A Hybrid NLP & Semantic Knowledgebase approach for the Intelligent Exploration of Arabic Documents

Hussein Khalil
School of Science and Technology

A thesis submitted in partial fulfilment of the requirements of Nottingham Trent University for the degree of

Doctor of Philosophy

AUGUST 2017

Abstract

In the contemporary era, a colossal amount of information is published daily on the Web in the form of articles, documents, reviews, blogs and social media posts. As most of this data is available in the form of unstructured documents, it makes it challenging and time-consuming to extract non-trivial, previously unknown, and potentially useful knowledge from the published documents. Hence, extracting useful knowledge from unstructured text, i.e., Information Extraction, is becoming an increasingly significant aspect of knowledge discovery.

This work focuses on Information Extraction form Arabic unstructured text, which is an especially challenging task as Arabic is a highly inflectional and derivational language. The problem is compounded by the lack of mature tools and advanced research in Arabic Natural Language Processing (NLP) in comparison to European languages for instance.

The principal objective of this research work is presenting a comprehensive methodology for integrating domain knowledge with Natural Language Processing techniques that were proven effective in solving most classification problems in order to improve the Information extraction process form online unstructured data. The importance of NLP tools lies in that they play a key role in allowing semantic concept tagging of unstructured text, and so realize the Semantic Web. This work presents a novel rule-based approach that uses linguistic grammar-based techniques to extract Arabic composite names from Arabic text. Our approach uniquely exploits the genitive Arabic grammar rules; in particular, the rules regarding the identification of definite nouns (معرفة) and indefinite nouns (نكرة) to support the process of extracting composite names. Furthermore, this approach does not place any constraints on the length of the Arabic composite name. The results of our experiments show that there are improvement in recognizing Arabic composite names entity in the Arabic language text.

Our research also contributes a novel, knowledge-based approach to relation extraction from unstructured Arabic text, which is based on the principles of Functional Discourse Grammar (FDG). We further improve the approach by integrating it with Machine Learning relation classification, resulting in a hybrid relation extraction algorithm that can handle especially complex Arabic sentence structures. The accuracy of our relation classification efforts was extensively evaluated by means of experimental evaluation that evidenced the accuracy of

the FDG relation extraction approach and the improvement gained by the Machine Learning integration.

The essential NLP algorithms of entity recognition and relation extraction were deployed in a Semantic Knowledge-base that was built from the outset to model the knowledge of the problem domain. The semantic modelling of the knowledgebase aided improving the accuracy of the NLP algorithms by leveraging relevant domain knowledge published in Open Linked Datasets. Moreover, the extracted information was semantically tagged and inserted into the Semantic Knowledge-base, which facilitated building advanced rules to infer new interesting information from the extracted knowledge as well as utilising advanced query mechanisms for intelligently exploring the mined problem domain knowledge.

Copyright Statement

Publications

The papers which resulted from this research and published in international conferences Are Listed below:

- H. Khalil and T. Osman, "Challenges in information retrieval from unstructured Arabic data," in Proceedings of the 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, 2014, pp. 456-461.

- Hussein Khalil ,Taha Osman , Paul Bowden , Mohammed Miltan, "Extracting Arabic composite name using a knowledge driven approach," in 17th International Conference on Intelligent Text Processing and Computational Linguistic, April 3–9, 2016 • Konya, Turkey, 2016, .

Declaration of Authorship

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Acknowledgements

I am grateful to many, many people for help, advice, ideas, encouragement and practical support whilst I was working towards producing this thesis. I would like to express my sincere gratitude to my director of studies Dr Taha Osman, whom have spent an ample part of his life dissipating knowledge to students. His wide knowledge and logic way thinking has been of great value to me. His understanding, encouragement, motivation and personal guidance have provided an excellent basis.

I thankfully acknowledge the contributions of Dr Paul Bowden for supporting me. I am deeply beholden to my dear father who cannot see these moments in my life. I am also thankful to my father for teaching me the value of knowledge and education. There is no words in any natural language could be sufficient to thank my great mother for all she had done for me. So this thesis is dedicated to them. I am especially thankful to my wife and children for their unconditional love, for putting up with the countless hours I spent in my thesis work, and for being there when I needed here.

My heart-felt thanks to Dr Mohamed Emlitan for the private, hands-on tutorials in using Advanced Arabic grammar. I am grateful to my friends Abduladem Aljamel and Al zain Al zain, they were supported and encouraged me. To my many friends and family, you should know that your support and encouragement was worth more than I can express on thesis.

List of Acronyms

NLP	Natural Language Processing
SW	Semantic Web
SWT	Semantic Web Technologies
IE	Information Extraction
RDF	Resource Description Framework
GDP	Gross Domestic Product
LOD	Linked Open Data
KB	Knowledge-Base
POS	Part Of Speech
IR	Information Retrieval
OWL	Web Ontology Language
SWRL	Semantic Web Rule Language
XML	eXtensible Mark-up Language
FDG	Function Discourse Grammar
AENER	Arabic Economic Named Entity Recognition
FKBR	Financial Knowledge-base Recommender
AGR	Arabic Genitive Rule

Contents

List of Figure

List of table

Chapter 1

1. Introduction

1.1. Motivation

The volume of information made publicly available in the web has been growing quickly with increase in the number of internet users. According to the Internet World Stats, the number of Internet users exceeded 3,631,124,813 up until the writing of this article. This information covers a plethora of domains in business, education, entertainment, politics, sports, etc., which presents an opportunity to exploit this information to inform a variety of applications, such as recommender systems, sentiment analysis and advanced data exploration engines. However, as most of the information published in the web is in unstructured format, this necessitates the use of Information Extraction (IE) and retrieval techniques to mine relevant facts and events from the unstructured text. This technique proposes innovative methods for querying, organising, and analysing data by relying on the semantics of structured data and the richness of unstructured data [1]. With the huge volume of text on the web emanating from different public resources, such as media, blogs, books, journal, magazine articles, expert opinions and personal experiences, it is becoming increasingly difficult for a human to analyse the text contents manually.

Most of the published Arabic information is in unstructured (raw) format; in addition, the Arabic text poses many challenges. Arabic language is highly inflectional and derivational language, which makes text mining a complex task. The development of Arabic information extraction applications is a quite recent event. For this opportunity to be exploited, one needs an advanced information extraction technique.

This work focused on IE from web documents that are written in Arabic, which is considered extremely challenging as the Arabic text poses a number of linguistic issues that have influenced the development of language processing tools, such as short

vowels, absence of capital letters, complex morphology. Moreover, it is orthographic with diacritics, and is highly inflectional and derivational, etc. According to A. Farghaly and K. Shaalan [2], "Arabic is a highly-inflected language which has a rich and complex morphological system. The Arabic word takes more than one word form to represent it. That includes root, prefixes, suffixes and clitics. Arabic discretisation, defined as the full representation of short vowels, is considered one of the major challenges to most Arabic NLP tasks". Therefore, one believes that IE efforts can benefit from assistance of problem domain knowledge. This can help IE techniques to more accurately extract useful knowledge from unstructured documents; i.e. the understanding of the domains key concepts and their interrelations.

To exploit domain knowledge in proving IE techniques, it is crucial to comprehensively model this knowledge in an automated manner. The knowledge representation is the technique utilised to encode knowledge in an intelligent system's knowledge-base. The main purpose of knowledge representation is to represent knowledge in computer-extractable form, such as helping intelligent system to perform well. In order to carry out this task, one of the most popular tools for modelling is the Semantic Web (SW) technologies because Semantic Web information is stored and represented in the RDF, which is an official and expressive method to define the semantics of concepts and the relationships between them.

The Semantic Web technology is ideally placed to model knowledge in a machine comprehensive way to be exploited in improving IE techniques because its nature allows to tag the concepts, tag the relations and infer a new set of knowledge.

1.2. **Problem statement**

The common problem for web data extraction tools is the structure of data on the website. When data is written as a plain text without utilising the classifiers, it is very

complex to identify what those text sections represent. One can possibly classify the challenges that face handling the web data as 1) quantity, 2) complex, and 3) diverse.

The web data has been published in different languages. One of these unstructured data is Arabic language data. It should be pointed out that most of the published Arabic information is in an unstructured (raw) format. In addition, the Arabic text poses many challenges, such as the highly inflectional and derivational language which makes text mining a complex task. This complexity causes limitation in the current Arabic IE research efforts in terms of handling the huge growth of Arabic data on the web. Therefore, there is an urgent need for technologies and tools to address the relevant information.

The principal objective of this research endeavour is to develop a framework that will intelligently automate the management of an Arabic unstructured data on the web by employing a Semantic Web technique to improve the intelligent exploration of unstructured documents written in Arabic, especially in the financial domain. The economic domain will be used as a case study to help implement and test our problem statement.

1.3. **Economic domain as a use case**

One of the major sectors to take advantage and gain access to high technologies tools and their interconnection over the last few years is the financial sector. It should be pointed out that the finance and economy domain represents a conceptually rich area, with information characterised by the sheer volume, complexity, and value of the business product. On a day-to-day basis, people produce a large amount of valuable information. In addition, it contains a lot of named entities, especially the Arabic composite names. practically in the context of complex sentences. Furthermore, it is a new domain for Arabic IE research as other researchers have worked on other domains. However, the task of its processing is a painstaking and time-consuming one. There

should be effective filtering, search, and browsing measures by information consumers when dealing with material most pertinent to their business profile. They should also go through them effectively [3].

In recent time, it has been made possible for traders and investors to gain access to a vast quantity of information, as provided by the various agencies, companies, governmental departments, historical archives and private institutions that collate, process and organise diverse kinds of data. Using the internet, it has also been possible for users to liaise in real-time with news providers, which has immensely narrowed the gap between the actual occurrence of news and its realisation into a final product by the operator.

Economic information may be classified as the information about the indicator (index) that is the typical form of expression of quantitative data for economic information, stock market information that provides real-time, objective market information and industry sources. The information extraction systems using these types of information on the web as the resources enable the extraction of relevant information to be fed into models for analysis of financial and operational risk and other business intelligence applications such as decision-making systems and recommender systems [3].

In Arabic economic domain, there are a number of Arabic a news has published their data on the web.

1.4. **Original contributions to knowledge**

The work described in this thesis has provided the following contributions to the knowledge

Contribution 1: The main contribution is a novel framework that presents a comprehensive methodology that utilises the huge Arabic data on the web in order to support intelligent exploration of semantic knowledge-base (see section 4.2).

Contribution 2: Building A Modern Standard Arabic Corpus especially for the economic domain (see section 4.3).

Contribution 3: Composing several Arabic NE gazetteers related to the economic domain collected by different Arabic resources (see section 5.4.3).

Contribution 4: Design and implementation of a set of syntactical rules and patterns by using JAPE rules to extract and classify NEs from Arabic economic documents (see section 5.4.4).

Contribution 5: A new algorithm for recognising Arabic composite named entities that uses advanced Arabic genitive rules (Definite Nouns "الاسم المعرفة" and Indefinite Nouns "الاسم النكرة") (see chapter 6).

Contribution 6: Adopting the principles of Functional Discourse Grammar, an advanced version of Functional grammar, as the basis for building a novel approach to relation extraction from Arabic natural texts (see section 7.3).

Contribution 7: The development of a novel Hybrid Approach combining Machine Learning and a Rule-Based approach for extraction Arabic high order relations (see section 7.4).

1.5. **Motivation scenario**

In recent times, the number of webpages has seen a dramatic increase and is continuing to do so on an unprecedented scale. Several kinds of information have been published on the internet in different languages, including the Arabic language. One of the major types of information on the Web relates to economy. There is extensive economic information on different topics, including information related to stock market and shares, commodities such as gold and oil, as well as other information related to economic indicators, such as GDP and inflation figures. Most of this information has been published on separate pages, allowing users to navigate through several pages in search for useful knowledge about the economic factors or aspects of a specific country or specific item. Regarding Arabic information, the amount of data has been on the increase in a plethora of topics, including economics as in the www.fxnewstoday.ae website, and investment and trading-secrets websites. These websites are pertinent to the economic domain, which will be the focus of this study. It should be mentioned that most of this information is represented in unstructured text, tabular and chart formats.

In the present digital space, many internet users seek information from the web, which is an increasingly prospective space for users to gain important information and ideas that can in turn help them take the correct decisions. Due to the huge influx of data every day, the need for large data analytics has become essential. Additionally, the use of an open source framework has become imperative in order to increase the effectiveness of the recommendation system in different domains in spite of the popularity of financial intelligent systems as a robust solution for supporting the users to take the right decision.

In this study, a framework for financial intelligent system is proposed. The main target of this framework is to automatically collect different types of texts from several Arabic economic online news websites and analyse them to extract useful knowledge, as well as populate this data into the knowledge-base. An intelligent technology is employed on that knowledge-base to generate new information, which can then be used to serve users with different economic queries depending on its relevance to the user's situation.

In this section, the scenarios that illustrate how the IE application software can mediate between unstructured texts and structured information of different types of users has been illustrated in the Figure 1. The framework uses the huge unstructured financial data on the web in order to improve the intelligent exploration by collecting the information from different Arabic online news covering several economic areas, such as the information about the country, economic indicators, stock market information, information about products and currency of the country. This data will be examined in order to extract useful information, with the Semantic Web technology being used to represent this information into ontology. Several economic rules that use the existing information will be adopted in the ontology to describe the relationship between the economic information in the ontology in order to infer new knowledge that will give the users an overview about the state of the country and the state of the stock market.

The framework aims to support several types of users, such as investors, analysts and journalists who may request several queries.

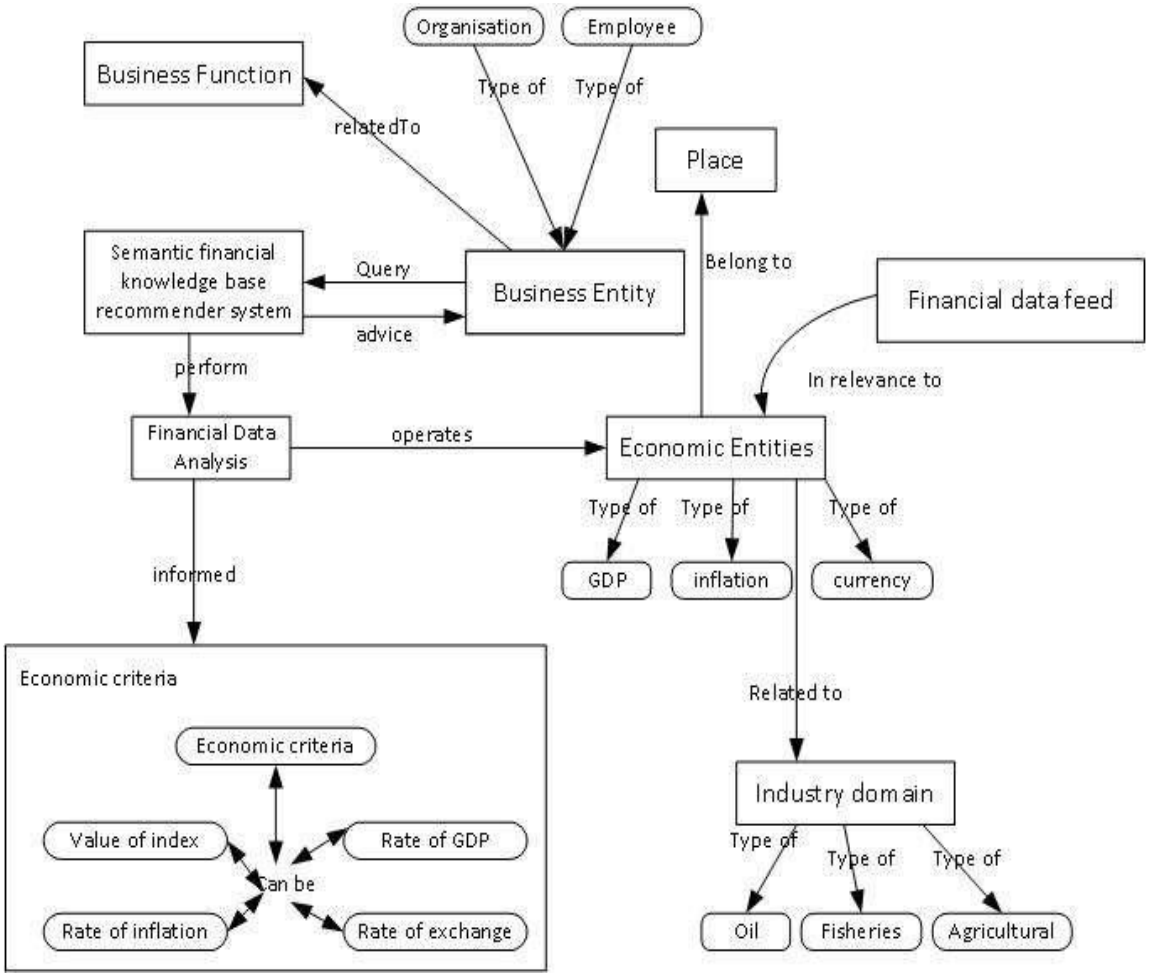


Figure 1: The motivation scenario of financial intelligence system

1.6. **Research Questions**

To sustain our main thesis, we articulate our argumentation around seven central research questions which we have attempted to answer:

Question 1: How can Arabic grammar rules be exploited to improve the recognition of Arabic composite names?

Question 2: Can the ontology-based approach be used to improve the state of art of Arabic relation extraction?

Question 3: Can exploitation of the structure of Arabic grammar be useful in improving the current state of the art relation extraction?

Question 4: If a knowledge-base approach if adopted for IE, can classification technique that are based on ML play a positive role in the IE process?

Question 5: Given the complex structure of the Arabic, do the linguistic features impact on the relation classification in the Arabic sentence structure?

Question 6: Can the adoption of Semantic Web standards for modelling of the domain knowledge improve the information extraction processes and subsequent intelligently exploring the extracted information?

Question 7: How can the gap between the human language and formalised knowledge be bridged to improve in the intelligent exploration of the unstructured document?

1.7. **The aim of study and objectives**

This study aims to build a comprehensive framework for an intelligent financial system in order to exploit the huge volume of Arabic data on the web to improve the intelligent exploration task. To achieve our aim, we will apply the following objectives.

1 Making sure that the retrieved Arabic documents are appropriate for the modern standard Arabic written styles and specific in the economic domain.

2 Modelling the domain knowledge in order to make the domain understandable by machines by constructing an ontology that is covering key concepts and events of the domain knowledge.

3 Using the NLP tools to extract the useful information from the Arabic unstructured data on the web in order to inject this information into the ontology

4 Utilizing the Linked Open Data to enrich knowledge-base.

5 Employing the Semantic Web technology in order to generate new knowledge from the existing information in the ontology in order to improve the intelligent exploration task.

6 Implementing an application prototype that exploits the semantic Knowledge-Base to deliver recommender system for user.

7 Evaluating the proposed framework by applying a specific use-case scenario.

1.8. **Research methodology**

The purpose of this section is to highlight the research methodology that is followed in this study. The main motivation underpinning this study is to go beyond a mere word-level analysis of the Arabic text in the web and present new financial intelligence framework that allows for a more efficient passage from unstructured textual information to structured machine-possible data in order to improve the intelligent exploration in a specific domain. In this section, the research methodology of this study has been described in following steps:

1. Literature Survey

- The first step in this methodology is the literature survey, which is the most important stage in the project because it aims to ensure the originality of contribution and obviate any recurrence of work in an existing area. In addition, it aids in illustrating the motivation and in establishing the roadmap for the research as well as in discovering and determining the approach used. Moreover, it will help in terms of studying and identifying the most relevant techniques, including information extraction approaches, ontology-based approaches and semantic based approaches, etc.
- The literature survey of all the relevant fields was an often-repeated process throughout the course of the PhD as it was taken to be a fundamental input

parameter to the requirement analysis, tuning and improvement phase. This is quite important due to the rapid progress in this area of research.

