Clauses

Finding clause markers To separate a sentence into clauses, the sentence is marked with clause boundary markers. At this stage, the markers are candidate clause boundaries. The markers can be recognised by the patterns shown in table 9.17.

A pattern is composed of two kinds of elements. One kind is a literal string, like the ";" in pattern 1. The other kind is a tag, expressed as $< tag >$, like the $< CC >$ in pattern 3. A literal element in a pattern can only be matched to the same word in a sentence, whereas a tag element can match any word with the same tag. All the patterns are composed of just one element. However, for complicted sentence markers like "in order to", one element is not enough. The way the pattern is designed can by used to process complicated sentence markers.

At the stage of finding clause markers for a sentence, all the patterns in table 9.17 are applied to the sentence to find as many candidate markers as possible. For ease of processing, the sentence is segmented into clauses. Applying the patterns to sentences 9.10 and 9.11 in figure 9.6, the following markers are found and their positions recorded, forming a series of marker ranges. For sentence 9.10, the ranges will be:

```
1 (He) 7 (because) 14 (,) 15(or) 16(,) 32 (and) 35 (which) 38 (.)
1 (If) 10 (,) 18 (;) 19 (and) 20 (,) 21 (if) 28 (,) 36 (.)
```

Table 9.17: Patterns Used to Identify Candidate Clause Markers

No	Pattern	Explanation	From Example
1	;	literal	
2	,	literal	
3	$< CC >$	coordinating conjunction	and, or
4	$< CB >$	adversative coordinating conjunction	but
5	$< CS >$	subordinating conjunction	if, because, unless, so for
6	$< CSA >$	"as" used as a conjunction	as
7	$< CSN >$	"than" as a conjunction	than
8	$< CST >$	"that" as a conjunction	that
9	$< CSW >$	"whether" as a conjunction	whether
10	$< DDQ >$	wh-determiner	which, that
11	$< DDQGE >$	wh- determiner, genitive	whose
12	$< DDQV >$	wh- ever determiner	whichever, whatever
13	$< PNQO >$	objective wh-pronoun	whom
14	$< PNQS >$	subjective wh-pronoun	who
15	$< PNQV >$	wh-ever pronoun	whoever
16	$< RGQ >$	wh- degree adverb	how
17	$< RGQV >$	wh-ever degree adverb	however
18	$< RRQ >$	wh- general adverb	where, when, why, how
19	$< RRQV >$	wh-ever general adverb	wherewver, whenever

Marking the Clauses At this stage, each clause is marked with its start position and end position in the sentence. Each clause contains only one finite verb. This marking process starts from the first range and checks if there is a finite verb in this range (may contain more than one finite verbs). If there is a finite verb in the current range, this range forms a clause, otherwise it is merged with the last range whether the last is a clause or not.

After all the clauses are processed, further checks are done to make sure each clause contains only one finite verb. If a clause contains more than one, this range of the clause will be split into as many ranges as the number of finite verbs in the clause. After this stage, each clause contains only on finite verb.

Applying the aforementioned procedure to sentences 9.10 and 9.11 in figure 9.6, marked clauses are shown in figure 9.5.

The marker that separates two clauses will be the end of one and the beginning of the other. The ranges of the clauses are saved for later use.

It can be seen from the examples that the marking algorithm can successfully mark the boundaries of compound sentences and complicated sentences. Note that the infinitive "to reveal" in sentence 10 and "to represent" in sentence 11 are not marked as separate clauses, because they are not finite verbs.

```
1,6 He may seek a summary dismissal because
7,15 because the opponent evidently has no case, or
16,34, through discovery proceedings he may force the other side to reveal more fully the issues and facts on
35,38 which it relies.

1,19 If incidental disputes concerning procedure have to be litigated
10,20, he is likely to conduct the proceedings; and,
21,27 if the procedure includes a pretrial conference
28,36, he is likely to represent the client.
```

Figure 9.5: Clause makers for sentences in 9.10 and 9.11 in figure 9.6

9.4.4 Finding Subjects and Objects

A clause usually has it own subject which is usually the noun phrase preceding the verb. Thus after marking the boundaries of clauses, it is simple to find the subject for finite verbs. This is the reason for marking clause boundaries. The clause boundary ranges and the range of verb groups recorded in the process of marking clause boundaries and finding verb groups are used to locate the clause in which the verb group appears. The verb group ranges recorded in the process of finding noun phrases and verb group ranges can be used to find noun phrase preceding the verb group.

There are some clauses which have not got their own subjects, such as a relative clauses and the second clause of a multiple verb sentence. By checking whether the first word of a clause has no subject within it ("that", "who" or "which"), we can find out if it is a relative clause. For this kind of clause, the subject for the verb is the antecedent of the clause.

If it is not a relative clause and is not the first clause, it will take as its subject, the subject of the nearest finite verb on its left which has a subject. The assumption is that it is a multiple verb sentences as in “the man is eating, drinking and talking”.

If a verb still has no subject after the above process, the nearest noun phrase on its left is assigned.

If a non-finite verb has a noun phrase before it and it is not the subject of the clause, this noun phrase is assigned to the non-finite verb as its subject. If there is no such noun group, the subject of the clause should be the subject of the non-finite verb. If it still has no subject after the above process, again the nearest noun phrase on the left is assigned.

9.4.5 Test Results of Extraction Using the Algorithm

This algorithm is tested using the same set of sentences as in the parse-tree based algorithm. After the pre-processing, i.e. finding the subjects and objects for verbs, the elements for a verb have been found. This stage just checks if the subject and object of a verb both contain a keyword. If it passes the check, this will be an extraction.

Table 9.18 shows the results of the algorithm applied to the two sentence sets given to human extractors. Correct extractions from the education and the law domains are listed in tables 9.19 and 9.20 respectively. Full extractions can be found in lists 1 and 2 in appendix J.

Table 9.18: Relation Extraction Results from Tagger-based Method on the Two Domains

Domain	Targets	Extractions	Correct Extractions	Recall	Precision
Education	14	48	5	36%	10%
Law	23	47	12	52%	26%

Table 9.19: Correct Extractions of Tagger-base Method from the Education Domain

No	Extraction
1	is currently being implemented (Mediated Learning, colleges and universities around the 22 country)
2	is (Mediated Learning, a form of technology-mediated instruction)
3	complete (a significant impact on the number of students, a course or sequence of courses)
4	can take on (instructors, students per class)
5	enter (students, college)

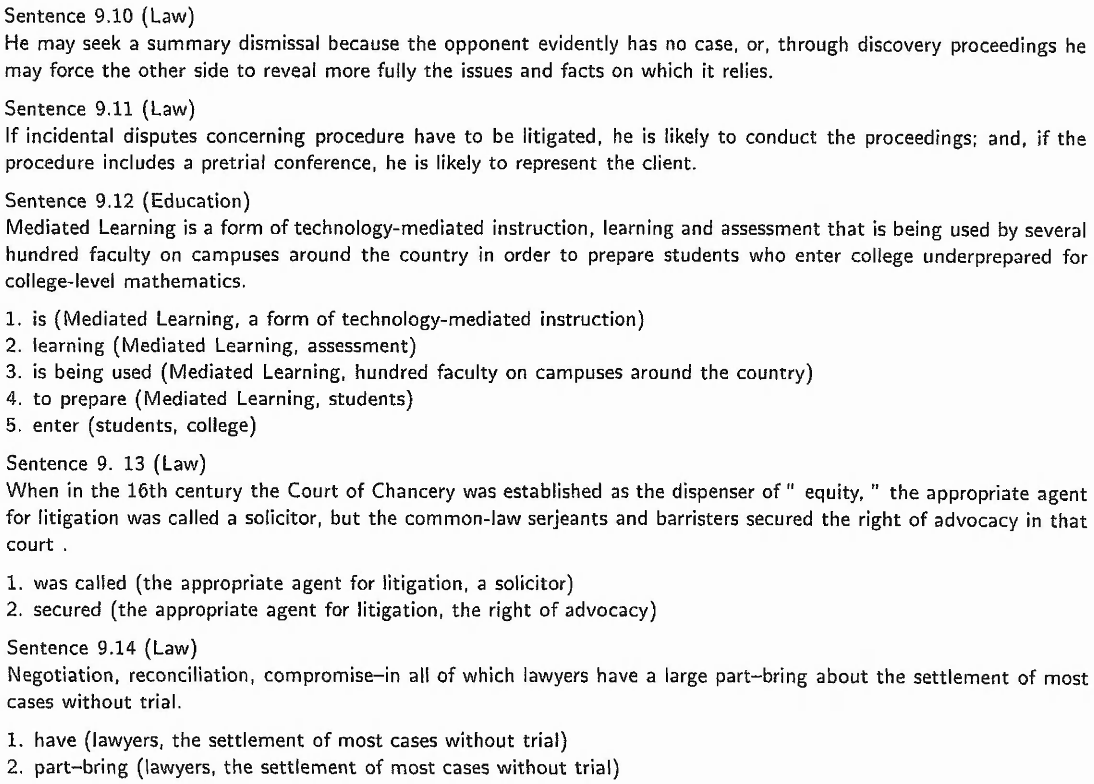
Table 9.20: Correct Extractions of Tagger-base Method from the Law Domain

No	Extraction
01	is to apply (The primary function of the profession and practice of law, the law in specific cases)
02	investigates (a lawyer, the facts and the evidence)
03	conferring (a lawyer, his client)
04	interviewing (a lawyer, witnesses)
05	suggest (Continental lawyers, lines of factual inquiry to the judge)
06	argue (Continental lawyers, the law)
07	is helping to particularize (the lawyer, the legal rights of the parties)
08	attended (the procurator, the formal aspects of litigation)
09	gave (Advocates, direct advice to clients)
10	presented (Advocates, oral arguments in court)
11	swallowed up (the procurators, the advocates)
12	is to fight (The duty of the advocate, the rights of his client)

9.4.6 Discussion of Test Results

We can see that the recall improves, however, precision drops dramatically. As stated in the evaluation section of this approach, "correct" here means that the subject of the extraction contains the subject keyword of a human extraction and the object contains the object keyword of the same extraction.

There are five extractions suggested for sentence 9.12 in figure 9.6. The first is a correct extraction. The second one is wrong. The third and fifth are correct extractions, but they are not in the human extraction list. Should the last two be considered as correct? This is a hard question. The fourth express the aim of employing the "mediated learning", it may not always true. This is the same



Sentence 9.10 (Law)
He may seek a summary dismissal because the opponent evidently has no case, or, through discovery proceedings he may force the other side to reveal more fully the issues and facts on which it relies.

Sentence 9.11 (Law)
If incidental disputes concerning procedure have to be litigated, he is likely to conduct the proceedings; and, if the procedure includes a pretrial conference, he is likely to represent the client.

Sentence 9.12 (Education)
Mediated Learning is a form of technology-mediated instruction, learning and assessment that is being used by several hundred faculty on campuses around the country in order to prepare students who enter college underprepared for college-level mathematics.

1. is (Mediated Learning, a form of technology-mediated instruction)
2. learning (Mediated Learning, assessment)
3. is being used (Mediated Learning, hundred faculty on campuses around the country)
4. to prepare (Mediated Learning, students)
5. enter (students, college)

Sentence 9. 13 (Law)
When in the 16th century the Court of Chancery was established as the dispenser of " equity, " the appropriate agent for litigation was called a solicitor, but the common-law serjeants and barristers secured the right of advocacy in that court .

1. was called (the appropriate agent for litigation, a solicitor)
2. secured (the appropriate agent for litigation, the right of advocacy)

Sentence 9.14 (Law)
Negotiation, reconciliation, compromise–in all of which lawyers have a large part–bring about the settlement of most cases without trial.

1. have (lawyers, the settlement of most cases without trial)
2. part–bring (lawyers, the settlement of most cases without trial)

Figure 9.6: Sentences Used in the Discussion of Tagger Based Relation Extraction Test Result

problem as in the extraction $loses(alawyer, hisclient'scase)$ which happens again in this approach. This shows the difficulty of relation extraction. Another example of this kind is in the law domain sentence 9.13. The first extraction is in a sentence where the event happened at a certain time. If we can extract time at the same time, it would be correct. Because the verbs in the relation are in the past tense, this gives the reader of the extraction an impression that these events happened some time in the past, although the exact time is unknown from this extraction. From another point of view, they should be considered as correct. These extractions are not included as true when calculating the extraction measures. There are other similar examples.

If such extractions are considered as correct, the results would be as shown in table 9.21.

Table 9.21: Re-calculated Results from Tagger-based Method on Two Domains

Domain	Targets	Extractions	Correct Extractions	Recall	Precision
Education	24	48	14	58%	29%
Law	26	47	15	58%	32%

The low precision is from the wrong assignment of subjects and objects for verbs. This is the major work of the extraction algorithm. Finding subjects and objects in this approach is not as fine-tuned as in the parse tree based method. It is because no grammatical functions are available. However, there is still much that can be done to improve it. The rules for finding subjects for infinitive and participles can be used to improve the accuracy for finding subjects. The following issues are not very hard to process, such as one verb with multi-subjects and identification of passive voice. The voice of all forms of verbs has already been identified, but this information was not used in the extraction. A simple exchange of subject and object will eliminate the false extractions involved in the passive voice. However, other problems are very hard to solve, e.g. relative clauses with the relative pronoun omitted and anaphoric resolution.

Another major problem of this approach is again the number of elements a verb has in a sentence. The target is three-tuple relations: subject, verb and object. Therefore the design of the algorithms are biased towards this kind of relation. This forces a verb to have an object, whereas in some case it does not have one. However, in some cases it has two. This biased design introduces false extractions. This problem was also evident in the parse tree based approach, but it is worse in the tagger-based approach. The reason is that in the parse tree based approach, the parser finds the object based on a more detailed analysis of the sentence. In that method, the extraction algorithm does not need to find objects for verbs, it simply consults the parse tree to find out if a verb has an object. If it has, the algorithm uses it. If it does not have one, the algorithm does not extract this verb. In the tagger-based method, an object is forced for all verbs, using a simple strategy because there is no other source of information. Thus, further work into the use of the tagger-based approach should make the use of verb subcategorisation a priority.

Table 9.21 shows that as with the parser-based approach, the performance on the law domain is better than on the education domain. In both cases, this is ascribed to the difference in the types of document (encyclopedia and academic paper respectively).

One interesting aspect of the extractions from this approach is that they are more understandable and keep the meaning in the original sentence.

The CLAWS tagger produces about 3-4% errors. This is a minor error rate, but some false extractions are the results of error from the tagger. An example of this is shown in sentence 9.14 and the two extractions from it. The tagger tagged "part-bring" as a single verb instead of a noun ("part") and a verb ("bring") separated by a hyphen. The correct tags would result in $have(lawyers, a - lager - part)$ and $bring(lawyers, the - settlement - of - most - cases - without - trial)$ when the algorithm is applied. The former one would not be extracted and the latter is a valid extraction.

9.4.7 Baseline of the Tagger-based Approach

The baseline is to take the nearest left noun as the subject of a finite verb and the nearest right noun as the object of the verb. If both nouns are keywords, these two nouns and the verb are extracted. Applying this simple process to the sentences used in the section 9.4.5 produce the baseline results shown in table 9.22. Extractions are also listed in lists 3 and 4 in appendix J.

Comparing the baseline results with the results in table 9.21, it can be seen that identifying subjects and objects for verbs in tagger-based approaches increases recall and precision significantly. The precision measure is increased by 14% for the education domain and by 15% for the law domain.

Table 9.22: Baseline Results of Tagger-based Method on Two Domains

Domain	Targets	Extractions	Correct Extractions	Recall	Precision
Education	14	33	5	36%	15%
Law	23	41	6	26%	15%

Recall is increased by 22% for the education domain and by 32% for law domain.

9.5 Summary

9.5.1 Evaluation and Comparison of the Two Methods

This work on relation extraction research is novel. No attempt to extract named verb-noun relations has been reported. As indicated in the literature review, work in this area is mostly on taxonomy relations or extracting a handful of pre-defined named relations. As a new research direction, 52% recall and 26% precision for the law document is an encouraging start.

In the attempt to extract relations from text, two approaches have been investigated. Both have advantages and disadvantages. The parse tree based method produces high precision and low recall. However, this is not the failure of the method itself. The main reason is parser failure. This reflects the current limitation of the parser technique. It is not advanced enough to process texts in the real world. Three different parsers were tested and none of them was suitable. Even so, the parse tree is still a great help in finding objects for verbs. A problem of this method is that it contains too much detail not needed by the extraction task and lacks high level information, such as phrase or constituent level information, which are of most interest for relation extraction. From the experiments it seems that shallow parsing is more suitable for the relation extraction task. A disadvantage of the parse tree based approach is that it requires more time to process. However, this is not a major problem.

The tagger-based approach produces high recall and low precision. This method is more robust than the parser based method. However, extracting relations based on tagged text is similar to writing a parser. There are too many linguistic details to be handled by the extractor.

The style of writing may have great impact on the results of relation extraction. The experiments have shown that both of the approaches perform better for one of the documents compared with the other.

9.5.2 Further Research Issues

A more robust parser should be used in further research. Parse-tree based method needs two kinds of information from the parser: the parse tree (for the functions of words in a sentence); and the tags of words in sentences. The basic information a parse provides is the function of words in a sentence. To find out the function of a word in a sentence, a parser has to find out the part of speech of the word, i.e. the tag of the word. It is reasonable to assume that these two kinds of information are provided by any parsers. That is to say, this method will work with other parsers.

Further research for relation extraction should also include more rigorous experiments at the phrase level. Using verb sub-categorisation information in the relation extraction should be a high priority. Other further research issues are reference resolution within and between sentences and semantic analysis.

Chapter 10

Conclusions and Further Work

10.1 Conclusions on Concept Extraction

TCE is a new research aim [239, 240, 241]. No similar attempt has been reported. It is different from IE and IR. IE concentrates on specific events in certain domains. In terms of extraction targets, TCE is broader. The purpose of IR is to describe documents i.e. it is more interested in the documents themselves, not the thematic concepts in the documents.

As well as the focus of automatically generating knowledge bases, TCs and relations between TCs can be used in various other applications, such as intelligent IR, Internet search based on content and document summary. It can also contribute to IE and IR.

Applying ANN to automatic concept extraction is novel. In TCE using ANN, the scope of the domain is in effect determined by the selection of seed words and keywords that are used in the ANN training. The number of seed words is always intended to be very small, if not just one. This word(s) defines the heart or centre of the domain. Thus only one word is required, though more than one is possible. In the case of the education and the law domain, the number of seed words was one. The ideal number of training words (concepts) is variable, depending on how precisely or closely the domain is to be defined. It is a list agreed by the domain authors. In the education case it was 111, including both keywords and non-keywords. However, the main point is that the evaluation of performance involves words that are classified according to the same criterion as the classification in the training data. For example, in the 'farm animals' domain, if the two keywords used for training are 'goat' and 'pig', it is reasonable to expect 'cow' to be classified as 'farm animal', but not 'trout'. Whereas, if 'salmon' was in the training as a keyword, 'trout' should then be recognised as a keyword. In summary, there is no fixed number or proportion of training words (concepts). It is simply a question of semantic consistency between training and testing.

It has been shown that thematic concepts can be automatically extracted from text using an ANN plus a lexical semantic resource. Results in the education domain show good natural and pure generalisation for non-keywords at 84% and 82% respectively and reasonable generalisation for keywords (62% for natural and 47% for pure). The baseline comparison for the ANN method in section 4.4.8 shows that the ANN method adds value to the external lexicon. The CWC measure shows that both the ANN and stemming methods work much better than chance.

A stemming analysis method which uses a different kind of information from the same lexicon has also been attempted at sense-level and word-level. Sense level stemming analysis produces 89%, 76% and 77% for KWs, NKWs and overall respectively. The measures on word level are 94%, 49% and 60%.

As concluded in section 5.2.3.3, the ANN alone produces best result for NKWs and overall. Word level stemming analysis alone is the best for identifying KWs, while sense level analysis provides the most balanced results between KWs and NKWs. Since ANN and stemming analysis use different kinds of information, they are not directly comparable.

The stemming analysis reveals that the word definition information in WordNet is a good information source, independent of the path information. Research into whether the word definition information can be used with the ANN to improve its performance would be useful.

The portability of the ANN and stemming methods are tested on a new domain (Law). The results show that the ANN itself in not transferable. However, the ANN method is transferable with consistent performance between domains. The stemming analysis approach also transfers reasonably well between domains, although not as well as the ANN method.

The performance of the ANN approaches should be improved by using sense-level path information, definition information and finer grain category information.

In TCE, an external semantic lexicon is used to provide semantic information for differentiating between KWs and NKWs. Two kinds of information are used in training the ANN: category and path. The overall hypothesis of using paths between a word and the seed word(s) is that KWs should have close relationships that are distinct and characteristic of the seed word-KW relationship. From the weight analysis, this hypothesis proves correct. One aspect of the close relationship between KWs and the seed word is the path length. The shorter a path, the more likely it is to be a PIP, which encourages a word to be classified as a KW. However, path length in not the only feature of the close relationship, because the baseline experiment in chapter 4 has already shown that identifying KWs simply based on path length does not work.

The hypothesis of using category information is that words from the same category of a KW should also be classified KWs. This also proves correct. However, because the categories used are crude, the category information draws more false KWs than the path information.

Weight analysis also reveals that different relations from WordNet influence the ANN in different ways. Synonym, holonym and hyponym are more helpful in identifying KWs, while coordinate and coordinatee often mislead the ANN. Intuitively, meronym should also provide positive information to the ANN. However, no conclusion can be made on this kind of relation because there are no examples of this relation in the training data (it is a sparse relation in WordNet).

The ANN was trained using all the orderings of paths. All the orderings are also used in the test. It is found that all the test patterns of a word with different ordering have the same result, i.e.

all the test patterns are identified as key patterns or all are identified as non-key patterns. Two conclusions can be drawn. Firstly, the purpose of using multi-training patterns for a single word to avoid the ANN being biased to a specific ordering of paths is fulfilled. Secondly, in an online version of the algorithm, only one test pattern for a word with any ordering of paths is enough to classify the word. It is not necessary to test all the test patterns.

Although WordNet is a valuable online lexicon, it has shown some serious limitations. Firstly, it lacks of sense level path information. Secondly, there is no close relation between the word definition and the path information. Thirdly, some of the relationships between words are not completely realised. For example, information on meronyms and holonyms is sparse. These suggest that another lexicon should be considered in future research.

10.2 Conclusions on Relation Extraction

This work on relation extraction research is novel. A very important relation type, named verb relation, is targeted. No attempt to extract named verb-noun relations has been reported. As indicated in the literature review, work in this area is mostly on taxonomy relations or extracting a handful of pre-defined named relations. As a new research direction, 52% recall and 26% precision for the law document is an encouraging start.

In the attempt to extract relations from text, two approaches have been investigated. Both have advantages and disadvantages. The parse tree based method produces high precision and low recall. However, this is not the failure of the method itself. The main reason is parser failure. This reflects the current limitation of the parser techniques is not advanced enough to process texts in the real world. We tested three different kinds of parsers, none of them is suitable. Even so, the parse tree is still gives a great help in the effort to find objects for verbs. A problem of this method is that it contains too much detail not needed by the extraction task and lacks high level information, such as phrase or constituent level information, which are of most interest for relation extraction. From the experiments it seems that shallow parsing is more suitable for the relation extraction tasks. A disadvantage of the parse tree based approach is that it requires more time to process. However, this is not a major problem.

The tagger-based approach produces high recall and low precision. Extracting relations based on tagged text is similar to writing a parser. There are too many linguistic details to be handled by the extractor.

The style of writing may have great impact on the results of relation extraction. The experiments have shown that both of the approaches perform better for one of the documents compared with the other.

10.3 Future Work

It is already known that taking only single nouns as concept extraction targets is sometime not accurate and sufficient enough. Phrases must be considered to accurately express concepts. Thus further research issues in concept extraction should include extracting key phrases not just keywords.

It is also seen that category information is important to ANN in recognising keywords because this information helping ANN to group patterns into different categories. However, the 25 top-level categories used are too coarse. Using fine granulated category information is thus worth more investigation.

From the comparison of results from sense level and word level stemming analysis, it is obvious that sense level path would be more accurate that word level path and would be help the ANN in differentiating keywords and non-keywords. However, the limitation of WordNet has prevented the concept extraction experiment using sense level paths. Thus, employing another semantic lexicon is another future research issue in concept extraction. An ideal lexicon for future work should have rich sense level information. These kinds of information would provide more accurate relationship between KWs and the seed word(s), and would help the ANN to reject coordinate/coordinatee relationships as they often provide misleading information. Rich holonym and meronym relations would also help in improving performance.

The major problem with parse-tree based RE is the parser. It failed to parse most of the sentences from both domains. Thus, a suitable parser should be used in future research. The parser employed for future research should be robust enough to parse most of the sentences for different domains. It should not be sensitive to different writing styles. Being able to parse at phrase level is also important because TCE and RE at phrase level is one of the major tasks in future research. A robust partial parser should be suitable.

All the information needed (the function of a word in the sentence and part of speech tag of a word) by the parse-tree based method should be available from any parser. Thus, the parse-tree based RE should work with any reasonable parser.

Further research for relation extraction should include more rigorous experiments at the phrase level. Using verb sub-categorisation information in the relation extraction should be a high priority. Other further research issues are reference resolution within and between sentences and semantic analysis.

Bibliography

[1] Abney, S. P., Schapire, R. E. and Singer, Y. 1999. *Boosting Applied to Tagging and PP Attachment.* In Peoceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99), College Park, MD, 38-45. ACL.

[2] Abney, S. P. 1991. *Parsing by Chunks.* In Berwick, R. C., Abney, S. P. and Tenny, C. (eds.), Principle-Based Parsing: Computation and Psycholinguistics, 257-278. Kluwer, Dordrecht.

[3] Agichtein, E. and Gravano, L. *Snowball: Extracting Relations from Large Plain Text Collections.* Proceeding of the 5th ACM international conference on Digital Libraires, June 2000.

[4] Agichtein, E., Eskin, E. and Gravano, L. *Combining Strategies for Extracting relations from text collections.* Proceedings of the 2000 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD 2000), May 2000.

[5] Agirre, E. 1999. *Formalization of Concept-Relatedness Using Ontologies: Conceptual Density.* Ph.D Thesis, Department of Computer Science, the University of Basque Country, Spain. 1999

[6] Agirre, E. and Rigau, G. 1996. *Word sense disambiguation using conceptual density.* In Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, 1996.

[7] Aliod, D., Berri, J. and Hess, M. 1998 *A real world implementation of answer extraction.* In Proceedings of the 9th International Workshop on Database and Expert Systems, Workshop: Natural Language and Information Systems (NLIS-98), 1998.

[8] Allen, J. 1995. *Natural Language Understanding.* Benjamin Cummings, Menlo Park, CA.

[9] Allen, R. B. and Riechen, M. E. 1988. *Anaphora and Reference in Connectionist Language Users.* In Proceedings of the international Computer Science Conference. Hong Kong.

[10] Alshawi, H. 1987. *Processing Dictionary Definitions with Phrasal Pattern Hierarchies.* Computational Linguistics 13 3-4 (1987).

[11] Andrade, M.A. and Valencia *A Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts,* In proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology 25-32. Halkidiki, Greece: AAAI Press, 1997.

[12] Aone, C., and Bennett, S.W. 1995. *Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies.* In Proceedings of the Thirty-third Annual Meeting of the Association for Computational Linguistics, 122-129. Somerset, N.j.: Association for Computational Linguistics.

[13] Argamon, S., Dagan, I. and Krymolowski, Y. 1998. *A Memory-Based Approach to Learning Shallow Natural Language Patterns.* In Proceedings of the International Conference on Computational Linguistic (in Year 1998).

[14] Armstrong-Warwick, S. 1993. *Preface. Computational Linguistics.* 19(1):iii-iv.

[15] Academic Systems Mediated Learning Library, *Mediated Learning: A New Model of Networked Instruction and Learning,* http://www.academic.com/library/articles/ mllibrary.html, Accessed on 24, May, 1999.

[16] Atwell E, Demetriou G, Hughes J, Schiffrin A, Souter C, and Wilcock S. 2000. *A comparative evaluation of modern English corpus grammatical annotation schemes.* ICAME Journal, volume 24, pages 7-23, International Computer Archive of Modern and medieval English, Bergen. ISSN: 0801-5775.

[17] Baker, J. K. 1979. *Trainable Grammars for Speech Recognition.* In Klatt, D. H. and Wolf, J. J. (eds.), Speech Communication Papers for the 97th Meeting of the Acoustical Society of American, 547-550.

[18] Bar-Hillel, L. 1964. *Language and Information.* Reading, Mass, Assidon-Wesley.

[19] Beeferman, D., Berger, A. and Lafferty, J. *Statistical Models for Text Segmentation.* Machine Learning, special issue on Natural Language Learning, C. Cardie and R. Mooney eds., 34(1-3), pp. 177-210, 1999.

[20] Beeferman, D., Berger, A. and Lafferty, J. *Text Segmentation Using Exponential Models.* Second Conference on Empirical Methods in Natural Language Processing. Providence, RI. (1997)

[21] Berg, G. 1992. *A Connectionist Parser with Recursive Sentence Structure and Lexical Disambiguation.* In Proceedings of the Tenth National Conference on Artificial Intelligence, 32-37, MIT Press, Cambridge, MA, 1992.

[22] Berland, M. and Charniak, E. *Finding Parts in Very Large Corpora.* Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pp 57-64 (1999)

[23] Berwick, R. 1985. *The Acquisition of Syntactic Knowledge.* MIT Press: Cambridge, Massachusetts.

[24] Blank, D. S., Meeden, L. A. and Marshall, J. B. 1992. *Exploring the symbolic/subsymbolic Continuum: a Case Study of RAAM.* In J. Dinsmore, (ed.), The symbolic and Connectionist Paradigms: Closing the Gap, 113-148, Erlbaum, Hillsdale, NJ, 1992.

[25] Blum, A. and Mitchell, T. 1998. *Combining Labeled and Unlabeled Data with Co-Training.* In Proceedings of the 1998 Conference on Computational Learning Theory, July 1998.

[26] Bowden, P. R. 1999. *Automatic Glossary Construction for Technical Papers.* PhD thesis. Computing Department, Nottingham Trent University, UK.

[27] Bowden, P. R. and Edwards, M. 1996 *Knowledge Extraction from Corpora for Pedagogical Applications.* TALC '96 (Teaching and Learning Corpora), Lancaster University, England, 9-12 August 1996

[28] Bowden, P. R., Halstead, P. and Rose, T. G. *Extracting Conceptual Knowledge from Text Using Explicit Relation Markers.* Proceedings of 9th European Knowledge Acquisition Workshop, Nottingham, United Kingdom, May 1996.

[29] Booth, B. 1985. *Revising CLAWS.* ICAME Journal 9:29¨C35.

[30] Brachman, R. J. 1979. *On the Epistemological Status of Semantic Networks.* In N. V. Findler (Ed.), Associative Networks: Representation and Use of Knowledge by Computer. New Yors: Academic Press, 3-50.

[31] Brandow, R., Mitze, K., and Rau, L.R. 1995. *The automatic condensation of electronic publications by sentence selection.* Information Processing and Management, 31 (5), 675-685.

[32] Brent, M. R. 1993. *From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax.* Computational Linguistics, 19(2): 243-262.

[33] Brill, E and Mooney R. *An Overview of Empirical Natural Language Processing.* AI Magazine, Volume 18, No. 4 Winter 1997.

[34] Brin, S. *Extracting Patterns and Relations from the World Wide Web.* Proceedings of the 1998 international workship on the Web and database (WebDB'98), Mar. 1998.

[35] Britannica.com *The Profession and Practice of Law,* Britannica Encyclopedia CD, Multimedia Edition, 1999.

[36] Brown, P., Della Pietra, s., Della Pietra, V. and Mercer, R. 1991. *Word sense disambiguation using statistical methods.* In Proceedings of the 29th Meeting of the Association for Computational Linguistics (ACL-91), pages 264-270, Berkley, C.A., 1991.

[37] Buchholz, S. and Daelemans, W. *Complex Answers: A Case Study using a WWW Question Answering System.* Natural Language Engineering, 2001.

[38] Buchholz, S., Veenstra, J. and Daelemans, W. *Cascaded grammatical relation assignment.* In Pascale Fung and Joe Zhou, editors, Proceedings of EMNLP/VLC-99, pages 239-246. ACL, 1999.

[39] Burnard, L. (ed.) 1995. *User Reference Guide for British National Corpus (Version 1.0).* Oxford University Computing Services.

[40] Byrd, R. and Ravin, Y. *Identifying and Extracting Relations from Text.* In NLDB'99, 4th International Conference on Application of Natural Language to Information Systems 1999.

[41] Califf, M. E. and Mooney, R. J. 1997. *Relational Learning of Pattern-Match Rules for Information Extraction.* In Proceedings of the ACL Workshop om Natural Language Learning, 9-15. Somerset, N.J.: Association for Computational Linguistics.

[42] Calzolari N., Bindi R. 1990. *Acquisition of lexical information from a large textual Italian corpus.* in Proceedings of COLING-90, Helsinky, Finland.

[43] Cardie, C. 1994. *Domain-specific Knowledge Acquisition for Conceptual Sentence Analysis.* Ph.D Thesis, University of Massachusetts, Amherst, MA.

[44] Cardie, C. 1993. *A Case-Based Approach to Knowledge Acquisition for Domain-Specific Sentence Analysis.* In Proceedings of the Eleventh National Conference on Artificial Intelligence, 798-803. Menlo Park, Calif.: American Association for Artificial Intelligence.

[45] Chalmers, D. J. 1990. *Syntactic Transformations on Dictributed Representations.* Connection Science, 2(1 and 2): 53-62,1990.

[46] Charniak, E. *Statistical Techniques for Natural Language Parsing.* AI Magazine, Volume 18, No. 4 Winter 1997.

[47] Chomsky, N. 1965. *Aspects of the Theory of Syntax.* MIT Press: Cambridge, Massachusetts

[48] Chomsky, N. 1957. *Syntactic Structure.* Mouton: The Hague, Holland.

[49] Christiansen, M. H. and Chater, N. 1999. *Toward a Connectionist Model of Recursion in Human Linguistic Performance.* Cognitive Science, 23, 157-205.

[50] Christiansen, M. H. 1994. *Infinite Language, Finite Minds: connectionism, Learning and Linguistic Structure.* Ph.D Thesis, University of Edinburgh.

[51] Christiansen, M. H. and Chater, N. 1994. *Generalization and Connectionist Language Learning.* Mind and Language 9: 273-287.

[52] Church, K., and Mercer, R. L. 1993. *Introduction to the Special Issue on Computational Linguistics Using Large Corpora.* Computational Linguistics 19(1):1-24.

[53] Church K.W., Gale W., Hanks P., Hindle D. 1989. *Parsing, word associations and typical predicate-argument relations.* in Proceedings of the International Workshop on Parsing Technologies, Carnegie Mellon University, Pittsburg, PA, pp. 103-112.

[54] Cimino, J. J. and Barnett, G. O. *Automatic Knowledge Acquisition from MEDLINE,* Center for Medical Informatics, Columbia University Columbia-Presbyterian Medical Center, New York, New York Laboratory of Computer Science, Harvard Medical School Massachusetts General Hospital, Boston, Massachusetts

[55] Collier, R. 1994. *An Historical Overview of Natural LanguAGE Processing Systems that Learn.* Artificial Intelligence Review 8: 17-54, 1994.

[56] Collins, M. J. *A new statistical parser based on bigram lexical dependencies.* In 34th Annual Meeting of the Association for Computational Linguistics. University of California, Santa Cruz, California, USA, June 1996.

[57] Connolly, D., Burger, J. D. and Day, D. S. 1994. *A Machine Learning Approach to Anaphoric Reference.* In Proceedings of the International Confernece on New Methods in Language Processing (NeMLaP). ACL.

[58] Cottrell, G. W. and Plunkett, K. 1994. *Acquiring the Mapping from Meaning to Sounds.* Connection Science, 6: 379-412, 1994.

[59] Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K. and Slattery, S. 2000. *Learning to Construct Knowledge Bases from the World Wide Web.* Artificial Intelligence, 118(1-2): 69-113.

[60] Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K. and Slattery, S.. *Learning to Extract Symbolic Knowledge from World Wide Web.* Proceedings of 15th National Conference on Artificial Intelligence (AAAI-98)

[61] Croft, W. B. and Turtle H. R. 1992. *Text Retrieval and Inference.* In P. S. Jacobs (eds.) Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval. ISBN 0-8058-1189-3.