2. **Requirement analysis**
    - The main aims of the early phase of this work is to build a model architecture that is appropriate for fulfilling the objectives of this research, which includes information extraction tasks and the Semantic Web techniques.
    - In order to achieve the main objectives in giving an adequate answer(s) for research motivation and questions, specifications methodologies have been identified and tools considered during the course of the research have been thoroughly analysed, examined and refined.
    - Because of a massive Arabic data on the web is used in this research to extract useful information from online unstructured data, we investigated several NLP techniques and the Semantic Web technologies to be applied to improve the information extraction process and the intelligent explorations.
3. **Implementation of data-sets and application**

The datasets have been collected from several Arabic news resources that are related our problem domain. There are different methods have used to collection datasets such as RSS (Rich Site Summary) and Manual Data Collection. The system is developed by adapting the required tools and techniques to realise and implement the proposed framework.

4. **Experimental Evaluation**

The framework has been evaluated by applying a specific use-case scenario. A simple prototype was built to demonstrate how the Arabic semantic knowledge-base can be intelligently explored. The prototype allowed us to review all the stages of the framework and to ensure that the framework can answer the research questions for this research, as well as providing an excellent opportunity to evaluate all the framework phases by testing the failure processes and effects analysis

1.9. **Research tasks**

The main tasks of this research are summarised as follows:

1- To construct a domain specific semantic knowledge-base by:
   i. Creating a taxonomy (classification) ontology that describes the key concepts and relations for a specific domain. It will be utilised as the use-base for intelligent information retrieval.
   ii. Using the ontology to build the skeleton of a semantic knowledge-base (KB) and inform the information gazetteers of the corresponding NLP engine.

2- To collect web published corpora that is represented at the diverse information published about selected domain.

3- Transform unstructured text in the corpora into structured data that can be used to enrich the KB. This involves:
   i. Using linguistic processing to process unstructured text content, extract the terms that are discussed within the text, and develop several rules to extract the Arabic named entities.
   ii. Organising captured terms into a semantic graph using their class types.
   iii. Extracting new relations for terms that are not mentioned in text to improve the KB and to also disambiguate some words mentioned in the text. This can have done by utilising structured data published in the Linked Open Data (LOD) dataset, such as DBpedia.
   iv. Populating the knowledge-base with classified entities and extracted relations.

4- To deploy semantic intelligence to further infer new knowledge from structured information in the KB. This is done using first order logic to build sophisticated object relations between the semantic graph entities and more explicitly by devising semantic rules that infer new knowledge from the data by means of auto-classification and using necessary and sufficient conditions.

5- To implement a prototype that utilises the semantic KB to deliver an intelligent document analysis system and evaluate the contribution of the hybrid approach to information extraction.

1.10. **Thesis Structure**

After touching on the objectives of this research, the structure of this thesis will be described briefly in the content of each chapter.

Chapter 2 provides an account of the basic characteristics of the Arabic language and challenges of morphology, especially those that effected on the information extraction tasks.

Chapter 3 sheds some light on the current work in information extraction from unstructured text by mentioning several Arabic information efforts.

Chapter 4 highlights the main steps conducted in this research toward explaining the research methodology and giving an overview of the ontology building.

Chapter 5 gives an overview of the NER task, covering previous work and spotlighting the main technique and methods used to identify the important named entities.

Chapter 6 explains a novel approach for ANER by using Arabic language genitives rules in order to solve the problem of Arabic composite names. It includes an evaluation and discussion of the experimental results.

Chapter 7 presents a new algorithm by using a rule-based approach for extraction the semantic relations from unstructured text by using Function Discourse Grammar. It includes an overview about functional discourse grammar, the algorithm for relation extraction and an evaluation and discussion of the experimental results.

Chapter 8 explains the building of Arabic ontology semantic knowledge-base steps. It includes an overview about Semantic Web technology, a review of several Arabic

Semantic Web efforts, a population of a semantic knowledge-base, and an intelligent interrogation of the semantic knowledge base. And implement a prototype in order to evaluate the contribution of the hybrid approach to information extraction.

Chapter 2

2. Arabic Language processing

2.1. **Introduction to the Arabic language**

The origins of the Arabic language go back to pre-Islamic Arabia, where the tribes spoke local Arabic dialects. Arabic is the native language of nearly 500 million people located in 23 different countries, and is the largest member of the Semitic language family. It is also the language of the Holy Quran, the main religious text of over 1.6 billion Muslims worldwide. The other native languages that are also spoken in the middle East include Berber, Kurdish and Mahri.

The Arabic language can be divided into different categories of Arabic: Classical Arabic (CA) (العربية الفصحى), Modern Standard Arabic (MSA) (العربية الحديثة) and the Arabic Dialects (AD) (العربية العامية).

The CA originates from the Arabic classic language which is the language of the Qur'an and other early Islamic literature. The MSA is the language that is used in the writing of all Arabic books, newspapers, official and business-related documents, therefore, Modern Standard Arabic is the language of literature and the media. There is an extensive amount of information that has been published on the web which is written by MSA.

The AD varies regionally from one Arabic speaking country to another. It is considered has been an important and challenging problem for Arabic language processing, especially for social media text analysis and machine translation. There are several Arabic social media texts have mixed forms and several variations especially between AD and MSA [4].

With over 200,000 Arabic websites on the Internet[4], most of the published Arabic information is in an unstructured (raw) format. In addition, the Arabic text poses many

challenges that have influenced the development of language processing tools, such as short vowels, absence of capital letters, complex morphology; moreover, it is orthographic with diacritics, and is highly inflectional and derivational etc. According to Farghaly et al. [2] the Arabic language is the main language of most of the Middle Eastern countries. The global significance of the Arabic language in today's world is the increased presence in the Middle East in our daily news.

The Arabic language is one of the richest semantic language that has specific words to describe a specific thing and Arabic script represents 8.9% of the world's languages. It has 28 letters, each one of which has several written forms depending on their position in the word (beginning, middle, or end) such as the letter 'ب' 'b' which is in the independent style has three forms in Arabic which are ' بـ'when it is written in beginning of a word, and 'b 'when it is written in the middle and ـبـ'when it is written at the end of a word.

The Arabic language differs from the English in terms of the direction of writing, it is written from right to left where the English is written from the left to right. Moreover, Arabic does not have the capital or small letters and does not support capitalisation features.

2.2. **General Arabic language challenges**

The Arabic text poses many challenges that have influenced the development of language processing tools. In this section, the major challenges faced in addressing Arabic language are discussed [5].

As mentioned previously, the Arabic language has specific particularity such as Arabic words have two genders, masculine "مذكر"and feminine "مؤنث", and three numbers, singular "مفرد", dual "مثنى" and Plural "جمع", and three linguistic cases, nominative "الرفع", accusative "النصب" , genitive "الجر"and A noun has several forms when it is a

subject “فاعل”, accusative when it is the object of a verb “مفعول”and the genitive when it is the object of a preposition "مجرور بحرف الجر".

The words are classified into three main cases of speech, nouns ,"اسماء"verbs افعال"and particles "ادوات". In this section, the Arabic language challenges have been classified into different types.

2.2.1. Complex morphology

Morphology is the branch of linguistics that deals with the internal form of words. The general definition of the morphology is the study of the form and pattern. In linguistic systems, the morphology analyser is used to analyse the word to extract the information about the word in term of class, number, data, gender etc.

Arabic morphological analysis is one of the main phases in Arabic Natural Language Processing [6], [7]. It is a rich and complex morphology; moreover, it is orthographic with diacritics, and is highly inflectional and derivational etc. According to Maloney et al in [6] defined Arabic as a highly-inflected language that has a rich and complex morphological system.

The Arabic word takes more than one word form to represent it, which includes root, prefixes, and suffixes. In table 1, the different forms of the letter “غ” is shown as it occurs in the different positions.

Table 1: different forms of the letter in Arabic language

Beginning	Middle	End	Separate
غـ	ـغـ	ـغ	غ

On the other hand, the Arabic language has a rich vocabulary and complex morphology, table 2 explains the morphology for a single Arabic word " **فعرفناهم**" (hence we knew them).

Table 2 :Complex Arabic language morphological

proclitic	Stem	Suffix	Enclitic
ف	عرف	نا	هم
hence	Knew	we	them

The morphological analyser used to give all possible meaning for the word. Table 3 shows the morphological analyser for the word **"بعد"**.

Table 3: Morphological analysis of the word

Word	Transliteration	POS TAG	Meaning
بعد	Ba'd	preposition	After
بعد	Ba'd	adverb	yet
بعد	Ba'd	adjective	following
بعد	Bu'd	noun	distance

Many approaches and algorithms were developed for text analysis in the fields of natural language processing for Arabic language to address Arabic morphological problems.

A. Alsaad and M. Abbod in [8] presented new algorithm to improve the root extract of the words. The algorithm relies on morphological analysis and linguistic constraints. The algorithm address the problems of infixes removal by removing prefixes, suffixes while comparing the word with a predefined list of patterns.

Al-Shalabi and M. Evens in [9] presented the new algorithm which works with trilateral roots. The algorithm relies on removing the longest possible prefix from the words. The position of the three letters of the root must lie some place in the first four or five letters of the remaining word. All the possible trigrams will be checked with the

first five letters of the remainder. The algorithm is test by preparing two files, the first one is a file of roots and second is a file of prefixes.

K. Shaalan and H. Raza in [10] presented an approach to address Arabic morphological analysis problem. The approach has been built based on Arabic morphological automaton. The technique that was used in this approach is a morphological database recognized using XMODEL language. The purpose of this approach is to use different type of information extraction applications such as syntactic and semantic analysis, information retrieval, machine translation and orthographical correction.

2.2.2. A Part-Of-Speech Tagger

A Part-Of-Speech Tagger (POS Tagger) is a part of software which concerns reading the text in some language and assigns parts of speech to each word such as noun, verb, adjective, etc [11]. It uses a default lexicon and ruleset, created by training an annotated corpus. The main goals of POS tagging are to assign all possible tags to each word in a text. There are several tagger systems that have been designed to address the Arabic text.

Stanford Part-Of-Speech tagger is based on the maximum-entropy model, which was originally developed for English at Stanford University [12], it is used to parse input data which are written in many languages such as Chinese, German and Arabic languages. In the last version, the author has been devolved the training models for different language including Arabic language. The accuracy value for the tested data is reached high as (96.42%) [13]. The Arabic Penn Treebank (ATB) has been used to train the tagger [14].

Khoja Arabic Part-Of-Speech Tagger considers one of the sever tagger systems which are designed to address the Arabic text. The tagger has been combined with two approaches; a statistical and rule-based approach as it is established to produce the best

results. The tagger has used the traditional Arabic grammatical theory to derive a tag set of 131 tags [15].

Four corpora have been built and used for testing, each corpus has the number of words such as the first corpora contains (59,040) words which are collected from the Saudi " AlJazirah" newspaper, other corpora contain (3,104) words and are collected from the Egyptian " Al-Ahram" newspaper.

J. H. Yousif and T. M. T. Sembok presented a new novel approach to handle the problem of Arabic POS tagger. The Support Vector Machines (SVMs) tagger has been used to solve this problem by utilising and implementing NeuroSolutions software. Several experiments have been conducted which aim to identify the correct POS tag for each word. The authors have reported the result archived high accuracy an of 91% and unbelievably high as 99.99% respectively. The first one uses a pattern-based approach to tag a word, without using a huge manually annotated lexicon [16].

2.2.3. Peculiarities in Arabic orthographic

Latin languages majorly utilize capitalisation to help with extracting proper names. This characteristic is not available in the Arabic language. The problem of extracting proper names is especially complex for Arabic language because the first letter of the word cannot be used to recognize the proper names. Shaalan and Raza in [7], [17] have mainly used the indicator such as person indicators "الرئيس" (the president) or "الملك" (the King) or company indicator "شركة" (Company) to solve this problem.

2.2.4. Ambiguity

The problem of ambiguity is one of the biggest challenges in NLP for many languages. One of these languages is the Arabic language. The internal diacritics in the Arabic

language is considered as an import characteristic, because if it is ignored in the word, this will lead to several types of ambiguity in Arabic text.

Arabic discretization, defined as the full representation of short vowels, is considered to be one of the major challenges to most Arabic NLP tasks. However, if discretization is not applied, then the phrase "كتب الولد في المدرسة" may take the meaning: "the books of a boy are in the school, or the boy wrote in the school". Therefore, discretization can improve clarifying the context of a sentence or paragraph but can also introduce discretization challenges in terms of associating distinct meaning to the same word, such as in the table 4 which illustrates the ambiguity in the Arabic language [28].

Table 4 :Ambiguity of Arabic language

Word 1 meaning	Word 2 meanings	Word 2 meanings
علم	عَلَمْ	عَلِمَ
Science (noun)	Flag (noun)	Knew (verb)

Almost all Arabic resources such as books, newspapers and internet websites do not contain the sign for short vowels.

2.2.5. Lack of Arabic language tools

A lack of tools is one of the major problems which faces the Arabic Information extraction applications. There are sets of tools which are designed to build the systems that process human language such as the English language and many European languages; some of the tools need some minor modifications to support the Arabic language.

This section presents some of the most important tools that exist today for Arabic IE Systems. In this section, we will focus on the NLP techniques such as the General Architecture for text engineering (GATE) and Semantic Web techniques such as Protégé and Jena.

One of these tools is The General Architecture of Text Engineering (GATE) tool. GATE is one of the most popular freely available NLP tools that deals with NLP technique; GATE developed at the University of Sheffield in 1996 as open source software. There are several Arabic works which have used GATE tool to achieve different IE tasks such as named entity and extraction relation [18]-[20].

GATE is an infrastructure for developing and deploying software components that process human languages. GATE has a set of components which are used for different purposes. The a Nearly-New IE (ANNIE) is one of the components in GATE. It contains set of processing resources such as: tokeniser, sentence splitter, POS tagger, gazetteer, finite state transducer. GATE supports several types of document formats: Plain text, HTML, SGML, XML, RTF, Email, PDF, Microsoft Word. To use GATE tool with Arabic NLP, for Arabic Text we need to modify the gazetteer list and JAPE (a Java Annotation Patterns Engine) rules.

2.2.6. Lack of Arabic resources

Language resources are essential for several works on computational methods to analyse and study languages. These resources are needed to support advancing the research in several fields such as natural language processing, machine learning, information retrieval and text analysis in general.

The acute lack of resources is major difficulties that Arabic linguistics researchers face [7] such as Arabic gazetteers, Arabic corpus for the specific domain such as economic domain. Many researchers' efforts [21] resorted to Linked Open Data (LOD) public data set to resolve the NLP ambiguity problem, most notably DBpedia [22], moreover it is considered as one of the recently emerging hot topics in the field of Semantic Web due to its importance in adding structure to Wikipedia and thus making it comprehensible to semantic software tools.

Unfortunately, this resource in the Arabic domain is still, to date, a prototype, Arabic DBpedia dataset requests to be enhanced and well managed to raise a number of extracted triples written in the Arabic language as a step to improve Arabic Semantic Web applications domain [23].

2.3. **Arabic economic domain challenges**

Finance domain is one of the important domains in the Web because contains many essential journals in economics and finance, as well as country economic information, and some country risk information, from several organisations. In addition, the domain of finance and economy is a conceptually rich domain where information is complex, massive in volume and an extremely valuable business product by itself. A huge amount of valuable information is produced world-wide every day, but its processing is a hard and time-consuming task.

In the previous section listed the many challenges that are faced for the Arabic language. However, there are difficulties and challenges that have been encountered that are related in the original text such as the grammatical mistakes and non-Arabic letters related in the economic domain. In the next subsection, we show some of problems which are noted in original text.

2.3.1. Latin words in the Arabic text

There are many foreign words used in Arabic documents that are written in Arabic letters. These words are classified by the POS tagger into Noun, Verb and particles. Table 5 shows some English names which are written by Arabic letters.

Table 5: English words written by Arabic letters

No	English word	Arabic word
1	S M N	أس أم إن
2	DRAKE & SCULL	دريك أند سكل
3	ALAFCO	ألافكو
4	JVC	جي في سي

2.3.2. Grammatical mistakes in the authored text

Some syntactic analysis errors were caused by grammatical mistakes in the authored text, such as "شركة ابيض اسمنت" (White Cement Company). The rules of the Arabic language do not allow three words or more to be joined together to compose an indefinite type (نكرة), which is necessary in order to recognize Arabic composite names.

2.3.3. A Part-Of-Speech Tagger (POS Tagger) problem

The research was impacted by problems associated with the shortcomings of the PoS tagger. The Stanford tagger were used to find the POS tags of words in the text. There are many problems encountered by it such as incorrectly tagging verbs as nouns. In other instances, the other errors, where words such as (" سهم ", " Share") is tagged as س "+ " هم where separated into the proclitic (س) and (هم) as a pronoun. In addition, when some words are translated from foreign language words, such as "جي في سي" (JVC), the Stanford tagger will deal with this word as three words.

2.4. **Summary**

With the Arabic Language characteristics and the challenges in Arabic Natural Language Processing described in this chapter, it was concluded that the Arabic language has its own special features and characteristics particularly regarding morphology. The Arabic language has a very complex morphology because of the derivational and inflectional nature of the language. In this section, a brief review is given of some of the most popular challenges that needed be addressed to improve the information extraction research from MSA language text.

In addition, I have reviewed other challenges such as a lack of Arabic resources and a lack of Arabic language tools. Finally, we have reviewed other challenges that are related to the original text such as the grammatical mistakes and non-Arabic letters related to the economic domain.

Chapter 3

3. Arabic information extraction

3.1. Introduction

There has recently been an increased amount of data on the web from different kinds of resources, such as news, government reports and academic research provided in an unstructured data format. As such, the extraction and analysis of this data for the purpose of obtaining the required information has become a difficult and time-consuming task. The huge value of this unstructured data has encouraged organisations to fund research and development in information extraction and data analytics solutions.

In the financial sector, there are several initiatives related to using the information extraction in the financial domain. For instance, Radzimski et al. [24] presented a framework called FLORA system that aims to transfer unstructured financial data into structured data and inject it into linked open data and then link it to the relevant linked open datasets in order to support the viability of a financial knowledge-base for financial data analysis framework. As proposed by Declerck, Thierry, et al. [25], the MONNET project aims to solve the cross-language information access problems for the public and financial sectors.

The first step towards gaining useful understanding from unstructured data is the extraction of relevant information. The IE is a technology based on analysing natural language in order to extract the specific information. The aim of IE research is to build applications that find and link relevant information from NL documents, while ignoring irrelevant information. One of the essential technologies for information extraction is Natural Language Processing (NLP), which is a field of computer science and linguistics that is concerned with the interactions between computers and human (natural) languages [26].

NLP tools are widely used to extract useful knowledge from unstructured documents, and many research studies utilise NLP knowledge. NLP has become widely spread around the world. There are three approaches to NLP (language models). First is the rule based approach that uses a predefined set of rules; the second approach is the statistical approach using probabilities technique; while the third approach is the hybrid approach that combines the two previous approaches.

NLP tasks take the unstructured texts as input and convert them into a structured format as output. The major tasks of NLP tasks are Named Entity Recognition (NER) and Relation task (RE).