[62] Cycorp, Inc, USA, *CYC Application Example Page,* http://www.cyc.com/applications.html, Accessed on 30, July, 1999.

[63] Daille, B. 1995. *Combined Approach for Terminology Extraction: Lexical Statistics and Linguistic Filtering.* UCREL papers no. 5., UCREL, Dept. of Linguistics and Modern English Language, University of Lancaster.

[64] Demetriou G and Atwell E. 2001. *A domain-independent semantic tagger for the study of meaning associations in English text.* In Harry Bunt, Ielka van der Sluis and Elias Thijsse (editors), Proceedings of the Fourth International Workshop on Computational Semantics (IWCS-4) pp.67-80. Tilburg, Netherlands. ISBN: 90-74029-16-7.

[65] Department of Cognitive Science, Princeton University, USA, *WordNet Home Page,* http://www.cogsci.princeton.edu/~wn/obtain/, Accessed on 2, June, 1999.

[66] Duda, R. O. and Hart, P. E. 1973. *Pattern Calssification and Scene Analysis.* John Wiley and Sons, New York.

[67] The EAGLES Lexicon Interest Group. 1996. *EAGLES Preliminary Recommendations on Semantic Encoding Interim Report.* http://www.ilc.pi.cnr.it/EAGLES96/rep2/rep2.html, Accessed on January 1, 2000.

[68] Elman, J. L. 1991. *Distributed Representations, Simple Recurrent Networks, and Grammatical Structure.* Machine Learning, 7: 195-225, 1991.

[69] Faure, D. and Nedellec, C. *A Corpus-based conceptual Clustering Method for Verb Frames and Ontology Acquisition.* LREC Workshop on Adapting Lexical and Corpus Resources to Sublanguage and Applications, Granada, Spain 1998

[70] Federici S., Montemagni S., Pirrelli V. 1997. *Inferring semantic similarity from Distributional Evidence: an Analogy-based Approach to Word Sense Disambiguation.* in Proceedings of the ACL/EACL Workshop on 'Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications.

[71] Federici S., Montemagni S., Pirrelli V. 1996. *Example-based word sense disambiguation: a paradigm-driven approach.* in Proceedings of the Seventh Euralex International Congress.

[72] Fellbaum, C. *WordNet: An Electronic Lexical Database.* MIT Press, 1998.

[73] Feng, C., T. Copeck, Stan Szpakowicz and Stan Matwin. *Semantic clustering: acquisition of partial ontologies from public domain lexical sources.* In Proceedings of the 7th Banff Knowledge Acquisition for Knowledge-Based Systems Workshop, 1993.

[74] Fiesler, E. and Beale, R. *Handbook of Neural Computation.* Institute of Physics and Oxford University Press, New York, ISBN: 0-7503-0312-3 and 0-7503-0413-8. 1996.

[75] Frank, E., Paynter, G. W., Witten, I. H., Gutwin, C. and Nevill-Manning, C. G. 1999. *Domain-specific Keyphrase Extraction.* Proceedings of the sixth International Joint Conference on Artificial Intelligence (ICJAI-99), pp. 668-673. California: Morgan Kaufmann.

[76] Frakes, W. B. and Baeza-Yates, R. *Information Retrieval,* Prentice Hall,1992.

[77] Freitag, D. 1998. *Multistrategy Learning for Information Extraction.* In Proceedings of the 15th International Conference on Machine Learning, Madison, WI, 161-169.

[78] Gale, Church and Yarowsky. 1992. *A Method for Disambiguating Dord Senses in a Corpus.* Computer and the Humanities, Vol 26, PP. 4215-439

[79] Garside, R. 1996. *The robust tagging of unrestricted text: the BNC experience.* In J. Thomas and M. Short (eds). Using corpora for language research: studies in the honour of Geoffrey Leech, 167¨C180. London: Longman.

[80] Gasser, M. E. 1988. *A Connectionist Model of Sequence Generation in a First and Second Language.* Technical Report, UCLA-AI-88-13, Artificial Intelligence Laboratory, Computer Science Department, University of California, Los Angeles.

[81] Ge, N., Hale, J. and Charniak, E. 1998. *A Statistical Approach to Anaphora Resolution.* In Proceedings of the Sixth Workshop on Very Large Corpora, ACL.

[82] Glasgow, B., Mandell, A., Binney, D., Ghemri, L. and Fisher, D. 1997. *MITA: An Information-extraction Approach to Analysis of Free-Form Text in Life Insurance Applications.* In Proceedings of the Ninth Conference on Innovative Applications of Artificial Intelligence, 992-999. Meolo Park, Calif.: American Association for Artificial Intelligence.

[83] Gold, E. 1967. *Language Identification in the Limit.* Information and Control 16: 447-474.

[84] Gomez, F., R. Hull, and C. Segami. (1994). *Acquiring Knowledge from Encyclopedic Texts.* Proceedings of the ACL 4th Conference on Applied Natural Language Processing, Stuttgart, Germany, 84-90. Also in the IEEE Int'l Journal of Artificial Intelligence Tools , 4(3), 1995, pp. 349-67.

[85] Graff, D. *North American News Text Corpus,* http://www.ldc.upenn.edu/Catalog/, LDC Catalog No.: LDC95T21. Accessed on 16, November, 2002

[86] Grice, H.P. 1975. *Logic and Conversation.* In Syntax and Semantics: Speech Acts, P. Cole and J. L. Morgan (Eds.), Academic Press, New York, 41-58.

[87] Haas, S. W. 1996. *Natural Language Processing: Toward Large-Scale Robust Systems.* In Annual Review of Information Science and Technology (ARIST), Volume 31, 1996. Martha E. Williams, Editor.

[88] Hahn, U. and Schnattinger, K. 1998. *Towards text knowledge engineering.* In AAAI '98 - Proceedings of the 15th National Conference on Artificial Intelligence. Madison, Wisconsin, July 26-30, 1998, pages 129–144, Cambridge/Menlo Park, 1998. MIT Press/AAAI Press.

[89] Hammerton, J., Osborne, M., Armstrong S. and Daelemans W. (eds.). *Special Issue of the Journal of Machine Learning Research: Machine Learning Approaches to Shallow Parsing.* Journal of Machine Learning Research 2, 2002.

[90] Hayes, P. J. And Mouradian, G. V. *Flexible parsing.* American Journal of Computational Linguistics 7 4 (1981)

[91] Hearst, M. and Plaunt, C. *Subtopic Structuring for Full-Length Document Access.* Proceedings of the 16th Annual International ACM/SIGIR Conference, Pittsburgh, PA. 1993.

[92] Hearst, M. A. 1992. *Automatic Acquisition of Hyponyms from Large Text Corpora.* Proceedings of the 14th International Conference on Computational Linguistics, Nantes Grance, July 1992.

[93] Henry, K and Francis, W. N. 1967. *Computational analysis of present-day English.* Providence, RI: Brown University Press.

[94] Hill, J. C. 1983. *A Computational Model of Language Acquisition of the Two Year Old.* Cognition and Brain Theory 6(3): 287-317.

[95] Hindle D., Rooth M. 1993. *Structural Ambiguity and Lexical Relations.* Computational Linguistics, vol. 19, n.1, March 1993, pp. 103-120.

[96] Hinton, G. E. 1981. *Implementing Semantic Networks in Parallel Hardware.* In G. E. Hinton and J. A. Anderson (eds.), Parallel Models of Associative Memory, Erlbaum, Hillsdale, NJ, 1981.

[97] Holowczak, R. D., and Adam, N. R. 1997. *Information Extraction-Based Multiple-Category Document Classification for Global Legal Information Network.* In Proceedings of the Ninth

Conference on Innovative Applications of Artificial Intelligence, 992-999. Meolo Park, Calif.: American Association for Artificial Intelligence.

[98] Huffman, S. 1996. *Learning Information-Extraction Patterns from Examples.* In Symbolic, Connectionist, and Statistical Approaches to Learning for Natural Language Processing, eds. S. Wermter, E. Riloff, and G. Scheler, 246-260. Lecture Notes in Artificial Intelligence Series. New York: Springer.

[99] Hutchins J. and Sommers, H. 1992. *Introduction to Machine Translation.* Academic Press, 1992.

[100] Jackenoff, R. 1983. *Semantics and Cognition.* MIT Press: Cambridge, Massachusetts, and London, England.

[101] Jacobs, P. S. 1992. *Joining Statistics with NLP for Text Categorization.* In Proceedings of the 3rd Conference on Applied Natural Language Processing, March 31-April 3 1992. Trento, Italy.

[102] Jacquemin, C. and Royaute, J. 1994. *Retrieving Terms and Their Variants in a Lexicalized Unification-Based Framework.* In Proceedings of the Association for Computing Machinery Special Interest Group on Information Retrieval (ACM/SIGIR) 17th Annual International conference on research and Development in Information Retrieval; July 3-6, 1994. 132-141. ISBN: 0-387-19889-X.

[103] Jain, A. N.1991. *Parsing Complex Sentences with Structured Connectionist Networks.* Neural Computation, 3: 110-120, 1991.

[104] Jang, D. H., and Myaeng, S. H. 1997. *Development of a document summarization system for effec-tive information services.* RIAO 97 Conference Proceedings: Computer-Assisted Information Searching on Internet, pp. 101-111. Montreal, Canada.

[105] Japan Key Technology Center, Japan, *The EDR Electronic Dictionary,* http://www.iijnet.or.jp/edr/, Accessed on 20, Augest, 2001.

[106] Jensen, K. and Biont J. L. 1987. *Disambiguating Prepositional Phrase Attachments by Using On-line Dictionary Definitions.* American Journal of Computational Linguistics, 13(3):251-260.

[107] Joachims, T. 1996. *A probabilistic analysis of Rocchio algorithm with TFIDF for text categorization.* Computer Science Technical Report CMU-CS-96-118, Carnegie Mellon University.

[108] Jobbins, A. C. and Evett, L. J. 1999. *Sementing Documents Using Multiple Lexical Features.* Fifth International Conference on Document Analysis and Recognition 20 - 22 September. Bangalore, India.

[109] Jones, R., McCallum, A,. Nigam, K., and Riloff, E. 1999. *Bootstrapping for Text Learning Tasks.* From the IJCAI-99 Workshop on Text Mining: Foundations, Techniques, and Applications

[110] Justeson, J. S. and Katz, S. M. 1995. *Technical Termniology: Some Linguistic Properties and an Algorithm for Identification in Text.* Natural Language Engineering 1:9-27.

[111] Karen, L. F. R. 1990. *Identification of Topical Entities in Discourse: a Connectionist Approach to Attentional Mechanisms in Language.* Connection Science, 2: 103-122, 1990.

[112] Karlsson, F., Voutilainen, A., Heikkila, J. and Anttila, A. (Eds), *Constraint Grammar.* Mouton de Gruyter, Berlin. 1995.

[113] Kaszkiel, M. and Zobel, J. *Passage Retrieval Revisited.* ACM SIGIR '97, pp. 178-185

[114] Kehler, A. 1997. *Probabilistic Coreference in Information Extraction.* In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP-97), Providence, RI, 163-173.

[115] Kelly, K. L. 1967. *Early Syntactic Acquisition.* Technical Report Number P3179, Rand Corporation, Santa Monica, California.

[116] Kim, J. T. and Moldovan, D. I. 1995. *Acquisition of Linguistic Patterns for Knowledge-Based Information Extraction.* IEEE Transactions on Knowledge and Data Engineering 7(5): 713-724.

[117] Knight, K. *Automatic Knowledge Acquisition for Machine Translation.* AI Magazine, Volume 18, No. 4 Winter 1997.

[118] Kolmogorov, A. N. 1956. *On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition.* Dokl. Akad. Nauk USSR 114 953-6.

[119] Kozima, Hideki (1993). *Text Segmentation Based on Similarity Between Words.* In Proceedings of the Association for Computational Linguistics.

[120] Krulwich, B., and Burkey, C. 1996. *Learning user information interests through the extraction of semantically significant phrases,* In M. Hearst and H. Hirsh, editors, AAAI 1996 Spring Symposiumon Machine Learning in Information Access. California: AAAI Press.

[121] Kwasny, S. C. and Kalman B. L. 1995. *Tail-recursive Distributed Representations and Simple Recurrent Networks.* Connection Science, 7: 61-80, 1995.

[122] Lang, K. 1995. *Newsweeder: Learning to filter netnews.* In Prieditis and Russel (Eds.), Proceedings of the 12th International Conference on Machine Learning (pp. 331-339). San Francisco: Morgan Kaufmann Publishers, 1995.

[123] Lauriston, A. 1994. *Automatic Recognition of Complex Terms: Problems and the TERMINO solution.* Terminology 1, 1:147-170.

[124] Leant, D.B. 1990. *Building Large Knowledge-based Systems: Respresentation and Interface in the CYC Project.* Addison-Wesley, Reading, MA, 1990

[125] Lebowitz, M. 1983. *Generalisation From Natural Language Text.* Cognitive Science 7: 1-40.

[126] Lee J.H., Kim M.H., Lee Y.J. 1993. *Information Retrieval based on conceptual distance in IS-A hierarchies.* Journal of Documentation, 49(2), June 1993, pp. 188-207.

[127] Leech, G., Garside, R. and Bryant, M. 1994. *CLAWS4: The Tagging of the British National Corpus.* Procs. 15th International Conference on Computational Linguistics (COLING'94), Kytot, Japan, pp. 622-628.

[128] W. Lehnert, C. Cardie, D.Fisher, J. MCCarthy, E. Riloff, and S. Soderland. *Evaluating an Information Extraction System.* Journal of Integrated Computer-Aided Engineering,1(6),1994.

[129] Leland Systems Home Page. *Link Grammar Parser version 4.1,* http://www.link.cs.cmu.edu/link/, Accessed on 22, November, 2000.

[130] Leung, C.-H., and Kan, W.-K. 1997. *A statistical learning approach to automatic indexing of controlled index terms,* Journal of the American Society for Information Science, 48 (1), 55-66.

[131] Lewis, D. D. and Sparch Jones, K. 1996. *Natural Language Processing for Information Retrieval.* Communications of the ACM. 1996 January; 39(1): 92-101. ISSN: 0001-0782.

[132] Lewis, D. 1991. *Representation and learning in informal retrieval.* Ph.D thesis, (COINS Technical Report 91-93), Department of Computer and Information Science, University of Massachusetts, 1991.

[133] Litman, D.J. 1996. *Cue Phrase Classification Using Machine Learning.* Journal of Artificial Intelligence Research 5:53-95.

[134] MacWhinney, B. 1987. *The Competition Model.* In MacWhinney, B. (ed.) Mechanisms of Language Acquisition, 249-308. Lawrence Erlbaum: Hillsdale, NJ.

[135] Maedche, A. and Staab, S. *Discovering Conceptual Relations from Text,* Proceedings of the 14th European Conference on Artificial Intelligence, IOS Press, Amsterdam, 2000.

[136] Manning, C. D. 1993. *Automatic Acquisition of a Large Subcategorization Dictionary from Corpora.* In Proceedings of the Annual Conference of the Association for Computational Linguistics Society, 1993. Columbus, Oiho, 235-242, ACL.

[137] Marcus M., Santorini B. and Marcinkiewicz M. 1993. *Building a large annotated corpus of English: The Penn treebank.* Computational Linguistics, vol. 19, 1993, p.313-330.

[138] Markowitz, J., Ahlswede, T, and Evens M. 1986. *Semantically Significant Patterns in Dictionary Definitions.* In Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics, P112-119.

[139] McCarthy, J. F. and Lehnert, W. G. 1995. *Using Decision Trees for Coreference Resolution.* In Proceedings of the Internationl Joint Conference on Artificial Intelligence, 1995. Montreal, Canada, 1050-1055.

[140] McClelland, J. L. and Kawamoto. A. H. 1986. *Mechanisms of Sentence Processing: Assigning Roles to Constituents.* In J. L. McClelland and D. E. Rumelhart, (eds.), Parallel Distributed

Processing: Explorations in the Microstructure of Cognition, Volumn 2: Psychological and Biological Models, 272-325. MIT Press, Cambridge, MA, 1986.

[141] Mcmahon, J and Smith F. *A Review of Statistical Language Processing Techniques.* Artificial Intelligence Review 12:347-391, 1998

[142] Merialdo, B 1994. *Tagging English Text with a Probabilistic Model.* Computational Linguistics 20(2):155-172.

[143] Michalski, R. 1992. *Understanding the Nature of Learning: Issues and Research Directions.* In Michalski, R. S., Carbonnell, J. G. and Mitchell, T. M. (eds.) Machine Learning, an Artificial Intelligence Approach 2: 3-25. Morgan Kaufmann: Los Altos, California.

[144] Miikkulainen, R. 1997. *Naturla Language Processing with Subsymbolic Neural Network.* in Brown A. (ed.) Neural Network Perspectives on Cognition and Adaptive Robots. Institute of Physics Publishing. 1997, 120-139. Bristol, UK.

[145] Miikkulainen, R. 1993. *Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon and Memory..* MIT Press, Cambridge, MA, 1993.

[146] Minsky, M. A. 1975. *A Framework for Representing Knowledge.* In the Psychology of Computer Vision, P. Winston (ed.), McGraw-Hill, Now York, 211-77.

[147] Mitchell, T. 1999. *The Role of Unlabeled Data in Supervised Learning.* In Proceedings of the Sixth International Colloquium on Cognitive Science, San Sebastian, Spain, 1999 (invited paper).

[148] Mitchell, T. 1997. *Machine Learning,* McGraw-Hill International Editions, 1997.

[149] Mooney, R. J. 1996. *Comparative Experiments on Disambiguating Word Senses: An Illustration of the Role of Bias in Machine Learning.* In Proceedings of the Conference on Empirical Methods in Natural LAnguage Processing (EMNLP-96), Philadelohia, PA, 82-91.

[150] Mooney, R. 1985. *Generalising Explanations of Narratives into Schemata.* Technical Report T-147, Coordinated Science Laboratory, University of Illinois, Urbana.

[151] Morik, K., Wrobel, S., Kietz, J. U. and Emde, W. *Knowledge Acquisition and Machine Learning: Theory, Methods, and Applications,* London Etc.: Academic Press, 1993.

[152] Morin, E. *Automatic Acquisition of Semantic Relations between Terms from Technical Corpora.* Proceedings of the 5th International Congress on Terminology and Knowledge Engineering TKE'99 1999.

[153] Munoz, M., Punyakanok, V., Roth, D. and Zimak, D. 1999. *A Learning Approach to Shallow Parsing.* In Peoceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99), College Park, MD, 168-178. ACL.

[154] Munro, P., Cosic, C. and Tabasko, M. 1991. *A Network for Encoding, Decoding and Translating Locative Prepositions.* Connection Science, 3: 225-240, 1991

[155] Nagao M. 1992. *Some Rationales and Methodologies for Example-Based Approach.* in Proceedings of International Workshop on Fundamental Research for the Future Gen- eration of Natural Language Processing, 30-31 July 1992, Manchester, UK, pp. 82-94.

[156] Nakagawa, H. 1997. *Extraction of Index Words from Manuals.* Procs RIAO'97, Montreal.

[157] Nakamura, J. and Nagao M. 1988. *Extraction of Semantic Information from an Ordinary English Dictionary and its Evaluation.* In Proceeding of the 12th Internatioanl Conference on Computational Linguistics. P459-464, Budapest.

[158] Ng, H.T., and Lee, H.B. 1996. *Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exampler-Based Approach.* In Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics, 40-47. Somerset, N.j.: Association for Computational Linguistics.

[159] Ng, H.T., and Zelle, J. *Corpus-Based Approaches to Semantic Interpretation in Natural Language Processing.* AI Magazine, Volume 18, No. 4 Winter 1997.

[160] Nirenburg S., Domashnev C., Grannes D.I. 1993. *Two Approaches to Matching in Example-Based Machine Translation.* in Proceedings of TMI-93, pp. 47-57.

[161] Owen, M. 1987. *Evaluating automatic grammatical tagging of text.* ICAME Journal 11:18¨C26.

[162] Otha, Y., Yamamoto, Y., Okazaki, T., Uchiyama, I. and Takagi, T. *Automatic construction of knowledge base from biological papers,* Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology, 218-225. Halkidiki, Greece: AAAI Press. 1997.

[163] Paice, C. D. and Jones, P. A. 1993. *The Identification of Important Concepts in Highly Structured Technical Papers.* In SIGIR'93: Proceedings of the Association for Computing Machinery Special Interest Group on Information Retrieval (ACM/SIGIR) 16th Annual International conference on research and Development in Information Retrieval; June 27-July 1, 1993. 69-78. ISBN: 0-89791-605-0.

[164] Pereira F., Tishby N. 1992. *Distributional Similarity, Phase Transitions and Hierarchical Clustering.* Working Notes, Fall Symposium Series, AAAI, pp. 108-112.

[165] Pinker, S. 1979. *Formal Models for Language Learning.* Cognition 7: 217-283.

[166] Pollack, J. B. 1990. *Recursive Distributed Representations.* Artificial Intelligence 46: 77-105.

[167] M Edwards, H Powell, D Palmer Brown, *A Hypermedia-based Tutoring and Knowledge Engineering Systems.* Proceedings of ED-MEDIA '95, World Conf on Educational Multimedia and Hypermedia, held in Graz, Austria, pp 199-204, Ed Hermann Maurer, ISBN 1-880094-15-0, 1995

[168] *Proceedings of the Third Message Understanding Conference (MUC-3).* Morgan Kaufmann, May,1991

[169] *Proceedings of the Fourth Message Understanding Conference (MUC-4).* Morgan Kaufmann, June 1992.

[170] *Proceedings of the Fifth Message Understanding Conference (MUC-5).* Baltimore, MD, Morgan Kaufmann,August 1994.

[171] *Proceedings of the Sixth Message Understanding Conference (MUC-6).* Columbia, MD, Morgan Kaufmann, November 1995.

[172] *Proceedings of the Seventh Message Understanding Conference (MUC-7).* Washington, D.C., Morgan Kaufmann, April 1998.

[173] Quinlan, J. R. and Cameron-Jones, R. M. 1993. *FOIL: A Midterm Report.* In Proc. Of the 12th European Conf. On Machine Learning.

[174] Quinlan, J. R. 1986. *Induction of Decision Trees.* Machine Learning, 1, 81-106.

[175] Quirk, Randolph, Greenbaum, S., Leech, G, and Svartvik, H. *A Comprehensive Grammar of the English Language.* Longman, Harcourt. 1985

[176] Rada R., Hafedh M., Bicknell E. and Blettner M. 1989. *Development and application of a metric on semantic nets.* IEEE Transactions on System, Man, and Cybernetics, 19(1):17-30.

[177] Ramshaw, L. A. and Marcus, M. P. 1995. *Text Chunking Using Transformation-Based Learning.* In Proceeding of the Third Annual Workshop on Very Large Corpora. 82-94. ACL.

[178] Resnik, P. 1995. *Using information content to evaluate semantic similarity in a taxonomy.* In Proceedings of IJCAI.

[179] Resnik, P. *Selection and Information: A Case-based Aproaches to Lexical Relationships.* Ph.D Dissettation, University of Pennsylania, 1993.

[180] Ribas Framis F. 1994. *An Experiment on Learning Appropriate Selectional Restrictions from a Parsed Corpus.* in Proceedings of COLING-94, Kyoto, Japan, pp. 769-774.

[181] van Rijsbergen, C. J. 1979. *Information Retrieval.* Butterworths, 1979.

[182] Riloff, E. and Jones, R. 1999. *Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping.* In Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99) , 1999, pp. 474-479.

[183] Riloff, E. and Schmelzenbach, M. 1998. *An Empirical Approach to Conceptual Case Frame Acquisition.* In Proceedings of the Sixth Workshop on Very Large Corpora, 1998.

[184] Riloff E. 1996. *An Empirical Study of Automated Dictionary Construction for Information Extraction in Three Domains.* AI journal, Vol. 85 August 1996.

[185] Riloff E. *Automatically Constructing a Dictionary for Information-Extraction Tasks.* In Proceedings of the Eleventh National Conference on Artificial Intelligence, 811-816. Menlo Park, Calif.: Amercian Association for Artificial Intelligence.

[186] Robertson, S. E. 1977. *The probability Ranking Principle in IR.* Journal of Documentation, 33(4): 294-304, December 1977.

[187] Rocchio, J. 1971. *Relevance Feedback in Information Retrieval.* In the SMART Retrieval System: Experiments in Automatic Document Processing, Chapter 14, 313-323. Englewood Cliffs, NJ: Prentice-Hall.

[188] Roland, D. *Predicting Verb Subcategorization from the Semantic Context.* Preceding the Verb CUNY sentence Processing Conference, New York, 2002.

[189] Rose, T. G. and Evett, L. J. 1995. *The Use of Context in Cursive Script Recognition.* Machine Vision and Application (1995) 8:241-248.

[190] Rumelhart, D. E., Hinton, G. E. and Williams, R. J. 1986. *Learning internal representations by error propagation.* In D. E. Rumelhart and J. L. McClelland (eds.), Parallel Distributed Processing, 318-362, Cambridge MA: MIT Press.

[191] Salton, G., Allan, J., and Buckley, C. *Approaches to Passage Retrieval in Full Text Information Systems.* ACM SIGIR'93

[192] Salton, G. and Bukley C. *Automatic Text Structuring Experiments.* Text-based Intelligent Systems—Current Research and Practice in Information Extraction and Retrieval Edited by Paul S. Jacobs, ISBN 0-8058-1189-3. 1992

[193] Salton, G. 1991. *Developments in Automatic Text Retrieval.* Science, 253, 974-979.

[194] Salton, G. and McGill, M. J. 1983. *Introduction to modern Information Retrieval.* McGraw-Hill.

[195] Schank, R. and Abelson, R. 1977. *Scripts, Plans, Goals, and Understanding.* Lawrence Erlbaum Assoc. Hillsdale, NJ.

[196] Sekimizu, T., Park, H. and Tsujii, J. *Identifying the interaction between genes and gene products based on frequently seen verbs in medline abstracts.* 1998.

[197] Sekine, S. and Grishman, R. *A Corpus-based Probabilistic Grammar with Only Two Non-terminals.* Fourth International Workshop on Parsing Technology (1995)

[198] Sekine S., Carroll J.J., Ananiadou S., Tsujii J. 1992. *Linguistic Knowledge Generator.* in Proceedings of COLING-92.

[199] Servan-Schreiber, D., Cleeremans, A. and McClelland, J. L. 1991. *Graded State Machines: The Representation of Temporal Contingencies in Simple Recurrent Networks.* Machine Learning, 7: 161-197, 1991.

[200] Sharkey, N. E. and Sharkey A. J. C. 1994. *A Modular Design for Connectionist Parsing.* In M. F. J. Drossaers and A. Nijholt, (ed.) Twente Workshop on LanguageTechnology 3: Connectoionism and Natural Language Processing, 87-96, Enschede, 1992. Department of Computer Science, University of Twente.

[201] Sharkey, N. E. 1988. *A PDP System for Goal-Plan Decision.* In Trappl, R. (ed.) Cybernetics and Systems, 1031-1038. Kluwer Academic: Dordrecht, The Netherlands.

[202] Siskind, J. M. 1990. *Acquiring Core Meanings of Words, Represented as Jackendoff-Style Conceptual Structures, from Correlated Streams of Linguistic and Non-Linguistic Input.* In Proceedings of the Twenty-eighth Annual Meeting of the Association for Computational Linguistics, 143-156. Pennsylvania: Association for Computational Linguistics, University of Pittsburgh.

[203] Skut, W. and Brants, T. *Chunk tagger: statistical recognition of noun phrases.* In ESSLLI-1998 Workshop on Automated Acquisition of Syntax and Parsing, 1998.

[204] Smadja F. 1993. *Retrieving Collocations from Text: Xtract.* Computational Linguistics, vol. 19, n.1, March 1993, pp. 143-177.

[205] Small, S. L., Cottrell, G. W. and Shastri, L. 1982. *Towards Connectionist Parsing.* In Proceedings of the Natural Conference on Artificial Intelligence. Pittsburgh, Pennsylvania: AAAI.

[206] Soderland, S. 1997. *Learning to Extract Text-Based Information from the World Wide Web.* In Proceedings of the Third International Confernece on Knowledge Discovery and Data Mining (KDD-97), 251-254. Menlo Park, Calif: AAAI Press.

[207] Soderland, S., Fisher, D., Aseltine, J. and Lenhert, W. *CRYSTAL: Inducing a conceptual dictionary.* In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 1995.

[208] Sowa, J. F. 1993. *Lexical Structures and Conceptual Structures.* In J. Pustejovsky (ed.), Semantics and the Lexicon, 223-262. Kluwer Academic Publisher, Netherlands.

[209] Srihari, R. and Li, W. *Information extraction supported question answering.* In Proceedings of TREC 8, 1999.

[210] Steier, A. M., and Belew, R. K. 1993. *Exporting phrases: A statistical analysis of topical language,* In R. Casey and B. Croft, editors, Second Symposium on Document Analysis and Information Retrieval, pp. 179-190.

[211] Stolcke A. *Linguistic Knowledge and Empirical Methods in Speech Recognition.* AI Magazine, Volume 18, No. 4 Winter 1997.

[212] Strzalkowski, T. 1995. *Information retrieval using robust language processing.* In AAAI Spring Symposium on Representation and Aquisition of Lexical Information, pages 104-111, Stanford, 1995.

[213] St.John, M. F. 1990. *The Theory Gestalt: A Model of Knowledge-Intensive Processes in Text Comprehension.* Cognitive Science, 16: 271-306, 1992.

[214] St.John, M. F. and McClelland, J. L. 1990. *Learning and Applying Contextual Constraints in Sentence Comprehension.* Artificial Intelligence, 46: 217-258, 1990.

[215] Tabor, E., Juliano, C. and Tanenhaus, M. K. 1997. *Parsing in a Dynamical System: an Attractor-Based Account of the Interaction ofLexical and Structural Constraints in Sentence Processing.* Language and Cognitive Processs, 12, 211-271.

[216] Tepper, J., Powell. H and Palmer-Brown, D. 2002. *A Corpus-based Connectionist Architecture for Large-scale Natural LAnguage Parsing.* Connection Science, Vol. 14, No. 2, 2002, 93-144

[217] Tepper, J. 2000. *Corpus-based Connectionist Parsing.* Ph.D Thesis, Compuiting Department, the Nottingham Trent University, UK. February 2000.

[218] Tepper, J., Powell. H. and Palmer-Brown, D. *Ambiguity Resolution in a Connectionist Parser.* The Cognitive Science of Natural Language Processing, July 5-7 1995, Editor A I C Monaghan, Natural Language Group. 1995a.

[219] Tepper, J., Powell, H. and Palmer-Brown, D. 1995. *Integrating Symbolic and Subsymbolic Architecture for Parsing Arithmetic Expressions and Natural Language Sentences.* Proceeding of 3rd SNN Neural Network Symposium, Nijmegen, Sept 1995, pp 81-84, Eds Bert Kappen and Stan Gielen, ISBN 3-540-19992-6. 1995b.

[220] Thompson, C. A., Mooney, R. J., anf Tang, L. R. 1997. *Learning to Parse Natural Language Database Queries into Logical Form.* In Proceedings of the ML-97 Workshop on Automata Induction, Grammatical Inference, and Language Acquisition. Somerset, N.J.: Association for Computational Linguistics.

[221] Thomson, A. J. and Martinet, A. V. *A practical English Grammar.* 4th edition, Oxford University Press, ISBN 0194313425, 1986.

[222] Touretzky, D. S. 1991. *Connectionism and Compositional Semantics.* In J. A. Barnden and J. B. Pollack, Editors, High-Level Connectionist Models, Volume 1 of Advances in Connectionist and Neural Computation Theory, 17-31. Ablex, Norwood, NJ, 1991.

[223] Tseng, Y. H. *Multilingual Keyword Extraction for Term Suggestion* SIGIR'98, Melbourne, Australia, 24-28 August 1998.

[224] Tsujii J., Ananiadou S., Arad I., Sekine S. 1992. *Linguistic Knowledge Acquisition from Corpora.* in Proceedings of 'International Workshop on Fundamental Research for the Future Generation of Natural Language Processing, 30-31 July 1992, Manchester, UK, pp. 61-81.

[225] Turney, P. D. 2000. *Learning Algorithms for Keyphrase Extraction.* Information Retrieval, 2(4).

[226] Wahlster, W. editor. *Verbmobil: Foundations of Speech-to-Speech Translation.* Springer, 2000.

[227] Waltz, D. L. and Pollack, J. B. 1985. *Massively Parallel Parsing: a Strongly Interactive Model of Natural Language Interpretation.* Cognitive Science 9: 51-74

[228] Waltz, D. L. 1982. *The State of the Art in Natural-Language Understanding.* In W. G. Lehnert and M. H. Ringle (Ed.), Strategies for Natural Language Processing. Lawrence Erlbaum Associates, Publishers. Hillsdale, New Jersey.

[229] Waltz, D.L. 1978. *An English Language Question-answering System for A LArge Relational Database.* Communications of the Association for Computing Machinery 21(7):526-539.

[230] P. D. Wasserman, 1993. *Advanced methods in neural computing.* New York: Van Nostrand Reinhold.

[231] Wiemer-Hastings, P., Graesser, A. and Wiemer-Hastings, K. *Inferring the Meanings of Verbs from Context.* Proceeding of the 20th Annual Conference of the Cognitive Science Society, 1998

[232] Wilks, Y. 1993. *Providing Machine Tractable Dictionary Tools.* Semantics and the Lexicon, J. Pustejovsky (ed.), 341-401.

[233] Wilks, Y. 1973. *Preference Semantics.* Memo AIM-206, Artificial Intelligence Laboratory, Stanford University, Stanford, California.

[234] Winograd, T. 1972. *Understandinf Natural Language.* Academic Press. New York, NY.

[235] Woods, W.A., Kaplan, R.M. and Nash-Webber, B. 1972. *The Lunar Sciences Natural Language Information System: Final Report.* BBN Report 2378, Bolt Beranek and Newman, Inc. Cambride, MA.

[236] Xu, F., Kurz, D., Piskorski, J. and Schmeier S. 2002. *A Domain Adaptive Approach to Automatic Acquisition of Domain Relevant Terms and their Relations with Bootstrapping.* In Proceedings of LREC 2002, the third international conference on language resources and evaluation, Las Palmas, canary island, Spain, May 2002.

[237] Zernik, U. 1987. *Language Acquisition: Learning a Hierarchy of Phrase.* In Proceeding of the Tenth International Joint Confernece on Artificial Intelligence, 125-132. Milan, Italy: Morgan Kaufmann.

[238] Zeman, D and Sarkar, A. 2000. *Learning Verb Subcategorization from Corpora: Counting Frame Subsets.* Proceedings of LREC 2000, pp. 227-233.

[239] Zhang, S., Powell, H., Plamer-Brown, D. and Eveet, L. *Methods for Concept Extraction Using ANNs and Stemming Analysis and Their Portability across Domains,* the Proceedings of The Second Workshop on Natural Language Processing and Neural Networks (NLPNN2001). November, 2001, Tokyo, Japan.

[240] Zhang, S., Powell, H. and Plamer-Brown, D. *Keyword Extraction from Stemming and Sense Information by Neural Networks,* the proceedings of the year 2000 International Conference on Artificial Intelligence (IC-AI'2000). 2000.

[241] Zhang, S., Powell, H. and Plamer-Brown, D. *Keyword Extraction Using Neural Networks.* Proceedings of the Tenth Meeting Computational Linguistics in the Netherlands.

Appendix A

Documents Used for the Experiments

A.1 Education Document: First Three Pages

MEDIATED LEARNING: A NEW MODEL OF NETWORKED INSTRUCTION AND LEARNING

INTRODUCTION

The range of possibilities for employing technology in instruction is so great that it boggles the mind. Because this is the case for many higher-education policymakers, one of the major objectives of this essay is to clarify the important issues for those in leadership roles in higher education, especially those who find themselves perplexed by talk of technology, teaching, learning and the future. I intend to embark upon this clarification by describing the model developed by my colleagues and me to improve the mathematics teaching and learning enterprise, particularly for entry-level college students.