Recently, the web information extraction has been one important research field able to discover massive amounts of relational data from the web. The IE is a technology based on analysing NL in order to identify useful information. There is a difference between the traditional information extraction systems and web information extraction systems in both methods and goals. The traditional information extraction systems focus on squeezing as much juice as possible from small corpora, while the web information extraction systems focus on domain independent extraction from relatively simple sentences, and rely on the redundancy of the Web to provide large quantities of information [27].

For the web information extraction, the Web is worth the trouble because it allows the possibility of succeeding an elusive goal in Artificial Intelligence (AI) especially in the domain knowledge. Therefore, the availability of strong, flexible Information Extraction (IE) systems that transform the Web pages into structures data such as a relational database and knowledge-base will become a major necessity [28].

In the English language, there are several efforts in terms of web information extraction [29], [30]; however, in the Arabic language, there are still more efforts required to improve the information extraction tasks and see how the output of these tasks can be

used to build comprehensive applications such as the recommendation systems, health care systems, and decision making systems.

In the following section, an overview of the several endeavours in Arabic information extraction systems will be presented.

3.2. **Mining knowledge from text using information extraction**

IE Systems offer tools for constructing high-performance, multilingual, adaptive, and platform-independent NLP applications. There are several IE systems that have been developed to analyse the natural language in order to extract useful information for various purposes [25], [31], [32]. In Arabic, building IE systems is still limited because of several challenges, as already mentioned in the previous chapter.

In this section, a literature review is provided on the relevant research efforts carried out on the Arabic information extraction systems. In this review, one aims to classify the IE efforts into two categories; the first is the traditional IE systems that are focused on the systems that have used the local corpus, including commercial works; while the second is the web IE systems that have utilised a huge number of documents on the web.

3.2.1. The traditional information extraction systems

In this section, a number of Arabic studies that focused on building the information extraction systems based on the local corpus will be provided.

- Sakhr system

The Sakhr Project [33] is one of the popular Arabic software systems for analysing Arabic documents and extracting terms from unstructured text. This program helps companies to transform basic business processes by identifying important data in large quantities of texts and then extracting the most essential details for utilize in the organisation or company. Sakhr Software company is a major and market leader in

advanced Arabic language technology and solutions; and with more than 35 years as a market leader in research and development in Arabic computational linguistics, "it has successfully transformed its research in NLP into industry-first commercial software and solutions. Governments and enterprises in multiple industries across the Arab region and beyond use Sakhr award-winning technology to handle any Arabic content in this digital age. They provide leading solutions for Arabic, including Machine Translation, Optical Character Recognition, Speech Technology, Knowledge Management, Advanced Research Services and Professional Translation and Localisation" [34].

The Sakhr's company has special in a number of platforms, including a knowledge management solution suite called ArabDox, which enables classifying, indexing, organising, storing and retrieving documents in different language such as Arabic, English and French. It also offers enterprises an integrated solution to address increasing amounts of information from structured and unstructured sources.

" Sakhr's Knowledge Management Solutions has successfully provided over 12,000 users and a repository of 10 million documents with full integration within a Microsoft environment. Governmental entities and organisations grappling with massive paper documents and archiving requirements rely on ArabDox to automate their processes, safeguard sensitive documents, and reduce paper waste. Selected customers include Arab Bank for Economic Development, Qatar Embassy in the US, and Abu Dhabi Tourism Authority" [33].

- Maknaz system

The Maknaz Project [35] aims to build a list of references to descriptors or indexing terms in the information system, and has been used by a number of research and business knowledge-based applications. The goal behind the Maknaz Project is to develop an extensive Arabic dictionary that would cover all fields of knowledge. It can be said that the Maknaz Project [35] is the expanded Thesaurus (Maknaz), with an

authored list of descriptors or indexing terms integrated into an information system application. Most Maknaz data is entered manually and has been developed on several stages. In 1996, the first hard copy of the Expanded Thesaurus (Maknaz) was published. In 2001, the first electronic version (1.0), and in 2013, the Expanded Thesaurus (Maknaz) web version (2.0) were introduced. These measures allowed for the search of the descriptor and all its relations records through the retrieval search criteria as an essential foundation for retrieving information and to meet the requirements and needs of computerised libraries and information centres.

A proposal was put forward by Helmy and Daud [36] in terms of a new application that aims to identify ways to enhance the polarity text classification and information extraction in general. The Arabic Hadith Narration has been used as the domain study. In this study, the researchers tried to address opinion mining (sentiment analysis) as an example of a problem in the Arabic Hadith Narration which has been chosen for the purpose of information extraction. In addition, the authors attempted to determine the properties of a language that can have the precision of a sentiment analysis.

Daoun [37] produced a new Arabic information application which is used for posting and searching for Arabic documents by using information extraction techniques. The classified ads through SMS (CATS) application using the Short Message Service (SMS) is a SMS based classified selling and buying platform. This application aims to allow SMS users to post and search for categorised ads in the Arabic language.

3.2.2. The web information extraction systems

Several systems have gone further afield by using the huge data on the web to create information extraction systems. Saleh et al. [38] proposed an Arabic semantic annotation tool called AraTation for semantically annotating Arabic news on web documents. This was accomplished over two stages; the first is for the Arabic Information Extraction (IE) in order to recognise named entities, while the second

refers to ‘Semantic Annotation’ that maps the extracted entities to the related ontological instances. The annotated documents are saved in an RDF form so that it can be reused and machine processable on the web. As has been described earlier, the annotation process cannot be accomplished without an ontology, used for mapping instances, with its concepts. The authors in [38] found that the best tool would be to build their own domain ontology, which was created using protégé-OWL editor.

Alruily et al. [39] proposed a web-based system that aims to gather news reports from Arabic newspaper websites in order to be able to discover the information relating to future events, such as event type, location and date. The reports have been stored in a web-based structured database which enables users to browse it freely.

In addition, the overwhelming majority of Arabic research works published in the IE area has tackled specific tasks that are related to extracting Arabic named entity task [40], [20], [38]. Other research efforts have focused on the Arabic relation extraction task [41], [42]. However, it does not provide a comprehensive framework that can deliver the expected outcomes, especially in the economic domain.

As already mentioned previously, it seems that the Arabic language is currently facing multiple issues and pressures under the information revolution. On the one hand, there is a strong competition from English being a monopoly in language programming, with data exchange codes originally designed for English, in addition to the methods of storing and retrieving information, and the same information on the Internet, which along with English and most research studies, references and periodicals are available in English. On this basis, the research is geared towards building a comprehensive Arabic information extraction framework that takes into account the knowledge of a specific problem domain. This is to enhance the accuracy of information extraction and to improve an exploitation of the information extraction tool to inform the application that may in turn benefit from the structured information.

In this study, a comprehensive Arabic information extraction framework is proposed to improve the intelligent exploration and identification of the Arabic documents available online. The novelty of our framework is that it presents a comprehensive methodology that utilises the huge Arabic data on the web in order to support intelligent semantic knowledge-base applications, such as recommender systems and decision-making systems. The major aim of the IE system design is to extract the information relevant to an economic domain to support the intelligent exploration of documents that are authored in the Arabic language.

3.3. **Summary**

In the previous section, a review of different Arabic efforts has been provided in the field of information extraction, including commercial works, along with other efforts focused on some specific Arabic information extraction issues. It can thus be concluded that information extraction of the Arabic language is already in a major crisis which threatens the existence and future of Arabic in this era of globalisation, Internet culture, information technology and knowledge economy. One may argue that it is not fair for this language to be considered unworthy of being a language of science and knowledge. One baffling fact that speaks volumes about this ordeal is the current state of the Arab countries and the Arab League who have been ominously silent or lacking in interest or willpower to stir the issue. It should also be pointed out that most of the initiatives are being carried on outside of the Arab world, not to mention a shortage of the appropriate tools and foundations needed to push the efforts of automated treatment of the Arabic language. The absence of a modern Arabic dictionary that benefits from automated processing and information technology is another issue to be overcome. This research is one of a few attempts in terms of seeking to improve the research within the field of intelligent information extraction systems in the Arabic domain.

Chapter 4

4. A holistic Knowledge-based Approach for Arabic Information Extraction

4.1. Introduction

The motivation for developing the financial intelligent framework is the data overload problems existing on the Web, especially for the Arabic economic domain.

Today and for the future, people will continue to publish information in an unstructured format. Therefore, to benefit from the advantage of Semantic Web technology, we need to transfer the unstructured information to structured information and we cannot do it without linguistic processes.

In this study, we will develop an original framework relying on several components which will be introduced progressively. A holistic knowledge-based approach methodology is presented to make use of a huge amount of the data on the Web in order to improve intelligent exploration in the Arabic domain. Information extraction characteristics integrated with Semantic Web technology have been used to build this methodology. This framework for modelling and benefiting from domain knowledge is to model and inquire of this domain knowledge by using knowledge-based systems.

4.2. A comprehensive framework for an Arabic Information Extraction

In this chapter, we propose a framework to automatically annotate domain specific information from large amounts of data (e.g. websites) and using this information in order to improve the intelligent exploration of Arabic unstructured text. The framework is based on a combination of NLP techniques and knowledge-based technology. The Fundamental NLP techniques and proprietary grammar rules to semantically tag Arabic text for a specific domain (Economy was our use case), will be followed by knowledge inference in an appropriately modelled semantic knowledge-base to discover a new fact in the mined text. We also demonstrate how a public semantic dataset from the Linked Open Data cloud can be used to resolve any ambiguity in the data.

The methodology presented below is applicable to other domains and only needs a one-off effort in building the semantic model of the domain knowledge, that is, engineering the semantic ontology that conceptualizes the domain's key terms and relations and identifying public datasets providing grounding facts about the domain's key events. A general framework is depicted in the figure 2, which relies on five phrases as follows:

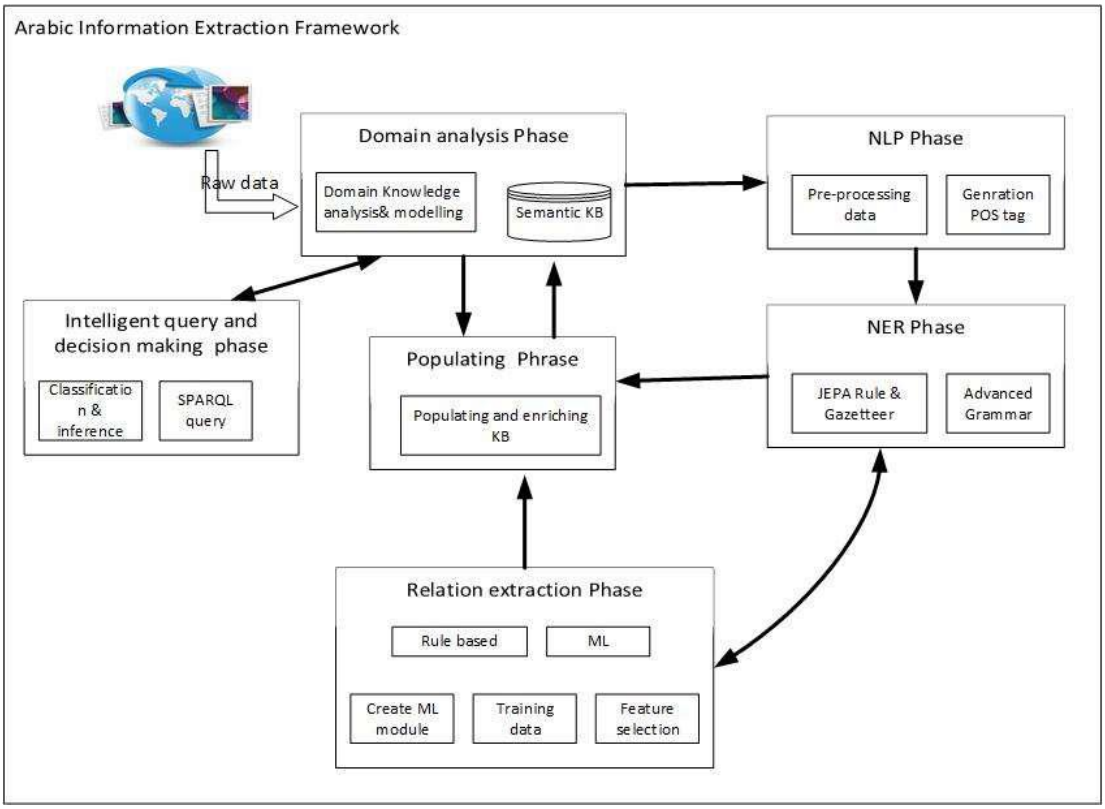


Figure 2: The phases of the proposing a framework

- Domain analysis Phase

This phase concerns analysis of the domain knowledge and constructing the knowledge map and then translating it into a formal semantic model or ontology. This phase involves adapting to the specific domain information, and implementing system resources such as lexica, knowledge-bases etc. in order to understand particular processing in the target domain and then constructing a high-quality knowledge-base relying on the design and development of well-structured ontologies based on the concepts within the domain and the relationships between those concepts.

- NLP Phase

Pre-processing data is an important step in any successful IE system that includes processing to prepare data for the next phase. Typically, this involves the process of converting the raw data from its original source into the format which is suitable for applying core mining operations.

The pre-processing data step is divided into two tasks; the first task is concerned with removing punctuations and diacritics to make the data ready for applying major IE techniques. The second task is to generate the POS tag list.

- Named entity recognise Phase

NER is a core IE task that is the essence of IE technology. The rule-based approach is concerned with extracting several Arabic named entities from Arabic text such as Location, Organisations, Date and Number. Arabic gazetteers and JAPE (a Java Annotation Patterns Engine) rules have been used to extract Arabic proper names from unstructured documents and Linked Open Data (LOD) has been used to enrich the gazetteers. The advanced Arabic grammar rules will be employed to improve the Arabic NER task.

- Relation extraction Phase

The relation extraction task is one of the phases of the framework that typically follows the NER task. The Arabic semantic relation extraction task concerns extracting different types of relations between the entities from the Arabic documents. In this study, the rule-based approach has been used to recognise the Arabic semantic relations by applying advanced Arabic grammar from a specific domain (the economic domain). An ML approach with the rule-based approach have been utilised to improve the relation extraction tasks that are not covered by the rule-based approach alone, thus building a new hybrid approach with relation classification including features such as subset selection, building training data sets and creating the ML relation classifier models.

- Populating Phase

In this phase, populating information into the knowledge-base will be accomplished in two ways: the first one is populating the entities and the relation triples which have been extracted from unstructured text into the KB, and secondly is populating the triples that are retrieved from LOD to enrich the KB.

- Intelligent query and decision-making phase

In this phase, we will use Semantic Web technologies for two purposes in order to improve the intelligent exploration of the semantic knowledge-base. The first purpose is in applying reasoning techniques on the resulting knowledge base in order to extract new and interesting information. And the second is using SPARQL queries to extract the required information from the semantic knowledge-base in order to improve intelligent exploration.

4.3. **Gathering Domain Specific Data**

There are two types of data; structured data and unstructured data. The structured data is stored in a database. Unstructured data (or unstructured information) refers to information that either does not have a pre-defined data model or does not fit well into relational tables; examples of this include email messages, word processing documents, web pages and many other kinds of business documents. Indeed, the majority of data in the real world is unstructured data. The main resources for this unstructured data are the Web. The massively large number of documents available on the Web were created by an equally diverse base of people, and as such represent unstructured data. The Web consists of several types of unstructured data domains such as health, historical and financial information. This data is represented in various languages, with the Arabic language being one of these languages. Recently, Arabic data on the Web has increased especially in the financial domain. The massive amount of information available on the Web especially in the economic domain and the variety of this data and its containment of a vast amount of information can assist in effective informed decision making.

As mentioned previously, the language resources and tools are important for the development of any information extraction applications. Before starting to build any corpus, we needed to determine what kind of corpus to build. The corpus is a significant resource for several types of language research particularly at the grammatical and lexical levels [43]. There are two types of corpus; first is a corpus of texts in a specific domain and the second is a general corpus that contains texts from a wide variety of different domains. People have always needed to derive understanding of knowledge from information to make better, real-time, smarter and fact-based decisions.

As mentioned previously, the main purpose of the framework is in building a semantic knowledge-base for improving intelligent exploration tasks based on a specific domain. We decided to build a specific corpus that covers a specific domain, and we have chosen the 'Economic' domain to build the ARB_ECON corpus, because the economic domain is one of the important domains on the Web and it contains a lot of entities and events. Moreover, several organisations have used this information to support their decision for multiple purposes.

An ARB_ECON corpus which contains a huge collection of Arabic economic news on the Web, describing various types of economic operations has been created. This documented news is collected from various Arabic news websites on the Web, and as a result, there is variety regarding the writing styles employed for describing the economics therein. This corpus will be made available online for other researchers to conduct further studies on the Arabic economic domain. In this section, the main steps are detailed for building the ARB_ECON corpus starting with data collected from Arabic economic news on the Internet and ending with cleaning the data and assigning the part-of-speech (POS) tags. As far as it is known, Arabic lexical resources are missing, in particular, a corpus which is annotated with the relations between Arabic entities [44].

4.3.1. Data collection

The ARB_ECON corpus is utilised to implement our proposed framework in this research. This corpus is extracted from different online Arabic documents related to the economic domain. The documents were published in different formats such as PDF and HTML on the Web, which require them to be converted into plain texts. There are several processing tasks that have been applied on the contents of these documents including a normalised process such as punctuation, hyphenation, quotation marks and spelling. We have extracted more than 1300 documents describing different types of economic operations from different Arabic news websites on the Web. In this work, we adopted two methods to retrieve documents from online Arabic news sources for this corpus. The first method is the RSS (Rich Site Summary) feeds [45]. It is a format for delivering regularly changing web content. RSS files consist of both static information about the feed, as well as dynamic items. Nonetheless, the online news web pages consist of navigational elements, templates, and advertisements in addition to the actual news contents. To detect the news contents and

remove undesirable texts such as advertising elements in these websites, we employed an open source Java API library to covert the websites into plain text files. However, there are some Arabic news websites that are not supported by RSS technique and therefore, we have manually retrieved these online news documents by using their URL. Then, we have used the same open resource Java library to convert HTML files to plain text files. The ARB_ECON corpus is an electronic corpus of Modern Standard Arabic; it contains 189,290 words. Table 6 shows a sample of the Arabic websites that were used to build the ARB_ECON corpus.

Table 6: List of Arabic website are using to build ARB_ECON corpus

Resources
http://www.fxnewstoday.ae/
http://sa.investing.com/
https://www.icn.com/ar/
http://www.aljazeera.net/ebusiness
http://www.alborsanews.com
http://www.bbc.com/arabic/business

4.3.2. Data characteristic

The ARB_ECON corpus is a reliable economic information source, comprising abundant information written in modern Arabic. It has several characteristics such as it is representative of different Arabic news and it is written in various styles. A characteristic of the ARB_ECON corpus is that it has been classified into two categories: comprehensive information and document style.

- Comprehensive information

The comprehensive information refers to the types of information in the corpus covering many types of information. The ARB_ECON corpus consists of several types of information

such as: stock market information, economic indicator information, resource information and location information. This information has been written in modern Arabic.

- Document style

The text corpus has been collected from different websites that represent different authoring styles. For instance, documents published on websites related to the financial domain has a special (bulletin – type) style compared to other general economic news websites.

Figure 3 illustrates the different types of writing styles in Arabic news. For example, in the Fxnewstoday, Investing and Alborsa news the style of the writing is different between these Arabic news sites based on the structuring of the writing and the information quantity. In the Fxnewstoday news site the structuring and quantity of information is better compared to other news such as the investing news site.