Our approach, called Mediated Learning, which is currently being implemented in colleges and universities around the country, has grown out of a willingness to rethink the very nature and character of that enterprise. It has been developed based on the most promising strategies and theories for employing technology to improve the effectiveness, efficiency, flexibility and responsiveness of the instructional enterprise. We devised this new model of instruction, learning and assessment with the understanding that expert instructors and instructional designers could use new authoring technologies to create flexible, interactive and adaptive computer-mediated instructional materials incorporating standard media formats, including text, hypertext, graphics, animation, simulations, digital audio and digital video. When supported by integrated multimedia database and networking technologies, these programmed materials and the learning activities they afford could be continuously and dynamically ordered to provide instructors and their students task-specific learning assistance, as well as real-time information on learner performance over local area networks of personal computers. This information on learner progress and achievement could be analyzed by instructors and their students to plan possible future courses of action. It could also be used by instructors to analyze possible relationships between students?lesson-navigation patterns and their academic achievement. That same information could be used by students to analyze their own learning proclivities and by instructional designers to inform continuous improvement efforts. Employed in this fashion, authoring, networking and database technologies could provide crucial support for instruction that would be more faculty-guided, more learner-paced and more individualized than is possible in most traditional classrooms, where the whole-class curriculum, lock-step teaching method predominates, especially in the case of lower-division undergraduate courses taken by entry-level students.

Developing Mediated Learning and its range of supporting technologies so that it capitalizes on those strengths of technology-mediated education and integrating it into classroom environments has permitted instructors and their students to spend more time together engaged in high-impact cognitive and developmental activities, and less time on activities that minimally impact cognitive development and learning.

As Mediated Learning becomes recognizable as a unique model of technology-mediated instruction, learning and assessment, the central question facing faculty and administrators is: Is this model cost-effective? Even if technology helps students learn more proficiently, can we afford to invest in it? The purpose of this paper is to address these questions and provide a new framework for analyzing the cost-effectiveness of technology-mediated instruction. In summary, the central point of this paper is that the key to fiscal management in higher education is not cost-cutting; it is improved learner-productivity and increased student academic achievement. By enhancing the learning environment, increasing student passing rates and developing new sources of revenue, colleges and universities can recover their investments in technology, increase support for faculty and generate new net resources.

The Need for New Methods for Determining Cost-Effectiveness

Roger E. Levien and associates, in their prescient examination of the economic and policy issues accompanying the employment of computer-assisted instruction in higher education, noted that cost-effective is a "deceptively simple" term "which has frequently been misapplied in the examination of instructional computer use" (1972, p. 436). Frequently, the misapplication of cost and effectiveness measures stems from unclear definitions and unsound metrics, such as the imposition of one set of metrics for evaluating technology-mediated instruction and another for assessing traditional lecture-based instruction. Problems also result from reluctance of some faculty to lend any level of legitimacy to the topic of cost-effectiveness on the grounds that to do so would constitute an open invitation to non-educators to apply uninformed and inappropriate evaluation and assessment measures to the higher education teaching and learning enterprise. The caution of the Levien team, therefore, is well founded; the pedagogical, policy and political nuances and complexities of the higher education teaching and learning enterprise are not easily captured by standard cost-accounting methodologies and they remain highly problematic two decades later. More recent articles addressing the relationship between technology-mediated instruction and instructional productivity by Gifford (1991), Green and Gilbert (1995a, 1995b), Levin (1991) and Plater (1995) make it clear that researchers may be no closer to establishing consistent metrics for comparing the cost-effectiveness of traditional and technology-mediated instruction than when the Levien study was published in 1972 under the auspices of the Carnegie Commission on Higher Education.

Still, there is a pressing need for evaluation and assessment metrics for comparing traditional lecture-centered instruction and technology-mediated instruction models, especially in light of the fierce pressure on higher education institutions to justify new instructional initiatives in terms of their impact on the economics of the teaching and learning enterprise. Mindful of this necessity, and of Levien's admonition against inappropriately applying the term "cost-effective," this paper presents a framework for comparing the relative cost-effectiveness of Mediated Learning to that of traditional lecture-centered instruction. Mediated Learning is a form of technology-mediated instruction, learning and assessment that is being used by several hundred faculty on campuses around the country in order to prepare students who enter college underprepared for college-level mathematics. To date, this method has been utilized to develop a series of introductory mathematics courses called Interactive Algebra for College Students I, II and III. Together, these courses comprise the first sub-series in the larger collection called Interactive Mathematics. (Because Interactive Algebra III was just released in August, 1996, all analyses in this paper will focus on implementations of Interactive Algebra I and II.) The Mediated Learning approach enables faculty to improve the instructional environment by reconfiguring the traditional, lecture-centered classroom. Faculty use specially constructed, interactive, multimedia lessons delivered via computers, which enable instructors and students to teach and learn more effectively and efficiently.

In Section 1, "Mediated Learning: A New Model of Instruction and Learning," a history of Mediated Learning's development, as well as the unique instructional environment facilitated by Mediated Learning and its components, are briefly introduced, and reasons for implementing Mediated Learning, aside from its cost-effectiveness, are discussed. Far from being simply an economically sound alternative to traditional instruction, Mediated Learning also yields pedagogical benefits, which although not the focus of this paper, need to be articulated in order to demonstrate that this technology-mediated approach is desirable for many reasons, not merely monetary ones. Section 2, "Findings on the Effectiveness of Mediated Learning," provides findings being reported by campuses that have been implementing Interactive Algebra I and II. One school's data, because it is fairly extensive, is analyzed in some detail and demonstrates that, in most cases, a greater percentage of students who take one or two Mediated Learning courses before enrolling in Precalculus earn grades of C and above in Precalculus than their peers who take either no algebra at college or traditional Beginning and/or Intermediate Algebra. In Section 3, "Impact of Student Completion Rates on

Enrollment Patterns," simulations are presented in which incremental changes in student completion rates are shown to impact student enrollment patterns significantly. The conclusion is that increases in completion rates caused by the implementation of technology-mediated instruction can have a significant impact on the number of students that successfully complete a course or sequence of courses. In Section 4, "Mediated Learning's Effect on Instructor Workloads," the analysis continues with an examination of the nature and character of instructional work in traditional instructional settings and in Mediated Learning environments. The conclusion is that in Mediated Learning settings, instructors can take on more students per class without increasing their instructional workload. Section 5, "The Economics of Mediated Learning," uses the information from Sections 3 and 4 to indicate that Mediated Learning need not increase instructional costs in those settings where it is implemented and is, in fact, more cost effective. Finally, the paper's conclusions are summarized in Section 6, "Conclusion."

1. MEDIATED LEARNING: A NEW MODEL OF INSTRUCTION AND LEARNING

Mathematics was selected as the first discipline for which Mediated Learning lessons would be developed in order to help solve one of the most vexing and persistent problems confronting the higher education mathematics community the issue of addressing effectively, efficiently and flexibly the learning assistance needs of entry-level students who arrive on college campuses underprepared for college-level studies in mathematics. The dimension of this challenge, manifested in the rapidly increasing enrollment of students placed in remedial and developmental mathematics courses, is documented in the latest survey of enrollment trends in mathematics courses conducted by the Mathematical Association of America (MAA) (Alberts, Loftsgaarden, Rung and Watkins, 1992). Table 1: Mathematics Enrollment in Four-Year Colleges and Universities: 1970 to 1990

Definitions: Remedial courses include Arithmetic, General Mathematics (Basic Skills), High School Elementary Algebra and High School Intermediate Algebra. Precalculus courses include College Algebra, Trigonometry, Combined College Algebra and Trigonometry, Elementary Function Precalculus, Mathematics for Liberal Arts Students, Finite Mathematics, Business Mathematics, Mathematics for Elementary School Teachers, Analytical Geometry and other Precalculus courses. Calculus courses include Mainstream Calculus I, II, III, etc., Non-Mainstream Calculus I, II, III, etc., Differential Equations, Discrete Mathematics, Introduction to Mathematical Logic, Linear Algebra and Other Calculus courses. Source: Alberts et. al., Appendix I. As depicted in Table 1, above, in the case of four-year colleges, between the fall of 1970 and the fall of 1990, enrollment in remedial mathematics courses increased nearly three times faster than enrollment in calculus courses, nominally the entry-level college mathematics course for well-prepared entry-level college students. Also note that more than one-third of the total increase in enrollment in mathematics courses between 1970 and 1990 (434,000) was due to increased demand for remedial courses (162,000). In fact, in four-year colleges, enrollment in remedial mathematics courses climbed to 16 percent of the total mathematics enrollment in the fall of 1990 from eight percent in the fall of 1970. Table 2: Enrollment in Mathematics Courses in Two-Year Colleges: 1970 to 1990

Definitions, see Table 1 on previous page. Source: Alberts et. al. Table 1, p. 4. In the case of two-year colleges, and very much reflecting their traditional mission as open-entry institutions, the challenge of appropriately addressing the instructional support needs of underprepared students is even more formidable, as evidenced by the data depicted in Table 2, above. Between the fall of 1970 and the fall of 1990, the enrollment increase in remedial and precalculus mathematics courses alone (644,000) was greater than the total enrollment growth in all mathematics courses in all four-year colleges and universities (434,000). Moreover more than 77 percent of the total increase in enrollment in mathematics courses between 1970 and 1990 (686,000) was due to increased demand for remedial courses (533,000). In two-year colleges, enrollment in remedial mathematics courses climbed to 58 percent of the total mathematics enrollment in the fall of 1990 from 34 percent in the fall of 1970.

Disturbing as these data in Tables 1 and 2 might be, they seriously underestimate the number of underprepared mathematics students attending the nation colleges and universities, particularly in larger industrial states with diverse college student populations, such as California and New York, where the California State University (CSU) and City University of New York (CUNY) campuses are located, respectively. The MAA survey does not account for students taking remedial and precalculus courses outside of mathematics departments, say, in departments specializing in remedial developmental mathematics or in programs headquartered in learning centers specifically organized to address the learning assistance needs of underprepared students. Nor does the MAA survey disaggregate student enrollment data down to the state level, nor by student language origins. However, because of California and New

York State large college student populations and their high level of student diversity, it is difficult to imagine that a breakdown of the MAA data for California or New York would have uncovered enrollment patterns and trends significantly dissimilar from those depicted in Tables 1 and 2. If anything, these states?large and growing population of non-native English speakers might lead one to conjecture reasonably that the enrollment patterns and trends depicted in Tables 1 and 2 actually underestimate the numbers of students in these states who arrive on college campuses underprepared for college-level studies in mathematics.

A.2 Law Document: First Three Pages

The Profession and Practice of Law

Introduction

The primary function of the profession and practice of law is to apply the law in specific cases–to individualize it. This function is manifest in the work of the advocate and the judge in the process of trying and deciding cases. In Anglo-American systems a lawyer investigates the facts and the evidence by conferring with his client, interviewing witnesses, and reviewing documents. He may seek a summary dismissal because the opponent evidently has no case, or, through discovery proceedings he may force the other side to reveal more fully the issues and facts on which it relies. At the trial he introduces evidence, objects to improper evidence from the other side, and advances partisan positions on questions of law and of fact. In continental European countries the judge has greater responsibility for investigation of the facts. At trial he plays an active role in taking evidence, questioning witnesses, and framing the issues. Continental lawyers suggest lines of factual inquiry to the judge and, like their Anglo-American counterparts, advance legal theories and argue the law in accord with the interests of their clients. In either system, if a lawyer loses his client's case, he may seek a new trial or relief in an appellate court.

Even controversies that are not resolved in court require the aid of lawyers. Negotiation, reconciliation, compromise–in all of which lawyers have a large part–bring about the settlement of most cases without trial.

The profession also applies and utilizes the law in the less dramatic setting of the office. The lawyer as counselor and negotiator may aid in shaping a transaction so as to avoid disputes or legal difficulties in the future, or so as to achieve advantages for his client, such as the minimization of taxes. The law gives to private persons extensive but not unlimited power to arrange and determine their legal rights in many matters and in various ways, such as through wills, contracts, leases, or corporate bylaws. In structuring these arrangements the lawyer is helping to particularize the legal rights of the parties.

Another field of legal work, which has developed rapidly in the 20th century, is the representation of clients before administrative commissions, commissions of inquiry, and, in some countries, legislative committees. This development has been a result of the increase of government regulation of economic life.

A lawyer has several loyalties in his work, including loyalty to his client, to the administration of justice, to the community, to his associates in practice, and to himself–whether to his economic interests or to his ethical standards. These diverse and at times competing loyalties must be reconciled with wisdom. It is the purpose of the standards of the profession to effect the reconciliation. (M.A.Gl.)

Legal profession

One definition of the legal profession is "the vocation based on expertness in the law and its application." This simple definition may be best, despite the fact that in some countries there are several professions and even some occupations (e.g., police service) that require this expertness but may not be considered to be within the "legal profession" at all.

HISTORY

Distinct legal systems emerged relatively early in history, but legal professions of size and importance are relatively modern. There is not the slightest trace in ancient times of a distinct legal profession in the modern sense. The earliest known legal specialist was the judge, and he was only a part-time specialist. The chief, prince, or king of small societies discharged the judicial function as part of the general role of political leader. As his power spread, he delegated the function, though not to legal specialists; in the secular stages of the early systems, legal duties were taken over by royal officials who were "generalists." In the wake of powerful religious or quasi-religious movements priests or wise men often judged or advised the judges, a situation that persisted in Muslim countries and in China until the 20th century AD. It may be suspected that in some of these cases specialized legal aid to the ordinary citizen did exist, but at levels of social status below the notice of chroniclers or tomb inscriptions and perhaps without benefit of official approval.

Classical beginnings of a legal profession.

A distinct class of legal specialists other than judges first emerged in the Greco-Roman civilization, and as with the law itself, the main contribution was from Rome in the period from 200 BC to AD 600. In the early stages of both Greece and Rome, as later among the German tribes who overran the Roman Empire, there was a prejudice against the idea of specialists in law being generally available for fee. The assumption was that the citizen knew the customary law and would apply it in transactions or in litigation personally with advice from kinsmen. As the law became more complex, men prominent in public life–usually patricians–found it necessary to acquire legal knowledge, and some acquired a reputation as experts. Often they also spent periods serving as magistrates and in Rome as priests of the official religion, having special powers in matters of family law. Among the German tribes noble experts were allowed to assist in litigation, not in a partisan fashion but as interpreters (Vorsprecher) for those unready of speech who wished to present a case. The peculiar system of development of the early Roman law, by annual edict and by the extension of trial formulas, gave the Roman patrician legal expert an influential position; he became the jurisconsult, the first nonofficial lawyer to be regarded with social approbation, but he owed this partly to the fact that he did not attempt to act as an advocate at trial–a function left to the separate class of orators–and was prohibited from receiving fees. (see also Index: Germanic law)

The modern legal professional, earning his living by fee-paid legal services, first became clearly visible in the later Roman Empire when the fiction that a jurisconsult received only gifts was abandoned and when at the same time the permissible fees were regulated. Changes in the methods of trial and other legal developments caused the jurisconsult to disappear in time. The orator, who now was required to obtain legal training, became the advocate. A subordinate legal agent of the classical system, the procurator, who attended to the formal aspects of litigation, took on added importance because later imperial legal procedure depended largely on written documents drawn by procurators. The jurisconsults had been important as teachers and writers on law; with their decline this function passed to government-conducted law schools at Rome, Constantinople, and Berytus and to their salaried professors. There was also a humbler class of paid legal documentary experts, the tabelliones, useful in nonlitigious transactions.

Medieval Europe.

This late Roman pattern of legal organization profoundly influenced the Europe that began to arise after the barbarian invasions from AD 1000 on; and even during the invasions the methods of Roman imperial administration never ceased to exist in some parts of southern France and in central Italy. The Christian Church, which became the official Roman imperial church after AD 381, developed its own canon law, courts, and practitioners and followed the general outline of later Roman legal organization. Because of its success among the invaders the church was in a position to establish its jurisdiction in many matters of family law and inheritance. Hence both the idea of a legal profession and the method of its operation retained sufficient force to offset Germanic and feudal objections to legal representation. After the revival of learning in the 12th century, in particular the revived study of Roman law at Bologna, the influence of the late Roman professional system was greatly strengthened. (see also Index: medieval law, Christianity, Roman Catholicism)

From then on every country in continental Europe acquired, by various stages and with numerous local variations, a legal profession in which four main constituents could be observed. Procurators attended to the formal and especially the documentary steps in litigation. Advocates, who usually were university graduates in Romanist learning, gave direct advice to clients and to procurators and presented oral arguments in court. Among a miscellany of legal scribes the notaries acquired importance because, in addition to being drafting experts, they also provided officially recognized document authentication and archives. University teachers of law took over the main task of explaining and of adapting the mixture of Roman law and Germanic custom that produced the modern laws of the major European countries and continued to dominate in the scholarly interpretation of the law even after the 19th-century codifications. The relative importance of these classes varied enormously from place to place and from century to century. At times the teaching doctors almost supplanted the advocates; in some courts the procurators swallowed up the advocates and in others the converse occurred; only the notaries managed to survive with little change. (see also Index: notary)

England after the Conquest.

England after the Norman Conquest also was influenced by Roman example, and the clerics who staffed the Norman and Plantagenet monarchies and who provided the earliest of their judges enabled the notion of a legal profession and especially of litigious representation to be accepted. Only in the ecclesiastical and admiralty courts, however, did procurators (proctors) and doctors of the civil and canon laws become established as practitioners. The native "common law" was developed by a specialized legal society, the Inns of Court in London; there, through lectures and apprentice training, men acquired admission to practice before the royal courts. More particularly, they could become serjeants–the most dignified of the advocates, from whom alone after about 1300 the royal judges were appointed. Various agents for litigation resembling procurators also became known. The "attorneys," authorized by legislation, at first shared the life of the Inns with the "apprentices" in advocacy, who themselves in time acquired the title of barristers. Indeed there were cases of men working as both barristers and attorneys. When in the 16th century the Court of Chancery was established as the dispenser of "equity," the appropriate agent for litigation was called a solicitor, but the common-law serjeants and barristers secured the right of advocacy in that court. It was not until the 17th century that the attorneys and solicitors were expelled from the Inns and the division between advocate and attorney became rigid, and not until the 18th century that the barristers accepted a rule that they would function only on the engagement of an attorney–not directly for the client. Other types of legal agents also developed in England, but in the 19th century all of the nonbarristers were brought under the one name, solicitor. The order of serjeants was wound up, leaving only barristers, of whom the most senior could be made Queen's (or King's) Counsel. (see also Index: United Kingdom, English law)

In its final development the English legal profession thus bore a resemblance to the European–particularly to that of northern France, where the parlements (courts) had a corporate life and apprentice training not unlike that of the Inns. But there were four significant differences between England and the Continent. No distinct class of university teachers and commentators on the national law developed in England. Development of the law took place chiefly through precedent based on the reported judgments of the courts, rather than through legislation. The continental monarchies also developed a system of career judicial office, in which the young university licentiate went straight into government service, whereas in England appointment of judges from the senior practicing profession remained the settled practice. In addition, the division between barristers and solicitors ultimately became much more rigid in England than did the division between the advocate and procurator in Europe, and Europe never adopted an equivalent of the English practice requiring a barrister to be employed by a solicitor; both the procurator and the advocate were separately and directly employed by the client. England never developed the profession of notary, so that the whole burden of transactional work fell on those who are now the solicitors, with legal advice from the bar. (see also Index: French law)

Worldwide legal profession.

The main patterns both of law and of legal practice were exported by the continental European powers and England to their overseas colonies and possessions, and most of the noncolonial countries of the rest of the world imitated one or the other system. Thus the Romano-Germanic practices (frequently called civil law) became the norm for Scandinavia, Scotland, Latin America, and most of the Muslim countries of the Middle East, for French-speaking areas and Portuguese and Spanish Africa, and for Japan, Thailand, and the former French parts of Southeast Asia. They have also influenced practice in what are now the socialist countries of eastern Europe. The English system provided the model for English-speaking North America, for most former English colonies in Africa, including South Africa, for most of the Indian subcontinent, and for Malaysia, Australia, and New Zealand. The original model has undergone considerable modification by both the countries of export and the countries of reception. In particular, the specialization of procurator-advocate and solicitor-barrister has tended to be replaced by a "fused" profession of legal practitioners qualified to perform both functions and usually doing so. Such a fusion occurred gradually in

Germany between the 16th and 18th centuries. It has taken place more recently in France except before the courts of appeal and, while the division still formally exists in Italy, it is no longer of practical importance. In Latin America the fused profession is general. Notaries as a separate specialized branch of the profession exist, however, in most civil-law countries.

CHARACTERISTICS OF THE PROFESSION

Social role.

The legal profession has always had an ambiguous social position. Leading lawyers have usually been socially prominent and respected–the sections of the profession so favoured varying with the general structure of the law in the particular community. The family status of early Roman jurisconsults may have been more important than their legal expertise in securing such a position, but by the time of the principate it was their legal eminence that made them respected. The English serjeants lived magnificently, especially in Elizabethan times, and the French Ordre des Avocats was established (14th century AD) by feudal aristocrats in circumstances reminiscent of early Rome–including an insistence on receiving gifts rather than fees. The early Italian doctors of the civil and canon law (12th-15th centuries) were revered throughout Europe. In England and the countries influenced by its system the highest prestige gradually came to be concentrated on the judges rather than on the order of serjeants, of which they were members, and the judges of high-level courts remain the only legal class in the liberal capitalist common-law countries of today to command great respect.

Appendix B

Data on Word-Level Path Experiments

There are three ways in deciding the Part of Speech (POS) of a word, i.e. POS tagging, parsing and looking-up in a dictionary. We chose to look up a word in WordNet. A word in the document is deemed to be a noun if it exists in WordNet with the part of speech of a noun.

Words followed by an asterisk (*) are keywords.

B.1 Nouns in the Education Document

B.1.1 All Nouns in the Education Document

ability, academic*, academic-year*, access, accordance, account, accounting, achievement, acquisition, act, acting, action, active, activity, actuality, ad, addition, address, adept, administration, admission, admonition, advance, advantage, advice, affect, aggregate, aid, al, algebra, alignment, allocation, alternative, america, american, amount, an, analysis, animation, answer, anxiety, appendix, application, approach, architecture, are, area, argument, arithmetic, article, arts, as, aside, aspect, assessment*, assistance, association, assumption, at, attempt, attending, attention, attractiveness, attribute, audio, august, auspices, author, authority, availability, average, b, baccalaureate*, background, basic, basis, be, beginning, behavior, behind, being, benefit, best, better, bit, body, bottom, breakdown, bridge, brief, broad, brown, bruce, budget, build, burden, business, c, calculus, calendar, california, campus*, can, capability, capacity, cardinal, career, carnegie, case, category, cause, caution, center, central, cf, chain, challenge, chancellor, change, character, chart, chi, chief, choice, chosen, circumstances, citizenship, city, clarification, class*, classroom*, clear, coaching*, cognition, cohort, collection, college*, college-level*, collins, colony, column, combination, combine, commission, commitment, common, commonplace, communication, community, comparative, compare, comparing, comparison, competence*, compilation, completion, complexity, comprehension*, computer, concentrate, concept, concern, conclusion, condition, conduct, con-founding, confronting, conjecture, consideration, consonant, construct, content, context, contract, contribution, control, core, cost, cost-accounting, cost-cutting, council, country, course*, cover, covering, creativity, credit, current, curriculum*, curve, cut, cycle, d, data, database, date, day, dealing, debate, decade, deciding, decline, decrease, degree*, delivery, demand, department, dependence, dependent, depth, description, design, designing, detail, detailing, determination, developing, development, deviation, devising, difference, differential, dimension, direction, directive, disappointing, discipline, discussion, diversity, do, document, doing, domain, dominant, down, draw, drawing, drift, drop, due, e, ease, economics, editor, education*, effect, effectiveness, efficacy, efficiency, effort, eight, elements, eligibility, elm, emphasis, employ, employment, end, engagement, engine, engineering, english, enough, enrollment*, entering, enterprise, entire, environment, equipment, equivalent, essay, essential, evaluation, evidence, exam*, examination*, example, exception, exclusive, exercise, exhibit, experiment, expert, expertise, exposure, extension, extent, external, extra, f, facing, fact, factor, faculty, failing, failure, fall, familiar, fashion, fast, feedback, feel, felt, few, figure, final, find, first, five, flexibility, florida, flourish, flow, focus, following, force, form, format, former, found, four, fourth, framework, free, freedom, frustration, function, fundamentals, funding, future, g, gap, gateway, gauge, general, geometry, gilbert, give, given, giving, go, goal, gold, good, governed, government, grade*, grading*, gradual, graduate*, graduation*, graph, graphics, green, grounds, group, growing, growth, guiding, half, hands, have, he, health, hearing, help, helping, here, high, high-quality, higher-education*, highlight, history, hold, homework*, hours, human, human-, hundred, hypertext, hypothesis, i, identification, ii, iii, impact, implement, implementation, importance, imposition, impossible, improvement, in, inability, incarnation, increase, independent, indicative, individual, influence, information, initial, input, inquiry, insensitivity, instance, institute, institution, instruction*, instructor*, instrumental, integrating, intensive, intent, interaction, intermediate, internet, introduction, investing, investment, invitation, issue, james, john, keeping, key, kind, knowing, knowledge*, knowledge-base, labor, laboratory, lager, language, large, last, lead, leadership, learner*, learning*, least, lecture*, lecturer*, lecturing*, leeds, left, legitimacy, less, lessening, lesson*, letter, level, leverage, liberal, lieu, life, light, line, links, lisp, literature, little, living, local, logic, low, lower, mainstream, major, majority, make, make-up, making, management, manifest, manner, mastering, mastery, material, mathematics, matter, may, mean, meaning, measure, media, meet, member, menu, mere, method, michigan, mid-term, middle, might, mind, minimum, minority, minus, minutes, misapplication, mission, mistake, mode, model, modeling, monitor, monitoring, more, most, move, much, multimedia, multiple, nation, national, nature, necessary, necessity, need, net, network, nine, no, notable, notation, note, notebook, notion, now, number, object, objective, obligation, occurrence, offer, one, one-, one-third, open, opportunity, option, order, organization, organizer, original, out, outline, outside, over, overall, p, pace, pacing, page, pains, paper, parallel, part, particular, pass, passing, past, percent, percentage, performance, period, permit, persistence, personal, perspective, place, plan, planning, play, plus, point, policy, polytechnic*, population, positive, possibility, possible, posture, potential, praise, predecessor, predisposition, preliminary, preparation, prerequisite, present, presentation, preserves, preserving, press, pressing, pressure, price, principal, principle, prior, probability, problem, problem-solving, process, productivity, professional, proficiency, program, programming, progress, progression, proportion, prospect, providing, provisions, public, purpose, put, quality, quantity, quarters, question, quick, quiz, raise, raising, range, rate, reach, readiness, ready, real, reason, recall, record, recording, reducing, reduction, reference, regular, regulating, reimbursement, rejection, relationship, relative, release, reluctance, remains, remediation*, repeat, repeating, repertoire, requirement, research, resource, response, responsibility, responsiveness, rest, result, retention, rethink, revenue, review, right, rise, role, routine, row, run, rung, sacrifice, salary, salient, sample, savings, say, schedule, scheme, school*, science, scope, scores, second, section, see, seeing, semester*, sense, separate, sequence, series, serve, session, set, setting, settling, seven, sharing, shift, short, shortening, shrink, significance, simple, situation, six, sixty, size, small, so, social, software, solution, sort, sound, source, specialist, specific, spending, spite, spring, square, staff, standard, stands, start, state, static, stay, steady, step, still, structure, student*, study*, studying*, subject*, subpopulation, success, succession, suffering, sum, summary, summer, support, supporting, survey, swelling, syllabus*, system, table, tactics, tailor, take, taking, talk, task, teach*, teacher*, teaching*, team, technology, ten, term*, terminal, test, test-bed, testing, texas, text, textbook*, then, theory, there, things, think, thinking, third, thought, three, time, tool, top, topic, total, training*, transfer, transformation, transmission, trigonometry, true, try, tuition*, tutor*, tutoring*, two, two-, two-thirds, type, u, underestimate, undergraduate*, underscore, understanding, unfolding, university*, upper, use, using, valuable, value, variable, variety, vehicle, video, view, virtue, w, want, washington, ways, week, well, while, who, whole, why, will, willingness, wish, withdrawal, won, words, work, working, workload, worry, worth, x, years, yield, yielding, york, zero

B.1.2 Nouns in Training Set

a, academic*, accounting, achievement, action, address, advantage, algebra, alignment, amount, an, approach, are, area, as, assistance, at, auspices, basis, be, being, better, calculus, campus*, can, career, carnegie, case, circumstances, city, class*, classroom*, clear, college*, commission, computer, concept, conclusion, core, cost, course*, data, description, detail, development, do, dominant, education*, effect, emphasis, enrollment*, enterprise, environment, essential, exercise, fact, faculty, failure, familiar, feedback, figure, first, format, four, given, gold, grade*, group, hands, have, high, hours, identification, in, increase, information, instruction*, instructor*, learner*, learning*, lecture*, leeds, less, lieu, lisp, local, material, mathematics, model, more, most, multimedia, notion, number, obligation, one, organization, out, over, parallel, passing, percentage, performance, personal, population, principal, public, put, quantity, question, raising, range, rate, regular, repeat, research, result, review, salary, savings, section, semester*, sequence, serve, short, six, software, state, still, student*, success, support, table, take, teaching*, technology, term*, three, times, two, university*, value, week, who, york

B.1.3 Nouns in Test Set

[1] a, **ability**, academic*, **access**, **account**, achievement, **acquisition**, action, **activity**, **ad**, **addition**, address, **adept**, **administration**, **admission**, **admonition**, advantage, **advice**, algebra, alignment, **allocation**, **alternative**, **american**, amount, an, **analysis**, **answer**, **application**, approach, **architecture**, are, as, **aside**, **assessment***, assistance, **association**, at, **attending**,

[1] Nouns in boldface are nouns that exist only in test set, not in training set.

attribute, **author**, **average**, **baccalaureate***, **background**, **basic**, basis, be, **beginning**, **behavior**, **behind**, being, **benefit**, **best**, better, **breakdown**, **bridge**, **brown**, **c**, campus*, can, **capability**, **capacity**, case, **cause**, **center**, **change**, circumstances, class*, classroom*, **cognition**, **cohort**, college*, **college-level***, **collins**, **combination**, **community**, **comparing**, computer, conclusion, **conduct**, **confronting**, **consonant**, **content**, **contract**, cost, **council**, course*, **cover**, **creativity**, **credit**, **cycle**, data, **database**, **date**, **debate**, **deciding**, **degree***, **demand**, **department**, **dependent**, detail, **developing**, **difference**, **directive**, **discipline**, **diversity**, do, **down**, **drop**, **due**, e, **economics**, education*, effect, **effectiveness**, **efficacy**, **effort**, **eight**, **elm**, emphasis, **employ**, **employment**, **end**, **engine**, enrollment*, enterprise, environment, **equipment**, **equivalent**, **essay**, essential, **evaluation**, **evidence**, **exam***, **examination***, **example**, **exception**, exercise, **experiment**, **exposure**, **extension**, **extra**, f, fact, faculty, **failing**, **fall**, **fashion**, feedback, **few**, figure, **final**, **find**, first, **five**, **focus**, **following**, **form**, format, **found**, four, **framework**, **free**, **freedom**, **function**, **funding**, **future**, **gauge**, given, **governed**, grade*, **grading***, **graduation***, **graph**, **grounds**, group, **guiding**, **half**, hand, have, **he**, **health**, **help**, **here**, high, **high-quality**, **higher-education***, **hold**, **homework***, hours, **hypothesis**, **i**, ii, **impact**, **implement**, **implementation**, **improvement**, in, increase, **independent**, **individual**, information, **initial**, **insensitivity**, **institute**, **institution**, instruction*, instructor*, **intensive**, **intermediate**, **introduction**, **issue**, **keeping**, **kind**, **knowledge***, **kw**, **language**, **large**, **last**, **lead**, **leadership**, learner*, learning*, **least**, lecture*, **left**, less, **lesson***, **level**, **leverage**, **light**, **line**, **links**, lisp, **low**, **lower**, **major**, **make**, **making**, **management**, **mastering**, **mastery**, material, mathematics, **may**, **meaning**, **means**, **measure**, **meet**, **mid-term**, **might**, **mind**, **minority**, **minutes**, model, **monitor**, **monitoring**, more, most, **move**, **much**, multimedia, **national**, **nature**, **necessity**, **need**, **net**, **no**, **note**, **notebook**, **now**, numbers, **occurrence**, **offer**, one, **one-**, **one-third**, **opportunity**, **option**, **order**, organization, **original**, out, **outline**, **outside**, over, **overall**, **pacing**, **page**, **paper**, **part**, **pass**, passing, **percent**, percentage, performance, **period**, **persistence**, personal, **plan**, **planning**, **play**, **plus**, **point**, **policy**, **polytechnic***, **possible**, **potential**, **predecessor**, **preparation**, **present**, **pressing**, **pressure**, **prior**, **probability**, **problem-solving**, **process**, **productivity**, **program**, **programming**, **progress**, **progression**, **providing**, put, **quality**, question, **quick**, **quiz**, **raise**, range, rate, **readiness**, **ready**, **real**, **record**, **reducing**, **regulating**, **relationship**, **repeating**, **requirement**, research, result, **retention**, **revenue**, review, **right**, **rise**, **role**, **routine**, **run**, savings, **say**, **schedule**, **scheme**, **school***, **scope**, **scores**, **second**, section, **seeing**, semester*, sequence, **session**, **set**, **setting**, **shift**, short, **shortening**, **shrink**, six, **sixty**, **small**, **so**, **social**, **sort**, **sound**, **specific**, **standard**, **stands**, **start**, state, **static**, **stay**, **steady**, still, student*, **study***, **subject***, **subpopulation**, success, **summer**, support, **supporting**, **survey**, **swelling**, **syllabus***, **system**, table, **tailor**, take, **taking**, **talk**, **teach***, teaching*, technology, **terminal**, terms*, **test**, **testing**, **text**, **textbook***, **then**, **theory**, **there**, **thinking**, three, time, **tool**, **total**, **true**, **tutoring***, two, **two-**, **type**, **u**, **understanding**, university*, **upper**, **use**, **using**, **valuable**, value, **vehicle**, **w**, **want**, **way**, week, **well**, **while**, who, **whole**, **will**, **wish**, **withdrawal**, **work**, **working**, **workload**, **years**, **yield**, york