4.3.3. Corpus specification

We have been using freely available Arabic websites to build our resources. We collected a set of digital newspapers related to our case study (the economic domain). The text corpus was collected from different websites that represent different authoring styles. For instance, documents published on websites related to several economic fields have a special (bulletin – type) style compared to other more general economic news websites. Our corpus contains more than 1300 news articles. After applying the NLP tasks, we extracted many sentences and each sentence contained many words. Table 7 shows the numbers of documents, sentences and words in the ARB_ECON dataset.

Table 7: The specification of the ARB_ECON dataset

Describe	Number
Document	1300
Sentences	6055
words	189290

4.4. **Domain Analysis and Conceptualization**

To develop the domain model, domain analysis is needed to understand the target of the domain and specification of the language used in that domain [46]. The domain model or ontology is considered a significant stage in developing knowledge-based systems. Moreover, to build a knowledge model, requires a precise understanding of the aim and purpose of the target domain model.

Several studies have discussed the methodologies to build the ontologies. In [47], D. Jones reported several methodologies for building ontologies, such as Toronto Virtual Enterprise (TOVE) which is based on experiences in the development of TOVE, and the Enterprise Model Approach which relies on several stages to build the ontology, which he sees as important processes for any comprehensive methodology [48]. The IDEF5 method is designed to assist in the building, modification and maintenance of ontologies [49]. A simple method to construct the ontology is presented in [50] which consists of the following steps:

- Definition of the basic domain concepts

  The boundary, scope, and vocabulary are considered the main component steps which can be used to build the domain architecture.
- Description of the domain data

Determines the variables, constants and parameters which are used to support the functions and state of the domain system or a family of domain systems.

- Identification of relationships and constraints among domain concepts, data, and functions within the domain.

This section focuses on domain analysis and the conceptualization level stages that are considered as the infrastructure needed to create a taxonomy (classification) ontology and describes the key concepts and relations for a specific domain such as the economic domain as in our case study. The methodology has been categorised into the following steps:

1- Determine the domain and scope of the ontology

2- Consider reusing existing ontologies, documents and experts

3- Enumerate important terms in the ontology

4- Capture the domain knowledge using concept maps

- Define the classes and the class hierarchy
- Define the properties of classes

5- Define the facets of the slots

In the next subsections, we present in detail our methodology for building a domain model.

4.4.1. Determine the domain and scope of the knowledge-base

The first phase involves determining the domain and data source, and also the purpose and the scope of the ontology. There are some questions which should be answered at this stage including: What is the purpose of the ontology? What domain will the ontology cover? Moreover, for what sorts of questions should the information in the ontology be able to provide answers? [51].

A domain model is utilised to find common characteristics and variations between a family of software systems in a given application domain. From the domain model, a target system can be generated by designing the domain model given the requirements of the target system. Thus, a target system engineer can develop the requirements for a target system using the domain model specified previously by a domain analysis, and does not have to perform a full systems analysis every time a new target system has to be constructed [52].

In this section, the main processes and tasks which define the domain and scope of the knowledge-based domain will be discussed. The purpose of this study is to design and

develop an ontology in the area of the economic domain that will be used to support the Semantic Web recommendation system. This knowledge-base will be exploited in the Arabic Recommender Semantic Web application to support Arabic users taking decisions regarding their business, for example, which investment characteristics should they consider when choosing an investment area, or what is the best value for shares in a specific sector of the stock market? Who will use the ontology? For examples, the ontology could be used by investors, journalists or analysts and the source of terms could be obtained from economic websites, experts or linked open data.

4.4.2. Consider reusing existing ontologies, documents and experts

To build the knowledge-base, we need to understand the domain knowledge of the financial domain including using the existing knowledge-base and reading the important documents related within the domain ontology such as documents containing information about stock market operations, and discussion with the expert who have domain knowledge are considered as key factors to develop high-quality ontologies.

- Experts have considered the ontology elicitation process to be an important task because they have the experience and knowledge about a specific domain. In this work, there are some staff members from the University of Misurata that are experts in the financial domain and were involved in this research to define the familiar terms in the economic domain. It is recommended that in the initial stages of constructing the ontology it is important to identify the terms (vocabulary) that represent the ontology [53]. This will aid in scoping the domain, reaching an agreement and building the class hierarchy. Moreover, identifying the terms will broaden the understanding of the economic domain.
- Documents: many of the documents relate to the financial domain rules and several techniques have been reviewed in order to build the knowledge-base. Also, it is used to select the candidate terms and to find out the important vocabulary related to the financial domain such as concepts of the domain and the relations between these concepts, which are important for both the researcher and the domain, in addition to understanding the domain knowledge.
- Existing ontologies: there are several ontologies online that cover diverse parts of economic domains. It would be of value if these ontologies were reprocessed to obtain further information in regards to our domain. Several ontologies have been

collected that facilitate finding numerous entities and the relations between them such as, for example, a financial ontology and a vCard ontology for describing people and organisations [54], [55].

4.4.3. Enumerate important terms in the ontology

This involves writing down the list of important terms which could be useful for using in the knowledge-base before creating the classes, individuals, properties or any other defined term in the knowledge-base domain. In this work, the information that was collected from the previous stage (Domain Analysis) that is related to the economic domain and the operations that describe the domain process have been used to create a list of terms. These terms have been classified as nouns which is the basis for class names and verbs which refer to the relations between classes.

For example, nouns such as ("السهم" share, "المؤشر "Index, "سوق الاوراق المالية "stock market, "مدينة"city, "الدولة "Country, "المنتج"Industry, "" مؤشرات اقتصادية economic indicators etc.), and verbs such as ("ارتفع "Increase, "انخفض"Decrease, "انتاج"make, "ينتمي"belong. These terms can refer to several relations between these terms. Table 8 shows that the relation statement is that a property holds between two individuals such as the property (belongto) creates releation between (Alkhalij Share) as subject and (Saudi Stock Exchange) as object.

Table 8: Sample of relation between classes in the economic domain

Subject	Relation	Object
سهم الخليج Alkhalij Share	belongto	بورصة الاسهم السعودية Saudi Stock Exchange
بورصة الاسهم السعودية Saudi Stock Exchange	winNumberOfPoints	32
سهم الخليج Alkhalij Share	hasPrice	12 ريال

4.4.4. Capturing the domain knowledge using concept maps

The knowledge derivation phase plays an important role in forming a general idea about our domain. Many researchers focus on different knowledge derivation approaches that help the

definition and understanding of specific domains within the scope of the system. The concept map is one of these approaches which started to emerge within the pedagogical sciences at the end of the 1970s [56].

Concept mapping is a general method that can be used to help any individual or group to describe their ideas about any topic in a pictorial form. Moreover, concept maps and ontologies are very similar to each other, especially structurally. Importantly, concept maps provide a human-centred interface to display the structure, content, and scope of an ontology. Also, it aims to represent a specific structure of the knowledge-base in various domains, and the analysis specifies domain-dependent knowledge required to develop an application.

There are many steps that have been applied to design the concept map based on the domain analysis stage. The following steps will explain the stages of designing the concept maps.

- Define the concepts and the concept hierarchy

There are several approaches which have been used to build the concepts hierarchy such as Top down, Bottom up and Combined. In this work, the top down approach has been used to build the class hierarchy. The initial stages of designing the concept map is done by determining the set of major concepts which are considered as important key terms in the economic domain. These have been installed as the super-concepts. The super-concepts play a crucial role in designing the concept map.

There are a number of terms that have been selected such as Market Entities, Business Entities, Place, Economic indicators, Natural Resources and Currency. In addition, the sub-concept has been determined as a branch from the super concept. Table 9 shows a sample of super-concepts and sub-concepts which are represented in the concept map. The “are” relationship should be used consistently over a tree. E.g. in the example below the lower element "Type of company" should the subclass of " Business Function ".

Table 9: List of the subclasses in the CMAP

No	Super concept	Relation ship	Sub concept
01	Business Function	Are	Type of company
02	Business Function	Are	Key Person
03	Business Function	Are	Share Holder
04	Market Entities	Are	Stock Market
05	Market Entities	Are	Index
06	Market Entities	Are	Share

07	Market Entities	Are	Sector

- Define the relation between concepts

This step represents the relations between the concepts. Table 10 shows how the relations are represented between the classes.

Table 10: List of the relations between the concepts in the CMAP

No	Concept	Relation	Concept
01	Person	Birth place	Country
02	Person	Nationality	Country
03	Country	Has currency	Currency
04	Person	Employee in	Organization
05	Stock market	Located in	Country
06	Share	Related in	Index
07	Natural Resources	Product By	Country

- Define the instance of the concepts

This step determines the instance of the concepts which are represented by the attributes for each class. Table 11 shows a sample of the classes and attributes of each class.

Table 11: List of the instance of concepts in the CMAP

No	Concept	Attribute of class
01	Person	Name
02	Person	Birth date
03	Index	Number of point
04	Index	Open date
05	Country	Rate of exchange
06	Country	Population
07	Place	Name

Figure 4 shows the concept map which represents the general concept map for the Arabic economic ontology.

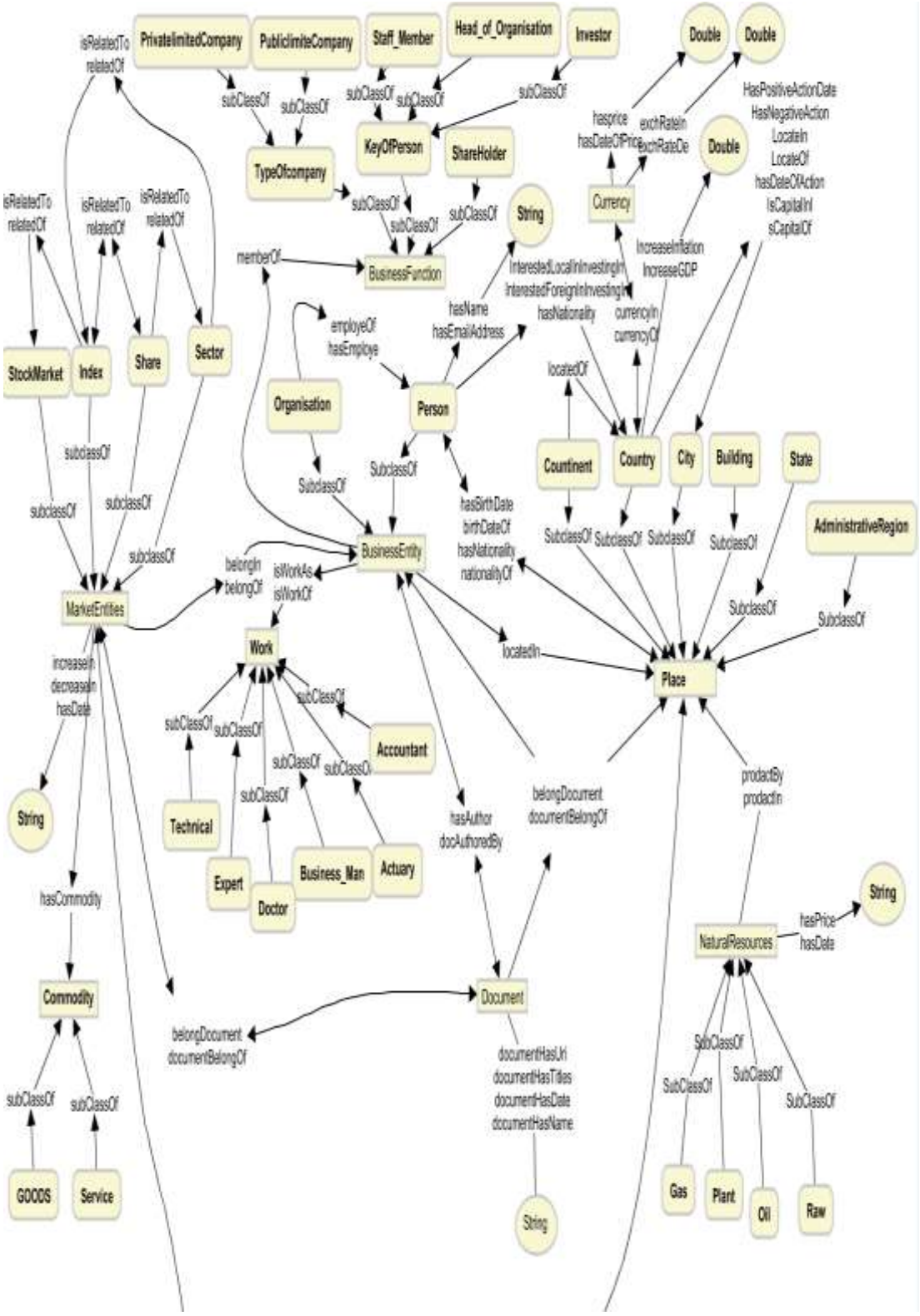


Figure 3: The CMAP for the economic ontology

4.5. **Semantic modelling of domain knowledge**

Knowledge refers to the information about a domain. This knowledge can be utilised to solve problems in the domain. Knowledge representation is the technique utilised to encode knowledge in an intelligent system's knowledge-base. The main purpose of knowledge representation is to represent knowledge in computer-tractable form, for example, it can be used to help intelligent systems perform well. This will require one of the most popular tools for modelling domain knowledge, which is a Semantic Web (SW) tool, because Semantic Web data is stored and represented in the RDF, which is an official and expressive method to define the semantics of concepts and the relationships between them.

The Semantic Web represents the technologies and methods that allow machines to read, understand and retrieve the meaning of the specific semantics or information on the Internet. Semantic Web technologies have proven successful in multiple domains, such as medicine and e-commerce applications. For the Semantic Web application, ontology development and population are essential functions. The ontology is the backbone of the Semantic Web application [57].

There are several standards for the Semantic Web such as Resource Description Framework (RDF), RDF Schema (RDFS) and Ontology Web Language (OWL).

RDF is a framework for representing information in the knowledge-base as the triple. It is represented as the subject, predicate and object. For example, in the triplet (" Mohammed is an investor " ,"محمد هو مستثمر ") ' Mohammed' is the subject, 'is a' the predicate which means the property rdf:type, and 'investor' the object. The RDF(S) is used to add additional information for RDF data such as the notion of hierarchy (e.g., local investor template rdfs: subClassOf investor).

The Web Ontology Language (OWL), is a Semantic Web language designed to represent rich and complex knowledge about things, and as an ontology language is principally concerned with describing terminology that can be used in RDF documents, i.e., classes and properties. Most ontology languages have some mechanism for specifying a taxonomy of the classes. In OWL, taxonomies can be specified for both classes and properties. The root standard of an OWL language is an RDF language. This is due to all OWL documents being

RDF documents and thereby providing some degree of compatibility between the two standards [58].

An object of an rdf:type statement, in RDF, is essentially a class, that is, it represents a set of resources. It can be clearly shown in OWL, that this resource is a class by stating that it is of rdf:type owl:Class.

Two types of properties can be defined by OWL and these are object properties and datatype properties. Relationships between pairs of resources are determined by object properties, while datatype properties determine a relation between a resource and a data type value; they are comparable to the notion of attributes in some formalisms. There are several OWL and RDF terms that can be used to describe properties.

The domain and range of a property are established by using the rdfs:domain and rdfs:range properties. It is indicated by the rdfs:domain of a property that the subject of any statement using the property is a member of the class it specifies. Furthermore, the rdfs:range of a property indicates that the object of any statement using the property is a member of the class or datatype it specifies. Although these properties may seem straightforward, these should be used carefully to avoid misinterpretation. OWL can be used to define some restricted rules that allow one to infer new knowledge based on the existing information by using ontology reasoning.

4.5.1. Translating the domain knowledge map into a semantic ontology

Significant similarities exist between concept maps and ontologies; especially the ontologies coded in RDF, which are represented through triples (subject, predicate and object) while CMs use the scheme structure (concept, link-word and concept). Considering that the OWL language is an extension of RDF, the integration between CMs and OWL ontologies can be put forward. However, knowledge in OWL is expressed as classes, subclasses, properties, relations, instances and axioms while in CMs this formal and explicit specification does not exist and has to be inferred [56].

The knowledge modelling should include reasoning and inference capabilities so as to infer new knowledge based on existing information in the knowledge-base.

Figure 5 shows the mapping between concept map and ontologies. In this work, a concept map has been exploited in order to perform the informal modelling stage of building an ontology. The concept map is the chart which represent the nodes as concepts. These nodes connect by arcs which represent the relations between them.

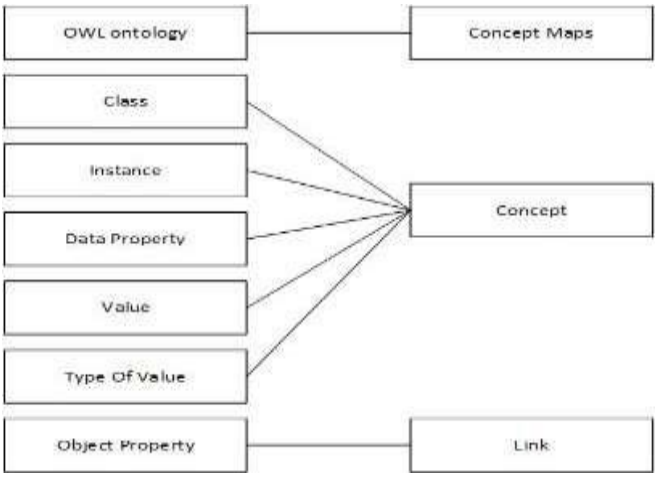


Figure 4: Correspondence of main ontology elements to concept map elements [59]

4.5.1.1. Define the classes and the class hierarchy

This step defines the classes (concepts) used in our ontology domain. These classes are not selected randomly, but they are selected depending on the domain. As mentioned previously, the Protégé tool has been used to create the ontology. In Protégé the class 'Thing' is considered as the root class which represents the set containing all individuals. Because of this, each class is a subclass of Thing, and it could be a superclass of other classes in the ontology. Many terms have been selected as the superclass based on the importance of the terms in the domain. Below is an image of the class hierarchy from Protégé. Figure 6 shows the tree of classes in the ontology.

As shown in Figure 6, there are two types of classes, the superclass and the subclass. An example of the superclass is Business Entity which contains important concepts in the domain such as organisation. This superclass displays all the instances below it, all instances and subclasses below it. Person class contains the most general classification of a person. For the superclass Location, it is anticipated that subclasses will be used when classifying places. In addition, all locations can be viewable via this class.

The superclass Business Function is used to describe the function of the Business Entity related to economic items, such that Business Function can have a key of person, or type of company.

The superclass Economic Event is a generic class which may include economic operations such as Increase, Decrease and Profit. The superclass Market Entities is a specific class which includes the main classes in the stock market domain such as Share, Index and Sector.

Figure 5 The tree of classes in the ontology

4.5.1.2. Define the properties of classes and individuals

The property in the ontology is classified into three types: data properties, object properties and annotation properties. The object properties are defined to link individuals to other individuals. The data property contains information about the class it is assigned to, without any relation to others. For example, all individuals in the person class have a data property (hasBirthdate, hasEmailAddress) defining the birth date and email address. In the next subsection, two types of properties are illustrated:

i. Object properties

The object properties (relations) define the relation between classes and is an essential task to come up with the ontology. The object properties in the ontology play an important role in connecting the members of classes (individuals) of the ontology in our economic ontology domain.

The following table explains the types of properties that link the individuals. Table 12 and Figure 7 show a list of object properties.

Table 12 : List of Object Properties

Name of property	Domain	Range	Characteristic
hasCapital	Country	City	Functional
locatedIn	Country	City	Symmetric
locatedIn	Country	Continent	Inversal functional
hascurrency	Country	Currency	Symmetric
linkedToContinent	Country	Continent	
linkedToEconomyGroup	Country	EconomicGroup	Symmetric
ShowToBuy	Broken	Share	
ShowToBuy	Broken	StockMarcket	
Hasaddress	Person	Address	

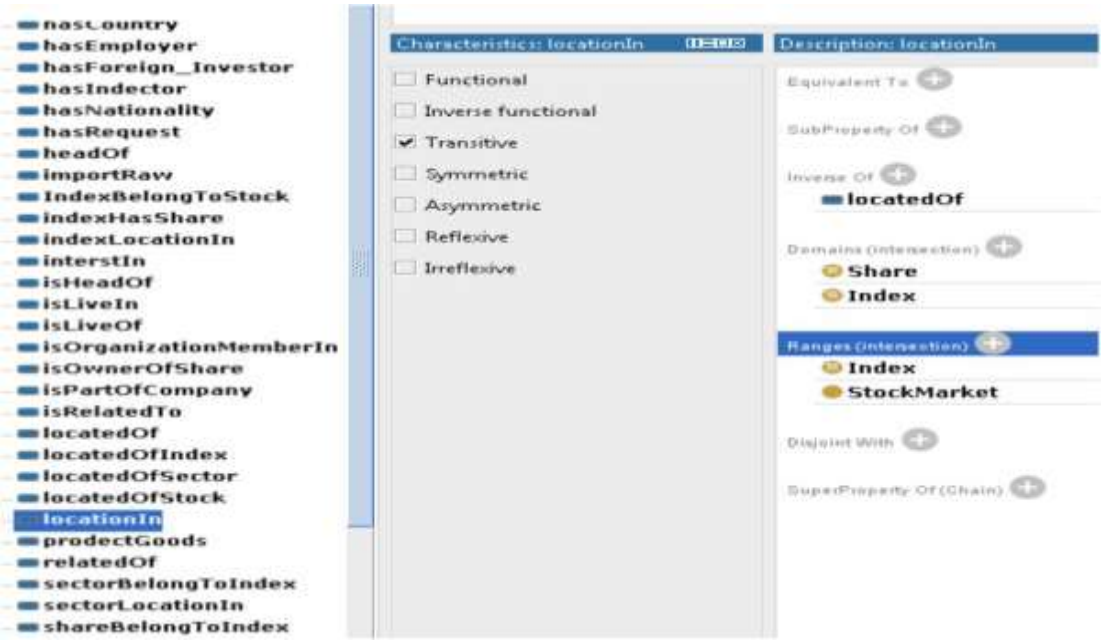


Figure 6: A screen shot of representing the Object Property in the ontology

ii. Data type properties

This DatatypeProperty links the individual with an XML schema data type value or an RDF literal (e.g., strings, numbers, datetypes, etc.). Table 13 shows a list of data type properties. E.g. in the table below the DatatypeProperty " hasLanguage " links between the individual " Country" and XML schema data type value such as "Libya"

Table 13 : List of Date type properties

Name of property	Domain	Range
hasLanguage	Country	String
hasPopulation	Country	String
increaseBy	Share	String
decreaseBy	Share	String
hasvalue	Share	String
hasCloseTime	Session	Date
hasNumberOfTrades	Session	Int
hasNumberOfshares	Session	Int

iii. Property characteristics

There are many types of property characteristics in object properties. One of these characteristics is Inverse function characteristics, for example the property

"countryHasCapital" where domain is country and range is city, meaning that if the city "cityIsCapitalOf " country then the country "countryHasCapital" city. Figure 8 shows the

how the Inverse property is implemented in Protégé.

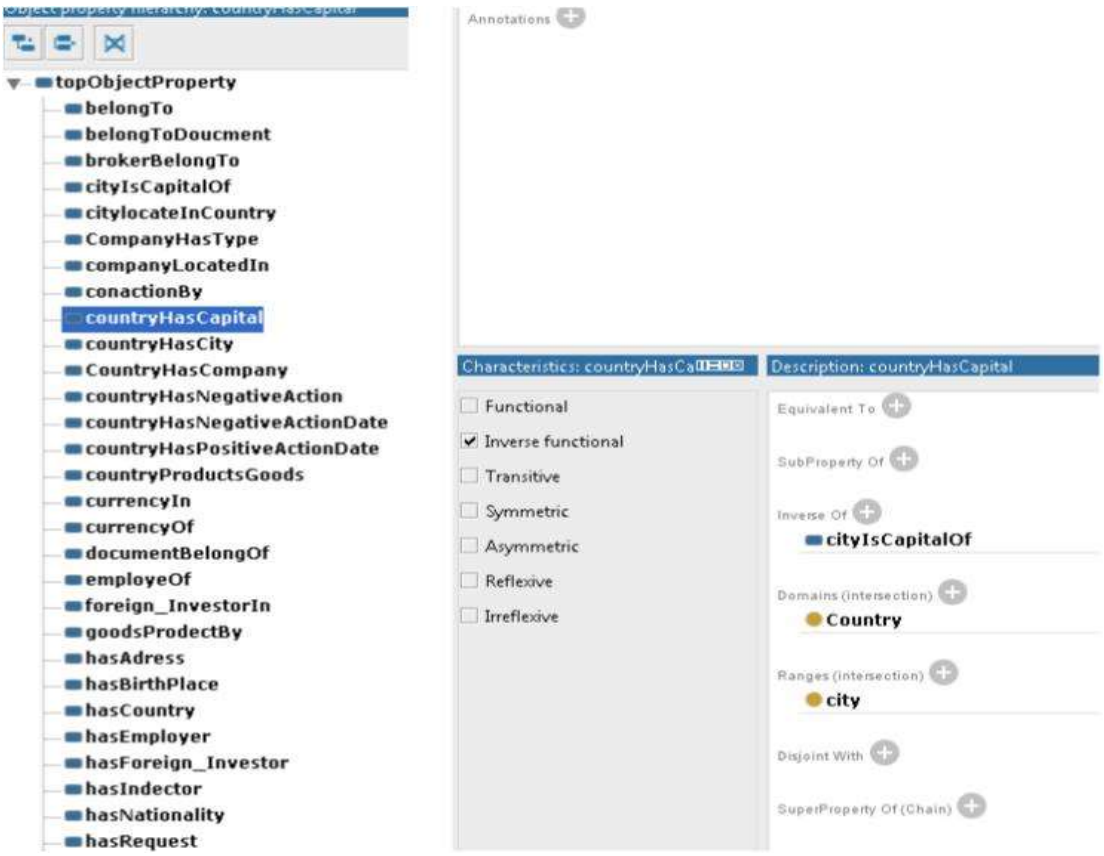


Figure 7: A screen shot of representing Inverse property characteristics

The property characteristics are used to add additional information to a specific property. The OWL language has been used to define several property characteristics including transitive and functional inverse relations. The transitive property can be applied to infer additional information about individuals in the ontology. For example, the transitive property characteristic can be applied to implement the relation between two individuals such as:

("" سهم الخليج , Golf Share) belongTo (" مؤشر دبي المالي", Dubai Financial Market)

(" مؤشر دبي المالي", Dubai Financial Market) belongTo (" بورصة دبي", Dubai Stock Exchange)

Then the transitive property will infer a new relation:

(" سهم الخليج " , Golf Share) belongTo (" بورصة دبي", Dubai Stock Exchange)

There are other types of property characteristics such as cardinality restriction which have been used to control the values for a property, such as restricting the minimum or maximum numerical values of a property. In addition, there are other types of property characteristics called domain and renege. These properties can be utilised to achieve both consistency checking of an ontology as well as for reasoning by OWL.

4.6. **Summary**

We have illustrated in this chapter the phases of a comprehensive framework for an Arabic information extraction methodology and this methodology is described very well. We have described two components: the first component concerns gathering of domain specific data. In this section, we have built a specific Arabic corpus that covers the economic domain which is called ARB_ECON corpus. The ARB_ECON corpus contains a huge collection of Arabic economic news on the Web, describing various types of economic operations. The second component concerns the building of an ontology. A taxonomy (classification) ontology has been created that describes the key concepts and relations for a specific domain. It will be utilised as the use-base for intelligent information retrieval.

Chapter 5

5. Developing rules for Arabic named entity recognition

5.1. Introduction

Proper nouns are an essential source of information in a text for extracting contents, recognising a topic in a text, or detecting relevant documents in information extraction systems. So far, they have accounted for a big percentage of the unknown words in a text. NER plays an important role in extracting proper names from unstructured text, which is essential for information extraction tasks. It was initiated in the Message Understanding Conferences (MUC) which influenced IE research in the US in the 1990s [60]. NER aims to detect proper nouns in a text as being a person name, organisation, location, date, time, monetary value, percentage, or "none-of-the-above". NER is important for several NLP systems such as question answering, information retrieval and machine translation. NER systems, which are important for any language in an open-domain text, allow for the identification of proper names. These entities represent 10% of the English and French newspaper articles [61]. A number of research studies have addressed this problem in several languages, but it seems that only in the Arabic language does one note the limited research efforts focusing on NER in Arabic texts because of the lack of resources and the little headway made in Arabic NLP in general [62].

In this chapter, the rule-based approach has been used to recognise the Arabic NER in Arabic economic domains, referred to as AENER pipeline for Modern Arabic texts, such as 'Organisation', 'Name', 'Location', 'Number', 'Date', 'Time', 'Price', which will be extracted from unstructured texts.

5.2. Natural language processing

NLP is a field of computer science that focuses on the communication between computers and humans. NLP techniques are utilised to analyse documents and provide a method for computers to understand human language. In particular, recent advances in the field of NLP

appear to give a robust, efficient and high-coverage shallow text processing techniques, as opposed to deep linguistic analysis, and these have contributed to the spread in the deployment of IE techniques in real-world applications for processing of vast amounts of textual data.

NER is considered a crucial pre-processing step in many NLP applications [63]. Recently, the NLP in Arabic has been receiving more attention, especially in linguistic analysis using Arabic. Current research in Arabic NLP suffers from a lack of coordination between many Arabic groups, which is due to the repetition of work, leading to the creation of non-standard research environments [64]. This section presents an overview of a number of annotation tools used to analyse the significant amount of information in Arabic texts.

5.2.1. Review of the natural language processing tools

As mentioned in the previous chapters, the Arabic language has been found lacking in terms of information extraction tools. In this section, some of the most important tools available today in support of the Arabic NLP will be presented. An NLP toolkit usually includes several tools for various computational linguistics problems, such as a tokenising, part-of-speech, named entity recognition and parsing. Several popular NLP toolkits will also be briefly presented.

5.2.1.1. Natural language toolkit – NLTK

NLTK is a leading platform created by Python to work with the human language. It consists of several components and data sets and deals with symbolic and statistical NLP. The initial type is designed based on rules and a deterministic approach for achievement analysis and tagging while the latter is a probabilistic approach. It is considered one of the most famous of NLP toolkits, being in use in over 200 countries. It has become very common among teachers and researchers around the world.

5.2.1.2. GATE tool

GATE is one of the best current free available software tools dealing with NLP techniques. It was developed at the University of Sheffield in 1996 as open source software. Many NLP applications use GATE, including IE, in multiple languages and media. There are several Arabic works that have used the GATE tool to achieve different IE tasks, such as named

entity and extraction relation [18]-[20]. GATE is an infrastructure for developing and deploying software components that process human language, with a set of components that are used for different purposes. The A Nearly-New IE (ANNIE) is one of the components in GATE. It also supports several types of document formats: Plain text, HTML, SGML, XML, RTF, Email, PDF and Microsoft Word. To use the GATE tool with Arabic NLP for Arabic text, the gazetteer list and JAPE rules need to be modified. GATE is a useful tool for scientists and developers in three ways; first, it has helped them to determine an architecture or organise a structure for language processing software; second, the GATE tool has the framework and a class library which it uses to implement the architecture and can be utilised to establish language processing ability in different applications; third, by providing a development environment built on top of the framework and made up of convenient graphical tools for developing components [65].

- Information extraction system within GATE

GATE has an information extraction component set called ANNIE (A Nearly-New IE) system. ANNIE depends on finite state algorithms and the JAPE language. The ANNIE component pipeline is shown in Figure 9. ANNIE consists of different resources. In other words, ANNIE was designed to be used with different applications, on several kinds of texts and for many different purposes.

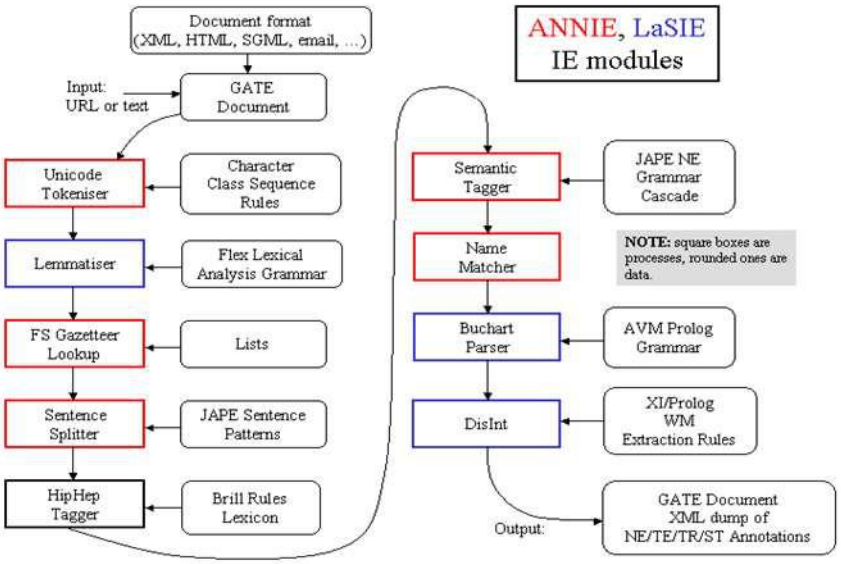


Figure 8: ANNIE component [65]

It contains a set of processing resources, such as tokeniser, sentence splitter, POS tagger, gazetteer, finite state transducer.

- Processing resources

Language resources (LRs) are data components, such as dictionaries, documents, lexicons and corpora. This is considered as an essential tool for creating applications.

5.2.1.3. The Stanford CoreNLP natural language processing toolkit

A Stanford is a natural language parser that practitioners use to analyse the grammatical structure of sentences, such as creating a phrase by grouping the words together, as well as words that are the subject or object of a verb. It is probabilistic parsers that utilise knowledge of language received from hand-parsed sentences to try to produce the most possible analysis of new sentences. These statistical parsers still make some mistakes, but commonly work rather well. Their development was one of the biggest breakthroughs in natural language processing in the 1990s. A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in a different language and assigns parts of speech to each word (and other tokens), such as nouns, verbs, adjectives, etc., although generally computational applications use more fine-grained POS tags like 'noun-plural'.

The Stanford tagger is based on the maximum-entropy model originally developed for English at Stanford University, but it now supports many languages, including Arabic, for which it claims 96.42% accuracy. Several Arabic studies have encouraged the use of the Stanford parser [66], [67].

5.2.1.4. NooJ

NooJ is the linguistic environment-based tool on .NET platform that consents for a computational linguistic analysis. It is an open resource tool and supports the use of many languages, including Arabic. The tool involves three modules; namely, corpus handling, lexicon and grammar development that are integrated into a single intuitive graphical user interface. An important feature of NooJ is that these modules are impeccably integrated and

are internally implemented in a finite state technology [68]. Many researchers have depended on NooJ in Arabic NER research [4], [69], [70].

5.2.1.5. Khoja Arabic Part-Of-Speech tagger

This tagger has been released by Khoja. It has been designed by means of integrating between statically and rule-based approaches, using a tag set of 131 tags that are extracted from the Arabic grammar theory [15].

5.2.1.6. AlKhalil Morpho Sys

Alkhalil Morpho system is a morphological analyser of standard Arabic words. The system attempts to integrate between Arabic morphological rules and linguistic resources. The second version was realised in 2016 to correct errors in the database of the first version [71].

5.2.1.7. AMIRA tool

AMIRA is an Arabic text processing toolkit. It is designed based on the SVM approach that is applied using Yamaha. It is a free resource and designed to implement several functionalities, such as tokenisation, POS tagging and Base Phase Chunking (BPC). The BPC is one of the special characteristics of AMIRA and is concerned with identifying the syntactic phrases, such as ‘noun phrases’ and ‘verb phrases’.

5.2.1.8. AraTation tool

AraTation, which is also a desktop application, is an Arabic semantic tool utilised to annotate the Arabic news on the Web. The JAVA program language and OWL ontologies have been used to construct the AraTation tool in order to produce the RDF triple.

As previously mentioned, the GATE tool is considered one of the popular tools as in the infrastructure for the information extraction applications. The GATE tool supports many languages, including the Arabic language. Many Arabic researchers have used the GATE to build Arabic IE applications [72]-[74]. As a result, there are many Arabic applications with far more insightful results. In this research, GATE will be used to build the application that processes unstructured data on the Web. In addition, the Stanford tagger has been chosen because of its availability, portability and good support, with many researchers and authors claiming that it performs very well on Arabic [67]. Furthermore, there is an integration between the GATE and Stanford tagger which could lead to a perfect processing.

5.3. **Related work**

NER is a basic prerequisite for information extraction. It is important to extract the names from unstructured text. Many applications have used NER to extract lists of names with the English language being the primary focus; but in the case of Arabic, more work is required. There are many difficulties facing Arabic natural language processing due to factors like morphological richness, short vowels, absence of capital letters and lack of resources such as tagged corpus and the gazetteers list [75]. However, there are a number of research projects aimed at tackling these problems.

There are three mains approaches in named entity recognition; rule-based, machine learning and the hybrid approach. The rule-based approach relies on hand crafted linguistic rules and using the linguistic contextual clues and indicators to extract the NEs, which presents acceptable results with the specific domain [76]. The machine learning approach requires a set of features and training dataset to classify possible NEs. Finally, the hybrid approach is a mix between the two approaches. In this section, various works that were published in Arabic NER will be discussed.

5.3.1. Rule-based approach

The rule-based approach is based on the set of rules that are predefined manually by humans in order to extract the Arabic NE. These rules are utilised to build specific patterns to extract different types of NEs, such as location, proper names and places etc. The patterns are built from grammatical, syntactic and orthographic features. In addition, a list of dictionaries is sometimes used to speed up the recognition process [77]. Dictionaries are indicating lists utilised to map terms to certain categories or types for NER, in the most commonly adopted sense. They contain reference entity names that are labelled by pre-defined categories relevant to the task. For example, a location gazetteer may be utilised as background knowledge to support recognising location entities. The dictionary list is a useful approach with an unchangeable domain, which focuses on the NEs, such as city, country, time, numbers, and locations that do not need changing. However, there are limitations to using this approach in this study because the resources will be sought from the Web. These resources consist of different Arabic NEs, such as company names, share names and index

names that are continuously changing. There are companies that are registered every day; therefore, relying on a fixed dictionary list is not possible and a dynamic dictionary is thus a requisite.

Zayed, El-Beltagy and Haggag [78] present a new approach to Arabic person names recognition. A dictionary of Arabic named entity types was used to label Arabic names. Moreover, four rules were developed to find the names of persons according to linguistic information about the names. The approach was applied in three domains, namely economics, sport and politics. F-measure was used to compute performance in each domain, where the quoted scores were high, registering 92.04 for the economic domain, 92.66 for sport, and 90.43 for the political domain. However, the author used Arabic dictionaries to recognise Arabic person names. The gazetteers approach is very helpful; the NER approach in particular, when dealing with the fixed resources, such as the Quran resource; however, when it comes to resources that are constantly changing, the gazetteer approach will not be effective because dictionaries do not provide an exhaustive list of names.