B.2 Nouns in the Law Document

B.2.1 All Nouns in the Law Document

abandonment, ability, absolute, academic, accident, accord, account, accused*, acquiring, act*, acting, action*, active, activity, ad, addition, administration, admiralty, admission, advance, advancement, advantage, advertising, advice, adviser, advocacy*, advocate*, africa, age, agent, agreement, aid, aim, al, am, america, american, amount, an, analogy, anglo-american, annual, appeal*, appearance*, application, appointment, apprentice, apprenticeship, approach, approaching, approbation, approval, aptitude, archives, are, area, argument, aristotle, arrangement, arthur, as, ascending, asia, aside, aspect, assessor*, assist, assistance, assistant, association, assumption, at, athens, attainment, attempt, attendance, attending, attention, attitude, attorney*, attrition, australia, authentication, author, authority, autonomy, availability, average, award, awareness, baby, back, backbone, background, balance, bar*, barbarian, barrister*, basic, basin, basis, beginning, behalf, behaviour, behind, being, belief, beneficiary, benefit, best, better, bibliography, blight, body, bologna, boom, bordeaux, bore, branch, bringing, britain, british, bulk, burden, business, buyer, buying, cambridge, can, canada, canadian, candidate, canon, capacity, capitalist, cappelletti, career, carry, case*, case-law*, casebook*, category, catholicism, cause, central, century, certainty, chancery*, change, characteristic, chief, china, chinese, choice, christian, christianity, christopher, church, circumstances, citizen, civil-law*, civilization, claim, class, classic, classroom, clerk, clerkship, client*, close, code, collection, college, collision, columbia, columbus, combination, coming, command, commerce, commercial, commission*, committee, common, common-law*, commonwealth, communication, communism, communist, community, comparative, compensation*, completion, complex, complexity, compulsive, concentrate, conception, concern, conduct, conference, conflict, conflict-of-interest, confucian, confucianism, connection, conquest, consciousness, consent, consequence, conservative, consideration, constantinople, constitutional, contemporary, context, continent, contingent, contract*, contrast, contribution, control, controlling, convenience, convention, converse, conviction*, core, corporation, corpus, cost, council, counsel*, counselor*, counterpart, country, courage, course, court*, cover, craft, creativity, credit, crime*, criminal*, criticism, cross, crown, curriculum, custom, cut, daily, damages, danger, days, deal, dealing, death, deciding, decision, decline, defeat, defendant*, defense*, deference, definition, degree, delay, delivery, demand, departed, department, dependent, des, desire, details, determination, deterrence, developing, development, dictation, difficulty, direction, director, disadvantage, disagreement, disbarment*, discipline, disclosure, discovery, discussion, dismissal, dispenser, dispute, disqualification, dissertation, distinction, distribution, district, distrust, divergence, diversity, division, divorce, do, doctorate, doctrine, document, documentary, doing, dominant, doubt, down, draft, drafting, draw, drawback, drinker, driver, driving, duration, duty, east, economics, edict, editor, education, effect, efficacy, election, elements, elizabethan, emile, eminence, emphasis, emphasizing, empire, employ, employee, employment, enactment*, end, enemy, engagement, england, english, enough, entering, entity, entrance, entrant, entry, equal, equipment, equity, equivalent, escape, essential, establishment, estate, ethics, etiquette, europe, european, evidence*, evolution, examination, example, exception, execution, exemption, exercise, existence, expansion, experience, expert, expertise, expertness, export, expounding, expulsion, extension, extent, f, face, fact, failure, fair, fairness, faithful, falling, familiar, family, farm, fashion, favour, feature, federal, federation, fee, feeling, fell, felt, few, fiction, field, fifth, fight, filing, final, finance, find, fine, first, five, focus, focusing, following, force, form, former, found, foundation, founding, four, fourth, framing, france, free, freedom, french, function, fusion, future, g, gain, general, genesis, german, germanic, germany, give, given, go, goal, good, governing, government, graduate, graduation, greece, group, growing, growth, guard, guidance, guide, guilt*, half, hall, hand, harmony, harvard, have, he, help, helping, henry, hierarchy, high, hire, history, hold, honours, hostility, idea, identification, ii, impact, imperfection, imperial, importance, imprisonment*, in, incident, incidental, income, incompetent, incorporation, increase, independence, independent, index, india, indian, individual, individualism, industry, infancy, inference, influence, information, inheritance, initiative, injury, inn, innocent, inquiry, insistence, institute, instruction, instrument, insurance, intellectual, intent, interchange, interest, intermediate, internship, interpretation, introduction, investigating, investigation, ireland, issue, italian, italy, j, jail, january, japan, japanese, jay, job, john, joint, joseph, journal, judge*, judgeship*, judgment*, judiciary*, junior, jurisdiction*, jurisprudence*, jurist*, jury*, justice*, justification, k, keep, king, kingdom, knowledge, korea, l, labour, land, language, large, last, lateral, latin, latter, law*, lawyer*, lay, lead, leader, leading, learning, least, leave, leaving, lecture, left, legality*, legislation*, legislature*, lenin, less, level, liberal, library, licentiate, life, lifetime, light, limit, listening, literature, litigant*, litigation*, little, living, local, logic, london, long, look, lost, low, lower, loyalty, lund, m, magistracy*, magistrate*, main, maintenance, major, make, making, malaysia, malpractice, management, mandatory, manifest, manner, mark, mary, master, material, matter, may, means, measure, mediterranean, meet, member, memory, men, mentality, mere, method, methodology, middle, middle-class, might, minimization, minimum, ministry, miscellany, misconduct, mistrust, mixture, mobility, mode, model, moderate, moderation, modern, modification, morals, more, most, motive, move, much, murphy, muslim, name, narrow, nation, national, native, nature, nd, necessary, necessity, need, negligence, negotiation, negotiator, netherlands, nigeria, no, noble, nordic, norm, normal, norman, north, northern, notary*, notice, notion, now, nowhere, number, objection, objective, obligation, observance, obstacle, offense*, offer, offering, office, officer, official, offset, one, open, opening, operation, opinion, opponent, oppression, oral, orator, order, ordinary, organization, organs, origin, original, out, outcome, outgrowth, outline, outlook, outside, over, ownership, oxford, pakistan, panorama, papers, pardon, paris, part, participation, particular, partisan, partnership, party, pass, passenger, passing, path, patient, patrician, pattern, pay, payment, pedagogy, people, percent, percentage, performance, period, periodical, permit, persecution, person, personal, philosopher, pierre, pioneer, place, plan, planning, plantagenet, plato, play, plebeian, point, police, policy, portuguese, position, possible, postgraduate, potential, power, practice, practitioner, precedent*, prejudice, premises, preparation, prerequisite, present, press, prestige, pretrial*, prevention, primary, prince, principle, prior, priority, private, problem, procedure, proceedings*, proceeds, process, procurator*, produce, product, profession, professional, professor, profits, program, prohibition, promotion, proof, property, proportion, prosecution*, prosecutor*, providing, province, provisions, public, publishing, punishment*, pupil, purchase, purpose, pursued, put, putting, qualification, qualifying, quantity, quarterly, queens, question, questioning, quota, raise, range, rarity, rate, ratio, ray, re, reading, ready, real, reason, reasoning, receiving, reception, recognition, reconciliation, record, recovery, reexamination, reference, reform, regard, regime, regular, regulating*, regulation*, rehabilitation, relation, relationship, relative, relief, religion, religious, remains, remove, report, representation, republic, reputation, repute, requirement, research, resemblance, resolve, resolving, resort, respect, responsibility, rest, restraint, restriction, result, return, revenue, review, revival, revolution, reward, rhetoric, right*, rigidity, rise, risk, road, role, roman, rome, room, royal, rule, ruling, russia, s, scandinavia, scanty, scholarship, school, science, scope, scotland, scottish, screening, second, secondary, section, see, seek, seeking, selection, self-interest, seller, seminar, senate, senior, sense, separate, series, servant, serve, service, serving, set, setting, settlement, settling, shaping, share, sharing, short, shortage, side, significance, simple, singapore, single, sir, situation, six, size, skill, small, so, social, socialist, society, solicitation*, solicitor*, someone, something, source, south, southeast, soviet, spain, spanish, special, specialist, specialization, specific, speech, spirit, spread, stable, staff, stage, stake, standing, staple, start, starting, state, statement, status, statute*, stay, steps, still, stone, story, straight, stress, structure, student, study, style, subcontinent, subject, subordinate, substantive, success, sum, summary, supervision, supplement, support, survey, sweden, swedish, syllabus, system, t, tactics, take, taking, task, tax, taxation, teach, teacher, teaching, tell, telling, temple, tendency, tension, terms, test, testimony*, testing, text, textbook, th, thailand, theme, then, theory, there, thesis, think, thinking, third, third-, thomas, thought, three, times, title, today, tomb, tone, tort, total, trace, trade, tradition, traffic, training, transaction, transfer, transition, transmitting, treatment, trend, trial*, tribunal*, true, truth, turn, two, type, undergraduate, understanding, undoing, uniform, uniformity, union, university, unknown, unrest, us, use, using, vacancy, validity, variety, vest, view, vocation, voluntary, wait, wake, wales, war, washington, wave, ways, well, west, western, while, who, whole, why, will, willingness, winding, wisdom, wise, wish, witness*, words, work, working, world, worse, wound, writing, years, york, young, zealand

B.2.2 Nouns in Training Set

advocacy*, advocate*, age, agent, aid, aim, american, amount, an, anglo-american, appointment, are, aristotle, as, association, at, attending, attention, attorney*, authority, bar*, barrister*, be, body, british, business, cambridge, can, case*, century, civil-law*, client*, code, common, common-law*, commonwealth, comparative, conduct, conflict, control, convention, counselor*, court*, criminal*, details, development, disclosure, divorce, education, employment, england, english, entry, establishment, ethics, europe, examination, example, existence, fact, family, first, following, founding, france, function, general, given, government, growing, hand, harvard, he, hold, idea, importance, in, increase, independence, index, individualism, inheritance, introduction, ireland, japan, journal, judge*, justice*, knowledge, law*, lawyer*, leader, life, lifetime, may, more, most, much, necessary, negotiation, norm, northern, notary*, obligation, official, original, oxford, panorama, pardon, person, place, plato, possible, power, practice, precedent*, premises, present, principle, produce, profession, professional, prosecutor*, providing, public, publishing, quarterly, ratio, reason, regulation*, relation, representation, republic, respect, role, roman, school, scottish, secondary, see, seek, sense, serve, services, setting, so, social, socialist, solicitor*, spirit, spread, state, statute*, step, student, study, subordinate, system, task, teaching, testimony*, there, third, third-, times, title, training, type, university, wales, war, ways, while, who

B.2.3 Nouns in Test Set

a, ability, academic, account, acquiring, action*, addition, administration, admission, advertising, advice, adviser, advocacy*, advocate*, africa, agent, agreement, aid, america, american, an, anglo-american, annual, appearance*, application, appointment, apprenticeship, approach, are, argument, arthur, as, ascending, asia, assist, assistance, assistant, association, at, athens, attainment, attitude, attorney*, autonomy, awareness, backbone, bar*, barrister*, basic, be, behind, being, better, body, bologna, bordeaux, branch, bringing, britain, british, business, buyer, buying, can, candidate, canon, career, carry, case*, case-law*, category, cause, central, century, chancery*, change, chief, china, circumstances, claim, client*, college, commercial, common, common-law*, commonwealth, communication, comparative, completion, complex, complexity, conception, conduct, conference, conflict, consent, constitutional, contemporary, continent, contingent, control, convenience, converse, conviction*, counselor*, country, course, court*, creativity, crime*, criminal*, criticism, cross, damages, dealing, death, deciding, decision, decline, defense*, deference, degree, delivery, department, detail, developing, development, direction, disagreement, disbarment*, discipline, discussion, dismissal, dispenser, disqualification, district, divergence, division, do, doctorate, doctrine, dominant, duration, duty, east, economics, editor, education, effect, elements, emphasis, employee, employment, enactment*, end, enemy, england, english, entering, entity, entrance, entry, equity, ethics, europe, european, evidence*, examination, example, exception, execution, expansion, experience, expert, extent, fact, falling, familiar, family, fashion, federal, feeling, felt, few, fifth, filing, finance, first, five, following, force, form, former, four, fourth, france, french, function, general, genesis, german, germanic, germany, given, good, government, graduate, growth, hall, hands, harmony, have, he, help, high, history, identification, imperial, importance, in, independence, independent, index, india, individual, influence, information, inn, institute, instrument, insurance, intent, interest, intermediate, introduction, italian, italy, j, jail, japan, japanese, jay, joint, judge*, judgment*, judiciary*, junior, jurisprudence*, justice*, k, kingdom, knowledge, kw, land, language, large, last, latin, law*, lawyer*, lay, lead, leading, learning, least, lecture, legislation*, legislature*, lenin, less, library, licentiate, light, limit, litigant*, litigation*, little, local, london, long, look, lower, magistrate*, main, make, making, malpractice, manifest, master, material, may, member, memory, men, method, middle, mode, moderation, modern, more, most, much, murphy, muslim, narrow, national, necessary, need, negotiation, negotiator, netherlands, noble, notary*, now, number, objection, offer, offering, office, officer, official, offset, one, operation, opinion, ordinary, organ, original, out, outcome, outgrowth, outside, over, oxford, particular, partisan, parts, pass, passing, pattern, payment, pedagogy, period, permit, person, place, police, portuguese, position, possible, power, practice, practitioner, premises, preparation, present, prestige, primary, prior, priority, private, problem, procedure, process, procurator*, produce, profession, professional, professor, program, property, prosecution*, province, provisions, public, purpose, pursued, putting, qualification, qualifying, question, raise, ready, real, reasoning, reconciliation, record, reference, regard, relation, religion, religious, report, representation, reputation, repute, requirement, research, revenue, review, revival, revolution, right*, role, roman, rome, royal, russia, scandinavia, school, science, scotland, second, secondary, see, seek, seminar, senior, sense, series, servant, serve, service, serving, setting, settlement, settling, shaping, short, side, significance, singapore, sir, situation, so, social, socialist, society, solicitation*, solicitor*, something, source, southeast, spanish, special, specialization, specific, starting, state, status, steps, still, straight, stress, structure, student, study, subject, sum, summary, support, survey, swedish, syllabus, system, t, take, taking, teaching, temple, tendency, tension, terms, test, testimony*, th, thailand, then, there, third, thought, three, times, title, today, total, training, transaction, transmitting, treatment, trend, trial*, true, truth, two, uniform, university, us, use, using, wake, wales, washington, wave, way, well, western, while, who, will, winding, wisdom, wise, work, working, world, worse, years, young

Appendix C

Data on Sense-Level Path Experiments

C.1 Nouns in Training Set

[1] academic(1), accounting, achievement, action, address, advantage, algebra, alignment, amount, an, approach, are, area, as, assistance, at, auspices, basis, be, being, better, calculus, can, campus(1), career, carnegie, case, circumstances, city, class(2), classroom(1), clear, college(1, 2, 4), commission, computer, concept, conclusion, core, cost, course(1), data, description, detail, development, do, dominant, education(1, 4), effect, emphasis, enrollment, enterprise, environment, essential, exercise, fact, faculty, failure, familiar, feedback, figure, first(5), format, four, given, gold, grade(1), group, hands, have, high(6), hours, identification, in, increase, information, instruction(2, 3), instructor(1), learner(1), learning, lecture(3), leeds, less, lieu, lisp, local, material, mathematics, model, more, most, multimedia, notion, number, obligation, one, organization, out, over, parallel, passing, percentage, performance, personal, population, principal, public, put, quantity, question, raising, range, rate, regular, repeat, research, result, review, salary, savings, section, semester(1), sequence, serve, short, six, software, state, still, student(1), success, support, table, take, teaching(1, 3), technology, term, three, times, two, university(1, 2, 3), value, week, who, york

C.2 Nouns in Test Set

ability, academic(1), access, account, achievement, acquisition, action, activity ad, addition, address, adept, administration, admission, admonition, advantage, advice, algebra alignment, allocation, alternative, american, amount, analysis, answer, application approach, architecture, assessment, assistance, association, attending attribute, author, average, baccalaureate(2), background, basic, basis, be, beginning, behavior being, benefit, breakdown, bridge, brown, campus(1), capability capacity, case, cause, center, change, circumstances, class(2,4,6), classroom(1), cognition

[1] Words followed by brackets are keywords. Numbers in brackets are the numbers of the key senses.

cohort, college(1,2,4), college-level(1), collins, combination, community, comparing, computer conclusion, conduct, confronting, consonant, content, contract, cost, council, course(1) cover, creativity, credit(6), cycle, data, database, date, debate, deciding, degree(3), demand department, dependent, detail, developing, difference, directive, discipline, diversity drop, economics, education(1,4), effect, effectiveness, efficacy, effort elm, emphasis, employ, employment, end, engine, enrollment, enterprise, environment, equipment equivalent, essay, evaluation, evidence, exam(1), examination(5), example, exception exercise, experiment, exposure, extension, extra, fact, faculty(2), failing, fall, fashion, feedback figure, focus, following, form, format, framework freedom, function, funding, future, gauge, grade(1), grading, graduation(1,2) graph, grounds, group, guiding, half, hand, health, help, high-quality, higher-education(1) homework(1), hours, hypothesis, impact, implementation, improvement increase, individual, information, initial, insensitivity, institute institution, instruction(2,3), instructor(1), introduction issue, keeping, kind, knowledge, language, leadership, learner(1) learning, lecture(3), lesson, level, leverage, line, links, lisp making, management, mastering, mastery, material, mathematics meaning, means, measure, meet, mid-term(2,3), mind, minority, minutes, model, monitor, monitoring multimedia, nature, necessity, need, net, note, notebook numbers, occurrence, offer, one, one-third, opportunity, option, order, organization outline, overall, pacing, page, paper, part, pass, passing, percent percentage, performance, period, persistence, personal, plan, planning, play, point policy, polytechnic(1), predecessor, preparation(5), present, pressing pressure, probability, problem-solving, process, productivity, program(4), programming progress, progression, providing, quality, question, quick, quiz, raise, range, rate, readiness record, reducing, regulating, relationship, repeating, requirement, research result, retention, revenue, review, rise, role, routine, run, savings, say, schedule, scheme school(1,2,3,4,5), scope, scores, second, section, seeing, semester(1), sequence, session(2), set, setting shift, shortening, shrink, sort, sound, standard stands, start, state, stay, student(1), study(2), subject, subpopulation success, summer, support, supporting, survey, swelling, syllabus(1), system, table, tailor talk, teach, teaching(1,2), technology, terminal, terms, test(4), testing, text textbook(1), theory, thinking, time, tool, total, tutoring type, understanding, university(1,2,3), use, using, value, vehicle, way week, withdrawal, work, working, workload, years(4), york

Appendix D

Relation Word Set of Education

D.1 Sense 1 of Education

D.1.1 Synonym

instruction, teaching, pedagogy, educational-activity

D.1.2 Coordinate

variation, variance, space-walk, domesticity, operation, practice, pattern, diversion, recreation, cup-of-tea, bag, dish, game, turn, play, acting, playing, playacting, performing, liveliness, animation, burst, fit, cinch, picnic, snap, duck-soup, child's-play, pushover, walkover, piece-of-cake, work, works, deeds, service, occupation, business, line-of-work, line, occupation, writing, role, wrongdoing, misconduct, waste, wastefulness, dissipation, attempt, effort, endeavor, endeavour, try, control, controlling, protection, protecting, guarding, sensory-activity, training, preparation, grooming, representation, creation, dismantling, dismantlement, disassembly, puncture, puncturing, search, searching, hunt, hunting, use, usage, utilization, utilisation, employment, exercise, measurement, measuring, measure, mensuration, organization, organisation, arrangement, grouping, support, supporting, continuance, continuation, procedure, process, ceremony, ceremony, worship, energizing, activating, activation, presentation, concealment, concealing, hiding, secreting, location, locating, placement, position, positioning, emplacement, situating, provision, providing, supply, supplying, demand, pleasure, enjoyment, delectation, market, marketplace, preparation, readying, aid, assist, assistance, help, helping, support, behavior, behaviour, conduct, behavior, behaviour, leadership, leading, precession, precedence, solo, buzz, fun, sin, hell, release, outlet, last, perturbation, disturbance, timekeeping

D.1.3 Coordinatee

act, human-action, human-activity

D.1.4 Hypernym

activity

D.1.5 Hyponym

coeducation, course, course-of-study, course-of-instruction, class, elementary-education, extension, extension-service, university-extension, extracurricular-activity, higher-education, secondary-education, work-study-program,

D.1.6 Meronym

project, classroom-project, homework, prep, preparation, lesson,

D.1.7 Holomyn

Null.

D.2 Sense 4 of Education

D.2.1 Synonym

profession,

D.2.2 Coordinate

profession, learned-profession, literature, architecture, journalism, politics, technology, engineering,

D.2.3 Coordinatee

profession, occupation, business, line-of-work, line,

D.2.4 Hypernym

profession,

D.2.5 Hyponym

teaching, instruction, pedagogy,

D.2.6 Meronym

Null.

D.2.7 Holonym

Null.

Appendix E

Trained Weight Analysis: Data for Relation Level

E.1 Meronym relations

Table E.1 shows the weights connecting all the meronym bits in an input pattern to the second hidden node. No MWs and relation patterns can be extracted. The weights in table E.1 never changed during the training of the ANN. This is because of the lack of meronym relations in the training data (WordNet has not fully implemented the meronym relation). It is concluded that the meronym relation is not important to the ANN.

Table E.1: Weights from Meronym Relations to Hidden Node 2

Name	Location	Path1	Length
Meronym11	1	0.42	1
Meronym21	9	0.22	2
Meronym22	17	0.21	
Meronym31	25	-0.15	3
Meronym32	33	0.19	
Meronym33	41	0.46	
Meronym41	49	-0.25	4
Meronym42	57	-0.42	
Meronym43	65	-0.41	
Meronym44	72	0.07	

E.2 Hyponym relations

Table E.2 shows the weights connecting all the hyponym bits in an input pattern to the second hidden node. Three important relation patterns involving hyponym are extracted, as shown in table E.3.

Table E.2: Weights from Hyponym Relations to Hidden Node 2

Name	Location	Path1	Length
Hyponym11*	2	0.86	1
Hyponym21	10	0.29	2
Hyponym22	18	0.18	
Hyponym31	26	-0.11	3
Hyponym32**	34	2.06	
Hyponym33**	42	-1.58	
Hyponym41	50	0.27	4
Hyponym42	58	-0.32	
Hyponym43	66	0.05	
Hyponym44	74	-0.18	

Table E.3: Hyponym Relation Patterns

Patterns			Meaning
+Hyponym			A path composed of a hyponym of the seed word is a KP (i.e. a hyponym of the seed word tends to be a KW)
*	+Hyponym	*	A path of length 3 with a Hyponym as the second relation of the path is a strong indicator of a KP
*	*	-Hyponym	A path of length 3 with a Hyponym as the third relation of the path is a strong indicator of a NKP

E.3 Hypernym relations

Table E.4 shows the weights connecting all the hypernym bits in an input pattern to the second hidden node. No MWs and no relation patterns can be extracted.

E.4 Coordinatee relations

Weights connecting all the coordinatee bits in an input pattern to the second hidden node are shown in table E.5. Four relation patterns concerning coordinatee are identified, as shown in table

Table E.4: Weights from Hypernym Relations to Hidden Node 2

Name	Location	Path1	Length
Hypernym11	3	-0.11	1
Hypernym21	11	0.41	2
Hypernym22	19	-0.13	
Hypernym31	27	-0.11	3
Hypernym32	35	0.08	
Hypernym33	43	-0.37	
Hypernym41	51	-0.01	4
Hypernym42	59	0.48	
Hypernym43	67	-0.12	
Hypernym44	75	0.10	

E.6.

Table E.5: Weights from Coordinatee Relations to Hidden Node 2

Name	Location	Path1	Length
Coordinatee11	4	0.59	1
Coordinatee21**	12	-4.49	2
Coordinatee22**	20	-2.30	
Coordinatee31*	28	-1.02	3
Coordinatee32**	36	1.62	
Coordinatee33	44	0.14	
Coordinatee41	52	-0.04	4
Coordinatee42	60	-0.33	
Coordinatee43	68	-0.04	
Coordinatee44	76	0.35	

Table E.6: Coordinatee Relation Patterns

No.	Patterns			Meaning
1	-coordinatee	*		A path of length 2 with a coordinatee as the first relation of the path is a stromg indicator of a NKP
2	*	-coordinatee		A path of length 2 with a coordinatee as the second relation of the path is a strong indicator of a NKP
3	-coordinatee	*	*	A path of length 3 with a coordinatee as the first relation of the path is an indicator of a NKP
4	*	+coordinatee	*	A path of length 3 with a coordinatee as the second relation of the path is a stromg indicator of a KP

Patterns 1 and 3 mean that a noun having a coordinatee of the seed word is probably not a key path. Patterns 1 and 2 together mean that a path of length 2 with a coordinatee relation is unlikely a key path. Note Parttern 4 is positive. This probably means that the coordinatee is useful in connecting relations in the first and third relation.

The strong negative MWs tell us that the ANN mainly uses coordinatee to discount paths as being KPs.

E.5 Coordinate relations

Weights connecting all the coordinate bits in an input pattern to the second hidden node are shown in table E.7. One relation patterns concerning coordinate are identified, as shown in table E.8.

Table E.7: Weights from Coordinate Relations to Hidden Node 2

Name	Location	Path1	Length
Coordinate11*	5	-0.91	1
Coordinate21	13	0.47	2
Coordinate22	21	-0.69	
Coordinate31	29	-0.63	3
Coordinate32*	37	-0.84	
Coordinate33*	45	1.31	
Coordinate41	53	-0.26	4
Coordinate42	61	-0.51	
Coordinate43	69	-0.05	
Coordinate44	77	-0.36	

Table E.8: Coordinate Relation Patterns

No.	Patterns			Meaning
1	-coordinate			A path of length 1 with a coordinate as the relation is an indicator of a NKP
2	*	-coordinate		A path of length 3 with a coordinate as the second relation of the path is an indicator of a NKP
3	*	*	+coordinate	A path of length 3 with a coordinate as the third relation of the path is an indicator of a KP

E.6 Antonym relations

Table E.9 shows the weights connecting all the antonym bits in an input pattern to the second hidden node. No MWs and no relation patterns can be extracted.

Table E.9: Weights from Antonym Relations to Hidden Node 2

Name	Location	Path1	Length
Antonym11	6	-0.42	1
Antonym21	14	0.41	2
Antonym22	22	0.27	
Antonym31	30	-0.23	3
Antonym32	38	-0.36	
Antonym33	46	0.14	
Antonym41	54	-0.09	4
Antonym42	62	0.47	
Antonym43	70	-0.14	
Antonym44	78	0.29	

E.7 Synonym relations

Weights connecting all the synonym bits in an input pattern to the second hidden node are shown in table E.10. Five relation patterns concerning synonym are identified, as shown in table E.11.

Table E.10: Weights from Synonym Relations to Hidden Node 2

Name	Location	Path1	Length
Synonym11**	7	3.80	1
Synonym21*	15	0.89	2
Synonym22*	23	-1.05	
Synonym31**	31	-2.09	3
Synonym32**	39	-1.96	
Synonym33	47	-0.55	
Synonym41	55	0.69	4
Synonym42*	63	-0.81	
Synonym43*	71	-1.00	
Synonym44*	79	-0.86	

Table E.11: Synonym Relation Patterns

No.	Patterns				Meaning
1	+Synonym				A path of a synonym of the seed word is a KP (i.e. a synonym of the seed word tends to be a KW)
2	+Synonym	*			A path of length 2 with a synonym as the first relation of the path is an indicator of a KP
3	*	-synonym			A path of length 2 with a synonym as the second relation of the path is an indicator of a NKP
4	-synonym	*	*		A path of length 3 with a synonym as the first relation of the pathis a strong indicator of a NKP
5	*	-synonym	*		A path of length 3 with a synonym as the second relation of the path is a strong indicator of a NKP
6	*	-synonym	*	*	A path of length 4 with a synonym as the second relation of the path is an indicator of a NKP
7	*	*	-synonym	*	A path of length 4 with a synonym as the third relation of the path is an indicator of a NKP
8	*	*	*	-synonym	A path of length 4 with a synonym as the fourth relation of the path is an indicator of a NKP

Appendix F

Sentences Containing Two Keywords

Sentences used in the relation extraction experiments that contain two sentences are listed in this appendix.

F.1 Sentences from the Education Domain

Because this is the case for many higher-education policymakers, one of the major objectives of this essay is to clarify the important issues for those in leadership roles in higher education, especially those who find themselves perplexed by talk of technology, teaching, learning and the future.

I intend to embark upon this clarification by describing the model developed by my colleagues and me to improve the mathematics teaching and learning enterprise, particularly for entry-level college students.

Our approach, called Mediated Learning, which is currently being implemented in colleges and universities around the country, has grown out of a willingness to rethink the very nature and character of that enterprise.

We devised this new model of instruction, learning and assessment with the understanding that expert instructors and instructional designers could use new authoring technologies to create flexible, interactive and adaptive computer-mediated instructional materials incorporating standard media formats, including text, hypertext, graphics, animation, simulations, digital audio and digital video.

When supported by integrated multimedia database and networking technologies, these programmed materials and the learning activities they afford could be continuously and dynamically ordered to provide instructors and their students task-specific learning assistance, as well as real-time information on learner performance over local area networks of personal computers.

This information on learner progress and achievement could be analyzed by instructors and their students to plan possible future courses of action.

It could also be used by instructors to analyze possible relationships between students lesson-navigation patterns and their academic achievement.

That same information could be used by students to analyze their own learning proclivities and by instructional designers to inform continuous improvement efforts.

Employed in this fashion, authoring, networking and database technologies could provide crucial support for instruction that would be more faculty-guided, more learner-paced and more individualized than is possible in most traditional classrooms, where the whole-class curriculum, lock-step teaching method predominates, especially in the case of lower-division undergraduate courses taken by entry-level students.

Developing Mediated Learning and its range of supporting technologies so that it capitalizes on those strengths of technology-mediated education and integrating it into classroom environments has permitted instructors and their students to spend more time together engaged in high-impact cognitive and developmental activities, and less time on activities that minimally impact cognitive development and learning.

As Mediated Learning becomes recognizable as a unique model of technology-mediated instruction, learning and assessment, the central question facing faculty and administrators is : Is this model cost-effective?

Even if technology helps students learn more proficiently, can we afford to invest in it?

The purpose of this paper is to address these questions and provide a new framework for analyzing the cost-effectiveness of technology-mediated instruction.

In summary, the central point of this paper is that the key to fiscal management in higher education is not cost-cutting; it is improved learner-productivity and increased student academic achievement.

By enhancing the learning environment, increasing student passing rates and developing new sources of revenue, colleges and universities can recover their investments in technology, increase support for faculty and generate new net resources.

Roger E.Levien and associates, in their prescient examination of the economic and policy issues accompanying the employment of computer-assisted instruction in higher education, noted that cost-effective is a " deceptively simple " term " which has frequently been misapplied in the examination of instructional computer use " (1972, p.436).

Frequently, the misapplication of cost and effectiveness measures stems from unclear definitions and unsound metrics, such as the imposition of one set of metrics for evaluating technology-mediated instruction and another for assessing traditional lecture-based instruction.

Problems also result from reluctance of some faculty to lend any level of legitimacy to the topic of cost-effectiveness on the grounds that to do so would constitute an open invitation to non-educators to apply uninformed and inappropriate evaluation and assessment measures to the higher education teaching and learning enterprise.

The caution of the Levien team, therefore, is well founded; the pedagogical, policy and political nuances and complexities of the higher education teaching and learning enterprise are not easily captured by standard cost-accounting methodologies and they remain highly problematic two decades later.

More recent articles addressing the relationship between technology-mediated instruction and instructional productivity by Gifford (1991), Green and Gilbert (1995a, 1995b), Levin (1991) and Plater (1995) make it clear that researchers may be no closer to establishing consistent metrics for comparing the cost-effectiveness of traditional and technology-mediated instruction than when the Levien study was published in 1972 under the auspices of the Carnegie Commission on Higher Education.

Still, there is a pressing need for evaluation and assessment metrics for comparing traditional lecture-centered instruction and technology-mediated instruction models, especially in light of the fierce pressure on higher education institutions to justify new instructional initiatives in terms of their impact on the economics of the teaching and learning enterprise.

Mindful of this necessity, and of Leviens admonition against inappropriately applying the term " cost-effective, " this paper presents a framework for comparing the relative cost-effectiveness of Mediated Learning to that of traditional lecture-centered instruction.

Mediated Learning is a form of technology-mediated instruction, learning and assessment that is being used by several hundred faculty on campuses around the country in order to prepare students who enter college underprepared for college-level mathematics.

To date, this method has been utilized to develop a series of introductory mathematics courses called Interactive Algebra for College Students I, II and III.

The Mediated Learning approach enables faculty to improve the instructional environment by reconfiguring the traditional, lecture-centered classroom.

Faculty use specially constructed, interactive, multimedia lessons delivered via computers, which enable instructors and students to teach and learn more effectively and efficiently.

In Section 1, " Mediated Learning: A New Model of Instruction and Learning, " a history of Mediated Learnings development, as well as the unique instructional environment facilitated by Mediated Learning and its components, are briefly introduced, and reasons for implementing Mediated Learning, aside from its cost-effectiveness, are discussed.

Far from being simply an economically sound alternative to traditional instruction, Mediated Learning also yields pedagogical benefits, which although not the focus of this paper, need to be articulated in order to demonstrate that this technology-mediated approach is desirable for many reasons, not merely monetary ones.

Section 2, " Findings on the Effectiveness of Mediated Learning, " provides findings being reported by campuses that have been implementing Interactive Algebra I and II.

One schools data, because it is fairly extensive, is analyzed in some detail and demonstrates that, in most cases, a greater percentage of students who take one or two Mediated Learning courses before enrolling in Precalculus earn grades of C and above in Precalculus than their peers who take either no algebra at college or traditional Beginning and/or Intermediate Algebra.

In Section 3, " Impact of Student Completion Rates on Enrollment Patterns, " simulations are presented in which incremental changes in student completion rates are shown to impact student enrollment patterns significantly.

The conclusion is that increases in completion rates caused by the implementation of technology-mediated instruction can have a significant impact on the number of students that successfully complete a course or sequence of courses.

In Section 4, " Mediated Learnings Effect on Instructor Workloads, " the analysis continues with an examination of the nature and character of instructional work in traditional instructional settings and in Mediated Learning environments.

The conclusion is that in Mediated Learning settings, instructors can take on more students per class without increasing their instructional workload.

Section 5, " The Economics of Mediated Learning, " uses the information from Sections 3 and 4 to indicate that Mediated Learning need not increase instructional costs in those settings where it is implemented and is, in fact, more cost effective.

Mathematics was selected as the first discipline for which Mediated Learning lessons would be developed in order to help solve one of the most vexing and persistent problems confronting the higher education mathematics communitythe issue of addressing effectively, efficiently and flexibly the learning assistance needs of entry-level students who arrive on college campuses underprepared for college-level studies in mathematics.

The dimension of this challenge, manifested in the rapidly increasing enrollment of students placed in remedial and developmental mathematics courses, is documented in the latest survey of enrollment trends in mathematics courses conducted by the Mathematical Association of America (MAA) (Alberts, Loftsgaarden, Rung &; Watkins, 1992).

Definitions: Remedial courses include Arithmetic, General Mathematics (Basic Skills), High School Elementary Algebra and High School Intermediate Algebra.

Precalculus courses include College Algebra, Trigonometry, Combined College Algebra and Trigonometry, Elementary Function Precalculus, Mathematics for Liberal Arts Students, Finite Mathematics, Business Mathematics, Mathematics for Elementary School Teachers, Analytical Geometry and other Precalculus courses.

Calculus courses include Mainstream Calculus I, II, III, etc., Non-Mainstream Calculus I, II, III, etc., Differential Equations, Discrete Mathematics, Introduction to Mathematical Logic, Linear Algebra and Other Calculus courses.

As depicted in Table 1, above, in the case of four-year colleges, between the fall of 1970 and the fall of 1990, enrollment in remedial mathematics courses increased nearly three times faster than enrollment in calculus courses, nominally the entry-level college mathematics course for well-prepared entry-level college students.

Also note that more than one-third of the total increase in enrollment in mathematics courses between 1970 and 1990 (434,000) was due to increased demand for remedial courses (162,000).

In fact, in four-year colleges, enrollment in remedial mathematics courses climbed to 16 percent of the total mathematics enrollment in the fall of 1990 from eight percent in the fall of 1970.

In the case of two-year colleges, and very much reflecting their traditional mission as open-entry institutions, the challenge of appropriately addressing the instructional support needs of underprepared students is even more formidable, as evidenced by the data depicted in Table 2, above.

Between the fall of 1970 and the fall of 1990, the enrollment increase in remedial and precalculus mathematics courses alone (644,000) was greater than the total enrollment growth in all mathematics courses in all four-year colleges and universities (434,000).

Moreover more than 77 percent of the total increase in enrollment in mathematics courses between 1970 and 1990 (686,000) was due to increased demand for remedial courses (533,000).

In two-year colleges, enrollment in remedial mathematics courses climbed to 58 percent of the total mathematics enrollment in the fall of 1990 from 34 percent in the fall of 1970.

Disturbing as these data in Tables 1 and 2 might be, they seriously underestimate the number of underprepared mathematics students attending the nations colleges and universities, particularly in larger industrial states with diverse college student populations, such as California and New York, where the California State University (CSU) and City University of New York (CUNY) campuses are located, respectively.

The MAA survey does not account for students taking remedial and precalculus courses outside of mathematics departments, say, in departments specializing in remedial developmental mathematics or in programs headquartered in learning centers specifically organized to address the learning assistance needs of underprepared students.

Nor does the MAA survey disaggregate student enrollment data down to the state level, nor by student language origins.

However, because of Californias and New York States large college student populations and their high level of student diversity, it is difficult to imagine that a breakdown of the MAA data for California or New York would have uncovered enrollment patterns and trends significantly dissimilar from those depicted in Tables 1 and 2.

If anything, these states large and growing population of non-native English speakers might lead one to conjecture reasonably that the enrollment patterns and trends depicted in Tables 1 and 2 actually underestimate the numbers of students in these states who arrive on college campuses underprepared for college-level studies in mathematics.

The enrollment trends depicted in Tables 1 and 2 elucidate the challenge confronting college mathematics departments and underscore the policy implications of a larger cohort of underprepared students for the larger higher education community.

Courses in remedial and developmental mathematics, as well as entry-level, core mathematics courses, fall into the category of " high-stakes " courses.

Their cost to the university escalates if students are required to repeat them.

In terms of opportunity costs, failure in these courses can exact a high price.

F.2 Sentences from the Law Domain

The primary function of the profession and practice of law is to apply the law in specific cases–to individualize it.

This function is manifest in the work of the advocate and the judge in the process of trying and deciding cases.

In Anglo-American systems a lawyer investigates the facts and the evidence by conferring with his client, interviewing witnesses, and reviewing documents.

At the trial he introduces evidence, objects to improper evidence from the other side, and advances partisan positions on questions of law and of fact.

At trial he plays an active role in taking evidence, questioning witnesses, and framing the issues.

Continental lawyers suggest lines of factual inquiry to the judge and, like their Anglo-American counterparts, advance legal theories and argue the law in accord with the interests of their clients.

In either system, if a lawyer loses his client 's case, he may seek a new trial or relief in an appellate court.

Even controversies that are not resolved in court require the aid of lawyers.

Negotiation, reconciliation, compromise–in all of which lawyers have a large part–bring about the settlement of most cases without trial.

The lawyer as counselor and negotiator may aid in shaping a transaction so as to avoid disputes or legal difficulties in the future, or so as to achieve advantages for his client, such as the minimization of taxes.

The law gives to private persons extensive but not unlimited power to arrange and determine their legal rights in many matters and in various ways, such as through wills, contracts, leases, or corporate bylaws.

In structuring these arrangements the lawyer is helping to particularize the legal rights of the parties.

Another field of legal work, which has developed rapidly in the 20th century, is the representation of clients before administrative commissions, commissions of inquiry, and, in some countries, legislative committees.

A lawyer has several loyalties in his work, including loyalty to his client, to the administration of justice, to the community, to his associates in practice, and to himself–whether to his economic interests or to his ethical standards.

In the wake of powerful religious or quasi-religious movements priests or wise men often judged or advised the judges, a situation that persisted in Muslim countries and in China until the 20th century AD.

A distinct class of legal specialists other than judges first emerged in the Greco-Roman civilization, and as with the law itself, the main contribution was from Rome in the period from 200 BC to AD 600.

The assumption was that the citizen knew the customary law and would apply it in transactions or in litigation personally with advice from kinsmen.

Often they also spent periods serving as magistrates and in Rome as priests of the official religion, having special powers in matters of family law.

Among the German tribes noble experts were allowed to assist in litigation, not in a partisan fashion but as interpreters (Vorsprecher) for those unready of speech who wished to present a case.

The peculiar system of development of the early Roman law, by annual edict and by the extension of trial formulas, gave the Roman patrician legal expert an influential position; he became the jurisconsult, the first nonofficial lawyer to be regarded with social approbation, but he owed this partly to the fact that he did not attempt to act as an advocate at trial–a function left to the separate class of orators–and was prohibited from receiving fees.

A subordinate legal agent of the classical system, the procurator, who attended to the formal aspects of litigation, took on added importance because later imperial legal procedure depended largely on written documents drawn by procurators.

The jurisconsults had been important as teachers and writers on law; with their decline this function passed to government-conducted law schools at Rome, Constantinople, and Berytus and to their salaried professors.

The Christian Church, which became the official Roman imperial church after AD 381, developed its own canon law, courts, and practitioners and followed the general outline of later Roman legal organization.

Because of its success among the invaders the church was in a position to establish its jurisdiction in many matters of family law and inheritance.

Procurators attended to the formal and especially the documentary steps in litigation.

Advocates, who usually were university graduates in Romanist learning, gave direct advice to clients and to procurators and presented oral arguments in court.

University teachers of law took over the main task of explaining and of adapting the mixture of Roman law and Germanic custom that produced the modern laws of the major European countries and continued to dominate in the scholarly interpretation of the law even after the 19th-century codifications.

At times the teaching doctors almost supplanted the advocates; in some courts the procurators swallowed up the advocates and in others the converse occurred; only the notaries managed to survive with little change.

Only in the ecclesiastical and admiralty courts, however, did procurators (proctors) and doctors of the civil and canon laws become established as practitioners.

The native " common law " was developed by a specialized legal society, the Inns of Court in London; there, through lectures and apprentice training, men acquired admission to practice before the royal courts.

More particularly, they could become serjeants–the most dignified of the advocates, from whom alone after about 1300 the royal judges were appointed.

Various agents for litigation resembling procurators also became known.

The " attorneys, " authorized by legislation, at first shared the life of the Inns with the " apprentices " in advocacy, who themselves in time acquired the title of barristers.

Indeed there were cases of men working as both barristers and attorneys.

When in the 16th century the Court of Chancery was established as the dispenser of " equity, " the appropriate agent for litigation was called a solicitor, but the common-law serjeants and barristers secured the right of advocacy in that court.

It was not until the 17th century that the attorneys and solicitors were expelled from the Inns and the division between advocate and attorney became rigid, and not until the 18th century that the barristers accepted a rule that they would function only on the engagement of an attorney–not directly for the client.

The order of serjeants was wound up, leaving only barristers, of whom the most senior could be made Queen 's (or King 's) Counsel.

Development of the law took place chiefly through precedent based on the reported judgments of the courts, rather than through legislation.

In addition, the division between barristers and solicitors ultimately became much more rigid in England than did the division between the advocate and procurator in Europe, and Europe never adopted an equivalent of the English practice requiring a barrister to be employed by a solicitor; both the procurator and the advocate were separately and directly employed by the client.

England never developed the profession of notary, so that the whole burden of transactional work fell on those who are now the solicitors, with legal advice from the bar.

In particular, the specialization of procurator-advocate and solicitor-barrister has tended to be replaced by a " fused " profession of legal practitioners qualified to perform both functions and usually doing so.

It has taken place more recently in France except before the courts of appeal and, while the division still formally exists in Italy, it is no longer of practical importance.

Notaries as a separate specialized branch of the profession exist, however, in most civil-law countries.

Leading lawyers have usually been socially prominent and respected–the sections of the profession so favoured varying with the general structure of the law in the particular community.

The early Italian doctors of the civil and canon law (12th-15th centuries) were revered throughout Europe.

In England and the countries influenced by its system the highest prestige gradually came to be concentrated on the judges rather than on the order of serjeants, of which they were members, and the judges of high-level courts remain the only legal class in the liberal capitalist common-law countries of today to command great respect.

In the Romano-Germanic systems it is the notaries and the advocates who have come to be most trusted or admired, the judiciary being more closely identified with the civil service.

In a few cases this has been the consequence of a general hostility to the whole idea of law, China being the most important example.

In the Soviet Union the early leaders (1917-22) imagined that law and lawyers were the instruments of the ruling classes and that law would soon wither away in classless Communism.

Further experience persuaded these governments that there was room for " socialist legality " and for lawyers to serve it, but a degree of mistrust remains and the repute of the legal expert is lower than that of the political and technological expert.

Most lawyers are conservative because the law itself is predominantly intended to satisfy expectations arising from an inherited pattern of behaviour; in a particular social setting this tends to identify the lawyer with the established and successful classes and to make him seem an enemy to oppressed classes or " new men. "

Individual lawyers have, nevertheless, often been on the side of rebels : Robespierre and Lenin were both lawyers.

Thus many lawyers took the British side in the American Revolution, and even among the lawyers who took the other side the predominant influence was against any attempt to turn the political revolution into a social revolution.

Most people would like law to be so certain that its application is of equal certainty in all cases and so simple that any person of sense can see how it applies.

The legal function likely to be most distrusted by the average person, though it also produces some of the law 's heroes, is litigious advocacy, particularly in the criminal law.

The feeling against advocacy in the criminal law was so strong that, at least in the case of the more serious kinds of crime, a right to representation by a trained advocate was nowhere generally recognized until the 18th century AD.

Governments and the members of organized legal professions have from the infancy of the craft endeavoured to meet the basic problem of representation by a basic rule of professional ethics–that the dominant duty of the advocate is not to his client but to truth and the law.

Since the later Roman Empire, advocates have been required to take oaths to this effect, and lawyers are often technically classed as " officers of court. "

The duty of the advocate is to fight for the rights of his client, but only up to the point where an honourable person could fairly put the case on his own behalf.

Appendix G

Human Extractions

The these tables, S No. means Sentence Number.

G.1 Extractions from Extractor 1

Table G.1: Direct Relation Extractions for the Education Document from Extractor 1

S No.	Keyword1	Keyword2	Link
3	Learning	College	Implemented in
3	Learning	Universities	Implemented in
4	Instructors	Instruction	Create
4	Instructors	Learning	Create
4	Instructors	Assessment	Create
5	Learning activities	Instructors	Provided [learning assistance]
5	Learning activities	Students	Provided [learning assistance]
6	Learner [progress]	Instructor	Analyzed by
6	Learner [progress]	Students	Analyzed by
6	Instructors	Courses [of action]	Plan
6	Students	Courses [of action]	Plan
7	Instructors	Lesson [navigation pattern]	Analyze
7	Lesson [navigation pattern]	Academic [achievement] Relationship	between
8	Students	Learning [proclivities]	Analyze own
9	Undergraduate course	Students	Taken by
10	[mediated] learning	[technology mediated] education	Capitalises on
10	[mediated] learning	classroom	Integrated into
10	[mediated] learning	Instructor + students	Permitted to spend more time together
11	[mediated] learning	Instructor	Become recognisable as

Table G.1: Direct Relation Extractions for the Education Document from Extractor 1 (Continued)

S No.	Keyword1	Keyword2	Link
11	[mediated] learning	Learning	Become recognisable as
11	[mediated] learning	Assessment	Become recognisable as
14	[computer-assisted] instruction	[Higher] education	In
15	Lecture	Instruction	Based
19	Assessment [metrics]	Lecture [centered] instruction	For comparing
19	Assessment [metrics]	[technology-mediated] instruction	For comparing
21	[mediated] learning	[technology-mediated] learning	Form of
21	[mediated] learning	Learning	Form of
21	[mediated] learning	Assessment	Form of
21	[mediated] learning	Campuses	Used on
21	[mediated] learning	Students	Prepare
21	Students	College	Enter
21	students	College-level [mathematics]	Underprepared for
22	Course	College student	For
23	[mediated] learning	Lecture[-centered] classroom	Enables faculty to environment by reconfiguring
24	[mediated] lesson	instructor	Enable to teach
25	Instructor		Teach more effectively
25	Students		Learn more effectively
25	[mediated] learning	Instruction and learning	New model at
26	[traditional] instruction	[mediated] learning	Alternative to
28	Students	[mediated] learning on courses	Earn grades of C and above
29	Student [completion rates]	Student enrollment [patterns]	Impact on
30	Technology-mediated] instruction	Students	Significant impact and number of
31	[mediated] learning	Instructor [workload]	Effect on
32	Instructions	Students	Take on more
32	Students	Class	Per
34	[mediated] learning	[higher] education	Help solve problems in
34	Learning [assistance]	[entry level] students	Needs of
34	Students	College campuses	Arrive on
34	Students	College-level students	Underprepared for

Table G.1: Direct Relation Extractions for the Education Document from Extractor 1 (Continued)

S No.	Keyword1	Keyword2	Link
35	enrollment	Students	Of
35	Student	courses	Placed in
35	Enrollment	[mathematics] courses	Trends in
36	[remedial] courses	[high] school [elementary algebra]	Include
	[remedial] courses	[high] school [intermediate algebra]	Include
37	[Precalculus] courses	College [algebra]	Include
	[Precalculus] courses	[combined] College [algebra]	Include
	[Precalculus] courses	College [algebra]	Include
	[Precalculus] courses	College [algebra]	Include
	[Precalculus] courses	College [algebra]	Include
38	[calculus] course	[Other] calculus	Include
39	Enrollment	[remedial maths] courses	Include
	Enrollment	[calaulus] courses	Include
40	Enrollment	[mathematics] courses	Include
	Enrollment	[remedial] courses	Increased demand for
41	Enrollment	[remedial mathematics] courses	Climbed to 10%
42	[two year] colleges	[underprepared] students	Address the instructional support needs of
43	Enrollment	[precalculus maths] courses	Increase in
	Enrollment	[all mathematics] courses	Growth in
	Courses	College	In
	Courses	Universities	In
44	Enrollment	[mathematics] courses	Total increase in
	Enrollment	[remedial] courses	Increased demand for
45	Enrollment	[remedial mathematics] courses	Climbed 58%
46	Student	Colleges	Attend
	Student	Universities	Attend
	University	Campuses	Are located
47	students	[remedial + precalculus] courses	Taking
	Learning [centers]	Learning [assistance needs]	Address
50	Enrollment [patterns + trends]	[number of] students	Underestimate
	Students	College campuses	Arrive on
	Students	College-level students	Underprepared for
51	[underprepared] students	[higher] education [community]	Policy implications for
52	[entry-level, core mathematics] course	[high-states] courses	Fall into category

Table G.2: Indirect Relation Extractions for the Education Document from Extractor 1

S No.	Keyword1	Keyword2	Link
5	Learning [activities]	Instructors	Provided information [learner performance]
5	Learning [activities]	Learner	Provided information [learner performance]

Table G.3: Direct Relation Extractions for the Law Document from Extractor 1

S No.	Keyword1	Keyword2	Link
1	Law	[specific] cases	Applied in
2	Advocate	Cases	Try and decided
	Judge	Cases	Try and decided
3	Lawyer	Evidence	Investigates
6	Lawyers	Judge	Suggest lines of factual inquiry to
	Lawyers	Law	Argue
7	Lawyer	Client's case	Lose
8	Court	Lawyers	Requires the aid of
9	Cases	Trial	Without
10	Lawyers	Counselor	As
11	Law	[legal] right	Gives extensive
12	Lawyer	[legal] right	Help to particularize
14	Lawyer	Client	Loyalty to
	Lawyer	Justice	Administration of
18	Magistrate	[family] law	Have special powers in
24	Jurisdiction	[family] law	Establish in in many matters of
25	Procurators	Litigation	Attend to formal and documentary steps in
26	Advocate	Clients	Gave direct advice
	Advocate	Procurator	Gave direct advice
	Advocate	Court	Presented oral argument
28	Advocates	[some] courts	Supplanted in
	Procurators	Advocates	Swallowed up
29	Courts	Procurators	Established as
30	[common] law	[inns of] court	Developed by
31	Advocates	[royal] judges	Appointed from
33	Attorneys	Legislation	Authored
	[apprentice] advocacy	Barrister	Acquired the title of
35	Litigation	Solicitor	Agent for
	Barrister	Right [of] advocacy	Secured
	Barrister	Court	In that
36	Advocate	Attorney	Division between
	Barrister	Attorney	Function only on the engagement of
	Barrister	Client	Not directly for the

Table G.3: Direct Relation Extractions for the Law Document from Extractor 1 (Continued)

S No.	Keyword1	Keyword2	Link
37	Barrister	[Queens] counsel	Made
38	Law	Precedent	Took place through
	Law	Judgements	Based on reported of the courts
39	Barristers	Solicitors	Division between
	Advocate	Procurator	Division between
	Barrister	Solicitor	Employed by
	Procurator	Client	Employed by
	Advocate	Client	Employed by
40	Solicitor	Bar	With legal advice from
43	Notaries	Civil-law [countries]	Existing
46	Judges	Common-law [countries]	Only legal class in
50	[socialist] legality	lawyers	Serve it
54	Law	[all] cases	Applied with equal certainty
55	Law	[litigious] advocacy	Produces some heroes
56	Advocacy	Criminal law	Felling against
	Right	Advocate	Representation by
57	Advocate	Law	Dominant duty to
58	Lawyers	[officer of] court	Classed as
59	Advocate	rights	Fight for the

Table G.4: Indirect Relation Extractions for the Law Document from Extractor 1

S No.	Keyword1	Keyword2	Link
3	Lawyer	Client	Confers with
3	Lawyer	Witnesses	Interviews
6	Lawyers	Clients	In accord with the interest of
7	Lawyers	[new] trial	Seek a
7	Lawyer	[appellate] court	Seek relief in
10	Counselor	Client	Achieve advantage for
57	Advocate	Client	Not dominant duty to
59	Advocate	Client	His

G.2 Extractions from Extractor 2

Table G.5: Direct Relation Extractions for the Education Document from Extractor 2

S No.	Keyword1	Keyword2	Link
8	Students	Their learning	Analyse
21	Mediated learning	Instruction, learning and assessment	Is a form of
47	students	courses	take

Table G.6: Indirect Relation Extractions for the Education Document from Extractor 2

S No.	Keyword1	Keyword2	Link
3	Mediated learning	college	Being implemented in
3	Mediated learning	University	Being implemented in
5	Learning activities	Instructor + students	Provide assistance
24	Instructor		Teach
24	Students		Learn
30	Students	course	Complete
32	Instructor	students	Take on
34	Students	college	Arrive on
35	Students	courses	Are placed in
46	Students	Colleges	attend
50	Students	college	Arrive on

Table G.7: Direct Relation Extractions for the Law Document from Extractor 2

S No.	Keyword1	Keyword2	Link
3	lawyer	The evidence	Investigates
6	lawyers	judge	Suggest lines of inquiry to
12	lawyer	rights	Helps to particularize
25	procurators	litigation	Attended to the formal steps in
26	advocate	Client	Gave advice to
28	procurator	advocate	Swallowed up
59	advocate	rights	Fights for

Table G.8: Indirect Relation Extractions for the Law Document from Extractor 2

S No.	Keyword1	Keyword2	Link
1	law	cases	Is applied to
2	Judge	cases	Works in the process of trying and deciding
2	Advocate	cases	Works in the process of trying and deciding
3	lawyer	client	Confers with
3	lawyer	witnesses	Interviews
6	lawyer	The law	Argue
7	A lawyer	A trial	May seek
14	lawyer	client	Has loyalty to
21	procurator	litigation	Attended to formal aspects of
26	advocate	procurator	Gave advice to
26	Advocate	court	Presented oral arguments in
31	judges	advocates	Appointed from
37	Most mentor	Queen counsel	Is
39	Barrister	A solicitor	Is employed by
40	bar	solicitors	Gives legal advice to
57	advocate	law	Has a duty to the

G.3 Extractions from Extractor 3

Table G.9: Direct Relation Extractions for the Law Document from Extractor 3

S No.	Keyword1	Keyword2	Link
6	Lawyer	Judge	Suggest lines of factual inquiry to
12	Lawyer	Right	Helping to particularize
21	Procurator	Litigation	Attend to
25	Procurator	Litigation	Attended to (steps in)
26	Advocators	Clients	Gave direct advice to
26	Advocators	Procurators	Gave direct advice to
26	Advocators	Court	Presented oral arguments in
37	All barristers	Counsel	Are
38	Precedent	Judgement	Based on
39	Procurators	Client	Employed by
39	Advocators	Client	Employed by

Table G.10: Indirect Relation Extractions for the Law Document from Extractor 3

S No.	Keyword1	Keyword2	Link
3	Lawyer	Client	Confer
3	Lawyer	Witness	Interview
6	Lawyer	Law	Argue
10	Lawyer	Client	Achieve advantages
14	Lawyer	Client	Has loyalty to his
14	Lawyer	Justice	Has loyalty to
31	Judge	Advocators	Appointed from
35	The Chancery	A court	Is

G.4 Extractions from Extractor 4

Table G.11: Direct Relation Extractions for the Education Document from Extractor 4

S No.	Keyword1	Keyword2	Link
12	Key to fiscal management in higher education	Learner-productivity and increased student academic achievement	is
31	instructors	students	Take on
32	instructor	more students per class	Can take
36	Remedial courses	High school elementary algebra	include
36	Remedial courses	High school intermediate algebra	include
37	Precalculus courses	College algebra	Include
37	Precalculus courses	Combined college algebra	Include
37	Precalculus courses	Mathematics for liberal art students	Include
37	Precalculus courses	Mathematics for elementary school teachers	Include
37	Precalculus courses	Other precalculus courses	Include
38	calculus courses	Other calculus courses	Include
52	Course in remedial and development mathematics	The category of high-stakes The category of high-stakes	Fall into

Table G.12: Indirect Relation Extractions for the Education Document from Extractor 4

S No.	Keyword1	Keyword2	Link
5	Learning activities	Instructor and student	Provide task-specific learning assistance
6	Learner progress	Instructor and student	Could be analysed by
6	Instructors and students	Possible future course	Plan
8	students	Their own learning proclivities	Analyse
9	Undergraduate courses	Entry-level students	Taken by
13	College and university	Learning environment	Enhancing
13	College and university	Student passing rate	increasing
21	Mediated learning	Technology-mediated instruction, learning and assessment	Is the form of
21	Mediated learning	Several hundred faculty on campuses	Is used by
21	Mediated learning	student	Prepare
21	students	college	Enter
21	students	College-level mathematics	Underprepared for
24	lessons Instructors and students		Enable (to teach and learn)
24	Instructor and students		Teach and learn
28	students	Mediated learning courses	Take
34	Entry-level students	College campuses	Arrive on

Table G.13: Direct Relation Extractions for the Law Document from Extractor 4

S No.	Keyword1	Keyword2	Link
1	law	cases	apply
3	Lawyer	facts, evidences	Investigates
6	Lawyers	judge	Suggest lines of factual inquiry to
6	Lawyers	law	Argue
8	Controversies not resolved in court	The aid of lawyers	Require
11	Law	right	Determine
12	Lawyer	right	Is helping to particularize
17	Citizen (should be a KW?)	law	Know
25	Procurators	Litigation	Attended to the formal steps in
26	Advocates	client	Give direct advice to
26	Advocates	court	Presented oral argument in
28	Procurator	advocates	Swallowed up
35	Barristers	right	Secured
35	Agent for litigation	solicitor	Was called
39	barrister	Solicitor	Be employed by
39	Procurator, advocate	client	Were employed by
43	notaries	Civil-law country	Exist in
57	The duty of the advocates	His client	Is to
58	lawyers	Officers of the court	Are classed as
59	Duty of the advocates	Rights to his client	Is to right for

Table G.14: Indirect Relation Extractions for the Law Document from Extractor 4

S No.	Keyword1	Keyword2	Link
1	The primary function of law	law	Individualize
2	Advocate, judge	cases	Trying, deciding
3	lawyer	client	Conferring with
3	lawyer	witnesses	Interviewing
10	lawyer	Advantages for his client	Achieve
12	lawyer	Legal right	Particularise
14	lawyer	client	Has loyalty to
14	lawyer	Administration of justice	Has loyalty to
27	University teacher of law	The mixture of Roman law	Explaining, adapting
32	litigation	procurator	Resembling
32	legislation	attorneys	Authorised
50	lawyers	Socialist legality	Serve

Appendix H

Sentences Used in Relation Extraction Algorithm Design

H.1 Compound Sentence

We fished all day; we did not catch a thing.

We fished all day; however, we did not catch a thing.

We fished all day, but did not catch a thing.

We fished all day, but we did not catch a thing.

He washed the car and polished it.

He not only washed the car, but polished it as well.

You can wait here and i will get the car.

Jim speaks Spanish, but his wife speaks French.

He washed the car, but did not polish it.

She sold her house, but she cannot help regretting it.

He either speaks french, or understands it.

He neither speaks french, nor understands it.

He could not find his pen, so he wrote in pencil.

We rarely stay in hotels, for we can not afford it.

I found a bucket, put it in the sink and turned the tap on.

I took of my coat, searched all my pockets, but could not find my key.

The hotel was cheap but clean.

Does the price include breakfast only, or dinner as well?

Does the price include breakfat, or not?

H.2 Complex Sentence

The alarm was raised as soon as the fire was discovered.

If you are not good at figures, it is pointless to apply for a job in a bank.

To get into university, you has to pass a number of examinations.

Seeing the door open, the stranger entered the house.

Free trade agreements are always threatened when individual countries protect their own markets.

Free trade agreements are always threatened by imposing duties on imported goods.

Free trade agreements are always threatened to encourage their own industries.

The racing car went out of control before it hit the barrier.

When she got on the train, Mrs Tomkind realized she had made a dreadful mistake.

The racing car went out of control and hit the barrier several times before it came to a stop on a grassy bank.

He told me that the match had been cancelled.

Holiday resorts which are very crowded are not very pleasant.

However hard I try, I can not remember people's name.

H.3 The Complex Sentence: Noun Clauses

He told me about the cancellation of the match.

He told me that the match had been cancelled.

I know that the match will be cancelled.

That the match will be cancelled is now certain.

That money does not grow on trees should be obvious.

It is obvious that money does not grow on trees.

Everybody knows that money does not grow on trees.

Everybody knows money does not grow on trees.

The dealer told me how much he was prepared to pay for my car and that I could have the money without delay.

He boasted that he was successful.

He boasted about how successful he was.

The fact that his proposal makes sense should be recognized.

The idea that everyone should be required to vote by law is something I do not agree with.

We must face the fact that we might lose our deposit.

His love of literature was due to the fact that his mother read poetry to him when he was a child.

In spite of the fact that hotel prices have risen, the number of tourists is as great as ever.

Despite the fact that hotel prices have risen, the number of tourists is as great as ever.

I am arfaid that we have sold out of ticket.

I am arfaid we have sold out of ticket.

I do not believe she will arrive before 7.

I believe she will not arrive before 7.

I do not suppose you can help us.

I suppose you can not help us.

Whether he has signed the contract does not matter.

Whether he has signed the contract or not does not matter.

The question is whether he has signed the contract.

I want to know whether he has signed the contract.

I want to know if he has signed the contract.

I am concerned about whether he has signed the contract.

When he did it is a mystery.

The question is when he did it.

I wonder when he did it.

It depends on when he did it.

I am interested in when he did it.

What matters most is good health.

What made him do it?

I wonder what made him do it.

H.4 Complex Sentence: Relative Pronouns and Relative Clauses

Crowded holiday resorts are not very pleasant.

Holiday resorts which are crowded are not very pleasant.

What kind of government would be popular?

The government which promises to cut taxes.

The government which promises to cut taxes would be popular.

The government, which promises to cut taxes, will be popular.

He asked a lot of questions, which were none of his business, and generally managed to annoy everybody.

He asked a lot of questions which were none of his business and generally managed to annoy everybody.

He is the man who lives next door.

He is the man that lives next door.

This is the photo which shows my house.

This is the photo that shows my house.

He is the man whose car was stolen.

He is the man who I met.

He is the man whom I met.

He is the man that I met.

He is the man who I gave the money to.

He is the man whom I gave the money to.

He is the man that I gave the money to.

This is the pohto which I took.

This is the pohto that I took.

This is the pan which I boiled the milk in.

This is the pan that I boiled the milk in.

It was an agreement the details of which could not be altered.

She is the woman who lives next door.

She is the woman that lives next door.

A doctor examined the astronauts who returned from space today.

The astronauts, who are reported to be very cheerful, are expected to land on the moon shortly.

This is tha cat which caught the mouse.

This is tha cat that caught the mouse.

The tiles which fell of the roof caused serious damage.

The Thames, which is now clean enough to swim in, was polluted for over a hundred years.

They are the men whose cars were stolen.

This is the house whose windows were broken.

This is the house where the windows were broken.

It was an agreement the details of which could not be altered.

It was an agreement of which the details could not be altered.

The millionair whose son ran away from home a week ago has made a public appeal.

Sally Smiles, whose cosmetics company has been in the news a great deal recently, has resigned as director.

He is the man who I met on holiday.

He is the man that I met on holiday.

That energetic man we met on holiday works for the factory.

That energetic man who we met on holiday works for the factory.

The author of "Rebels", whom I met at a party last week, proved to be a well-known journalist.

This is the photo I took.

This is the photo that I took.

This is the photo which I took.

The shed we built in the garden last year has begin to rot.

The shed that we built in the garden last year has begin to rot.

The shed which we built in the garden last year has begin to rot.

The shed in our garden, which my father built many years ago has lasted for a long time.

He is the man to whom I gave the money.

He is the man whom I gave the money to.

He is the man who I gave the money to.

He is the man that I gave the money to.

They are the people I gave the money to.

There is hardly anybody he is afraid of.

The person to whom I complained is the manager.

The person whom I complained to is the manager.

The person that I complained to is the manager.

The person I complained to is the manager.

The hotel manager, to whom I complained about the service, refunded part of our bill.

The hotel manager, who I complained to about the service, refunded part of our bill.

The hotel manager, whom I complained to about the service, refunded part of our bill.

This is the pan in which I boiled the milk.

This is the pan that I boiled the milk in.

This is the pan which I boiled the milk in.

This is the pan I boiled the milk in.

These are the cats I gave the milk to.

The agency from which we bought our tickets is bankrupt.

The agency which we bought our tickets from is bankrupt.

The agency that we bought our tickets from is bankrupt.

The agency we bought our tickets from is bankrupt.

The travel agency, with which our company has been dealing for several years, has opened four new branches.

The travel agency, which our company has been dealing with for several years, has opened four new branches.

He is the man from whose house the pictures were stolen.

He is the man whose house the pictures were stolen from.

In 1980 he caught a serious illness from whose effects he still suffers.

In 1980 he caught a serious illness the effects of which he still suffers from.

Mr Jason Matthews, from whose collection of pictures a valuable Rembrandt was given to the nation, died last night.

1979 was the year in which my son was born.

1979 was the year my son was born.

1979 was the year when my son was born.

The summer of 1969, the year in which men first set foot on the moon, will never be forgotten.

The summer of 1969, the year men first set foot on the moon, will never be forgotten.

The summer of 1969, the year when men first set foot on the moon, will never be forgotten.

The summer of 1969, when men first set foot on the moon, will never be forgotten.

This is the place in which I grew up.

This is the place which I grew up in.

This is the place I grew up in.

This is the place where I grew up.

The Tower of London, in which so many people lost their lives, is now a tourist attraction.

The Tower of London, the place where so many people lost their lives, is now a tourist attraction.

The Tower of London, where so many people lost their lives, is now a tourist attraction.

That is the reason for which he dislikes me.

That is the reason he dislikes me.

That is why he dislikes me.

My success in business, the reason for which he dislikes me, has been due to hard work.

My success in business, the reason he dislikes me, has been due to hard work.

My success in business, the reason why he dislikes me, has been due to hard work.

I still remember the summer that we had the big drought.

I still remember the summer we had the big drought.

I still remember the summer when we had the big drought.

I still remember the summer during which we had the big drought.

I don't know any place that you can get a better exchange rate.

I don't know any place you can get a better exchange rate.

I don't know any place where you can get a better exchange rate.

I don't know any place at which you can get a better exchange rate.

That was not the reason that he lied to you.

That was not the reason he lied to you.

That was not the reason why he lied to you.

That was not the reason for which he lied to you.

My neighbour Mr Watkins never misses the opportunityto tell me the latest news.

Mr Watkins, a neighbour of mine, never misses the opportunityto tell me the latest news.

All that remains for me to do is to say goodbye.

Everything that can be done has been done.

I will do anything that I can.

I will do anything I can.

God bless this ship and all who sail in her.

It is the silliest argument that I ahve ever heard.

It is the silliest argument I ahve ever heard.

Bach was the greatest composer who has ever lived.

Both players, neither of whom reached the final, played well.

The treasure, some of which has been recovered, has been sent to the British Museum.

she married Joe, which surprised everyone.

she married Joe, this surprised everyone.

she married Joe, that surprised everyone.

I may have to work late, in which case I will telephone.

The speaker paused to examine his notes, at which point a loud crash was heard.

It is the only building which I have ever seen which is made entirely of glass.

It is the only building I have ever seen which is made entirely of glass.

H.5 Complex Sentence: Adverbial Clauses

I try hard, but I can never remember people's names.

However hard I try, I can never remember people's names.

Tell me as soon as he arrives.

You can sit where you like.

He spoke as if he meant business.

He went to bed because he felt ill.

You did not look very well when you got up this morning.

After she got married, Madeleine changed completely.

I pulled a muscle as I was lifting a heavy suitcase.

You can keep these records as long as you like.

Once you have seen one penguin, you have seen them all.

He has not stopped complaining since he got back from his holidays.

We always have to wait till the last customer has left.

We always have to wait until the last customer has left.

The Owens will move to a new flat when their baby is born.

Once we have decorated the house, we can move in.

When we have decorated the house, we can move in.

Now that we have decorated the house, we can move in.

The hotel receptionist wants to know when we will be checking out tomorrow morning.

I shall be on holiday till the end of September when I return to London.

You can not camp where you like these days.

You can not camp wherever you like these days.

You can not camp anywhere you like these days.

Everywhere Jenny goes she is mistaken for Princess Diana.

The church was built where there had once been a Roman temple.

With a special train ticket you can travel wherever in Europe for just over £100.

Type this again as I showed you a moment ago.

Type this again in the way I showed you.

This fish is not cooked as I like it.

This fish is not cooked in the way I like it.

This steak is cooked how I like it.

She is behaving the same way her elder sister used to.

I feel as if I am floating on the air.

I feel as though I am floating on the air.

It sounds as if ths situation will get worse.

It feels as though it is going to rain.

Lillian was trembling as if she had seen a ghost.

She acted as if she were mad.

As there was very little support, the dtrike was not successful.

Because there was very little support, the dtrike was not successful.

Since there was very little support, the dtrike was not successful.

I am afraid we do not stock refills for pens like yours because there is little demand for them.

As you can not type the letter yourself, you will have to ask Susan to do it for you.

Since you can not type the letter yourself, you will have to ask Susan to do it for you.

Jim is trying to find a place of his own because he wants to feel independent.

Although I felt sorry for him, I was secretly pleased that he was having difficulties.

Though I felt sorry for him, I was secretly pleased that he was having difficulties.

Even though I felt sorry for him, I was secretly pleased that he was having difficulties.

We intend to go to India, ecen if air fares go up again between now and the summer.

Much as I would like to help, there is not a lot I can do.

While I disapprove of what you say, I would defend to the death your right to say it.

However far it is, I intend to drive there tonight.

No matter where you go, you can not escape from youself.

Whatever I say, I seem to say the wrong thing.

No matter what I say, I seem to say the wrong thing.

However brilliant you are, you can not know everything.

However brilliant you may be, you can not know everything.

Whatever you may think, I am going ahead with my plans.

Unlikely as it sounds, what I am telling you is true.

Unlikely as it may sound, what I am telling you is true.

Beatuiful though the necklace was, we thought it was over-priced so we did not buy it.

Try as he might, he could not solve the problem.

I have arrived early so that I may get a good view of the procession.

I have arrived early in order that I may get a good view of the procession.

Let us spend a few moments in silence so that we remember those who died to preserve our freedom.

H.6 Infinitives

To study hard is your duty.

To help the poor is a noble deed.

To solve this problem will enable use to win the victory.

To correct the papers makes me busy all day.

I want to read some stories about great scientists.

He has decided to leave his job.

Do not try to swim across this river.

We are about to start for the west.

He has no desire but to rise in the world.

She desires nothing but to study abroad.

To do two things is to do nothing.

All you have to do is to study hard.

She seems to be out of sorts.

He told me to get up early.

We expect him to do his best.

I want you to come again tomorrow.

He taught me how to drive a car.

How to solve the problem is still unknown.

Our trouble is how to solve the problem.

I have no friend to help me.

A house to let is not easy to be found.

This is the best way to learn English.

He has no house to live in.

I have no pen to write with.

She wants some books to amuse herself with.

We eat to live, not live to eat.

He comes to see you.

You go to school in order to get knowledge.

He worked hard in order to succeed in the entrance examination.

I got up early so as to be in time for the first train.

He opened his lips as if to make some reply.

Teddy is making as though to leave the room.

I am very glad to see you.

She will be sorry to hear that.

I was surprised to see her there.

She wept to hear the news.

He works hard only to find fail again.

The went out, only to get wet.

He went there only to find her crying.

He must be a great fool to believe such a thing.

How rich she must be to wear such a splendid necklace.

We are ready to start.

Are you able to finish the work today?

She is likely to pass the examination.

English is very easy to learn.

This question is quite difficult to solve.

He is old enough to go to school.

He is strong enough to defeat his enemy.

She was kind enough to help us.

He run fast enough to catch you.

This room is large enough for us five to sleep in.

I am too tired to work any more.

This question is too difficult for me to answer.

Germs are too small to be seen with naked eye.

We were not too late to attend the meeting.

To be frank with you, I am against that plan.

To begin with, she was not very happy and was in poor health.

To make matters worse, her husband died soon after they had got married.

Strange to say, nobody wants to have the chance.

He seems to have been idle.

She appeared to have forgotten me.

Ted is said to have been elected monitor.

They wished to have come, but a heavy rain prevented them from coming.

I intended to have discuss the matter with you, but I had some guests then.

We expected to have succeeded, but we failed.

She desired to have married Frank,but her parents were against the marriage.

We saw him enter her room.

They are watching the children play ball in the yard.

I stood there listening to the rain patter on the tree.

We beheld the monster rise to the surface.

Who made her do that?

Let me know all the details.

I helped him finish the work.

He offered to help carry her basket.

Let us go talk to the other fellows.

Will not you come take a look at him?

Now we may be too late. the only thing to do is go right ahead.

She does nothing but cry all day.

I cannot but feel sorry for him.

H.7 Participles

The sleeping baby is my youngest brother.

The Chinese are peace-loving people.

The setting sun looks very beautiful.

Amy is the best_looking girl in our village.

The man standing over there is my uncle.

I threw a stone at the dog barking furiously.

The children playing baseballs down there are my classmates.

Members wishing to resign are requested to notify the secretary.

The policeman finding nothing wrong around here left in a hurry.