Elsebai, Meziane and Belkredim [20] developed and implemented systems to recognise person names in the Arabic language using a rule-based approach. The output of Buckwalter Arabic morphological analysis was used as an input to their system. They also used a set of keywords as indicators to phrases, which contained a person name. The system was evaluated when comparing it to the 'Person Name Entity Recognition' (PERA), by both including and excluding the gazetteers. The result was better than the PERA system. Precision, recall and f-measure were used to evaluate the quality of the system, scoring 93%, 86% and 89%, respectively. Even though, the authors succeeded in extracting the Arabic person names, they concentrated on using a list of words and phrases rather than using POS taggers, which is required to solve the ambiguity problem. The author used Arabic dictionaries to recognise Arabic person names, which is a limited approach, as previously highlighted.

Traboulsi [79] presented a new approach to extract an Arabic person name from Arabic counterparts using the local grammar approach. He defines local grammar as a way of describing syntactic restrictions of certain subsets of sentences, which are closed under some or all of the operations in the language. For extracting the Arabic person name, their method recognises function words that are clustered around Reporting Verbs (RV) in the new Arabic text. Three analysis methods were used to find an Arabic person name; namely, frequency,

collocation and concordance analyses. The author performed limited evaluation studies, which made it difficult to draw definitive conclusions about their achievement.

Galicia-Haro, Gelbukh and Bolshakov [80] presented an approach to identify and disambiguate groups of capitalised words. Their work focused on composite named entities in a Spanish text. They used extremely short lists, such as personal names, mainly Mexican cities, lists of telephone codes, and a list of POS-marked dictionaries. They used a set of heterogeneous knowledge (Local context, Linguistic knowledge, Heuristics and Statistics) to decide on splitting or joining groups with capitalised words. The results were obtained from 400 sentences that included several topics used to train the system. The overall results were 92.45% for precision and 90.88 for recall. However, grammar rules using capitalisation are not applicable to the Arabic language.

Aboaoga and Aziz [81] introduced a rule-based Arabic NER system which aims to extract the person names from Arabic text. The approach covered three domains: sports, politics and economics. The four linguistic rules have been devolved to extracting the person names. The system applied three processing steps to extract Arabic person names including; (1) pre-processing which consists of a number of tasks such as (tokenisation, data cleaning and sentence splitting); (2) automatic NE tagging where predefined lists of person names and keywords are used for person NE annotation and keyword annotation processes; and (3) applying the rules to the text in order to extract person names that do not exist in the built-in dictionaries. The system was evaluated by collecting several Arabic unstructured texts from online Arabic newspapers. The authors reported that the system achieved accuracy in the sport domain higher than that achieved in the politic and economic domains.

Wajdi Zaghouani [82] proposed an Arabic information extraction system called (RENAR) which aimed to extract different types of Arabic NE, such as person names, locations, organisations, date and numbers from different Arabic online news. The RENAR system relies on three main steps; 1) pre-processing; 2) lookup of full known names; and 3) recognition of unknown names by using local grammars and a set of dictionaries. The authors reported that the system had performed well and provided good results with different Arabic named entities, except for the organisation category where the result was low due to several challenges, such as the extended length of the name and limits of the gazetteers.

Btoush *et al.* [83] presented a new method to build a tool for Arabic PoS and Arabic NER for Arabic language. The rule-based approach was used to create this tool, which involved two components; the first component is the PoS tagging component consisting of two phases; i.e. lexicon phase and morphological phase. The second component is named entity detector and is designed based on several rules that are applied on the text to extract different Arabic named entities, such as location, person names and organisation. In the named entity detector component, three types of training datasets (location, person names and organisation) were used to recognise the Arabic names. The authors did not give more details about how they applied the named entity detector and experiments.

Feriel *et al.* [72] provided an approach to support the automated exploration and extraction of spatio-temporal information from unstructured texts in the Arabic language. This approach has integrated many systems and corpora, while the rule-based approach was utilised to identify, extract and combine spatial and temporal information from Arabic texts.

The GATE tool was used to build the model for automatically extracting patterns. There are three phases involved in this approach: Arabic gazetteer, text process and extraction and combination phases. The authors reported that this approach is efficient and performs satisfyingly well.

As for Al-shala *et al.* [84], they used the rule-based approach to build an Arabic NER algorithm for the extraction of Arabic proper nouns by employing the lexical trigger. The algorithm has identified several kinds of Arabic NEs such as person names, city, countries, locations and organisations. However, the authors have reported only person names. The heuristic rules are utilised to pre-process input data to clean and remove affixes.

Then, the person names connector as the internal evidence trigger was used to extract the person names NE, with the system demonstrating good results. The work focused on addressing the Arabic person names.

In the literature, several Arabic efforts utilising the rule-based approach to address the Arabic NER problems have been reviewed by using the grammar rules and dictionaries. The improvement of NER tools encouraged the researchers to use the rule-based approach, which compensates for the lack of Arabic resources.

The applications that are based on the rule-based approach may perform well with some specific domains, but with the general domains, they require the highest skilled labour, and the rules have to be changed based on the application domain.
Similarly, the dictionary can also recognise the NEs, such as person names, time, and locations. But, the dictionary lists cannot cover all the named entity in the specific domain; especially if one wants to identify new named entities that have not been seen before, such as the organisation's gazetteers because the new names of the organisations, such as company names, stock market names and shares names are increasing every day. Therefore, relying on a fixed dictionary list is not possible, and thus one needs to identify an appropriate method to extract these names.

5.3.2. Statistical and machine learning approach

The machine learning-based approach utilises the statistical module to recognise the specific type of proper names based on the features-based representation of the observed data [85]. The ML-based solutions rely on two crucial elements: features and annotate data.

AbdelRahman, Samir *et al.* [86] devolved a new approach integrating two machine learning techniques; namely, bootstrapping semi-supervised pattern recognition and Conditional Random Fields (CRF) classifier as a supervised technique. They utilised pattern and word semantic fields as CRF features. In addition, they applied a 6-fold cross- validation and found that their work outperformed previous CRF work. They used 15 features that can be categorised into three types: 1) unigram word features, 2) window gram features, and 3) bi-features.

Mohamed and Omer [87] presented a ML system that utilised the artificial neural network method to extract different types of Arabic NE, such as person names, locations and organisation. The system consists of three components. In the first component, they performed a number of linguistic tasks to pre-process the data, such as data cleaning, text tokenisation and POS tagging. The second component included the Romanisation of Arabic, while the third involved the application of the ANN classifier to the documents. They collected many articles of modern standard Arabic from different Arabic Web sources and used ANERcorp to evaluate the approach.

Salah and Zakaria [88] presented a comparative survey of ML approach for Arabic NE. In this survey, they highlighted a number of Arabic efforts in machine learning ANER works. They also used several factors to compare between these works, such as linguistic resource, entity type, domain, method, and performance. In addition, they examined the major challenges facing Arabic NE extraction when using the ML approaches.

The ML approach is the most successful method with the open domains; however, it requires a large size of corpus. It relies on the study of the featured positive and negative examples of NE over the huge collected text [89]. The dearth in Arabic resources to cover the most important NEs, such as person, organisation and location, caused the restricted use of this approach [84].

5.3.3. Hybrid approach

The hybrid approach is the integration between the rule-based approach and ML approach in order to improve overall performance. There are many Arabic efforts embarking upon the Arabic NEs through the use of the hybrid approach.

Meselhi *et al.* [90] adopted the hybrid approach to develop a novel Arabic NER approach using the rule-based and ML approaches. This approach aims to recognise three types of named entities: person name, location and organisation. The authors have reported that the approach performed well and yielded better results than when using other techniques, such as the rule-based approach or the ML approach.

Oudah and Shaalan in [91] integrated the rule-based and ML approaches to create a new hybrid approach to address specific Arabic NER tasks. Their approach was able to recognise 11 different types of Arabic entities, such as person, location, organisation, etc. Three different ML classifications were applied to evaluate the performance of hybrid systems. The authors claimed that the results outperformed the state-of-the-art Arabic NER in terms of accuracy, scoring f-measure results of 94.4 for the person, 90.1 for location, and 88.2 for the organisation. However, the approach did not consider composite names.

O. H. Zayed *et al.* [92] presented a new approach which was used to extract Arabic person names without using any Arabic parsers and taggers. The approach, which was based on the limited public Arabic dictionary, integrated a set of dictionaries and a statistical model based

on association which was used to extract the patterns that refer to the occurrence of Arabic person names. The benchmark dataset was used to evaluate the new approach.

Oudah and Shaalan [93] proposed a study which aimed to investigate the impact of language-independent and language-specific features on hybrid Arabic person name recognition. They presented the hybrid approach, which mixes rule-based with ML approaches, while the features space was categorised into six categories (person named entity tags predicted by the rule-based component, word-level features, POS features, morphological features, gazetteer features, and other contextual features). In the rule-based component, the GATE tool was used to extract the person names, while in the ML component, the ML decision trees algorithm applied the J48 classifier of WEKA. They used a number of datasets for testing and training to evaluate their system, including ACE (2003–2004) and ANERcorp datasets.

Meselhi *et al.* [90] presented a new approach by integrating a rule-based approach and a machine learning approach in order to improve the performance of Arabic NER. The system aimed to recognise three type of entities: person names, locations and organisations. In the rule based approach, the system applied two steps: firstly, recognising and classifying NEs in text through the exact matching with gazetteers entries and secondly based on a set of grammar rules that are implemented using JAPE. The ML approach, which used the output of the rule-based approach, was adopted as the input of the ML approach. The authors concluded that the hybrid approach provided better results than the rule-based approach and ML approach especially when the windows size is 3. One may agree with them in that the hybrid approach can achieve a better performance, but they still had to rely on the gazetteers list in the rule-based component to extract the NE, which caused the limitation for extracting the new entities not mentioned in the gazetteers. In addition, most of the researchers did not consider the length of the Arabic NE and the boundary of the NE.

However, the rule-based approach seems to achieve better results in specific domains, as the gazetteers can be adopted very precisely. It is also able to detect complex entities, as the rules can be tailored to meet nearly any requirement. However, if one has to deal with an unrestricted domain, it is better to choose the machine learning approach, as it would be inefficient to acquire and/or derive rules and gazetteers in this case.

In this section, different Arabic named NEs efforts have been surveyed, and a detailed analysis of the techniques applied and the results acquired have been provided. In addition, the limitations and advantages of each approach have been explored.

Starting with the rule-based approach, many researchers seem to have focused on addressing the Arabic NE by employed the linguistic rules and the dictionaries lists to recognise different types of Arabic proper nouns. Several attempts have been made [72], [83], [72] to deal with the problems by extracting different Arabic proper names such as location, city and organisation, etc. Similarly, other research studies focused on extracting the person names [79] and [81]. On the other hand, some researchers have used the ML approach to tackle Arabic NER problems [86], [87]. The ML approach relies on a large amount of annotated training data. Other works have adopted an integration between both approaches and developed a hybrid approach for the purpose of recognising the ANE and to improve the performance of Arabic NEs extraction tasks [92], [93].

The coverage of ML approaches is still rather restricted, and as yet, the extraction of Arabic named entities remains a major issue, even though there is room for improvement. The lack of Arabic resources, especially in the economic domain, is considered one of the main encountered challenges while attempting to improve research in this area. For this reason, many Arabic researchers have been using the rule-based approach to extract Arabic proper nouns.

In this study, the rule-based approach has been chosen to recognise Arabic named entity rather than the machine learning approach and the hybrid approach because of the unavailability and limited scope of Arabic resources required to cover the most significant Arabic named entities, such as organisations, location and person names [84]. Also, preparing the huge set of gazetteers and sometimes the large training sets can be a time-consuming task. In addition, the researcher is encouraged by the results obtained using several similar Arabic rule-based approaches, such as [82], [84], [94].

While working on the literature review, it was noted that most of the previous studies that adopted the rule-based approach focused on how to extract Arabic NE using various methods. Almost all of these methods relied on gazetteers and morphological aspects to annotate the entities, but without considering Arabic grammar in the syntactic analysis of the text. According to Zaghouani in [82], it is difficult to create rules to extract composite names, such as organisation and person names in Arabic. This is because the main challenge

lies in predicting the boundaries of the NE, particularly with long and composite NE [82], [95].

Although Arabic-named entity extraction using a different approach has been studied extensively, one can assert that the Arabic-named entity recognition grammar has not been as exhaustively investigated using advanced Arabic rules. This work makes the following contribution to the body of work on Arabic named entity recognition tasks:

- Adopting the Arabic grammar rules for extracting the Arabic composite names by utilising the Arabic genitive rule to build the several patterns used to extract the Arabic complex names, such as organisation names.
- Improving the Arabic gazetteers by updating the existing gazetteers and creating new gazetteers.
- Using the structured information on the Web, such as DBpedia, to enrich the Arabic gazetteers.
- Improving the body of Arabic NER research and contributing to bringing Arabic NER to the foreground.

In the rest of this chapter, a new Arabic pipeline approach called Arabic Economic NER (AENER) pipeline will be presented. It aims to extract and to enhance Arabic information extraction processing by improving Arabic gazetteers based on different methods, such as Arabic grammar and linked open data. In Chapter 6, a new approach will be presented by using advanced Arabic grammar to solve the problem of extracting Arabic composite names.

5.4. **The Arabic named entity recognition pipeline**

Unstructured data sources naturally need more complex processing to identify, extract, and map entities and events. This section presents NEs pipeline developed to handle unstructured Arabic documents called AENER pipeline. The pipeline architecture, which is described in Figure 10, is developed to recognise and extract specific named entities. It consists of two main stages; the first is linguistic pre-processing (compilation of POS tag list). In this stage, each document will be initially pre-processed to correct the textual mistakes as far as possible. Then, the input document will be segmented into sentences and words. Each word is analysed to extract its lexical features, such as part-of-speech tag, category and abstract

form. The second stage is the lexico-syntactic stage, which is based on the different types of gazetteers and JAPE rules to recognise the Arabic-named entities.

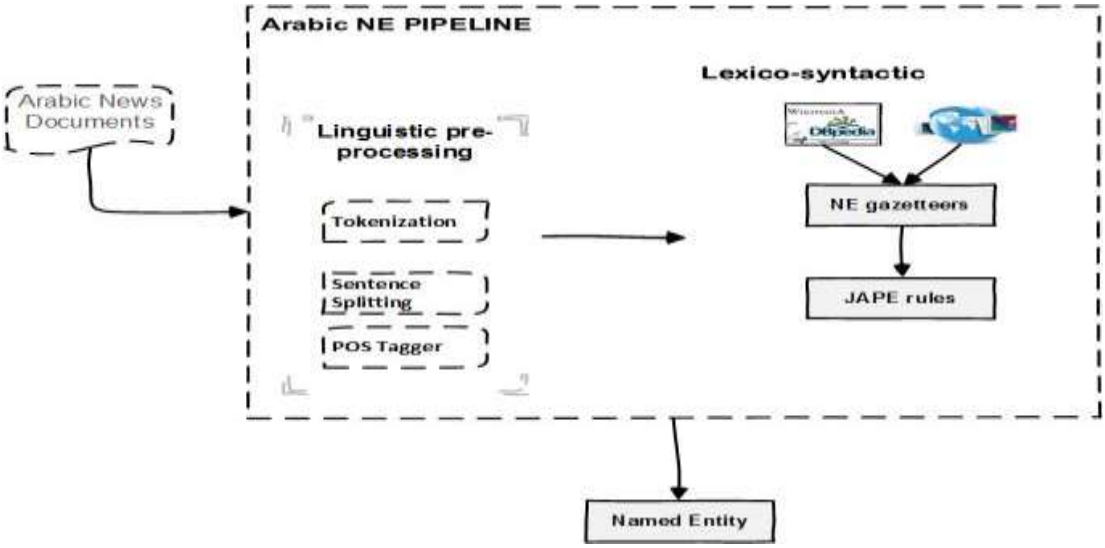


Figure 9 AENER pipeline architecture

5.4.1. Compilation of POS tag list stage

A compilation of a POS tag list plays the role of preparing the Arabic text by removing Arabic variations and diacritics, and then generating the POS list.

5.4.1.1. Removing Arabic variations and diacritics

Pre-processing is a way to improve the text by removing unnecessary elements. After converting our text corpus into UTF”8 encoding, in the first step, it is necessary to clean up the texts by removing punctuation marks, diacritics, numbers, non-Arabic letters [6], etc. In addition, the following steps illustrate the cleaning process of the text:

- Removing punctuation that is attached to the word (+ ،- ،/ ،* ،(،) ،_ ،& ،^ ،% ،$ ، # ،@ ،| ،\ ،> ،< ،{ ،} ،[،] ،× , ÷).
- Differentiating between forms of signs and commas

  (؛ ، ‘ ، “ ، ، ، : ، ٔ ، ٍ ، ٖ ، ٗ ، ٘ ، ، ´).

- Normalising some writing forms that include the following:
  “أ” “*Hamza*” “ لإ “ “ la’i “ ”لآ“ “la”.

  The reason for this normalisation is that all forms of *Hamza* are represented in dictionaries as one form, and people often misspell different forms of ‘*aleph*’.
- Removing non-Arabic words for example (JVC, Vodafone, etc.).
- Removing words attached to numbers (1020, 1243, ...).

5.4.1.2. A Part-Of-Speech tagging

This component is related to assigning POS tags to each word by using the Stanford tagger. It considers a POS Tagger used to assign parts of speech to each word (and other tokens), such as nouns, verbs, and adjectives [96]. As stated, the GATE tool is used as an IE tool in order to extract the Arabic named entity, with the GATE team devoloping the plugin, ‘Parser_Stanford’ and providing a PR (gate.stanford.Parser). The plugin is supplied with the unmodified jar file and one English data file obtained from Stanford. Stanford’s software itself is subject to the full GPL. The current versions of GATE tool do not consist of the plugin ‘Parser_Stanford’ to support the Arabic language [65]. Therefore, a JAVA code programme has been written using Stanford tagging to read each word in the document, and to obtain the POS tag and then save it into the list. The JAVA code has annotated more than 1300 documents to build a POS dictionary, which will be used during the processing of the ANER pipeline to assign the POS for each token in sentences. The POS dictionary consists of 21100 different words with each word assigned a POS. The list has been reviewed by a Language Engineering Expert to correct any incorrect analysis as output from Stanford Tagging, with around 6% of the list needing revision according to the Language Engineering Expert. Table 14 shows how the Stanford parser analyses the Arabic sentence. (e.g. the word "جاء " assigns as “VBD”)

Table 14: Example Stanford parser analyses the Arabic sentence

جاء سهم الأسمنت الوطنية مرتفعا بنسبة 12.01%			
Word	Translate	Tags	Description
جاء	come	VBD	VERB
سهم	Share	NN	Noun

الاسمنت	Cement	DTNN	Noun
الوطنية	National	DTNN	Noun
12.01		CC	Number

5.4.2. Lexico-syntactic stage

The lexico-syntactic stage plays a major role in this pipeline as it aims to extract the named entity from the different Arabic economic documents. However, the first stage in this pipeline is to extract the Arabic economic named entity from an unstructured text. There are sets of tools to extract entities in English as in many European languages, such as Stanford [97]. Some of the tools need some minor modifications to deal with the Arabic language. One of these tools is the General GATE tool.