The story was really interesting.

Then, another dog came running after the big one.

Many of us stood listening to his fine playing.

Mother goes shopping every other day.

I heard someone knocking at the door.

Soon I felt something creeping up my side.

Don't keep your friends waiting.

We caught him stealing.

It is burning hot.

She is an amazing fine girl.

The lost child was found dead in the trees this morning.

Did you pass the written examination?

The thief entered the room through the broken window.

The fallen leaves have covered the paths.

The book written by Mr. Wall sells very well.

The language spoken in Canada in English.

One of the slaves was a man descended from a noble family.

The beach covered with sand is suitable for swimming.

He is gone.

The old man sat surrounded by his grandchildren.

The returned utterly exhausted.

She looked much surprised when she heard the news.

You can hear English spoken in almost every large city in the world.

He could not make himself understood in English.

Did you see him punished by the teacher?

We found the old man killed in a traffic accident.

The wound have been carried to the hospitals.

The learned are to be given the best chance.

Walking along the street yesterday, I met a friend of mine.

Having been written in haste, the book has too many faults.

Born in better times, he would become famous.

Admitting what you say is true, I still think that you made a mistake.

Living as I do in a remote village, I seldom have visitors.

Standing as it does on a little hill, our home commands a fine view.

Being written in an easy style as it is, this book is suitable for beginners.

The weather being fine, we went on a picnic.

Spring having come, the trees begin to bud.

It being Sunday, there was no school.

The lesson ended, the boys rushed out of classroom to play ball.

H.8 Gerunds

Living in Paris is quite different from visiting it once in a while.

Wanting things makes us do useful work.

Talking a walk in the morning every day makes one healthier and happier.

Learning foreign language takes us a lot of time.

A loud knocking at the door was heard.

Her comings and goings are frequent.

My favorite sport is swimming.

One of the bad habits is eating too much.

My hobby is collecting stamps.

We enjoyed dancing and singing around the fire.

I remember seeing her somewhere before.

Would you mind opening the door?

I hate listening to his tedious lectures.

My father is fond of hiking on Sundays.

I am proud of being a Chinese.

He assured me of their being alive.

We talked about having a party next Sunday.

They praised him for winning the prize.

No one would have dreamed of there being such an old man.

There is no question of John's being able to do it.

We were surprised at her beauty being made so much of.

You will oblige me by all leaving the room.

I cannot think of anything else being wanted.

I remember my mother reading the story in my childhood.

She thanked him for having saved her child.

I am surprised to hear of her having commited suicide.

He scolds me for having neglected my duties.

The swimming pool is not large enough.

There is no sleeping car attached to this train.

Is there a smoking room in this theater?

We are going to buy a washing machine.

It is no use crying over spilt milk.

It is no good your trying to deceive me.

It is hard work keeping the grass green at this time of year.

It is worth while asking how far their education contributed to their success.

There is no use your telling me that you are going to be good.

There is no saying what will happen tomorrow.

There is no knowing when another war will break out.

There is no accounting for tastes.

It goes without saying that the proposal will not be accepted.

It goes without saying that hunger is the best sauce.

It goes without saying that such a diligent man as he will succeed.

I feel like eating a bowl of ice-cream.

He makes a point of calling on me on New Year's Day.

We were on the point of leaving when you called.

I am looking forward to seeing you.

These are flowers of my own growing.

Are these pictures of your on painting?

On arriving in Washington, we went to hotel.

On my getting into the train, it begin to move.

Upon reaching San Francisco, he started for New York.

Appendix I

UCREL CLAWS7 Tagset

Table I.1: CLAWS7 Tagset

Patterns	Meaning
APPGE	possessive pronoun, pre-nominal (e.g. my, your, our)
AT	article (e.g. the, no)
AT1	singular article (e.g. a, an, every)
BCL	before-clause marker (e.g. in order (that),in order (to))
CC	coordinating conjunction (e.g. and, or)
CCB	adversative coordinating conjunction (but)
CS	subordinating conjunction (e.g. if, because, unless, so, for)
CSA	as (as conjunction)
CSN	than (as conjunction)
CST	that (as conjunction)
CSW	whether (as conjunction)
DA	after-determiner or post-determiner capable of pronominal function (e.g. such, former, same)
DA1	singular after-determiner (e.g. little, much)
DA2	plural after-determiner (e.g. few, several, many)
DAR	comparative after-determiner (e.g. more, less, fewer)
DAT	superlative after-determiner (e.g. most, least, fewest)
DB	before determiner or pre-determiner capable of pronominal function (all, half)
DB2	plural before-determiner (both)
DD	determiner (capable of pronominal function) (e.g any, some)
DD1	singular determiner (e.g. this, that, another)
DD2	plural determiner (these,those)
DDQ	wh-determiner (which, what)
DDQGE	wh-determiner, genitive (whose)
DDQV	wh-ever determiner, (whichever, whatever)

Table I.1: CLAWS7 Tagset (continued)

Patterns	Meaning
EX	existential there
FO	formula
FU	unclassified word
FW	foreign word
GE	germanic genitive marker - (' or's)
IF	for (as preposition)
II	general preposition
IO	of (as preposition)
IW	with, without (as prepositions)
JJ	general adjective
JJR	general comparative adjective (e.g. older, better, stronger)
JJT	general superlative adjective (e.g. oldest, best, strongest)
JK	catenative adjective (able in be able to, willing in be willing to)
MC	cardinal number,neutral for number (two, three..)
MC1	singular cardinal number (one)
MC2	plural cardinal number (e.g. sixes, sevens)
MCGE	genitive cardinal number, neutral for number (two's, 100's)
MCMC	hyphenated number (40-50, 1770-1827)
MD	ordinal number (e.g. first, second, next, last)
MF	fraction,neutral for number (e.g. quarters, two-thirds)
ND1	singular noun of direction (e.g. north, southeast)
NN	common noun, neutral for number (e.g. sheep, cod, headquarters)
NN1	singular common noun (e.g. book, girl)
NN2	plural common noun (e.g. books, girls)
NNA	following noun of title (e.g. M.A.)
NNB	preceding noun of title (e.g. Mr., Prof.)
NNL1	singular locative noun (e.g. Island, Street)
NNL2	plural locative noun (e.g. Islands, Streets)
NNO	numeral noun, neutral for number (e.g. dozen, hundred)
NNO2	numeral noun, plural (e.g. hundreds, thousands)
NNT1	temporal noun, singular (e.g. day, week, year)
NNT2	temporal noun, plural (e.g. days, weeks, years)
NNU	unit of measurement, neutral for number (e.g. in, cc)
NNU1	singular unit of measurement (e.g. inch, centimetre)
NNU2	plural unit of measurement (e.g. ins., feet)
NP	proper noun, neutral for number (e.g. IBM, Andes)
NP1	singular proper noun (e.g. London, Jane, Frederick)
NP2	plural proper noun (e.g. Browns, Reagans, Koreas)
NPD1	singular weekday noun (e.g. Sunday)
NPD2	plural weekday noun (e.g. Sundays)

Table I.1: CLAWS7 Tagset (continued)

Patterns	Meaning
NPM1	singular month noun (e.g. October)
NPM2	plural month noun (e.g. Octobers)
PN	indefinite pronoun, neutral for number (none)
PN1	indefinite pronoun, singular (e.g. anyone, everything, nobody, one)
PNQO	objective wh-pronoun (whom)
PNQS	subjective wh-pronoun (who)
PNQV	wh-ever pronoun (whoever)
PNX1	reflexive indefinite pronoun (oneself)
PPGE	nominal possessive personal pronoun (e.g. mine, yours)
PPH1	3rd person sing. neuter personal pronoun (it)
PPHO1	3rd person sing. objective personal pronoun (him, her)
PPHO2	3rd person plural objective personal pronoun (them)
PPHS1	3rd person sing. subjective personal pronoun (he, she)
PPHS2	3rd person plural subjective personal pronoun (they)
PPIO1	1st person sing. objective personal pronoun (me)
PPIO2	1st person plural objective personal pronoun (us)
PPIS1	1st person sing. subjective personal pronoun (I)
PPIS2	1st person plural subjective personal pronoun (we)
PPX1	singular reflexive personal pronoun (e.g. yourself, itself)
PPX2	plural reflexive personal pronoun (e.g. yourselves, themselves)
PPY	2nd person personal pronoun (you)
RA	adverb, after nominal head (e.g. else, galore)
REX	adverb introducing appositional constructions (namely, e.g.)
RG	degree adverb (very, so, too)
RGQ	wh- degree adverb (how)
RGQV	wh-ever degree adverb (however)
RGR	comparative degree adverb (more, less)
RGT	superlative degree adverb (most, least)
RL	locative adverb (e.g. alongside, forward)
RP	prep. adverb, particle (e.g about, in)
RPK	prep. adv., catenative (about in be about to)
RR	general adverb
RRQ	wh- general adverb (where, when, why, how)
RRQV	wh-ever general adverb (wherever, whenever)
RRR	comparative general adverb (e.g. better, longer)
RRT	superlative general adverb (e.g. best, longest)
RT	quasi-nominal adverb of time (e.g. now, tomorrow)
TO	infinitive marker (to)
UH	interjection (e.g. oh, yes, um)

Table I.1: CLAWS7 Tagset (continued)

Patterns	Meaning
VB0	be, base form (finite i.e. imperative, subjunctive)
VBDR	were
VBDZ	was
VBG	being
VBI	be, infinitive (To be or not... It will be ..)
VBM	am
VBN	been
VBR	are
VBZ	is
VD0	do, base form (finite)
VDD	did
VDG	doing
VDI	do, infinitive (I may do... To do...)
VDN	done
VDZ	does
VH0	have, base form (finite)
VHD	had (past tense)
VHG	having
VHI	have, infinitive
VHN	had (past participle)
VHZ	has
VM	modal auxiliary (can, will, would, etc.)
VMK	modal catenative (ought, used)
VV0	base form of lexical verb (e.g. give, work)
VVD	past tense of lexical verb (e.g. gave, worked)
VVG	-ing participle of lexical verb (e.g. giving, working)
VVGK	-ing participle catenative (going in be going to)
VVI	infinitive (e.g. to give... It will work...)
VVN	past participle of lexical verb (e.g. given, worked)
VVNK	past participle catenative (e.g. bound in be bound to)
VVZ	-s form of lexical verb (e.g. gives, works)
XX	not, n't
ZZ1	singular letter of the alphabet (e.g. A,b)
ZZ2	plural letter of the alphabet (e.g. A's, b's)

Appendix J

Extractions of Tagger-based Method

List 1: Extractions from the Education Domain

No Extraction

1 is currently being implemented (Mediated Learning, colleges and universities around the 22 country)

2 is (Mediated Learning, a form of technology-mediated instruction)

3 complete (a significant impact on the number of students, a course or sequence of courses)

4 can take on (instructors, students per class)

5 enter (students, college)

6 could be analyzed (information on learner progress and achievement, instructors and their students)

7 becomes (Mediated Learning, a unique model of technology-mediated instruction)

8 enhancing (colleges and universities, the learning environment)

9 is being used (Mediated Learning, hundred faculty on campuses around the country)

10 to prepare (Mediated Learning, students)

11 to improve (The Mediated Learning approach, the instructional environment)

12 reconfiguring (The Mediated Learning approach, lecture-centered classroom)

13 being (Mediated Learning, sound alternative to traditional instruction)

14 include (Precalculus courses, College Algebra)

15 called (Mediated Learning, Mediated Learning)

16 to plan (information on learner progress and achievement, possible future courses of action)

17 would be (crucial support for instruction, most traditional classrooms)

18 is (crucial support for instruction, most traditional classrooms)

19 predominates (lock-step teaching method, the case of lower-division undergraduate courses)

20 has permitted (classroom environments, instructors and their students)

21 learning (Mediated Learning, assessment)

22 accompanying (the employment of computer-assisted instruction in higher education, the employment of computer-assisted instruction in higher education)

23 noted (the employment of computer-assisted instruction in higher education, term)

24 is (the employment of computer-assisted instruction in higher education, term)

25 has frequently been misapplied (the employment of computer-assisted instruction in higher education, the examination of instructional computer use)

26 was published (the Levien study, the auspices of the Carnegie Commission on Higher Education)

27 comparing (a pressing need for evaluation and assessment metrics, traditional lecture-centered 28 instruction and technology-mediated instruction models)

28 to justify (the fierce pressure on higher education institutions, new instructional initiatives)

29 learning (Mediated Learning, assessment)

30 facilitated (Mediated Learning and its components, Mediated Learning and its components)

31 implementing (Mediated Learning and its components, Mediated Learning)

32 being reported (Findings on the Effectiveness of Mediated Learning, campuses)

33 are shown (incremental changes in student completion rates, impact student enrollment patterns)

34 can have (the implementation of technology-mediated instruction, a significant impact on the number of students)

35 Effect (Mediated Learnings, Instructor Workloads)

36 to indicate (The Economics of Mediated Learning, Mediated Learning)

37 need not increase (Mediated Learning, instructional costs)

38 arrive (Mediated Learning lessons, college campuses)

39 manifested (developmental mathematics courses, increasing enrollment of students)

40 placed (developmental mathematics courses, developmental mathematics courses)

41 is documented (developmental mathematics courses, the latest survey of enrollment trends in mathematics courses)

42 was (the total increase in enrollment in mathematics courses, increased demand for remedial courses)

43 climbed (enrollment in remedial mathematics courses, percent of the total mathematics enrollment in the fall)

44 addressing (the instructional support needs of underprepared students, the instructional support 46needs of underprepared students)

45 was (precalculus mathematics courses, the total enrollment growth)

46 was (percent of the total increase in enrollment in mathematics courses, increased demand for remedial courses)

47 climbed (enrollment in remedial mathematics courses, percent of the total mathematics enrollment in the fall)

48 underscore (college mathematics departments, the policy implications of a larger cohort of 50 underprepared students for the larger higher education community)

49 escalates (Their cost to the university, students)

List 2: Extractions from the Law Domain

No Extraction

01 is to apply (The primary function of the profession and practice of law, the law in specific cases)

02 investigates (a lawyer, the facts and the evidence)

03 conferring (a lawyer, his client)

04 interviewing (a lawyer, witnesses)

05 suggest (Continental lawyers, lines of factual inquiry to the judge)

06 argue (Continental lawyers, the law)

07 is helping to particularize (the lawyer, the legal rights of the parties)

08 attended (the procurator, the formal aspects of litigation)

09 gave (Advocates, direct advice to clients)

10 presented (Advocates, oral arguments in court)

11 swallowed up (the procurators, the advocates)

12 is to fight (The duty of the advocate, the rights of his client)

13 was called (the appropriate agent for litigation, a solicitor)

14 has tended to be replaced (the specialization of procurator-advocate and solicitor-barrister, 1 profession of legal practitioners)

15 remain (the judges of high-level courts, the only legal class in the liberal capitalist common-law countries)

16 loses (a lawyer, his client 's case)

17 require (court, the aid of lawyers)

18 have (lawyers, the settlement of most cases without trial)

19 part–bring (lawyers, the settlement of most cases without trial)

20 may aid (counselor and negotiator, a transaction)

21 shaping (counselor and negotiator, a transaction)

22 to achieve (counselor and negotiator, advantages for his client)

23 to arrange (The law, their legal rights)

24 determine (The law, their legal rights)

25 drawn (the procurator, procurators)

26 attended (Procurators, the documentary steps in litigation)

27 explaining (University teachers of law, the mixture of Roman law and Germanic custom)

28 adapting (University teachers of law, the mixture of Roman law and Germanic custom)

29 produced (University teachers of law, the modern laws of the major European countries)

30 continued to dominate (the modern laws of the major European countries, the scholarly interpretation of the law)

31 did (admiralty courts, procurators)

32 become established (canon laws, practitioners)

33 resembling (procurators, procurators)

34 authorized (legislation, legislation)

35 working (cases of men, barristers and attorneys)

36 secured (the appropriate agent for litigation, the right of advocacy)

37 based (Development of the law, the reported judgments of the courts)

38 adopted (the division between barristers and solicitors, an equivalent of the English practice)

39 requiring (the division between barristers and solicitors, a barrister)

40 to be employed (the division between barristers and solicitors, a solicitor)

41 employed (the division between barristers and solicitors, the client)

42 developed (the whole burden of transactional work, the profession of notary)

43 are (the whole burden of transactional work, the solicitors)

44 favoured varying (Leading lawyers, the general structure of the law in the particular community)

45 has been (cases, the consequence of a general hostility to the whole idea of law)

46 are (Most lawyers, the law)

47 were (Individual lawyers, lawyers)

48 is (the law 's heroes, litigious advocacy)

49 was (The feeling against advocacy in the criminal law, serious kinds of crime)

50 is (the dominant duty of the advocate, his client)

51 are (lawyers, officers of court)

52 classed (advocates, officers of court)

List 3: Baseline Extractions from the Education Domain

1 + implemented(Learning, colleges)

2 + analyze(students, learning)

3 + complete(students, course)

4 + take(instructors, students)

5 + arrive(students, college)

6 is(Learning, colleges)

7 being(Learning, colleges)

8 learning(instruction, assessment)

9 plan(students, courses)

10 would(instruction, classrooms)

11 be(instruction, classrooms)

12 is(instruction, classrooms)

13 taken(courses, students)

14 integrating(education, classroom)

15 learning(instruction, assessment)

16 is(education, learner-productivity)

17 is(education, learner-productivity)

18 noted(education, term)

19 is(education, term)

20 has(term, examination)

21 been(term, examination)

22 misapplied(term, examination)

23 assessing(instruction, instruction)

24 learning(instruction, assessment)

25 enter(students, college)

26 take(students, Learning)

27 enrolling(courses, grades)

28 earn(courses, grades)

29 Effect(Learnings, Instructor)

30 can(instructors, students)

31 include(courses, College)

32 was(courses, enrollment)

33 escalates(university, students)

List 4: Baseline Extractions from the Law Domain

1 + conferring(evidence, client)

2 + interviewing(client, witnesses)

3 + particularize(lawyer, rights)

4 + swallowed(procurators, advocates)

5 + employed(barrister, solicitor)

6 + fight(advocate, rights)

7 is(law, law)

8 apply(law, law)

9 introduces(trial, evidence)

10 questioning(evidence, witnesses)

11 loses(lawyer, client)

12 may(case, trial)

13 seek(case, trial)

14 is(lawyer, rights)

15 helping(lawyer, rights)

16 would(law, transactions)

17 apply(law, transactions)

18 did(fact, advocate)

19 attempt(fact, advocate)

20 act(fact, advocate)

21 did(courts, procurators)

22 become(laws, practitioners)

23 established(laws, practitioners)

24 resembling(litigation, procurators)

25 authorized(attorneys, legislation)

26 was(litigation, solicitor)

27 called(litigation, solicitor)

28 secured(barristers, right)

29 based(precedent, judgments)

30 requiring(practice, barrister)

31 be(barrister, solicitor)

32 were(advocate, client)

33 employed(advocate, client)

34 have(advocates, judiciary)

35 come(advocates, judiciary)

36 be(advocates, judiciary)

37 admired(advocates, judiciary)

38 are(lawyers, law)

39 was(law, case)

40 is(advocate, client)

41 is(advocate, rights)

Appendix K

Glossary

Acronym	**Meaing**
AE	Answer Extraction
AIG	Automatic Index Generation
AKE	Automatic Knowledge Extraction
ANN	Artificial Neural Network
ARPA	The American Advanced Research Projects Agency
ATA	Automatic Terminology Acquisition
ART	Adaptive Resonance Theory
BIG	Geographical Information System
BNC	British National Corpus
CLAWS	the Constituent Likelihood Automatic Word-tagging System
CNLP	Connectionist Natural Language Porcessing
CWC	Comparison with Chance
ENLP	Empirical Natural Language Porcessing
FF-MLP	Feed-Forward Multi-Layer Percepton
IE	Information Extraction
IP	Important Path
IPP	Integrated Partial Parser
IR	Information Retrieval
KA	Knowledge Acquisition
KB	Knowledge Base
KEP	Knowledge Extraction Program
KP	Key Path
KW	Keyword
LBI	Learn by Intuition
The	Longman Dictionary of Contemporary English

Acronym	**Meaing**
LLA	Longman Language Activator
MLP	Multi-layer Perceptron
ML	Machine Learning
MRD	Machine Readable Dictionary
MT	Machine Translation
MUC	Message Understanding Conference
NG	Natural Generalisation
NIP	Negative Important Path
NP	Noun Phrase
NKW	Non-keywords
NKP	Non-key Path
NL	Natural Language
NLP	Natural Language Porcessing
NLU	Natural Language Understanding
NS	Noun Set
PG	Pure Generalisation
PIP	Positive Important Path
PTB	Penn Tree Bank
RAAM	Recursive Auto-Associative Memory
RE	Relation Extraction
RWS	Relation Word Set
SA	Stemming Analysis
SGML	Standard Generalized Markup Language
SRN	Simple Recurrent Network
TC	Thematic Concepts
TCE	Thematic Concept Extraction
TT	Technical Term
UCREL	University Center for Computer Corpus Research on Language
WSD	Word Sense Disambiguation
WWW	World Wide Web

Appendix L

Published Works

1. Zhang, S., Powell, H., Plamer-Brown, D. and Eveet, L. Methods for Concept Extraction Using ANNs and Stemming Analysis and Their Portability across Domains, the Proceedings of The Second Workshop on Natural Language Processing and Neural Networks (NLPNN2001). November, 2001, Tokyo, Japan.

2. Zhang, S., Powell, H. and Plamer-Brown, D. Keyword Extraction from Stemming and Sense Information by Neural Networks, the proceedings of the year 2000 International Conference on Artificial Intelligence (IC-AI'2000). 2000.

3. Zhang, S., Powell, H. and Plamer-Brown, D. Keyword Extraction Using Neural Networks. Proceedings of the Tenth Meeting Computational Linguistics in the Netherlands.

Methods for Concept Extraction Using ANNs and Stemming Analysis and Their Portability across Domains

Shaomin Zhang, Heather Powell and **Lindsay Evett**
Newton Building, Computing Department, Nottingham Trent University
Burton Street, Nottingham NG1 4BU, UK

Dominic Palmer-Brown
Trends in Cognitive Sciences, Elsevier Science London, 84 Theodbald's Rd, London,
WC1X 8RR

Abstract

A major concern for knowledge acquisition research is domain portability of the techniques developed. The research presented in this paper investigates domain independent techniques for automatic knowledge extraction from text, based on Artificial Neural Networks and Stemming Analysis. The knowledge is to be organised into a knowledge base. The techniques presented are aimed at the automatic identification of concepts (keywords).

Artificial Neural Networks (ANNs) are trained to recognise keywords on the basis of their sense information, stemming analysis and relationships to one or more seed words which are manually selected as indicative of the areas of knowledge required. The relationships are obtained from an electronic dictionary. Training data is generated using example keywords that humans have identified as being keywords associated with particular seed words. After training, the ANN can be used to extract keywords automatically from other documents. Stemming analysis on the definitions of senses of nouns has also been performed to enhance the ANN approach by trying to re-establish the relationships between the nouns in the definitions and the seed words. The definitions are also from the electronic dictionary. Results from ANN and stemming analysis are combined to take advantage of the two different methods.

Experiments on documents concerning education and law show that first, the results from the combination of ANNs and Stemming Analysis are better results than the results from either of them separately and second, these approaches are domain-portable.

Keywords: Knowledge Acquisition, Natural Language Processing, Neural Networks, Knowledge Base, Stemming Analysis

1 Introduction

This paper presents research into approaches for constructing knowledge bases by automatic knowledge extraction from text. It investigates the nature of semantic relations between concepts in text, specifically to determine the level and types of lexical information that are required to define the scope of the set of relations within a given topic. The existing techniques are successful in generating an index for information retrieval but they are difficult to use in thematic concept extraction. These techniques tend to produce a long list of concepts that covers nearly all the details in the text and this is not usually suitable for incorporation into a knowledge base. Using these methods to perform thematic concept extraction would result in high recall and very low precision. Because the extracted knowledge is to be put into knowledge bases, techniques that produce high precision are more suitable to avoid introducing false knowledge. The techniques presented are aimed at the automatic identification of concepts (keywords) within any given theme. The main objective of the research is to develop techniques for automatic knowledge extraction directly from plain text in electronic form, so that the extracted knowledge can be organised into a knowledge base.

The target knowledge base is used in a hyper-knowledge interaction environment called HyperTutor[7]. This uses a novel and generic formalism for structuring and interrogating hypermedia-based knowledge via a natural language interface. The system engages users in a dialogue with knowledge as well as allowing them to browse. It also has pedagogic features for tutoring. It employs semantic hyperlinks to represent knowledge.

HyperTutor is a generic environment, therefore generic Knowledge Acquisition (KA) techniques are required. It will be a great benefit to automate the knowledge acquisition process so that knowledge can be automatically extracted from text with minimum human involvement. This paper presents research into enabling the important concepts (keywords) in a domain to be automatically identified. The identification is based on one or more seed words which are provided by a human author to define the domain.

1.1 Related Research

One approach to solving the task of automatic knowledge acquisition is to fully understand the natural language text. This method, however, is beyond the capabilities of current natural language understanding (NLU) systems. The main reason for this is the complexity of natural language and the lack of appropriate linguistic theory to manage this complexity. It is difficult to build a grammar for a realistic subset of natural language[10]. In particular it is difficult to process exceptions.

Another approach to knowledge acquisition is Information Extraction (IE)[1]. IE aims to identify instances of a particular class of event or relationship in natural language text. Relevant arguments concerning events and relationships are extracted and encoded in a format suitable for incorporation into a database[5]. The construction of extraction patterns which are used by most IE systems to extract information is a time-consuming, knowledge-intensive and tedious task. Recently, there has been a trend in this field to attempt to construct the patterns for extraction automatically[11, 12].

Machine learning is also widely used in knowledge extraction research. Most researchers who employ this method consider knowledge extraction from text as a kind of text classification. Mitchell[8] proposed a general algorithm for learning to classify text based on a naive Bayes classifier. Detailed information about probabilistic machine learning approaches can be found in Joaxhims[4] and Lewis[6]. Information on NLP-based machine learning approaches can be found in Craven[2].

The approach taken here does not involve full NLU and so is potentially more tractable. However it also avoids the very domain-specific pattern-matching techniques of IE. It is a machine learning method based on artificial neural networks (ANNs). The benefits of ANNs are their abilities to generalise different information and learn from examples and most importantly, the compatibility with statistical and corpus-based NLP approaches. Our approach is novel in that although ANNs have been used in parsing[13, 14], there have been no similar application of ANNs in KA.

2 Identifying Keywords from Electronic Text

As mentioned, the main purpose of this research is to develop a knowledge acquisition front end for the knowledge representation formalism used in HyperTutor. The ultimate aim is to organise knowledge extracted into the same formalism. In this, knowledge is represented as a network of nodes interconnected by links where the nodes denote concepts and the links denote relationships between concepts. In each node, there is text relating to the node including some derived from the link relationships. In this paper, we refer to the names of nodes as keywords and are concerned with identifying them automatically as the first stage in a complete KA process.

We have developed techniques that identify keywords from electronic text. These techniques involve using ANN and Stemming Analysis. Results of applying them separately have been presented previously[17, 18]. This paper concentrates on the results of transferring them to a second domain.

2.1 Outline of the ANN approach

The approach taken is to train an ANN to differentiate between keywords and non-keywords based on an input representation of their relationships to a seed word which is defining the domain[17]. The relationships between each potential keyword and the seed word are obtained by searching the electronic semantic lexicon (WordNet[3]). Training data consists of input patterns for keyword and non-keyword examples where the keyword/non-keyword distinction has been judged by humans. Once trained the network should be able to recognise input patterns/relationships that correspond to keywords of the original seed word.

In order to test the feasibility of this approach the following steps were carried out with "education" as the seed word:

1. The nouns in documents relevant to the seed word domain are divided into three groups for training, testing and validation respectively. These are each judged as being keywords or non-keywords by humans. The nouns in the training set form the basis for the training data.
2. All training nouns and their relationships to seed words are identified automatically according to a universal (domain-independent) semantic lexicon. All the information for a noun (mainly derived from paths in WordNet between the noun and the seed word) is organised into a pattern that will be input to an ANN for training. The output target is 1 or 0 depending on whether the noun is a keyword of the seed word.
3. The ANN is trained.
4. The trained ANN is tested to see how well it can extract keywords from the test nouns. A threshold, T_1, of 0.5 is applied to the output to decide between keywords and non-keywords.

2.2 WordNet: The Semantic Lexicon

The semantic lexicon used is WordNet [3], an online lexical reference system. In WordNet, nouns, verbs, adjectives and adverbs are all organized into the smallest semantic unit: Synonym Set (called Synset in WordNet) which represent a single concept in English. The Synsets are interconnected by semantic relationships.

There are more than 57000 nouns in WordNet which are organised into about 48800 Synsets and are represented as a kind of semantic inheritance network. All nouns belong to one or more categories in the inheritance hierarchy but only one of the 25 top-level categories. An example of this hierarchy is shown in figure 1, from 'student', the lowest level, to 'entity', the highest. Each level in the hierarchy represents a category. There are 25 top-level categories in WordNet.

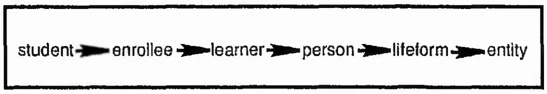


Figure 1: An example from the interitance hierarchy

There are seven semantic relationships that interconnect the noun Synsets in WordNet. They are synonym, antonym, hypernym, hyponym, meronym, holonym, coordinate.

2.3 Input Pattern Design

For each noun in the training document, there is an input-output pattern pair in the training data set. Each input pattern is composed of two parts. The first is the category information of the noun. This information should be useful because it is more likely for a noun within the same category as the seed words to be identified as a keyword. The twenty-five top-level categories in the inheritance hierarchy are used in this part, so there are twenty-five bits to represent category information. If the noun belongs to one of the top-level categories, the corresponding bit is set to 1, the remaining bits being set to 0.

The second part of the input pattern is more complicated. It represents the distance in WordNet between the noun and the seed words as well as the relationships between the words on the linking paths. A path from one noun to another is composed of all the nouns on the way and the relationship type between the adjacent nouns. For example, a path from "university" to "education" is shown in figure 2. (The intermediate words on the path are not represented on the input as the structural information about them in WordNet is confined to their relationships to other words.)

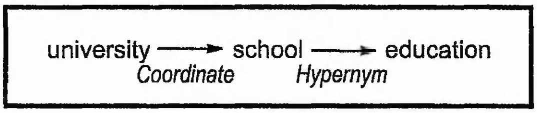


Figure 2: An example path in WordNet between 'university' and 'education'

The distance from university to education is 2. There are eight types of relationship. The second part of the input pattern contains the distance and relationship information of the shortest N paths up to maximum length of M. The criteria for choosing M and N are described later.

How are paths presented to an ANN? Suppose the maximum path length (M) is 4. A path will therefore have a length in the range 1 to 4. There are 4 fields, A to D, each representing one of the 4 path lengths. Each field contains sub-fields that allow the relationship type for each link on the path to be represented. A relationship type is represented using 8 bits. Each bit corresponds to one relationship i.e. like the classification coding only 1 bit is high at a time. The coding of 1 path with M=4 is shown in figure 3.

A	B	C	D
8 bits	16 bits	24 bits	32 bits

Figure 3: Bit pattern for a path

A denotes a path of length 1, B denotes a path of length 2, C a path of 3 and D, 4. If a path is of length 1, then the B, C and D fields are all set to 0. The A field is set according to the relationship i.e. the bit corresponding to the relevant relationship is set high. If the path is of length 2, then the A, C and D fields are all set to 0 and the B field is set according to the relationships in the path: the first 8 bits is used to represent the first relationship and the second 8 bits is used to represent the second relationship. The same principle applies to paths of length of 3 and 4.

For the example in figure 2, the length of the path is 2. The pattern of this path is shown in figure 4.

A	B		C	D
0...0,	00000100	00010000	0...0,	0...0
8 bits	*coordinate*	*hypernym*	24 bits	32 bits

Figure 4: Bit pattern of the path of length 2 in Figure 2

Up to N paths can be repeated, thus the total input pattern with M=4 is shown in figure 5.

The output pattern is one bit for the target which is either 1 for an example keyword or 0 for a non-keyword.

2.4 The Selection of Paths?

Nearly all nouns have more than one path to a seed word, so how many paths is enough for training purpose and which paths should be selected? The aim

category	path1	path2	path3	path4
	ABCD	ABCD	ABCD	ABCD
25 bits	80 bits	80 bits	80 bits	80 bits

Figure 5: Total input pattern of N paths with maximum length 4

is to present enough information for the network to learn the problem. This decides the choice of M and N. M should be large enough for all keywords in the training data have at least one path with a length equal to or shorter than M. If M is too small, some keywords will be presented to the ANN with no path information, which would give the network no information on which to base its selection.

Another requirement is to present enough information for there to be no contradictions in the training data. A contradiction occurs when two patterns have the same inputs and different outputs. If there are contradictions in the training data, the ANN will not be able to acquire the training data.

A contradiction may arise when two nouns belong to the same WordNet categories, have the same path to the seed word but one is classified as a keyword and the other a non-keyword. See Figure 6, where "week" and "semester" both belong to the same categories and have the same path to education. Identical path information can also be generated when the intermediate words are different between the two paths.

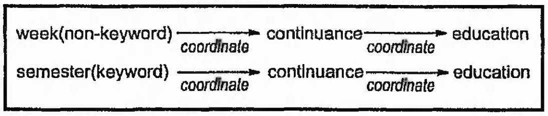


Figure 6: Total input pattern of N paths with maximum length 4

Therefore, one path for each noun is often not enough to distinguish between them. A combination of M and N is required such that there are no contradictions in the training data set. However, the M-N combination should also minimize the amount of training data. For the nouns that have more than N shortest path to choose from, the first N paths are chosen. For those that have less than N shortest paths, the path length is increased until N paths are found.

A further complication is that there is no systematic way of ordering paths on the input. Therefore, training data is generated with input patterns for all the possible ways of ordering the inputs. This aims to allow the network to recognise path features regardless of the order that they were found in when WordNet was searched, e.g. for N=3 paths, 6 training patterns containing the 6 permutations (together with the category information) are generated.

2.5 Data Preparation

The whole text was divided into three parts (training, testing and validation) based on the criteria that ideally unique keywords should be distributed evenly in the three parts, i.e. one third in each part. Unique keywords in the training set may also exist in the testing set and/or validation sets. Unique keywords in the testing set are those that do not occur in the training set, but may occur in the validation set. Unique keywords in the validation set are those that occur neither in training set nor in the test set.

[17, 18] presented results from the education domain. They showed that, the identification of keywords in a document is a complex task and that the use of an ANN to process path information improves the results for identifying keywords compared with simple path-based criteria. It was suggested that the method is limited by the richness of the path information in the dictionary.

2.6 Outline of the Stemming Analysis

Lack of sense level information (paths through WordNet between the senses of words) prevents us from applying the ANN techniques at the sense level although we believe this would give better results than at the word-level. This has led to the use of stemming analysis on sense definitions because we found that the definitions for the key senses of keywords usually have a strong verbal relevance to the seed words.

In order to test the applicability of definition and stemming analysis, the following steps were carried out:

1. Construct a relation word set (RWS) for each key sense of the seed word by combining all the synonyms, hypernyms, hyponyms, holonyms, meronyms and coordinatees of the sense. Merge the RWSs for different key senses of the seed word into a single RWS for the seed word by unioning them.
2. For a sense of a noun in the training data set (and testing data set), extract all nouns in the sense definition to form a noun set (NS).
3. Generate stemming information of all nouns in the NS against the RWS of the seed word.
4. Convert the sense-level stemming information into word-level stemming information. This is necessary for compatibility with the ANN approach which is limited to word level processing.

A RWS for the seed word was created for finding stemming information instead of the seed word itself because the seed word itself is normally too narrow to provide enough stemming information.

The similarity of all nouns in the NS to all nouns in the RWS are calculated and the stemming information of a sense is chosen as the highest of them. The result is a number between 0 and 1, which can be considered as the confidence in the sense being a key sense.

After the stemming analysis at the sense level has been done, the result of the analysis is converted to word-level by choosing the highest stemming similarity value of all the senses of a word. A threshold, T_2, is then applied to the similarity value to classify the word as a keyword or a non-keyword. Rose and Evett [22] reported similar work in using stemming analysis. Results of stemming analysis on education domain are shown in table 1.

Table 1: Results of Stemming Analysis in the Domain: Education

Data Set	Number	Identified	%
Total	344	215	0.63
Keywords	37	34	0.92
Non Keywords	307	181	0.59

2.7 Combination of the Two Approaches

The approaches were combined as fellows: suppose A_p is the output for a test pattern for a noun from the trained ANN and S_p is the stemming information for the same noun. If S_p is larger than or equal to the threshold, $(A_p + S_p)/2$ is the decision value. Otherwise, A_p is the decision value. This can be represented as formula 1, giving R_p, the decision value for the noun P. This means that the combination only applies to those words that are recognised as keywords by stemming analysis. If this is applied to all words, the words with low stemming information which are considered as non-keywords will distort the final output and thus the overall performance.

$$R_p = \begin{cases} (A_p + S_p)/2 & \text{if } S_p > T_2 \\ A_p & \text{otherwise} \end{cases} \tag{1}$$

If R_p is greater than 0.5, it is considered as a keyword. Otherwise, it is decided as a non-keyword. T_2 was 0.6.

2.8 Evaluation Method and Measures

To evaluate this new approach, new measures based on the concept of generalisation have been developed. Natural generalisation is the percentage of nouns in new text that are correctly categorised as keywords or non-keywords. Pure generalisation is the percentage of nouns with previously unseen input patterns in the new text that are correctly classified. Analogue versions of recall and precision measures commonly used in knowledge extraction research have also been developed to accommodate the ANN analogue outputs. Detailed description of these measures can be found in [17, 18].

2.9 Results for a Single Domain

Preliminary experiments have been performed with the domain defined by the seed word "education". The document chosen is a research paper entitled "Mediated Learning: A New Model of Networked Instruction and Learning" [24]. The human judges based their keyword classification on considering education in the sense of "education in a formal setting".

Table 2 shows the results from the ANN and the results from the combination of the two approaches.

Table 2: Results for the ANN approach alone and when combined with Stemming Analysis for the domain: "education"

Results of the Combination				
Data Set	**NG**	**PG**	**AR**	**AP**
Total	0.71	0.69	0.66	0.70
Keywords	0.77	0.56	0.71	0.77
Non Keywords	0.71	0.70	0.66	0.70
Results of Using ANN Only				
Data Set	**NG**	**PG**	**AR**	**AP**
Total	0.84	0.82	0.81	0.86
Keywords	0.62	0.47	0.59	0.63
Non Keywords	0.87	0.83	0.84	0.88

NG=Natural Generalisation PG=Pure Generalisation
AR= Analogue Recall AP= Analogue Precision

[17] pointed out the importance of pure generalisation because it measures the amount of induced knowledge. We take PG as the most important indication of ANN performance. We also want balanced results (i.e. as equal as possible) between keywords and non-keywords. It is easy to achieve a very high performance on keywords or on non-keywords, but for a useful system, similar performance for both is required.

The results show that, stemming analysis, combined with the ANN, improves the identification rate of keywords thereby creating a better balance. Although the measures for overall performance for keywords drop by 13%, the pure generalisation for keywords rise by 9%.

According to the criteria of achieving, high PG measure and balanced results between keywords and non-keywords, we conclude that results from the combination of ANN and stemming analysis are better than those from applying them separately.

3 Domain Portability Analysis

A major concern for knowledge acquisition researchers is domain portability of the techniques developed. Although the techniques performed well on education domain, their domain portability must be tested.

To test whether the approaches are portable across different domains, we carried out experiments on another domain: Law. The document is chosen from the encyclopaedia Britannica titled as "The Profession and Practice of Law"[21]. The seed word used is "law".

3.1 Levels of Portability

Portability experiments were carried out on two levels. One is to test the generality of the ANN, i.e. is an ANN trained on one domain portable to another domain? The other is to test the generality of the methods, i.e. are the ANN and stemming analysis methods portable to other domains?

3.2 Generic ANN for All Domains

The generic test is to input test data from a new domain into a trained ANN. Therefore, test data from the Law domain was input to the ANN trained on the Education domain. The results are shown in table 3. When compared with table 2, the results show measures for non-keywords are only slightly lower than those from using the same domain in training and test. However, measures for keywords are not good: only 60% (0.28/0.47) of the training results are portable.

It is clear that an ANN trained on a specific domain is not sufficient to be used in another domain without further processing, i.e. the ANN trained in one domain is usually not portable to another domain. However, the 60% portability suggests that different domains have something in common. It also suggests that training on more than one domain might incorporate enough generalisation to allow it to be portable to new domains.

Table 3: Results of running law test data on ANN trained using education data

Data Set	NG	PG	R	P
Total	0.74	0.74	0.71	0.86
Keywords	0.27	0.28	0.24	0.26
Non Keywords	0.77	0.77	0.74	0.79

3.3 Portability of the ANN method

Data from the Law domain was divided into training, testing and validation as before. An ANN was trained in the same way. The results are in table 4.

Comparing the results of ANN from Education and Law domains shown in table 2 and 4, it is clear that the performance of the ANN approach on the Law domain is as good as on the Education domain. Although the measures on Non-keywords are slightly worse than from the Education domain, the measures on keywords are all better. For example, Pure Generalisation in the Law domain is 7% higher than in the Education domain.

Table 4: Results for the ANN approach in domain: Law

Data Set	NG	PG	R	P
Total	0.80	0.75	0.77	0.82
Keywords	0.74	0.54	0.66	0.71
Non Keywords	0.80	0.76	0.78	0.83

Thus, we can conclude that the technique using ANN to identify keywords is domain portable.

3.4 Portability of the Stemming Method

Stemming analysis was carried out for the Law domain data, see table 5 for the results. The value of T_2 was again 0.6.

Table 5: Results of Stemming Analysis in the Domain: Law

Data Set	Number	Identified	%
Total	521	370	0.72
Keywords	41	19	0.54
Non Keywords	351	351	0.73

Comparing results of stemming analysis from the two domains (table 5 and table 1 respectively), it can be seen that the stemming analysis results on the Law domain are not as good as in the Education domain in terms of keyword identification, but it produced a good result for non-keywords. The percentage of identified keywords is only 54%, much worse than that of the Education domain, of 92%. However, the overall correctness of identification is better than from education domain. The low rate of keyword identification does not necessarily mean a failure of stemming analysis. Higher rates of non-keyword identification can help to reject spurious keywords recognised by the ANN.

By comparing the keywords, shown in table 6, from both domains, we suggest that the low keyword identification rate may be because the legal terms are more specialised and therefore not as well represented in WordNet which is intended as a general purpose lexicon. Conversely, educational terms are in more commonly use and are also applied outside the domain itself. Their representation and interconnectivity in WordNet is therefore richer.

Table 6: Keywords from education and law domain

Keywords from education domain
Academic, academic-year, assessment, baccalaureate, campus, class, classroom, coaching, college, college-level, competence, comprehension, course, curriculum, degree, education, enrollment, exam, examination, grade, grading, graduate, graduation, higher-education, homework, instruction, instructor, knowledge, learner, learning, lecture, lecturer, lecturing, lesson, polytechnic, remediation, school, semester, student, study, studying, subject, syllabus, teach, teacher, teaching, term, textbook, training, tuition, tutor, tutoring, undergraduate, university
Keywords from law domain
accused, act, action, advocacy, advocate, appeal, appearance, assessor, attorney, bar, barrister, case, case-law, casebook, chancery, civil-law, client, commission, common-law, compensation, contract, conviction, counsel, counselor, court, crime, criminal, defendant, defense, disbarment, enactment, evidence, guilt, imprisonment, judge, judgeship, judgement, judiciary, jurisdiction, jurisprudence, jurist, jury, justice, law, lawyer, legality, legislation, legislature, litigant, litigation, magistracy, magistrate, notary, offense, precedent, pretrial, proceeding, procurator, prosecution, prosecutor, punishment, regulating, regulation, right, solicitation, solicitor, statute, testimony, trial, tribunal, witness

3.5 Portability of ANN and Stemming Analysis Combined

The results from this combination in the new domain are shown in table 7. It can be seen that all the measures follow the same trends as in the Education domain, e.g. the pure generalisation for keywords increases by 11% compared with using ANN alone.

Results from stemming analysis in the law domain are worse than in the education domain, but the results from the ANN are better. The overall performance (by combining the two approaches) in the new domain is better than in the education domain.

4 Conclusions and Future Work

4.1 Conclusions

We have shown that concepts can be automatically extracted from text using an ANN and stemming analysis. Experiments on domains concerning education and law have shown that these approaches (ANN and stemming analysis) are portable across domains although stemming analysis alone is sensitive to differences in the nature of the domain keywords. The combination of ANN and stemming analysis produces the best results and was robust to differences in the characteristics of the keywords. Experiments to test the portability of trained ANNs between domains suggest that although training on one domain does not give good results for a different domain, there was sufficient crossover to suggest that it may be possible to generate a portable network using training data from several domains.

Table 7: Results for ANN approach alone and when combined with Stemming Analysis for the domain: "law"

Results of the Combination				
Data Set	**NG**	**PG**	**R**	**P**
Total	0.74	0.69	0.60	0.78
Keywords	0.80	0.65	0.68	0.84
Non Keywords	0.73	0.69	0.59	0.78
Results of Using ANN Only				
Data Set	**NG**	**PG**	**R**	**P**
Total	0.80	0.75	0.77	0.82
Keywords	0.74	0.54	0.66	0.71
Non Keywords	0.80	0.76	0.78	0.83

NG=Natural Generalisation PG=Pure Generalisation R=Recall P=Precision

Many other researchers [9, 15, 16] who extract keywords from text use information and probability theories aimed at providing keyword lists and/or glossaries for information retrieval. [19] extracts concepts from text based on some seed words, using Wordnet as electronic dictionary. Our approach is based on the semantic relationships between words. It is more appropriate for our final objective, i.e. to construct a knowledge base. One contribution of our work is the novel approach to using ANNs in knowledge acquisition, including the definition of an evaluation methodology which involves new measures of performance.

4.2 Future Work

We are currently investigating training an ANN on multiple domains to provide domain portability. Such an ANN would significantly reduce the human resource needed to identify keywords. Testing on a third domain is also under consideration to fully understand the nature of the transferability of stemming analysis. The unstableness of stemming analysis may result from the characteristics of the domain itself or from the electronic dictionary used. We have found some limitations of WordNet in the process of training ANNs. Some other researchers [20, 23] also reported similar problems. Thus, using the same techniques with another electronic dictionary is worth investigating. Future work

Table 6: Keywords from education and law domain

Keywords from education domain
Academic, academic-year, assessment, baccalaureate, campus, class, classroom, coaching, college, college-level, competence, comprehension, course, curriculum, degree, education, enrollment, exam, examination, grade, grading, graduate, graduation, higher-education, homework, instruction, instructor, knowledge, learner, learning, lecture, lecturer, lecturing, lesson, polytechnic, remediation, school, semester, student, study, studying, subject, syllabus, teach, teacher, teaching, term, textbook, training, tuition, tutor, tutoring, undergraduate, university
Keywords from law domain
accused, act, action, advocacy, advocate, appeal, appearance, assessor, attorney, bar, barrister, case, case-law, casebook, chancery, civil-law, client, commission, common-law, compensation, contract, conviction, counsel, counselor, court, crime, criminal, defendant, defense, disbarment, enactment, evidence, guilt, imprisonment, judge, judgeship, judgement, judiciary, jurisdiction, jurisprudence, jurist, jury, justice, law, lawyer, legality, legislation, legislature, litigant, litigation, magistracy, magistrate, notary, offense, precedent, pretrial, proceeding, procurator, prosecution, prosecutor, punishment, regulating, regulation, right, solicitation, solicitor, statute, testimony, trial, tribunal, witness

3.5 Portability of ANN and Stemming Analysis Combined

The results from this combination in the new domain are shown in table 7. It can be seen that all the measures follow the same trends as in the Education domain, e.g. the pure generalisation for keywords increases by 11% compared with using ANN alone.

Results from stemming analysis in the law domain are worse than in the education domain, but the results from the ANN are better. The overall performance (by combining the two approaches) in the new domain is better than in the education domain.

4 Conclusions and Future Work

4.1 Conclusions

We have shown that concepts can be automatically extracted from text using an ANN and stemming analysis. Experiments on domains concerning education and law have shown that these approaches (ANN and stemming analysis) are portable across domains although stemming analysis alone is sensitive to differences in the nature of the domain keywords. The combination of ANN and stemming analysis produces the best results and was robust to differences in the characteristics of the keywords. Experiments to test the portability of trained ANNs between domains suggest that although training on one domain does not give good results for a different domain, there was sufficient crossover to suggest that it may be possible to generate a portable network using training data from several domains.

Table 7: Results for ANN approach alone and when combined with Stemming Analysis for the domain: "law"

Results of the Combination				
Data Set	**NG**	**PG**	**R**	**P**
Total	0.74	0.69	0.60	0.78
Keywords	0.80	0.65	0.68	0.84
Non Keywords	0.73	0.69	0.59	0.78
Results of Using ANN Only				
Data Set	**NG**	**PG**	**R**	**P**
Total	0.80	0.75	0.77	0.82
Keywords	0.74	0.54	0.66	0.71
Non Keywords	0.80	0.76	0.78	0.83

NG=Natural Generalisation PG=Pure Generalisation
R=Recall P=Precision

Many other researchers [9, 15, 16] who extract keywords from text use information and probability theories aimed at providing keyword lists and/or glossaries for information retrieval. [19] extracts concepts from text based on some seed words, using Wordnet as electronic dictionary. Our approach is based on the semantic relationships between words. It is more appropriate for our final objective, i.e. to construct a knowledge base. One contribution of our work is the novel approach to using ANNs in knowledge acquisition, including the definition of an evaluation methodology which involves new measures of performance.

4.2 Future Work

We are currently investigating training an ANN on multiple domains to provide domain portability. Such an ANN would significantly reduce the human resource needed to identify keywords. Testing on a third domain is also under consideration to fully understand the nature of the transferability of stemming analysis. The unstableness of stemming analysis may result from the characteristics of the domain itself or from the electronic dictionary used. We have found some limitations of WordNet in the process of training ANNs. Some other researchers [20, 23] also reported similar problems. Thus, using the same techniques with another electronic dictionary is worth investigating. Future work also includes finding the definitions of concepts and the semantic relationships between concepts in order to construct the information in the final knowledge base.

References

[1] *Proceedings of the Sixth Message Understanding Conference (MUC-6),* Columbia, MD, Morgan Kaufmann, November 1995.

[2] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam and S. Slattery *Learning to construct Knowledge Base from the World Wide Web,* Artificial Intelligence, 1999.

[3] C. Fellbaum *WordNet: An Electronic Lexical Database,* MIT Press, 1998.

[4] T. Joachims *A probabilistic analysis of Rocchio algorithm with TFIDF for text categorization,* Computer Science Technical Report CMU-CS-96-118, Carnegie Mellon University.

[5] W. Lehnert,, C. Cardie, D.Fisher, J. MCCarthy, E. Riloff, and S. Soderland. *Evaluating an Information Extraction System,* Journal of Integrated Computer-Aided Engineering,1(6),1994.

[6] D. Lewis *Representation and learning in informal retrieval,* Ph.D thesis, (COINS Technical Report 91-93), Department of Computer and Information Science, University of Massachusetts, 1991.

[7] M. Edwards, H. Powell, D. Palmer-Brown, *A Comparative Evaluation of a Natural Language Exploration Tool within a Hypermedia Environment,* ICTAI'96: IEEE International Conference on Tools with Artificial Intelligence, Toulouse, France, Nov 1996

[8] T. Mitchell *Machine Learning,* McGraw-Hill International Editions, 1997.

[9] Y. Otha, Y. Yamamoto, T. Okazaki, I. Uchiyama, and T. Takagi *Automatic construction of knowledge base from biological papers,* Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology, 218-225. Halkidiki, Greece: AAAI Press. 1997.

[10] R. Quirk, S. Greenbaum, G. Leech, J. Svartvik *A comprehensive Grammar of the English Langiage,* Longman, 1985.

[11] E. Riloff *An Empirical Study of Automated Dictionary Construction for Information Extraction in Three Domains,* AI journal, Vol. 85 August 1996.

[12] S. Soderland, D. Fisher, J. Aseltine, and W. Lenhert *CRYSTAL: Inducing a conceptual dictionary,* In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 1995.

[13] J. Tepper, H. Powell, D. Palmer-Brown *Ambiguity Resolution in a Connectionist Parser,* The Cognitive Science of Natural Language Processing, July 5-7 1995, Editor A I C Monaghan, Natural Language Group. 1995a.

[14] J. Tepper, H. Powell, D. Palmer-Brown *Integrating Symbolic and Subsymbolic Architecture for Parsing Arithmetic Expressions and Natural Language Sentences,* Proceeding of 3rd SNN Neural Network Symposium, Nijmegen, Sept 1995, pp 81-84, Eds Bert Kappen and Stan Gielen, ISBN 3-540-19992-6. 1995b.

[15] P.D. Turney *Extraction of Keyphrase from Text: evaluation of Four Algorithms,* NRC Technical Report ERB-1051, National Research Council Canada, 1997.

[16] M. Weeber, and R. Vos 1998. *Extracting expert medical knowledge from texts,* In Working Notes of the Intelligent Data Analysis in Medicine and Pharmacology Workshop, 23-28.

[17] S. Zhang, H. Powell and D.Plamer-Brown, *Keyword Extraction Using Neural Networks,* Proceedings of the Tenth Meeting Computational Linguistics in the Netherlands,1999.

[18] S. Zhang, H. Powell and D.Plamer-Brown, *Keyword Extraction from Stemming and Sense Information by Neural Networks,* the proceedings of the year 2000 International Conference on Artificial Intelligence (IC-AI'2000). 2000.

[19] D. Moldovan and R. Girju, *Domain-Specific Knowledge Acquisition and Classification using WordNet,* The proceedings of the 13th International FLAIRS Conference (FLAIRS-2000), Orlando, Florida, May, 2000

[20] M. De Boni, *Use and Limitation of WordNet for QA,* http://www-users.cs.york.ac.uk/ ~mdeboni/research/wordnet_critique.html, Accessed on 20, April, 2001.

[21] Britannica.com *The Profession and Practice of Law,* Britannica Encyclopedia CD, Multimedia Edition, 1999.

[22] T.G. Rose, and L.J. Evett *The Use of Context in Cursive Script Recognition,* Machine Vision and Application (1995) 8:241-248. 1995

[23] A. Maedche, and S. Staab *Discovering Conceptual Relations from Text,* Proceedings of the 14th European Conference on Artificial Intelligence, IOS Press, Amsterdam, 2000

[24] Academic Systems Mediated Learning Library, *Mediated Learning: A New Model of Networked Instruction and Learning,*

http://www.academic.com/library/articles/
mllibrary.html, Accessed on 24, May, 1999.

Keyword Extraction from Stemming and Sense Information by Neural Networks

Shaomin Zhang and **Heather Powell**
Newton Building, Computing Department, Nottingham Trent University
Burton Street, Nottingham NG1 4BU, UK

Dominic Palmer-Brown
School of Computing, Faculty of Information and Engineering Systems
Beckett Park Campus, Leeds Metropolitan University
Leeds LS6 3QS, UK

Abstract

The research presented in this paper investigates domain independent techniques for automatic knowledge extraction from text. The knowledge is to be organised into a knowledge base. The techniques presented are aimed at the first stage: the automatic identification of keywords.

Artificial Neural Networks (ANNs) are trained to recognise keywords on the basis of their sense information, stemming analysis and relationships to one or more seed words which are manually selected as indicative of the areas of knowledge required. The relationships are obtained from an electronic dictionary. Training data is generated using example keywords that humans have identified as being keywords associated with particular seed words. After training, the ANN can be used to extract keywords automatically from other documents.

New measures, natural generalisation and pure generalisation, based on the concept of generalisation have been introduced. Recall and precision measures commonly used in knowledge extraction research have been adapted to suit the ANN-based approach. Experiments so far, on documents concerning education, show the encouraging result of this new approach.

Keywords: Knowledge Acquisition, Natural Language Processing, Neural Networks, Knowledge Base

1 Introduction

In this paper, we present research on knowledge extraction from text. The main objective of the research is to develop techniques for automatic knowledge extraction directly from plain text in electronic form, so that the extracted knowledge can be organised into a knowledge base.

The target knowledge base is used in a hyper-knowledge interaction environment called HyperTutor[8]. This uses a novel and generic formalism for structuring and interrogating hypermedia-based knowledge via a natural language interface. The system engages users in a dialogue with knowledge as well as allowing them to browse. It also has pedagogic features for tutoring. It employs semantic hyperlinks to represent knowledge. HyperTutor is a generic environment, therefore generic KA techniques are required. It will be a great benefit to automate the knowledge acquisition process so that knowledge can be automatically extract from text with minimum human involvement. This paper presents research into enabling the important concepts (keywords) in a domain to be automatically identified. The identification is based on one or more seed words which are provided by a human author to define the domain.

2 Related Work

The first conceivable approach for solving the task of automatic knowledge acquisition is to fully understand the natural language text. This method, however, is beyond the capabilities of current natural language understanding (NLU) systems. The main reason for this is the complexity of natural language and the lack of appropriate linguistic theory to manage this complexity. It is difficult to build a grammar for a realistic subset of natural language[11]. In particular it is difficult to process exceptions.

Another approach to knowledge acquisition is Information Extraction (IE)[1]. IE aims to identify instances of a particular class of event or relationship in natural language text. Relevant arguments concerning events and relationships are extracted and encoded in a format suitable for incorporation into a database [6]. The construction of extraction patterns which are used by most IE systems to extract information is a time-consuming, knowledge-intensive and tedious task. Recently, there has been a trend in this field to attempt to construct the patterns for extraction automatically [12, 13].

Machine learning is also widely used in knowledge extraction research. Most researchers who employ this method consider knowledge extraction from text as a kind of text classification. Mitchell [9] proposed a general algorithm for learning to classify text based on a naive Bayes classifier. Detailed information about probabilistic machine learning approaches can be found in Joaxhims [5] and Lewis [7]. Information on NLP-based machine learning approaches can be found in Craven[3].

The approach taken here does not involve full NLU and so is potentially more tractable. However it also avoids the very domain-specific pattern-matching techniques of IE. It is a machine learning method based on artificial neural networks (ANNs). The benefits of ANNs are their abilities to generalise, learn from examples and most importantly, their compatibility with statistical and corpus-based NLP approaches. Our approach is novel in that although ANNs have been used in parsing[14], there have been no similar application of ANNs in KA.

3 Previous Work

3.1 Introduction

As mentioned, the main purpose of this research is to develop a knowledge acquisition front end for the knowledge representation formalism used in HyperTutor[8]. The ultimate aim is to organise knowledge extracted into the same formalism. This represents knowledge as a network of nodes interconnected by links where the nodes denote concepts and the links denote relationships between concepts. In each node, there is text relating to the node including some derived from the link relationships. In this paper, we refer to the names of nodes as keywords and are concerned with identifying them automatically as the first stage in a complete KA process.

3.2 Outline of the approach

The approach taken is to train an ANN to differentiate between keywords and non-keywords based on an input representation of their relationships to a seed word which is defining the domain[19]. The relationships between each potential keyword and the seed word are obtained by searching an electronic semantic lexicon. Training data consists of input patterns for keyword and non-keyword examples where the keyword/non-keyword distinction has been judged by humans. Once trained the network should be able to recognise input patterns/relationships that correspond to keywords of the original seed word. It is hoped that what the network has learnt about what signifies a keyword relationship to the original seed word will be transferable to other seed words i.e. domain independent. However this is not evaluated here, as this work evaluates the approach for one domain.

In order to test the feasibility of this approach the following steps were carried out with education as the seed word:

1. The nouns in documents relevant to the seed word domain are divided into three groups for training, testing and validation respectively. These are each judged as being keywords or non-keywords by humans. The nouns in the training set form the basis for the training data.
2. All training nouns and their relationships to seed words are identified automatically according to a universal (domain-independent) semantic lexicon. All the information for a noun is organised into a pattern that will be input to an ANN for training. The output target is 1 or 0 depending on whether the noun is a keyword of the seed words.
3. The ANN is trained.
4. The trained ANN is tested to see how well it can extract keywords from the test nouns.

3.3 WordNet: The Semantic Lexicon

The semantic lexicon used is WordNet [4], an on-line lexical reference system. Concepts (called Synsets in WordNet) are represented as a kind of semantic inheritance network. All nouns belong to one or more categories in the inheritance hierarchy but only to one of the 25 top-level categories.

The Synsets are interconnected by seven semantic relationships. They are synonym, antonym, hypernym, hyponym, meronym, holonym, coordinate. For symmetry, we have introduced a new relationship called coordiantee which means "nouns that have the same hyponyms".

3.4 Input Patterns for the ANN

For each noun in the training document, there is an input-output pattern pair in the training data set. Each input pattern is composed of two parts. The first is the category information of the noun. There are twenty-five bits in this part corresponding to the twenty-five categories. If the noun belongs to one of the top-level categories, the corresponding bit is set to 1, the remaining bits being set to 0.

The second part of the input pattern represents the distance in WordNet between the noun and the seed words as well as the relationships between the words on the linking paths. An example path from "university" to "education" is shown in figure 1a. This part contains N shortest paths up to maximum length of M.

A combination of M and N is required such that there is enough information for keywords to be leant by the network and no contradictions in the training data set. Usually, M should be large enough for all keywords in the training data to have at least one path with a length shorter than or equal to M. However, the M-N combination should also minimize the amount of training data.

Suppose the maximum path length (M) is 4. A path will therefore have a length in the range 1 to 4. There are 4 fields, A to D, each representing one of the 4 path lengths. Each field contains sub-fields that allow the relationship type for each link on the path to be represented. A relationship type is represented using 8 bits. Each bit corresponds to one relationship i.e. like the classification coding only 1 bit is high at a time.

A denotes a path of length 1, B denotes a path of length 2, C a path of 3 and D, 4. If the path is of length 2, then the A, C and D fields are all set to 0 and the B field is set according to the relationships in the path: the first 8 bits are used to represent the first relationship and the second 8 bits are used to represent the second relationship. The same principle applies to paths of length of 1, 3 and 4. The pattern for the path in figure 1a is shown in figure 1b.

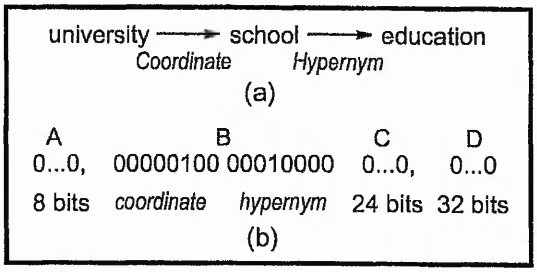


Figure 1: (a)a path from university to education (b) the bit pattern for the path in (a)

There are 80 bits per path. Up to N paths can be repeated, thus the total input pattern is 25 category bits plus N sets of 80 bits.

The output pattern is one bit for the target which is either 1 for an example keyword or 0 for a non-keyword.

Because there is no systematic way of ordering paths on the input, therefore, training data is generated with input patterns for all the possible ways of ordering the inputs. This aims to allow the network to recognise path features regardless of the order that they were found in WordNet when it was searched, e.g. for N=3 paths, 6 training patterns containing the 6 permutations (together with the category information) are generated.

3.5 Experiment

Experiments have been performed with the domain defined by the seed word "education". The document chosen is a research paper entitled "Mediated Learning: A New Model of Networked Instruction and Learning"[2]. It is comprised of 19672 words with 707 unique nouns occurring 8334 times. 54 words were identified as keywords. The human judges based their keyword classification on considering education in the sense of "education in a formal setting".

The whole text was divided into three parts based on the criteria that ideally unique keywords should be distributed evenly in the three parts, i.e. one third in each part. The result of the division is shown in Table 1.[1]

[1] Unique keywords in the training set may also exist in the testing set and/or validation sets. Unique keywords in the testing set are those that do not occur in

Table 1: Result of Document Division

Data Set	Nouns	Key-words	Unique Keywords
Whole	707	54	54
Training	111	19	19
Testing	344	39	20
Validating	687	50	15

The ANN architecture used is Feed-Forward with Backpropagation. A series of tests were carried out to establish M=4 and N=5 as the optimum combination for getting rid of the contradictions. The experiments also aimed to minimise the number of hidden-neurons on the training and test data which resulted in 2.

After trained, the network was presented with 41280 patterns in the test data set to see how well the network learnt the problem. The result of testing is shown in Table 2.

Table 2: Result of Testing

Word Type	Total	P	C
Total Nouns	344	41280	84%
Keywords	39	4680	62%
Non-Keywords	305	36600	87%
Unique Nouns	252	30240	82%
Unique Keywords	20	2400	47%
Unique Non-Keywords	232	27840	83%

P=Number Of Patterns(120 patterns per word)
C=Percentage of patterns identified correctly

4 Word Sense Information and Stemming Analysis

4.1 Word Sense Information

Word sense disambiguation (WSD) has long been considered to be able to increase the accuracy of natural language processing, e.g. information extraction [16], Machine Translation [17] and Parsing. This section presents experiments carried out to investigate the possibility using WSD to increase the rate of correct keyword extraction. This paper is not attempting to decide which sense a noun should take in a given context, but is attempting to use word sense information to increase the accuracy of keyword identification.

In WordNet, the relationships between nouns are actually the relationships between the senses of nouns. Therefore a noun should be identified as a keyword only when at least one of its senses is a key sense. If the identification is carried out via paths through WordNet then at least one sense of a keyword should have a qualifying path to one or more key senses of the seed word. In WordNet, "education" has six meanings, but only two of them are relevant to the domain definition used in our experiment. Some nouns may have paths to "education" but not to the senses that we are concerned with. These paths are spurious and should be excluded.

For each sense of a noun in the training set, there are patterns in the training data. This also applies to the testing data.

Before the generation of sense-level paths, all senses of the nouns in the document were carefully examined by three different people to identify the key senses.

Finding paths for a sense needs to be strictly based on the sense information. This leads to the discovery that WordNet has a lack of links between concepts. For example, sense two of "college" is defined as "An institution of higher education created to educate and grant degrees; often a part of a university" and yet has no path to either key sense of the seed word "education". It has no path to "higher education" either although "higher education" appears in its definition. However it has paths to "educational institution" but this is also not connected to "education" and "higher education". This is an aspect of the design of WordNet.

We therefore conclude that there is no close relationship between how the noun synsets are connected together in the WordNet hierarchy and the definitions of the noun synsets. This leads to the lack of paths based on sense information and hence the conclusion that keyword identification could not be carried out by just using the sense-level path information as had been hoped. There is a clear need for an extended form of WordNet.

4.2 Stemming analysis

By examining the sense definitions of the synsets it was found that words in the synset definitions of keywords usually have a strong

the training set, but may occur in the validation set. Unique keywords in the validating set are those that occur neither in training set nor in the test set.

verbal relevance to the seed word, as was seen for "college" above. The possibility of using of the definition information for keyword identification has therefore been investigated.

In utilising the sense definition, we chose stemming analysis instead of syntactic structure analysis because the former is much easier to perform and the result of it is easier to combine with the result of a neural network or to feed to a neural network as part of training patterns. The purpose of a stemming analysis is to identify when two words have the same root.

In order to test the applicability of definition and stemming analysis, the following steps were carried out:

1. Construct a relation word set (RWS) for each key sense of the seed word by combining all the synonyms, hypernyms, hyponyms, holonyms, meronyms and coordinatees of the sense. Merge the RWSs for different key senses of the seed word into a single RWS for the seed word by unioning them.

2. For a sense of a noun in the training data set (and testing data set), extract all nouns in the sense definition to form a noun set (NS).

3. Generate stemming information of all nouns in the NS against the RWS of the seed word.

4. Convert the sense-level stemming information into word-level stemming information. Because of the lack of sense-level information, word-level path information must be used and stemming and path information must be on the same level, thus this step is necessary.

A RWS for the seed word was created for finding stemming information instead of the seed word itself because the seed word itself is normally too narrow to provide enough stemming information.

4.3 Stemming Algorithms

There are four popular automatic approaches to stemming, namely affix removal, n-gram, table lookup and successor variety[18]. Table lookup is not an option currently because it will take a lot of time to build a comprehensive stem dictionary. Our experiments using n-gram stemming produced high assocition for nearly all keywords, however it also leads to a high level of false associations for non-key senses. The main reason that nouns are associated spuriously is that they have the same suffix which contributes much to the similarity measure used in the method. Therefore suffix removal has been combined with the n-gram method to keep the associations for key senses while decreasing the spurious associations of non-key senses. The result on the training and testing set is shown in table 3.

Table 3: Result of Stemming Analysis

Training Set						
Data Set	Sense Level			Word Level		
	N	**C**	**P**	**N**	**C**	**P**
Total	489	376	0.77	111	67	0.60
Key	27	24	0.89	18	17	0.94
NK	462	352	0.76	93	46	0.49
Testing Set						
Data Set	Sense Level			Word Level		
	N	**C**	**P**	**N**	**C**	**P**
Total	1353	1069	0.79	344	215	0.63
Key	52	43	0.83	37	34	0.92
NK	1301	1026	0.79	307	181	0.59

NK=Non-key; N=Number of Items; P=Percentage
C=Number of items with strong stemming information

The similarity of all nouns in the NS to all nouns in the RWS are calculated and the stemming information of a sense is chosen as the highest of them. The result is a number between 0 and 1, which can be considered as the confidence in the sense being a key sense.

After the stemming analysis at the sense level has been done, the result of the analysis is converted to word-level by choosing the highest stemming similarity of all the senses of a word.

5 Evaluation

5.1 Methodology

To evaluate this novel approach, we introduced new measures based on the concept of generalisation in ANN research and adapted recall and precision measures which are widely accepted in KA research. The two systems evaluate different aspects of the approach.

Natural generalisation (NG) is the percentage of nouns in the testing data that are correctly categorised as keywords or non-keywords. Therefore,

$$NG = \frac{N_{CorrectlyIdentifiedWordsInASet}}{N_{WordsInTheSet}}; \quad (1)$$

This can be evaluated for the total test set or evaluated seperately for keywords and non-keywords by substituting the set in the formula with total test set, keyword set or non-keyword set.

Pure generalisation (PG) was introduced to measure the amount of induced knowledge. It is the percentage of nouns with previously unseen input patterns in the testing data that are correctly classified. Again it can be applied to the total result and to keywords and non-keywords separately. It can be described as,

$$PG = \frac{N_{CorrectlyIdentifiedUniqueWordsInASet}}{N_{UniqueWordsInTheSet}}; \quad (2)$$

Another evaluation method widely use in KA research is recall and precision[6]. Suppose the target and actual output of a pattern P in the testing data set, TS, are T_p and A_p respectively, where T_p is either 1 or 0 and $0 <= A_p <= 1$. The formulae of recall (R) and precision (P) suitable for an ANN-based approach are:

$$R = \frac{2\sum_{p\in TS}|A_p - 0.5| * (1 - |T_p - A_p|)}{N_{WordsInTS}}; \quad (3)$$

$$P = \frac{\sum_{p\in TS}|A_p - 0.5| * (1 - |T_p - A_p|)}{\sum_{p\in TS}(|A_p - 0.5|)}; \quad (4)$$

Formulae (3) and (4) can also be used to evaluate recall and precision for keywords and non-keywords by replacing TS with TKW and $TNKW$ which means keyword set and non-keyword set respectively in the testing set.

5.2 Experiment results

Suppose A_p is the output of a testing pattern from the previous trained network (using the word-level path information because of the lack of path information on sense-level) obtained by running the trained network over the testing data and S_p is the stemming information for the noun representing this testing pattern, formula $(A_p + S_p)/2$ is used to calculate the final result as the decision value. This simple calculation produces the result in table 4.

Table 4: Experiment results

Using Stemming Information				
Data Set	**NG**	**PG**	**R**	**P**
Total	0.71	0.69	0.66	0.70
Keywords	0.77	0.56	0.71	0.77
Non Keywords	0.71	0.70	0.66	0.70
Previous Results				
Data Set	**NG**	**PG**	**R**	**P**
Total	0.84	0.82	0.81	0.86
Keywords	0.62	0.47	0.59	0.63
Non Keywords	0.87	0.83	0.84	0.88

Although the PG of 0.82 was high for the previous results, the identification of keywords was poor at 0.47. As identification of keywords is the main purpose of this KA stage, the incorporation of stemming information represents an improvement (0.56 for PG) whilst maintaining a high rejection rate (0.7) for non-keywords. The same pattern is seen for all of the other measures (NG, R and P).

6 Conclusions

We have shown that concepts can be automatically extracted from text using an ANN. Results in the education domain are encourging.