As stated in the above mentioned review, ANNIE pipeline is considered acceptable for extracting information from English texts, but for Arabic texts, one needs to modify the gazetteer list and JAPE (JAVA Annotation Patterns Engine) rules. In this work, the ANNIE pipeline has been used to extract the entities from different Arabic corpora that are related to the financial domain. There are many modification processes that have been performed to use the ANNIE pipeline with the Arabic language.

The implementation of this named entity recognition pipeline, including a number of tasks, predefine the list of gazetteers and add new names to it. The linguistic pre-processing stage aims to assign the POS of each taken and build rules to recognise NEs by using JAPE rules, which is a finite-state transducer.

5.4.2.1. Linguistic pre-processing

This stage contains three main components: Tokenisation, Sentence Splitter and POS Tagger. Each component is described as follows:

- Tokenisation

Here each word in the documents will be separated out into individual words that are identified by the blank spaces or a special character between them. This process, which is

performed using the GATE tool, is increasingly important in the linguistic pre-processing of unstructured documents and it should be applied at the beginning of the Arabic named entity pipeline.

- Sentence Splitting

The Sentence Splitter is also one of the important processes and works to split the input document into individual sentences.

- POS Tagging

The main purpose of the part-of-speech tagger (POS) is to assign each word in the text to a word class. The GATE tool does not support the Arabic POS tagger task. Therefore, the Stanford Tagger has been integrated with the GATE tool to assign the POS to each token in the text. A JAVA code has also been built to read each word in the Arabic text and assign the tagged POS, and then save it to the POS tagger list. The GATE tool using the JAPE rule adds the POS tagging of each token as features in the token. For example, the word ("الاخير", the last) is assigned as noun "DTNN" and then saved into the POS tagging list (See Figure 11).

Figure 10: A screen shot of the POS tagging list

The GATE tool will use this list to add the POS tagging for each token using the JAPE rule. Figure 12 shows the result of the integration between the Stanford tagger and GATE tool to assign the POS tagging for each word.

Figure 11: A screen shot showing assign the POS tagging for each word in the text

5.4.3. Building resources phase

This phase aims to collect different types of Arabic named entities which are related to the economic domain. The gazetteers list is one of the important resources that is used to recognise the named entities. The gazetteer lists define the list of the names into the plain text files, where each line has one entry name and each list contains a set of names, such as person names or locations, organisations, date, numbers, etc.

The gazetteers help to identify instances in the text. A set of gazetteers has been created and developed to recognise the different named entities in the text. These gazetteers were collected from different resources, such as Makens, an Arabic database of organisations, and Internet resources. Each gazetteer list has been broken down into two categories; the first category is the major category which is referred to as the main category, such as 'location', 'organisation', 'date', and 'person names'; while the second category is the minor category and refers to the subcategory from the main category, such as 'company names', 'share names', 'index names' and 'ports names', which are the subcategory from the 'organisation' class. The major and minor types, as well as features, will be added as a 'feature only lookup annotation' generated from a matching entry from the respective list.

As mentioned above, the ANNIE GATE consists of different parts; one of them is the ANNIE Gazetteer which consists of a set of existing gazetteers. According to the last version

of the GATE tool, it can be noted that Arabic gazetteers are very limited compared to other language gazetteers such as English gazetteers. Table 15 shows the comparison between Arabic gazetteers, English gazetteers and Russian language gazetteers based on (Version 8.3) of the GATE tool. Table 15 shows that the number of existing Arabic named entities in the ANNIE gazetteer of the GATE tool is very limited. The popularity of existing NER systems rely on the utilisation of gazetteers to improve the accuracy of the system [73].

Table 15: Compassion between the existing gazetteer in GATE within different language

N0	Name of the list of gazetteers	English gazetteer	Russian	Arabic Gazetteer
01	City	27650	901	211
02	Country	240	288	193
03	Currency	257	288	17
04	Date	150	2	89
05	Name	4439	1232	1734
06	Organisation	10749	42572	96
07	Number	131	32	106

In this work, many resources have been used to improve the Arabic gazetteers for the economic domain as in the current case study. The following are the resources that have been used to improve the gazetteers lists:

- Maknaz Resources

The Expanded Vocabulary (Maknaz) is a specialist list of descriptors or indexing terms integrated into an information system application. It is a crucial tool for indexers to use suitable terms or descriptors when describing the content of documents. It is also a fundamental tool for researchers to use for retrieval [30].

- Linked Open Data

The term LOD refers to a set of pieces of information that are linked together on the Web. It refers to data published on the Web so that it is machine-readable data [98]. The linked data relies on the documents which are represented in RDF format. There are several datasets saved as linked data such as FOAF and DBpedia datasets.

DBpedia is a community effort which aims to extract the structured information from Wikipedia for IE and IR. Moreover, DBpidia contains more than 4.5 million entities and more than 3 billion triples for different languages and domains, such as country, city, etc. Although the Arabic version of the RDF is not available in DBpedia, the English version has thus been used to extract the Arabic NE by using 'label property' (RDF: DBpedia: label), while the list of Arabic NEs has been reviewed manually.

The LOD Dbpedia has been used to improve several kinds of gazetteers, such as country, city, organisation, person name and location. Table 16 compares the number of named entity in the existing gazetteer to that in the updated gazetteer.

Table 16 : Compares between the number of NEs after and before updated the existing gazetteers by LOD

N0	Name of the list of gazetteers	Existing Arabic Gazetteers in GATE	Update Arabic Gazetteers in GATE
01	City	211	10338
02	Country	193	199
03	Currency	17	275
04	Date	89	89
05	Name	1734	13327
06	Organisation	96	10102
07	Company	0	4007
08	Port	0	1420
09	Stock Market	0	270

10	Share	0	787

5.4.4. Engineering of Arabic grammar rules for extracting ANER

The rule-based approach is a successful technique for the recognition of NE tasks. It is based on a set of human-crafted patterns to extract the named entities. In this study, a set of rules based on Arabic grammar was developed in order to extract ANE, while the rules were implemented using GATE's JAPE rule (JAVA Annotation Pattern Engine). JAPE gives a finite-state transduction over annotations based on regular expressions. JAPE is a version of CPSL (Common Pattern Specification Language).

A JAPE grammar consists of a set of phases, each of which consists of a set of pattern/action rules. The phases run sequentially and constitute a cascade of finite-state transducers over annotations. The left-hand-side (LHS) of the rules consists of an annotation pattern description. The right-hand-side (RHS) consists of annotation manipulation statements. Annotations matched on the LHS of a rule may be referred to the RHS by means of labels that are attached to pattern elements [73].

The main processing is carried out by gazetteer lists and a set of grammar rules. The JAPE rules are used to annotate the text and detect named (classified) entities, such as company name, stock market name, share name, etc. There are many rules that have been written for annotated entities. These rules use gazetteers as a lookup to annotate them. Table 17 lists a number of rules which have been created to extract the Arabic-named entities. Most of these rules have been created according to the economic domain as in our case study.

Table 17: List of JAPE rules

No	Category of rule	Name of rule	Discerption rules
01	Date	day.jape	Rule to annotate the day
02	Date	Year.jape	Rule to annotate a year
03	Date	Session_date.jape	Rule to annotate a date in a different format
04	Number	Percent.jape	Rule to annotate a percentage value
05	Number	valueofmony.jape	Rule to annotate a monetary value
06	Number	point.jape	Rule to extract a number of points
07	Number	valueOfTrade	Rule to annotate a number of trades
08	Number	NumberOfShares.jape	Rule to annotate a number of shares
09	Location	city.jape	Rule to annotate a city named entity
10	Location	Country.txt	Rule to annotate a country named entity
11	Location	continent.jape	Rule to annotate a continent named entity
12	Location	Port.jape	Rule to annotate a port named entity
13	Organisation	Share.jape	Rule to annotate a share named entity
14	Organisation	Index.jape	Rule to annotate an index named entity
15	Organisation	company.jape	Rule to annotate a company named entity
16	Organisation	Stock_market.jape	Rule to annotate a stock market named entity
17	Organisation	Sector.jape	Rule to annotate a sector named entity
18	Person	Person.jape	Rule to annotate a person name named entity

Figure 13 illustrates JAPE rule for extracting the city named entity. In this rule, the token "مدينة" (city) will be used as an indicator to annotate the next word. If the next word's kind of POS is NNP or DTNNP, the system will recognise the phrase as a city name. The system will add several features to the city, such as: kind="city", rule="EX_CITY", category="NNP".

```
Phase:exe_date
Input:  word Token
//note that we are using Lookup and Token both inside our rules.
Options: control = appelt

Rule: EX_CITY
(
( Token.string=="مدينة")
({word.kind =="NNP"}|{word.kind =="DTNNP"}):City
 )
 :Tag
-->
:City.city={kind="city",rule="EX_CITY",category="NNP"},
:City.word={kind="city",rule="EX_CITY",category="NNP"}
```

Figure 12: A screen shot of the JAPE rule to extract 'city' as named entity

Figure 14 shows the rule used to extract the percentage named entity from the Arabic text. The percentage value comes in different forms in the text, such as ("50%","50 من ","بالمائة" المائة"). As such, the rule will annotate the number first and checks the tokens that come after the number according to the percentage formats, and then annotate these tokens as the percentage named entity. Figure 15 shows the rules which deal with the date format. Also, the date in the Arabic text comes in a different format, such as "10/12/2015"," 20/2016/مايو". In this case, the rule will annotate the date value according to the date format which is identified in the rule. Figure 16 shows ANNIE pipeline for extracting the Arabic-named entity.

```
Phase:FindPersentage

Input:  Token
//note that we are using Token to extract the percentage value from unstractured text .
// ANNOTATE THE PERCENT VALUE
Options: control = appelt
Rule: Percentage
(
 (({Token.kind == number))
({Token.string=="."}|{Token.string==","})*
({Token.kind == number})*):percent1
({Token.string=="%"}|({Token.string=="من"}{Token.string=="المئة"})|({Token.string=="في"}{Token.string=="المئة"})
|{Token.string=="بالمئة"}|{Token.string=="بالمائة"}|({Token.string=="في"}{Token.string=="المئة"})
|({Token.string=="في"}{Token.string=="المئة"}))

):percent
-->
 :percent.percentage= {rule="Percentage ",kind="VOFP",type="percentage",category="number"}
```

Figure 13: The JAPE rule for extracting the Percentage named entity

```
Phase:exe_date
Input:  Lookup Token SpaceToken
//note that we are using Lookup and Token both inside our rules.
// TO ANNOTATE THE DATE
Options: control = appelt
Rule: INDEX_ALL
(
 ((({Token.string=="الموافق"})*|({Token.string=="يناير"})*)
((({Token.kind == number})
((({Token.string=="/"})*|({SpaceToken})*|({Token.string=="-"})*)
 ((({Lookup.majorType == "date",Lookup.minorType == "month"})*|({Token.kind ==
number})*)
((({Token.string=="/"})*|({SpaceToken})*)
(({Lookup.majorType == "date",Lookup.minorType == "month"})*
((({Token.string=="/"})*|({SpaceToken})*|({Token.string=="-"})*)
(({Token.string=="لعام"})*
((({Token.string=="/"})*|({SpaceToken})*|({Token.string=="-"})*)
(({Token.kind == number}))):Session_date
 )
 :Session_date1
-->
:Session_date.TimeSession={kind="TimeSession",category="Time",rule="INDEX_ALL"}
```

Figure 14:The JAPE rule for extracting the Date named entity

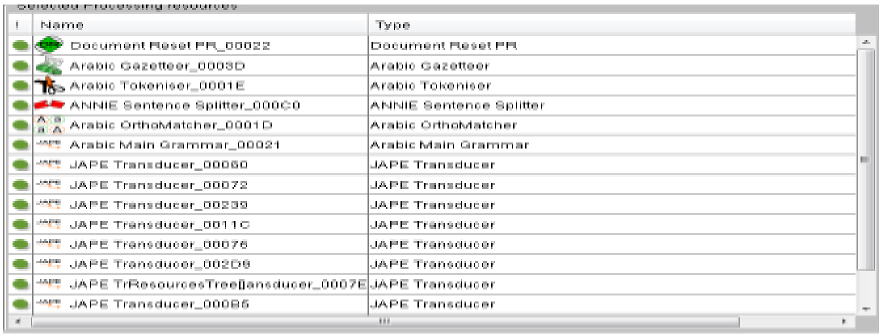


Figure 15: ANNIE pipeline for extracting Arabic named entity

The previous rules have been used to annotate the ANE based on the list of gazetteers and a set of rules, where the POS for each Arabic named entity is noun (NNP). In some cases, the ANE appears in the text in different forms that do not match the words in the gazetteer list. For example, the name of the country comes in an adjective form; i.e. adding a suffix to the general noun. In the Arabic grammar, prefixes and suffixes are referred to by appending a suffix to a noun. This suffix reflects the gender and plurality of the noun.

Table 18 shows an example of an Arabic country which appears in the text in an adjective form.

Table 18: Example showing the Country named entity in the text as adjectives word

كشف وزير التجارة الجزائري الهاشمي جعبوب أمس الاول ان بلاده قررت استيراد 3 ملايين طن من الاسمنت في الاشهر القليلة المقبلة لتلبية حاجات السوق المتزايدة.
Algerian Trade Minister El Hachemi revealed yesterday that his country has decided to import three million tons of cement in the following few months to meet the growing needs of the market.

In the sentence above, the country name ("Algerian"," الجزائري") is used as an adjective word by adding the letter ("ي") to the original word ("الجزائر"," Algeria"). The previous rule which is used to annotate the name of the country is difficult to annotate in this Arabic name. Therefore, an advanced rule is required to solve this problem.

- **Country Named Pattern**

As mentioned above, the names of the entities refer to the proper noun. In some cases, the named entity takes different forms in the text such as when the name of a country appears in the text as an adjective word as already shown. Therefore, to annotate the country names entity, one needs to apply an Arabic morphology rule to extract the proper names. In this study, this problem has been tackled by building an algorithm called COUN_ADJ pattern that aims to solve this problem by removing the attached clitics in the country names (e.g. "ي", "Ya", "ين", "Yn"). The JAPE rule has been used to solve this problem. Figure 17 details the mechanism for annotating the country name entity adjective.

- Start
- Reading each token in the sentences
- Removing the Prefixes and Suffixes from the word
- Match new words with the words in the country gazetteer
- If the new word matches one of the words in the gazetteer list, then:
    - Annotate the token as country
    - Adding a number of features to the token, such as (Original Name, String, Category and Kind)
- End

Figure 16: The mechanism steps for extracting the country names by COUN_ADJ algorithm

An example of applying the mechanism is illustrated in Figure 18, where the text contains the names of the countries that come in an adjective form, such as ("الجزائري",

"The Algerian"). The standard rules with the country list may not recognise the country name. Therefore, advanced rules are required to recognise this type of the named entity. As mentioned in Figure 18, the COUN_ADJ mechanism has addressed this problem by applying several steps.

- First, read the taken ("الجزائري", "The Algerian")
- Second, removing any attached clitics in the token, as in the ("*الجزائري*", "The Algerian"). For example, the "*ي*" (Ya) letter will be removed.
- Third, matching a new token ("*الجزائر*", "The Algeria") with the names in the country dictionary list.
- Fourth, if the new token is available in the list, then the system will recognise the word as country NE and add some features to the token, such as 'Original Name', 'String', 'Category' and 'Kind'. Figure 18 shows how the COUN_ADJ pattern addresses this problem using the GATE tool and Figure 19 shows the JAPE rule.

Figure 17 : COUN_ADJ algorithm with GATE to recognise the country names

Figure 18: COUN_ADJ algorithm implemented by JAPE for recognising the country names

5.5. **Results and evaluation**

The standard IE measures are used to evaluate the AENER pipeline include Precision, Recall and F-Measure. The precision and recall are the binary classifications which are used to measure the information extraction tasks, such as named entity and relation extraction task. “The precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percentage. Similarly, the recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually expressed as a percentage” [99].

In this work, an economic ontology is constructed specifically to serve the semantic search and information retrieval from the Arabic domain on the Web. For this aim, we focused on annotating the important ANEs which are related to the economic domain and serve my scenario; for instance, the location category NE (country name, city name) and organisation category NE (share, index, company, stoke market, bank, port) are annotated as the NE.

The main goal of this experiment was to evaluate the performance of the AENER pipeline for Arabic NEs. Three experiments have been conducted to evaluate the Arabic NE pipeline. The first experiment tested the common categories for named entities (location, organisation, numbers and dates), while the second experiment was designed to test the JAPE rules for recognising the country names, which comes as an adjective word when using Arabic grammar. The final experiment is intended to apply the Arabic NE pipeline on another

dataset.

As mentioned previously, the Arabic language lacks the linguistic resources, such as Arabic corpus for specific domain such as economic domain [84], [100].

Hence, the ARB_ECON corpus was used to apply the first experiment. Table 19 shows the performance measures (recall, precision and f-measure) for the common categories of NE (location, organisation, date and number), while Figure 20 shows the precision, recall and measure curve of the Arabic NE approach performance in extracting economic NEs when applied.

Table 19 : Recall, Precision and F-measure of evaluating the first experiment

Category	Recall	precision	F-measure
Location	0.95	1	0.99
Organisation	0.60	0.68	0.63
Date	0.82	0.98	0.89
Number	0.87	1	0.93
Industry	0.90	1	0.94
Country	0.68	0.82	0.74

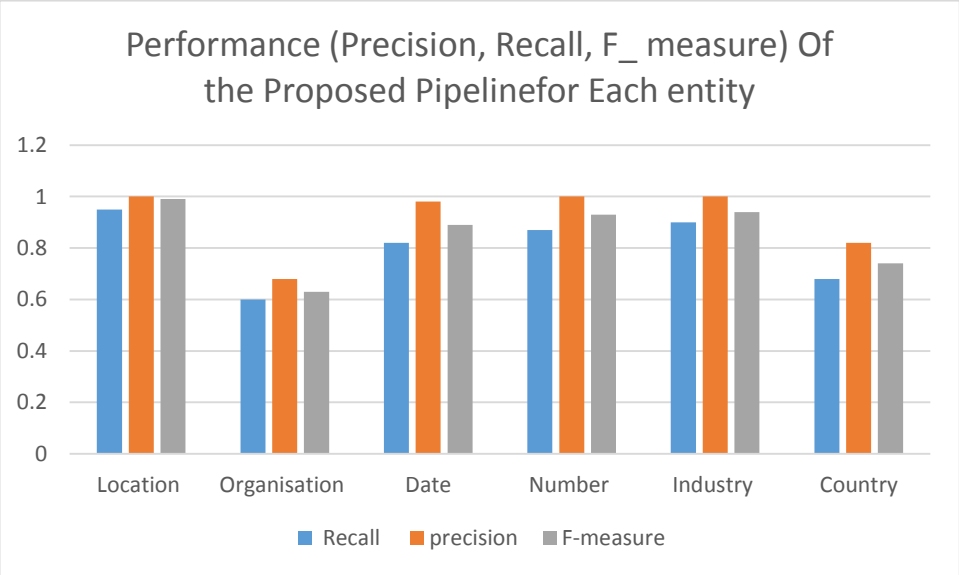


Figure 19 Precision, Recall and F-Measure curve of the Arabic NE approach applied on the ARB_ECON corpus

The second experiment aims to evaluate the performance of the COUN_ADJ algorithm for annotating the country when it appears in an adjective form. The performance measured by 'Precision', 'Recall' and 'F-measure' contacted to extract NE is promising, as shown in Table 20.