Many other researchers [10, 15] who extract keywords from text use information and probability theories aimed at providing keyword lists and/or glossaries for information retrieval. Our approach is based on the semantic relationships between words. It is more appropriate for our final objective, i.e. to construct a knowledge base. One contribution of our work is the novel approach to using ANNs in knowledge acquisition, including the definition of an evaluation methodology which involves new measures of performance.

We are currently investigating using a stemming dictionary to improve the accuracy of stemming analysis, and the pure generalisation of both keywords and non-keywords. Also under investigation is the use of a neural network as an analysis tool to exploit the stemming information. Section 5.2 presents initial results to give an idea of the feasibility of stemming analysis. Future work will investigate using a neural network to learn how to combine the output from the previous network and the stemming information. Further aims are to extract the definitions of concepts and the semantic relationships between concepts in order to construct the information in the final knowledge base.

References

[1] *Proceedings of the Sixth Message Understanding Conference (MUC-6),* Columbia, MD, Morgan Kaufmann, November 1995.

[2] Academic Systems Mediated Learning Library, *Mediated Learning: A New Model of Networked Instruction and Learning,* http://www.academic.com/library/articles/mllibrary.html, Accessed on 24, May, 1999.

[3] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam and S. Slattery *Learning to construct Knowledge Base from the World Wide Web,* Artificial Intelligence, 1999.

[4] C. Fellbaum *WordNet: An Electronic Lexical Database,* MIT Press, 1998.

[5] T. Joachims *A probabilistic analysis of Rocchio algorithm with TFIDF for text categorization,* Computer Science Technical Report CMU-CS-96-118, Carnegie Mellon University.

[6] W. Lehnert,, C. Cardie, D.Fisher, J. MCCarthy, E. Riloff, and S. Soderland. *Evaluating an Information Extraction System,* Journal of Integrated Computer-Aided Engineering,1(6),1994.

[7] D. Lewis *Representation and learning in informal retrieval,* Ph.D thesis, (COINS Technical Report 91-93), Department of Computer and Information Science, University of Massachusetts, 1991.

[8] H. Powell, D. Palmer-Brown, and J. Downs, G. Long, M. Edwards *Artificial Intelligence for Hypermedia Access: Issues in Knowledge Representation and Natural Language Processing,* Submitted to International Journal of Intelligent Systems.

[9] T. Mitchell *Machine Learning,* McGraw-Hill International Editions, 1997.

[10] Y. Otha, Y. Yamamoto, T. Okazaki, I. Uchiyama, and T. Takagi *Automatic construction of knowledge base from biological papers,* Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology, 218-225. Halkidiki, Greece: AAAI Press. 1997.

[11] R. Quirk, S. Greenbaum, G. Leech, J. Svartvik *A comprehensive Grammar of the English Langiage,* Longman, 1985.

[12] E. Riloff *An Empirical Study of Automated Dictionary Construction for Information Extraction in Three Domains,* AI journal, Vol. 85 August 1996.

[13] S. Soderland, D. Fisher, J. Aseltine, and W. Lenhert *CRYSTAL: Inducing a conceptual dictionary,* In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 1995.

[14] J. Tepper, H. Powell, D. Palmer-Brown *Ambiguity Resolution in a Connectionist Parser,* The Cognitive Science of Natural Language Processing, July 5-7 1995, Editor A I C Monaghan, Natural Language Group. 1995a.

[15] M. Weeber, and R. Vos 1998. *Extracting expert medical knowledge from texts,* In Working Notes of the Intelligent Data Analysis in Medicine and Pharmacology Workshop, 23-28.

[16] T. Strzalkowski, *Information retrieval using robust language processing,* AAAI Spring Symposium on Representation and Aquisition of Lexical Information, Stanford,104-111,1995.

[17] J. Hutchins and H. Sommers, *Introduction to Machine Translation,* Academic Press,1992.

[18] W. Frakes and R. Baeza-Yates, *Information Retrieval,* Prentice Hall,1992.

[19] S. Zhang and H. Powell and D.Plamer-Brown, *Keyword Extraction Using Neural Networks,* Proceedings of the Tenth Meeting Computational Linguistics in the Netherlands,1999.

Keyword Extraction Using Neural Networks

Shaomin Zhang and **Heather Powell**
shaomin.zhang@ntu.ac.uk hmp@doc.ntu.ac.uk
Newton Building, Computing Department, Nottingham Trent University
Burton Street, Nottingham, England, NG1 4BU

Dominic Palmer-Brown
d.Palmer-Brown@lmu.ac.uk
School of Computing, Faculty of Information and Engineering Systems,
Beckett Park Campus, Leeds Metropolitan University,
Leeds LS6 3QS, UK

Abstract

The research presented in this paper investigates domain independent techniques for automatic knowledge extraction from text. The knowledge is to be organised into a knowledge representation (KR) scheme. The techniques presented are aimed at the first stage: the automatic identification of keywords (any word closely associated with a particular domain as defined by one or more seed word). The aim is to discover any key concepts from any section of text given a small number of seed words associated with any domain.

Artificial Neural Networks (ANNs) are trained to recognise keywords on the basis of their relationships to one or more seed words which define a subject domain. The relationships are obtained from an electronic dictionary. Training data is generated using example keywords that humans have identified as being keywords associated with particular seed words. After training, the ANN can be used to extract keywords automatically from other documents.

To evaluate this new approach, new measures based on the concept of generalisation have been introduced. Also, analogue versions of recall and precision measures commonly used in knowledge extraction research have been developed to accommodate the ANN analogue outputs. Natural generalisation is the percentage of nouns in new text that are correctly categorised as keywords or non-keywords. Pure generalisation is the percentage of nouns with previously unseen input patterns in the new text that are correctly classified. Experiments so far, on documents concerning education show good natural and pure generalisation for non-keywords at 84% and 82% respectively and reasonable generalisation for keywords (62% for natural and 47% for pure). Results for recall and precision are, for keywords: 59%(analogue recall), 63%(analogue precision), 62%(binary recall), 38%(binary precision) and for non-keywords: 84%(analogue recall), 88%(analogue precision), 87%(binary recall), 95%(binary precision)

1 Introduction

In this paper, we present research on knowledge extraction from text. The main objective of the research is to develop techniques for automatic knowledge extraction directly from plain text in electronic form, so that the extracted knowledge can be organised into a knowledge representative scheme.

The target knowledge KR scheme is used in a hyper-knowledge interaction environment called HyperTutor(Powell et al., 1999). This uses a novel and generic formalism for structuring and interrogating hypermedia-based knowledge via a natural language interface. The system engages users in a dialogue with knowledge as well as allowing them to browse. It also has pedagogic features for tutoring. It employs an augmented semantic network to represent knowledge. An authoring environment called HyperLab is used by an author to organise their knowledge into the knowledge representation structure. The authoring system is a kind of knowledge acquisition tool: it can acquire knowledge via interaction with human experts. Knowledge acquisition (KA) is a difficult and time-consuming process. It will therefore be a great benefit to automate the knowledge acquisition process so that knowledge can be automatically extract from text with minimum human involvement. HyperTutor is a generic environment, therefore generic KA techniques are required. This paper presents research into enabling the important concepts (keywords) in a domain to be automatically identified. The identification is based on seed words which are provided by a human author to define the domain.

2 Related Work

The first conceivable approach to solve the task of automatic knowledge acquisition is to fully understand the natural language text. This method, however, is beyond the capabilities of current natural language understanding (NLU) systems. The main reason for this is the complexity of natural language and the lack of appropriate linguistic theory to manage this complexity. It is difficult to build a grammar for a realistic subset of natural language(Quirk et al., 1985). In particular it is difficult to process exceptions.

Another approach to knowledge acquisition is Information Extraction (IE)(MUC, 1991; MUC, 1992; MUC, 1994; MUC, 1995). IE aims to identify instances of a particular class of event or relationship in natural language text. Relevant arguments concerning events and relationships are extracted and encoded in a format suitable for incorporation into a database (Lehnert et al., 1994). Compared with full text understanding which attempts to extract and represent all information in the text explicitly, IE is only concerned with the facts related to a specific domain that has been decided before the extraction starts. Although IE is less comprehensive than full text understanding and puts more emphasis on the facts themselves than on the relationships between the facts, it is more feasible in practice than full text understanding. Almost all IE systems use a pattern-matching method, thus the first task when developing an IE system is to construct patterns which will be used to extract information. The quality and quantity of patterns strongly influence the resulting performance. Patterns construction is usually performed manually by human experts. It is a time-consuming, knowledge-intensive and tedious task. Recently, there has been a trend in this field to attempt to construct the patterns for extraction automatically (Riloff, 1996; Soderland et al., 1995).

Machine learning is also widely used in knowledge extraction research. Most researchers who employ this method consider knowledge extraction from text as a kind of text classification. Mitchell (1997) proposed a general algorithm for learning to classify text based on a naive Bayes classifier. Detailed information about probabilistic machine learning approaches can be found in Joaxhims (1996), Lang (1995) and Lewis (1991). Information on NLP-based machine learning approaches can be found in Craven(1997; 1999) and Solderland(1998).

The approach taken here does not involve full NLU and so is potentially more tractable. However it also avoids the very domain-specific pattern-matching techniques of IE. It is a machine learning method based on artificial neural networks (ANNs). The benefits of ANNs are their abilities to generalise different information and learn from examples and most importantly, the compatibility with statistical and corpus-based NLP approaches. Our approach is novel in that although ANNs have been used in parsing (Tepper et al., 1995a; Tepper et al., 1995b), there have been no similar application of ANNs in KA.

3 Keyword Extraction

3.1 Introduction

As mentioned, the main purpose of this research is to develop a knowledge acquisition front end for HyperTutor that uses a kind of knowledge representation formalism similar to a semantic network. The ultimate aim of this research is to organise knowledge extracted into the same formalism. It represents knowledge as a network of nodes interconnected by links where the nodes denote concepts and the links denote relationships between concepts. In each node, there is text relating to the node including some derived from the link relationships. In this paper, we refer to the names of nodes as keywords and are concerned with identifying them automatically as the first stage in a complete KA process.

3.2 Outline of the approach

The approach taken is to train an ANN to differentiate between keywords and non-keywords based on an input representation of their relationships to a seed word which is defining the domain. The relationships between each potential keyword and the seed word are obtained by searching an electronic semantic lexicon. Training data consists of input patterns for keyword and non-keyword examples where the keyword/non-keyword distinction has been judged by humans. Once trained the network should be able to recognise input patterns/relationships that correspond to keywords of the original seed word. It is hoped that what the network has learnt about what signifies a keyword relationship to the original seed word will be transferable to other seed words i.e. domain independent. However this is not evaluated here, as this work evaluates the approach for one domain.

In order to test the feasibility of this approach the following steps were carried out with education as the seed word:

1. The nouns in documents relevant to the seed word domain are divided into three groups for training, testing and validation respectively. These are each judged as being keywords or non-keywords by humans. The nouns in the training set form the basis for the training data.
2. All training nouns and their relationships to seed words are identified automatically according to a universal (domain-independent) semantic lexicon. All the information for a noun is organised into a pattern that will be input to an ANN for training. The output target is 1 or 0 depending on whether the noun is a keyword of the seed words.
3. The ANN is trained.

4. The trained ANN is tested to see how well it can extract keywords from the test nouns.

Sample documents are used to mimic the situation where an author is converting a document concerning a given domain into the Hypertutor knowledge representation scheme.

3.3 WordNet: The Semantic Lexicon

The semantic lexicon used is WordNet (Fellbaum, 1998), an on-line lexical reference system. In WordNet, nouns, verbs, adjectives and adverbs are all organized into the smallest semantic unit: Synonym Set (called Synset in WordNet) which represent a single concept in English. The Synsets are interconnected by semantic relationships.

There are more than 57000 nouns in WordNet (as WordNet is updated the exact number increases). Most of them are compound nouns and few are proper nouns. They are organised into about 48800 Synsets and are represented as a kind of semantic inheritance network. All nouns belong to one or more categories in the inheritance hierarchy but only one of the 25 top-level categories. An example of this hierarchy is shown in figure 1, from 'student', the lowest level, to 'entity', the highest. Each level in the hierarchy represents a category. There are 25 top-level categories in WordNet.

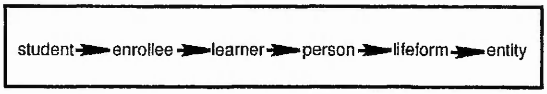


Figure 1: An example from the interitance hierarchy

There are seven semantic relationships that interconnect the noun Synsets in WordNet. They are synonym, antonym, hypernym, hyponym, meronym, holonym, coordinate. If X is a kind of Y then Y is a hypernym of X and X is a hyponym of Y. If X is a part of Y then X is a meronym of Y and Y is a holonym of X. Coordinate means words that have the same hypernym. For symmetry, we have introduced a new relationship called coordiantee which means "nouns that have the same hyponyms".

3.4 Input Patterns for the ANN

For each noun in the training document, there is an input-output pattern pair in the training data set. Each input pattern is composed of two parts. The first is the category information of the noun. This information should be useful because it is more likely for a noun within the same category as the seed words to be identified as a keyword. The twenty-five top-level categories in the inheritance hierarchy are used in this part, so there are twenty-five bits to represent category information. If the noun belongs to one of the top-level categories, the corresponding bit is set to 1, the remaining bits being set to 0.

The second part of the input pattern is more complicated. It represents the distance in WordNet between the noun and the seed words as well as the relationships between the words on the linking paths. A path from one noun to another is composed of all the nouns on the way and the relationship type between the adjacent nouns. For example, a path from "university" to "education" is shown in figure 2. (The intermediate words on the path are not represented on the input as the structural information about them in WordNet is confined to their relationships to other words.)

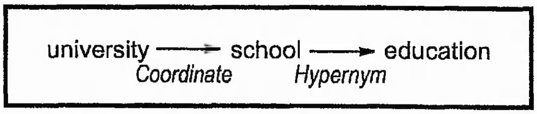


Figure 2: An example path in WordNet between 'university' and 'education'

The distance from university to education is 2. There are eight types of relationship. The second part of the input pattern contains the distance and relationship information of the shortest N paths up to maximum length of M. The criteria for choosing M and N are described later.

How are paths presented to an ANN? Suppose the maximum path length (M) is 4. A path will therefore have a length in the range 1 to 4. There are 4 fields, A to D, each representing one of the 4 path lengths. Each field contains sub-fields that allow the relationship type for each link on the path to be represented. A relationship type is represented using 8 bits. Each bit corresponds to one relationship i.e. like the classification coding only 1 bit is high at a time. The coding of 1 path with M=4 is shown in figure 3.

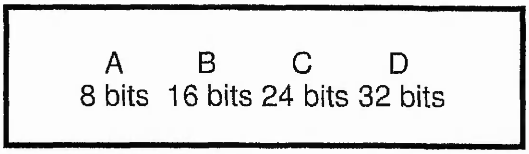


Figure 3: Bit pattern for a path

A denotes a path of length 1, B denotes a path of length 2, C a path of 3 and D, 4. If a path is of length 1, then the B, C and D fields are all set to 0. The A field is set according to the relationship i.e. the bit corresponding to the relevant relationship is set high. If the path is of length 2, then the A, C and D fields are all set to 0 and the B field is set according to the relationships in the path: the first 8 bits is used to represent the first relationship and the second 8 bits is used to represent the second relationship. The same principle applies to paths of length of 3 and 4.

For the example in figure 2, the length of the path is 2. The pattern of this path is shown in figure 4.

A	B		C	D
0...0,	00000100	00010000	0...0,	0...0
8 bits	*coordinate*	*hypernym*	24 bits	32 bits

Figure 4: Bit pattern of the path of length 2 in Figure 2

Up to N paths can be repeated, thus the total input pattern with M=4 is shown in figure 5.

category	**path1**	**path2**	**path3**	**path4**
	ABCD	ABCD	ABCD	ABCD
25 bits	80 bits	80 bits	80 bits	80 bits

Figure 5: Total input pattern of N paths with maximum length 4

The output pattern is one bit for the target which is either 1 for an example keyword or 0 for a non-keyword.

3.5 How Many and Which Paths?

Nearly all nouns have more than one path to a seed word, so how many paths is enough for training purpose and which paths should be selected? The aim is to present enough information for the network to learn the problem. This decides the choice of M and N. M should be large enough for all keywords in the training data have at least one path with a length equal to or shorter than M. If M is too small, some keywords will be presented to the ANN with no path information, which would give the network no information on which to base its selection.

Another requirement is to present enough information for there to be no contradictions in the training data. A contradiction occurs when two patterns have the same inputs and different outputs. If there are contradictions in the training data, the ANN will not be able to acquire the training data.

A contradiction may arise when two nouns belong to the same WordNet categories, have the same path to the seed word but one is classified as a keyword and the other a non-keyword. See Figure 6, where "week" and "semester" both belong to the same categories and have the same path to education. Identical path information can also be generated when the intermediate words are different between the two paths.

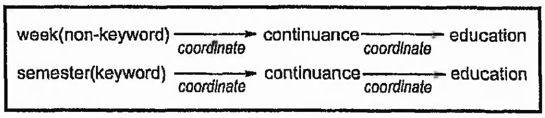


Figure 6: Total input pattern of N paths with maximum length 4

Therefore, one path for each noun is often not enough to distinguish between them. A combination of M and N is required such that there are no contradictions in the training data set. However, the M-N combination should also minimize the amount of training data. For the nouns that have more than N shortest path to choose from, the first N paths are chosen. For those that have less than N shortest paths, the path length is increased until N paths are found.

A further complication is that there is no systematic way of ordering paths on the input. Therefore, training data is generated with input patterns for all the possible ways of ordering the inputs. This aims to allow the network to recognise path features regardless of the order that they were found in when WordNet was searched, e.g. for N=3 paths, 6 training patterns containing the 6 permutations (together with the category information) are generated.

4 Experiment

4.1 Training

Preliminary experiments have been performed with the domain defined by the seed word "education". The document chosen is a research paper entitled "Mediated Learning: A New Model of Networked Instruction and Learning" (ASMLL, 1999). It is comprised of 19672 words with 707 unique nouns occurring 8334 times. 54 words were identified as keywords. The human judges based their keyword classification on considering education in the sense of "education in a formal setting".

The whole text was divided into three parts based on the criteria that ideally unique keywords should be distributed evenly in the three parts, i.e. one third in each part. The result of the division is shown in Table 1[1].

A series of tests were carried out to establish M=4 and N=5 as the optimum combination for getting rid of the contradictions. The ANN architecture used

[1] Unique keywords in the training set may also exist in the testing set and/or validation sets. Unique keywords in the testing set are those that do not occur in the training set, but may occur in the validation set. Unique keywords in the validation set are those that occur neither in training set nor in the test set.

Part	Nouns	Key-words	Unique Keywords
Whole document	707	54	54
Training set	111	19	19
Testing set	344	39	20
Validation set	687	50	15

Table 1: Result of document division

is Feed-Forward with Backpropagation. The initial weight range was set between {-0.5,0.5}, and the error threshold was set 0.2.

We used a pattern-oriented adaptive learning method based on learning errors (difference between the target and actual output) (Tepper et al., 1995a; Tepper et al., 1995b). Suppose the current learning rate and the learning error for pattern P are α and E respectively, then the new learning rate for P, α', will be:

$$\alpha' = \alpha + (1 - \alpha) * |E|; \qquad 0 < \alpha < 1$$

This method requires E to be in the range {-1,1}. The Sigmoid output satisfies the requirement. Our experiments show this is a very efficient learning method. The network using this method converges within 44 iterations while it needs more than 4110 iterations using a constant learning rate.

According to the representation scheme, for the 111 training nouns there should be 13320 (111*N! = 111*5!) training patterns. Patterns representing keywords were repeated in the training set to balance the number of keyword patterns and non-keyword patterns because without balancing the distribution of patterns in the training set is biased. The ratio of non-keywords to keywords is about 5. By duplicating all keyword patterns 5 times, the training data was balanced. The total extra patterns is 19*5!*(5-1)=9120. Thus altogether 22440 training patterns were represented to the network and the network learnt all the patterns in 44 iterations.

A series of experiments were performed to minimise the number of hidden-neurons. We used a method similar to binary search to find the minimum number of hidden-neurons. First, the number of hidden-neurons was set large enough (e.g. 40) so that the problem can be leant by the network. Then, the number was halved (20) and the network was trained again. If the network can not learn the problem with this number of hidden-neurons, the number was set to half of the sum of the two number (30). If the network can learn the problem, the lower number was half-reduced again (10 this time). By using this method, the minimum number of hidden-neurons was found to be 2.

4.2 Training Results

After trained, the network was presented with 41280 (344*5!) patterns in the test data set to see how well the network learnt the problem. We used a threshold of 0.5 to classify a tested pattern, i.e. if the output of a tested pattern is larger than 0.5, it was classified as a keyword. If the output is less than 0.5, it was classified as a non-keyword. The result of testing is shown in Table 2.

Word Type	Total	P	C
Total Nouns	344	41280	84%
Keywords	39	4680	62%
Non-Keywords	305	36600	87%
Unique Nouns	252	30240	82%
Unique Keywords	20	2400	47%
Unique Non-Keywords	232	27840	83%

P=Number Of Patterns(120 patterns per word)
C=Percentage of patterns identified correctly

Table 2: Result of testing

5 Evaluation

5.1 Introduction to Methodologies

Basic neural network theory tells us that if a problem is linear, it can be solved without the use of hidden neurons, i.e. with a single layer of connections between input neurons and output neurons. In this case, hidden neurons are required to solve the problem to any reasonable level of accuracy. We therefore know that the problem is non-linear and non-trivial.

To evaluate this novel approach, we introduced new measures based on the concept of generalisation in ANN research and recall and precision widely accepted in KA research. The most basic measure (natural generalisation) states what proportion of nouns are correctly classified (as keyword and non-keyword) in the test text. Standard binary recall and precision measures are also applied together with more sophisticated measures, developed to give a more detailed picture of performance (pure generalisation and analogue measures of recall and precision).

As stated, both binary and analogue recall and precision metrics are used. In traditional information retrieval, recall and precision are binary metrics. The analogue nature of the ANN output and the desire to have a single overall performance measure that is unbiased according to the ratio of keywords to non keywords, has led to the development of novel analogue measures of recall and precision.

Generalisation is appropriate to evaluate ANN results, however the linguistic problem domain suggests two types of generalisation, pure and natural. Pure generalisation evaluates the effectiveness of the ANN learning of the problem in terms of its ability to classify unseen patterns and is commonly used in ANN research. Natural generalisation evaluates the effectiveness in terms of the classification of unseen text. This is more appropriate for evaluating the overall ability of the trained network in performing the text processing task.

5.2 Generalisation: Natural and Pure

Generalisation refers to how well a network performs with new data sets after training. The ability to generalise is the main reason that ANNs attract researchers. Generalisation refers to the ability to learn not only by memory but more importantly, by induction. Therefore generalisation forms the basis of the evaluation of ANNs.

Definition	Symbol
Number of keywords patterns in testing data	N_{kw}
Number of non-keywords patterns in testing data	N_{nkw}
Number of unique keywords patterns in testing data	N_{ukw}
Number of unique non-keywords patterns in testing data	N_{unkw}
Number of patterns identified as keyword patterns	N_{ikw}
Number of patterns identified as non-keyword patterns	N_{inkw}
Number of patterns correctly identified as keyword patterns	N_{ickw}
Number of patterns correctly identified as non-keyword patterns	N_{icnkw}
Number of unique patterns correctly identified as keyword patterns	N_{icukw}
Number of unique patterns correctly identified as non-keyword patterns	N_{icunkw}

Table 3: Definitions of symbols

As previously mentioned, two types, Natural and Pure, were defined. Natural generalisation (NG) is the percentage of nouns in the testing data that are correctly categorised as keywords or non-keywords. This can be evaluated for the total test set or evaluated separately for keywords and non-keywords. Therefore (refer to table 3 for the symbols used),

$$NG_{total} = \frac{N_{ickw} + N_{icnkw}}{N_{kw} + N_{nkw}}; \quad (1)$$

$$NG_{kw} = \frac{N_{ickw}}{N_{kw}}; \quad (2)$$

$$NG_{nkw} = \frac{N_{icnkw}}{N_{nkw}}; \quad (3)$$

This is indicative of the overall performance on unseen text, but in terms of neural network learning may include data that is repeated from the training set. This means that a component of natural generalisation may involve memorisation. NG alone is therefore not sufficient to fully evaluate the learning of an ANN. Let us consider an extreme situation: suppose all words in the test set had also occurred in the training set. Because the network can memorise all the patterns in the training data set (provided there are enough hidden neurons in the network), then all the patterns in the testing set will be identified correctly. Using NG to evaluate the performance of the network, could result in a very high score (1.0). But this says nothing about how much knowledge of new examples has been derived from the training examples. Thus the performance when the trained network is applied to new text is unknown. Pure generalisation (PG) was introduced to measure the amount of induced knowledge. PG is the percentage of nouns with previously unseen input patterns in the testing data that are correctly classified. Again it can be applied to the total result and to keywords and non-keywords separately. It can be described as,

$$PG_{total} = \frac{N_{icukw} + N_{icunkw}}{N_{ukw} + N_{unkw}}; \quad (4)$$

$$PG_{kw} = \frac{N_{icukw}}{N_{ukw}}; \quad (5)$$

$$PG_{nkw} = \frac{N_{icunkw}}{N_{unkw}}; \quad (6)$$

5.3 Recall and Precision: Binary and Analogue

Another evaluation method widely used in KA research is recall and precision (Lehnert et al., 1994). Recall measures the ratio of correct information ($N_{correct}$) extracted from the text against all the information (N_{all}) available in the text. Precision measures the ratio of correct information that was extracted against all the information extracted ($N_{extracted}$). Thus,

$$recall = \frac{N_{correct}}{N_{all}}; \quad (7)$$

$$precision = \frac{N_{correct}}{N_{extracted}}; \quad (8)$$

These are applicable to keywords and non-keywords separately and defined for keywords as

$$recall_{kw} = \frac{N_{ickw}}{N_{kw}}; \quad (9)$$

$$precision_{kw} = \frac{N_{ickw}}{N_{ikw}}; \quad (10)$$

and for non-keywords as

$$recall_{nkw} = \frac{N_{icnkw}}{N_{nkw}}; \qquad (11)$$

$$precision_{nkw} = \frac{N_{icnkw}}{N_{inkw}}; \qquad (12)$$

These measures are commonly used in knowledge extraction systems. However, they have two limitations. Firstly, they do not immediately provide an overall performance measure because they take no account of the ratio of keywords to non-keywords. Secondly, they do not accommodate the analogue nature of the ANN response which provides extra information about the level of confidence of the decisions. Therefore, we adapt the basic formulae (7 and 8) for recall and precision in the following way.

Suppose the target and actual output of a pattern P in the testing data set, TS, are T_p and A_p respectively, where T_p is either 1 or 0 and $0 <= A_p <= 1$.

Correctness is defined as decreasing in proportion to the output error, but also increasing in proportion to the deviation from 0.5, since that is the point of zero correctness. This gives a correctness scale of {0,1}. Thus

$$N_{correct_p} = 2|A_p - 0.5| * (1 - |T_p - A_p|)$$

Therefore $N_{correct}$ for all patterns is:

$$N_{correct} = 2 \sum_{p \in TS} |A_p - 0.5| * (1 - |T_p - A_p|)$$

Extraction is defined in terms of the decisiveness or responseiveness of the network i.e. its deviation from a natural response. Since Ap is in the range {0,1}, an output of 0.5 means the network does not make a response to P. So

$$N_{extracted_p} = 2 * |A_p - 0.5|$$

A coeffcient of 2 puts $N_{extracted_p}$ in the range of 0 and 1. Therefore, for all patterns $N_{extracted}$ is

$$N_{extracted} = 2 \sum_{p \in TS} (|A_p - 0.5|)$$

The number of patterns is the sum of number of keyword patterns and the number of non-keywords patterns, thus

$$N_{all} = N_{ikw} + N_{inkw}$$

Thus, we get the formulae of recall and precision suitable for an ANN-based approach:

$$R = \frac{2 \sum_{p \in TS} |A_p - 0.5| * (1 - |T_p - A_p|)}{N_{ikw} + N_{inkw}} \qquad (13)$$

$$P = \frac{\sum_{p \in TS} |A_p - 0.5| * (1 - |T_p - A_p|)}{\sum_{p \in TS}(|A_p - 0.5|)} \qquad (14)$$

5.4 Results for ANN System

Applying the above measures to our experimental results, we get the results in tables 4 ,5 and 6.

Data Set	NG	PG
Total	0.84	0.82
Keywords	0.62	0.47
Non Keywords	0.87	0.83

Table 4: Natural and Pure Generalisation

Data Set	Recall	Precision
Total	n/a	n/a
Keywords	0.62	0.38
Non Keywords	0.87	0.95

Table 5: Binary Recall and Precision

Data Set	Recall	Precision
Total	0.81	0.86
Keywords	0.59	0.63
Non Keywords	0.84	0.88

Table 6: Analogue Recall and Precision

5.5 Baseline Comparison

In order to evaluate the contribution of the ANN to the overall solution which combines the information from WordNet with the ANN processing, a simple method using just WordNet is used to give baseline results. Instead of evaluating the relationships along the paths between a word and the seed word, a simple decision rule is applied, i.e. that any word within N steps of the seed word is closely related to it and is therefore classified as a key word. This gives the results in Table 7 for comparison with the ANN-based method.

6 Conclusions

We have shown that concepts can be automatically extracted from text using an ANN. Results in the education domain show good natural and pure generalisation for non-keywords at 84% and 82% respectively and reasonable generalisation for keywords (62% for natural and 47% for pure). Under the standard measures used in the information extraction

Measure	No of Steps(N)				ANN
	1	2	3	4	
R_{kw}	0.31	0.39	0.49	0.64	0.62
P_{kw}	0.27	0.17	0.13	0.11	0.38
R_{nkw}	0.89	0.76	0.58	0.35	0.87
P_{nkw}	0.91	0.91	0.90	0.88	0.95
NG_{total}	0.83	0.72	0.57	0.38	0.84
NG_{kw}	0.31	0.39	0.49	0.64	0.62
NG_{nkw}	0.89	0.76	0.58	0.35	0.87

R=recall P=Precision

Table 7: Baseline and ANN Results

community, i.e. recall and precision, our results are encouraging.

The results in Table 7 show that the task of extracting keywords is complex. The simple 'distance from seed word' rule is inadequate: it fails to extract most keywords until the step size is so large that a high proportion of non-keywords are mistaken for keywords. The ANN approach is a significant improvement on this situation. To a precision of one decimal point, the results in table 7 show the ANN to equal or improve on every metric for all step sizes. On average, across the various metrics, the ANN is a significant improvement, irrespective of step size.

Several other works (Andrade and Valencia, 1997; Otha et al., 1997; Weeber and Vos, 1998) also extract keywords from text. All of them are based on information and probability theories aimed at providing keyword lists and/or glossaries for information retrieval. Our approach is based on the semantic relationships between words. It is more appropriate for our final objective, i.e. to construct a knowledge base. One contribution of our work is the novel approach to using ANNs in knowledge acquisition, including the definition of an evaluation methodology which involves new measures of performance. These new measures give a detailed picture of the strengths and weaknesses of the method's performance, and allow a clear comparison to be made with other methods.

Our approach does not require tagging, annotating or a domain-dependent lexicon. The only human involvement needed is identifying a seed word to define the domain and keywords for training purposes. The time-consuming and tedious process of preparing domain-dependent information for knowledge acquisition in a new domain, which is the major knowledge engineering bottleneck, is avoided. The generality of the approach across domains has yet to be evaluated. Future work will investigate this by applying a single network to multiple domains.

Although WordNet is a valuable online lexicon, it has shown some limitations. Firstly, there is no stemming information in it. Secondly, some of the relationships between words are not completely realised. For example, information on meronyms and holonyms is sparse. Thirdly, WordNet does not attempt to capture general or commonsense knowledge in the sense that some knowledge based systems do, e.g. CYC (Guha and Leant, 1990; Leant, 1990). However, we have not fully explored the potential of WordNet. The 25 top-level categories used to train the network could be extended one level down the inheritance hierarchy. CYC, the largest knowledge base in the world which contains commonsense knowledge, is a possible alternative source of the information we need.

The work presented here is the initial results of the first stage in the complete knowledge acquisition process. We are currently investigating using stemming information to improve pure generalisation of keywords. Nouns that have the same stem as a keyword will be treated as keywords. Word sense disambiguation is also under investigation. In WordNet, "education" has six meanings, but only two of them are relevant to the domain definition used in our experiment. Some nouns may have paths to "education" but not to the sense that we are concerned with. These paths are spurious. They may be removed by word sense disambiguation. Future work includes finding the definitions of concepts and the semantic relationships between concepts in order to construct the information in the final knowledge base.

References

M.A. Andrade and Valencia. 1997. A automatic annotation for biological sequences by extraction of keywords from medline abstracts. In *proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, pages 25–32, Halkidiki, Greece. AAAI Press.

ASMLL. 1999. Mediated learning: A new model of networked instruction and learning. World Wide Web, url-http://www.academic.com/library/articles/ mllibrary.html, May. Academic Systems Mediated Learning Library.

M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. 1997. Learning to extract symbolic knowledge from the world wide web. In *Proceedings of 15th national conference on Artificial Intelligence (AAAI-98)*.

M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. 1999. Learning to construct knowledge base from the world wide web. *Artificial Intelligence.*

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database.* MIT Press.

R.V. Guha and D.B. Leant. 1990. Cyc: a midterm report. *AI Magazine*, 11(3).

T. Joachims. 1996. A probabilistic analysis of rocchio algorithm with tfidf for text categorization. Computer Science Technical Report CMU-CS-96-118, Carnegie Mellon University.

K. Lang. 1995. Newsweeder: Learning to filter netnews. In *Proceedings of the 12th International Conference on Machine Learning, In Prieditis and Russel (Eds.)*, pages 331–339, San Francisco. Morgan Kaufmann Publishers.

D.B. Leant. 1990. *Building Large Knowledge-based Systems: Respresentation and Interface in the CYC Project.* Addison-Wesley, Reading, MA.

W. Lehnert, C. Cardie, D.Fisher, J. MCCarthy, E. Riloff, and S. Soderland. 1994. Evaluating an information extraction system. *Journal of Integrated Computer-Aided Engineering*, 1(6).

D. Lewis. 1991. *Representation and learning in informal retrieval (COINS Technical Report 91-93).* Ph.D. thesis, Department of Computer and Information Science, University of Massachusetts.

T. Mitchell. 1997. *Machine Learning.* McGraw-Hill International Editions.

MUC. 1991. *Proceedings of the Third Message Understanding Conference (MUC-3).* Morgan Kaufmann, May.

MUC. 1992. *Proceedings of the Third Message Understanding Conference (MUC-4).* Morgan Kaufmann, June.

MUC. 1994. *Proceedings of the Third Message Understanding Conference (MUC-5).* Baltimore, MD, August.

MUC. 1995. *Proceedings of the Third Message Understanding Conference (MUC-6).* Columbia, MD, November.

Y. Otha, Y. Yamamoto, T. Okazaki, I. Uchiyama, and T. Takagi. 1997. Automatic construction of knowledge base from biological papers. In *proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, pages 218–225. Halkidiki, Greece: AAAI Press.

H. Powell, D. Palmer-Brown, J. Downs, G. Long, and M. Edwards. 1999. Artificial intelligence for hypermedia access: Issues in knowledge representation and natural language processing. *Submitted to Knowledge and Information System Journal.*

R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik. 1985. *A comprehensive Grammar of the English Langiage.* Longman.

E. Riloff. 1996. An empirical study of automated dictionary construction for information extraction in three domains. *AI journal*, 85 August.

S. Soderland, D. Fisher, J. Aseltine, and W. Lenhert. 1995. Crystal: Inducing a conceptual dictionary. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence.*

S. Soderland. 1998. Learning to extract text-based information from the world wide web. In *Proceedings of third International conference on Knowledge Discovery and Data Mining (AAAI-98).*

J. Tepper, H. Powell, and D. Palmer-Brown. 1995a. Ambiguity resolution in a connectionist parser. In *The Cognitive Science of Natural Language Processing. Eidtor A I C Monaghan, Natural Language Group*, July 5-7.

J. Tepper, H. Powell, and D. Palmer-Brown. 1995b. Integrating symbolic and subsymbolic architecture for parsing arithmetic expressions and natural language sentences. In Bert Kappen and Stan Gielen, editors, *Proceeding of 3rd SNN Neural Network Symposium*, pages 81–84. Nijmegen, Sept 1995, ISBN 3-540-19992-6.

M. Weeber and R. Vos. 1998. Extracting expert medical knowledge from texts. In Working Notes of the Intelligent Data Analysis in Medicine and Pharmacology Workshop,23-28.