Table 20: performance of the COUN_ADJ algorithm

Category	Recall	precision	F-measure
Country	0.90	1	0.94

However, the evaluation results achieved a better result than the previous experiment when the COUN_ADJ algorithm was used to recognise the country named. The algorithm achieved the highest recall, precision and F-measure (0.90, 1, 0.94) for the country names.

The third experiment aims to apply the AENER pipeline on other types of datasets. In this experiment the pipeline was running on forty documents collected from the Linguistic Data Consortium LDC Web site. Table 21 reports on the performance measures in terms of recall, precision and F-measure in the LDC datasets.

Table 21: Recall, Precision and F-measure from the evaluation of the LDC datasets

Category	Recall	Precision	F-measure
Location	1	0.97	0.98
Organisation	0.67	0.57	0.63
Date	1	0.91	0.95
Number	1	0.58	0.73
Industry	1	1	1

Another experiment was conducted to evaluate the performance of the COUN_ADJ algorithm for annotating the country which appears in the adjective form in the LDC datasets. Figure 21 provides a comparison between the ARB_ECON corpus and the LDC datasets.

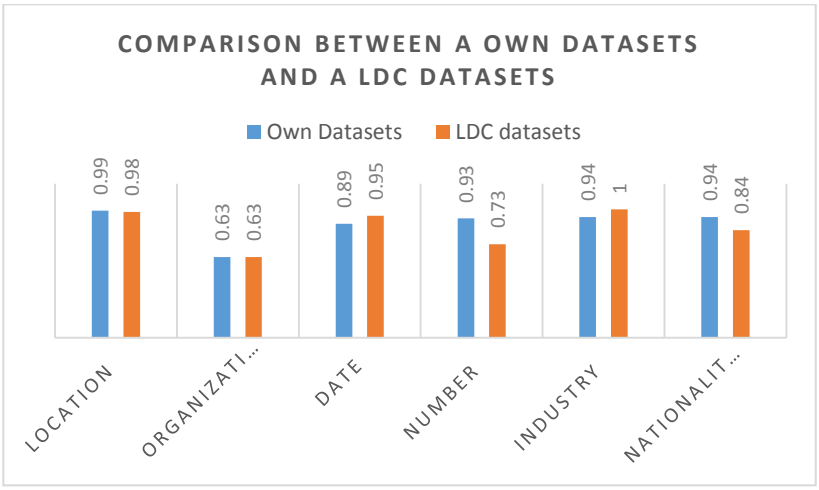


Figure 20: Comparison between ARB_ECON corpus and LDC datasets

5.5.1. Comparison of the current results with other results

Comparing the present findings with other works, one can shed light on the results of K. Shaalan and H. Raza (2008). They presented an approach to extract the ten most important Arabic named entities [101], and used the LDC corpus to evaluate their system. For this purpose, more than forty documents were collected from the same corpus in order to compare our results with theirs. The result of extraction some Arabic named entity in the current work, such as 'Location', 'Company', 'Date' and 'Price' have been used to compare their system against ours. The comparison results are shown in Table 22

Table 22: Comparison between ANE pipeline and another system in terms of F-measure

Class	ANE pipeline	Other System
Location	0.98	0.85
Organisation	0.63	0.83
Date	0.95	91.6
Number	0.90	98.6

5.5.2. Discussion of results

As mentioned in the previous section, the Arabic named entity pipeline has provided good results in some of the Arabic named entities in both datasets (the ARB_ECON corpus, LDC dataset), including location, date, number and industry. This can be explained by a good

coverage of gazetteers and building better JAPE rules that cover most of the named entity forms.

However, the AENER pipeline did not exhibit the same good performance results when applied to detect the organisations named entity. This can be attributed to three factors: the limited number of gazetteers, inconsistency of transcribing foreign organisation names in Arabic and the extended length of organisation name.

- **The limitation of gazetteers**

In some cases, our system was not able to extract the organisation NE, as shown with the F-measure from 0.63 (own dataset) and 0.63 (LDC dataset). A deeper analysis has revealed many cases where the name of the organisation is not listed in the gazetteers because these names can be a new organisation, which needs to be usually updated in the gazetteers. In addition, in some cases, the names were only partially extracted; for example, the name of the company saved in the gazetteer is "شركة قاريونس", "Garyounis Company"; however, in the text, it appeared as "شركة قاريونس لخدمات الحاسوب", "Garyounis Company for Computer Services". For this reason, our pipeline cannot annotate this name because the name in the gazetteer does not match the name in the text.

- **Inconsistency of the transcription of foreign organisation names in Arabic**

In some cases, transcribing foreign organisations' names in Arabic in the text does not match the name of the organisation in the list of the gazetteer; thus, the system cannot match the named entity. For example, the name could be written in the text ("شركة أي سي فليكس", "AC Flex Company"), but in gazetteer, it is written in a different way ("ش آي سي فلكس", "IC Flex Company").

- **The extended length and boundaries of organisation names**

The length and boundaries (the start and the end of NE) of the Arabic named entity is considered one of the main challenges that are faced in Arabic named entity recognitions [82], [84], [100]. However, composite names can be composed of different phrases, such as place or owner etc., and may also contain several words, representing a mixture of nouns, adjectives and particle, which makes the automatic identification of Arabic composite names more challenging. There are some studies that have attempted to solve this problem by using the gazetteers [82]. According to the researcher's knowledge, up until the time of writing this thesis, there has been no related work, especially in terms of addressing this particular problem using lexical rules. In this study, the advanced Arabic rule has been used to solve

this problem by using the Arabic grammar restrictions and definiteness and indefiniteness concepts; a novel approach that uses domain knowledge to formalise a set of syntactical rules and linguistic patterns to extract proper composite names from unstructured texts.

The initial approach has achieved an overall improvement for extracting the Arabic named entity, except when dealing with organisations' names as the composite names, where the initial approach is not enough to recognise the composite names.

Therefore, one needs to improve the existing pipeline by using the advanced Arabic grammar rules; in other words, the Arabic Genitive Rules should be employed to build several patterns in order to recognise Arabic composite names.

5.6. **Knowledge driven approach to IE from unstructured Arabic text**

A pipeline software for Arabic named entity recognition (ARNE) has been presented. It involved tokenisation, morphological analysis, a part of speech tagger in order to extract different Arabic named entities in relation to the economic domain. The Arabic dictionaries lists and Arabic grammar rules have been used to detect the Arabic NE.

However, and particularly in Arabic, named entities can consist of complex names, which necessitates devising a more sophisticated set of grammar rules to provide their recognition (extraction). In the following chapter, we will discuss the use of advanced pattern matching rules to extract the Arabic composite names.

5.7. **Summary**

At the start of this chapter, an overview of the most important efforts that are supported the Arabic NLP have been presented. In addition, a comprehensive literature review of Arabic NER has been presented in this chapter. The origin and applications of Named Entity Recognition are given along with the related work done in Named Entity Recognition for all three major approaches including rule-based, statistical machine based learning and hybrid approaches. In the rest of this chapter, a new Arabic pipeline approach called the Arabic Economic NER (AENER) pipeline has been presented. It aimed to extract and to enhance Arabic information extraction processing by improving Arabic gazetteers based on different methods, such as Arabic grammar and linked open data. The initial approach has achieved

an overall improvement for extracting the Arabic named entity, except when dealing with organisations' names as composite names, where the initial approach is not sufficient to recognise the composite names. In the next chapter, a new approach will be presented in order to solve the problem of composite names by using Arabic advanced grammar.

Chapter 6

6. A novel approach for A NER using language genitive rules

6.1. Introduction

As the argument in the previous chapter, the gazetteers and simple rule-based approach did not deliver to processing the composite names such as organisation name entity. The problem of extracting proper names is especially complex names in the Arabic language, because the first letter of the word cannot be used to recognise proper names. Saad & Ashour and Shaalan & Raza [7], [17] have mainly used indicator words, such as person indicators "الرئيس" (the president) or "الملك"(the King) or company indicator "شركة" (Company) to solve this problem. And the boundary and the length of the NE[82]. However, composite names can be composed of different phrases, such as place, or owner etc., and may contain several words, representing a mixture of nouns, adjectives and particles, which makes the automatic identification of Arabic composite names especially challenging.

In this study, we suggest a new approach to extracting Arabic composite names. Our approach has used advanced Arabic grammar rules that aim to classify the Arabic words as the definiteness and indefiniteness nouns and employ the genitive rules to build different patterns in order to recognize Arabic composite names. Indeed, Part of Speech (POS) tags are assigned for each word in the text, such as noun, verb, preposition, etc. [102]. We devised several syntactical rules to create linguistic patterns that match Arabic Genitive Rules (AGR) and composite organisation name patterns in financial and economic texts. We used recognition of Arabic organisation names as a use case, and the results of our initial experiments show high precision and recall scores.

6.2. Extracting Arabic composite names using a knowledge driven approach

The extracting Arabic composite names approach comprises two methods; (i) compilation of a POS tag list that is used to remove the non-essential symbols and generate the POS tag for each token, and (ii) lexico-syntactic analysis for extracting Arabic composite names. The extracted entities are injected into a semantic knowledgebase, which forms the basis for document analysis of the next phase of our research. Figure 22 shows the architecture of our approach to NER of Arabic composite names.

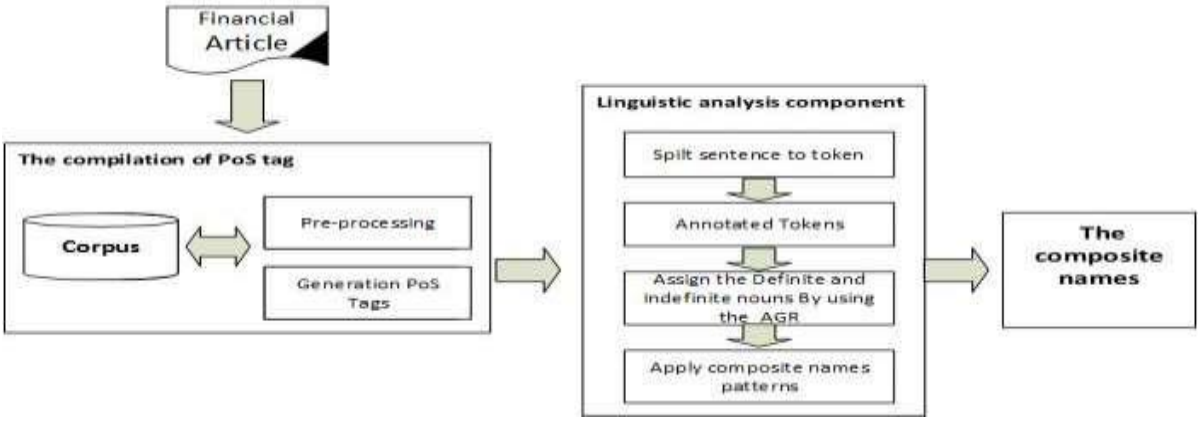


Figure 21 : Architecture of the extraction the Arabic composite named entity approach

6.2.1. Linguistic analysis for composite name extraction

In this component, we developed an application in the GATE tool, which makes use of lexicon syntactic pattern matching techniques to extract Arabic composite names. The component is implemented in three stages (text tokenisation and token annotation, grammar-based analysis, and pattern extraction). See section 5.4.2.1.

6.2.2. Grammar-based analysis to classify words as definite or indefinite

The grammar-based analysis stage is considered the most important stage. In this component, Arabic grammar rules are applied to extract Arabic composite names.

The genitive rules have been used to classify the words as definite and indefinite nouns with helps to identify a sequence that comprise the composite names of a company and ashore. The classification of pronouns into definite noun (الإسـم المعرفة) and indefinite noun (الإسـم النكرة) is considered by us to be key to extracting composite names. The definite noun is one that refers to a specific noun (person, animal, thing …. etc.) [102] . For example, (Mohammad) "محمد" / (the company) "الشـركة". There are many Arabic grammar rules for definiteness and indefiniteness, as discussed below.

Arabic grammar rules for definiteness are:

- The Proper Noun (العلم); example: ("محمد") Mohammed
- The Definite Noun with (the...) ("ال" الإسم المعرف بـ); example: ("البيت") the house

- The Possessive Pronoun (الضمير); example: (“سيارتهم”) their car
- The Relative Nouns (الأسماء الموصولة); example: (“”الذي) that
- The Demonstrative Nouns (أسماء الإشارة); example: ("هذا") this
- The Indefinite Noun added before a definite noun (المضاف الى معرفة); example: ("لخدمات الحاسوب") for computer services.

In Arabic grammar, the indefinite noun (الإسم النكرة), is one that refers to a common and non-particular noun (person, animal, thing, etc.). It can be given to any member under that category of nouns. For example, ("مدينة" a town); ("شارع" a street); ("دولة" a country). Based on the above, we devised rules to classify the tokens annotated at the previous stage into definite and indefinite nouns, based on the following conditions.

- If the kind of token is DTNN or DTNNS or DTJJ or DTJJS or NNP, then the token will be identified as DE.
- If the kind of token is NN, then the token will be identified as INDE.
- If the kind of token indicates non-noun; for example, verb or preposition, the system will reject this token. We also needed to devise new rules to classify the tokens associated with genitive articles, such as preposition and conjunction, as explained below.
- If a preposition is used, such as in “لخدمات” (for services), where the word " خدمات “(services) is combined with the preposition “ل” (for), so the word “ " خدماتwill be identified as indefinite. However, since it is combined with a preposition, it is identified as (INDEIN).
- Where a conjunctive (حرف عطف) “و” (and) is used to join two or more tokens together, such as in “لخدمات وصيانة” (for services and maintenance), our system will classify this phrase as in the above explanation. So “لخدمات” (for services) will be classified as INDEIN. Moreover, since the word “وصيانة” (maintenance) is combined within the previous word using a conjunction, it will also be identified as Indefinite (INDECC). Table 23 shows the abbreviations that are used in this work.

Table 23: List of Symbols in Stanford tagger

Symbol	Describe	Symbol	Describe
DT	Articles including ‘a’, ‘an’,	DTJJ	an adjective with a definite article attached
IN	preposition	DTJJS	a plural adjective with a definite article attached
JJ	Adjective	NN	noun - singular or mass
DTNN	with a definite article attached	NNP	proper noun
DTNNS	a plural noun with a definite article attached	NNPS	proper noun – plural
NNS	noun – plural	INDE	Indefiniteness
NP	proper noun – singular	INDECC	Indefiniteness word attached with conjunctive
NPS	proper noun – plural	INDEIN	Indefiniteness word attached with preposition
DE	Definiteness	CC	conjunction: ‘و’(and)

6.2.3. Pattern recognition to extract composite names

Information extraction can be defined as the extraction and acquisition of particular events of interest from text. At this stage, we create a set of linguistic patterns that are used to retrieve composite names from each unstructured text as a use case for composite Arabic name recognition. Our approach uses two types of patterns to extract the information. The first is used to construct phrases based on Arabic genitive rules. The second pattern is used to extract the composite name. Table 24 below shows the pattern recognition mechanism used. Here, genitive rules have enabled the extension of classification to include definite words within Arabic phrases.

Step 1, the algorithm will be assigned each word based of the POS tagger

Stage 2: the algorithm will be classified each word as Definite noun or Indefinite noun

Step 3: the algorithm will be assigned the "لخدمات " "for service" as the prepositional Phrase

Step 4: extracting the genitive phrase based on previous rule ("لخدمات الحاسوب","for computer services ") and assigned it as Definite noun

Step 5: extracting the composite name.

Table 24: Example showing the composite names pattern recognition mechanism

Steps	شركة الخليج لخدمات الحاسوب Gulf Company for Computer Services				
	الحاسوب	خدمات	ل	الخليج	شركة
	Computer	service	for	Alkaleg	Company
01	DTNN	NP	IN	DTNN	NP
02	Definite	Indefinite	preposition	Definite	Indictor
03	Definite	the Prepositional Phrase		Definite	Indictor
04	Definite (because add to genitive)			Definite	Indictor
05	لخدمات الحاسوب			الخليج	شركة

- Genitive Patterns for classifying definiteness with phrases

In this section, we extend our approach to classifying definite words within phrases using more complex genitive rules. The AGR is usually composed of two or more nouns that are semantically related and in a sequence; the genitive is considered one of several kinds of rules. We applied Arabic grammar rules to devise four patterns for classifying definiteness in noun phrases.

- The first pattern is used to extract the phrase that contains three words joined together by the conjunctive (**حرف عطف**) INDECC and first word is INDEIN; all these words came before DE, hence, the system classified the phrase as DE. Table 25 illustrated the mechanism of first pattern.

Table 25: Example showing the mechanism of the first pattern

شركة قاريونس لخدمات وصيانة وبرمجة الحاسوب Garyounis Company for Computer Services, Maintenance and Programming					
الحاسوب	وبرمجة	وصيانة	لخدمات	قاريونس	شركة
Company	Programming	Maintenance	for Services	Garyounis	Company
DE	INDECC	INDECC	INDE	INDE	Indictor
DE			INDEIN	INDE	Indictor

- The second pattern is used to extract the phrase that contains two words joined together by the conjunctive (**حرف عطف**) "و" (and), and the first word a preposition (INDEIN). Those words come before DE (الحاسوب), which our system classifies as DE. Table 26 illustrated the mechanism of second pattern.

Table 26: Example showing the mechanism of the second pattern

شركة قاريونس لخدمات وصيانة وبرمجة الحاسوب Garyounis Company for Computer Services, Maintenance				
الحاسوب	وصيانة	لخدمات	قاريونس	شركة
Company	Maintenance	for Services	Garyounis	Company
DE	INDECC	INDE	INDE	Indictor
DE	**DE**		INDE	Indictor

- The third pattern is used to extract the phrase that contains one word, which starts with the first letter as IN. This word that comes before the DE (الحاسوب), which our system classifies as DE. Table 27 illustrated the mechanism of third pattern.

Table 27: Example showing the mechanism of the third pattern

شركة قاريونس لخدمات الحاسوب Garyounis Company for Computer Services			
الحاسوب	لخدمات	قاريونس	شركة
Company	for Services	Garyounis	Company
DE	INDE	INDE	Indictor
DE		INDE	Indictor

- The fourth pattern is used to extract the phrase that contains one word that starts with the first letter as a conjunctive (**حرف و عطف**) (CC). This word comes before the definite word (**الحاسوب**); the phrase takes the type DE. Table 28 illustrated the mechanism of fourth pattern.

Table 28: Example illustrated the mechanism of the fourth pattern

شركة الانشاءات وخدمات النظافة Construction and cleaning services company			
النظافة	وخدمات	الانشاءات	شركة
cleaning	and Services	construction	Company
DE	INDE	INDE	Indictor
DE		INDE	Indictor

Now that each token has been correctly identified as a definite noun" **الاسم المعرفة**" or indefinite noun "**الاسم النكرة**", taking into consideration the genitive rules at phrase level. The next stage is to apply patterns that were devised to extract the actual composite names.

- Linguistic patterns to extract composite names

One of the greatest challenges in Arabic NER is the lack of capitalisation for proper nouns. Many workers attempted to solve this problem by using indicator words. In this work, we also use indicators, which are referred to as trigger words within our patterns, and are used for locating named entities and their semantic meanings in unstructured text.

For extraction of Arabic composite names from the unstructured text, we use two linguistic patterns that take into consideration the attachment of the definite article "ال" (the) to the Arabic composite names. The first pattern considers indicators that do not have a definite article attached, and the second pattern considers indicators that have a definite article attached.

- The first pattern is used to extract the composite names when indicated with "ال" (the), such as "الشركة" (the company), "المؤشر" (the share). In this pattern, to construct the composite name, all consecutive definite words(DE), succeeding the indicator word will be added to the composite name. This is illustrated in Figure 23.

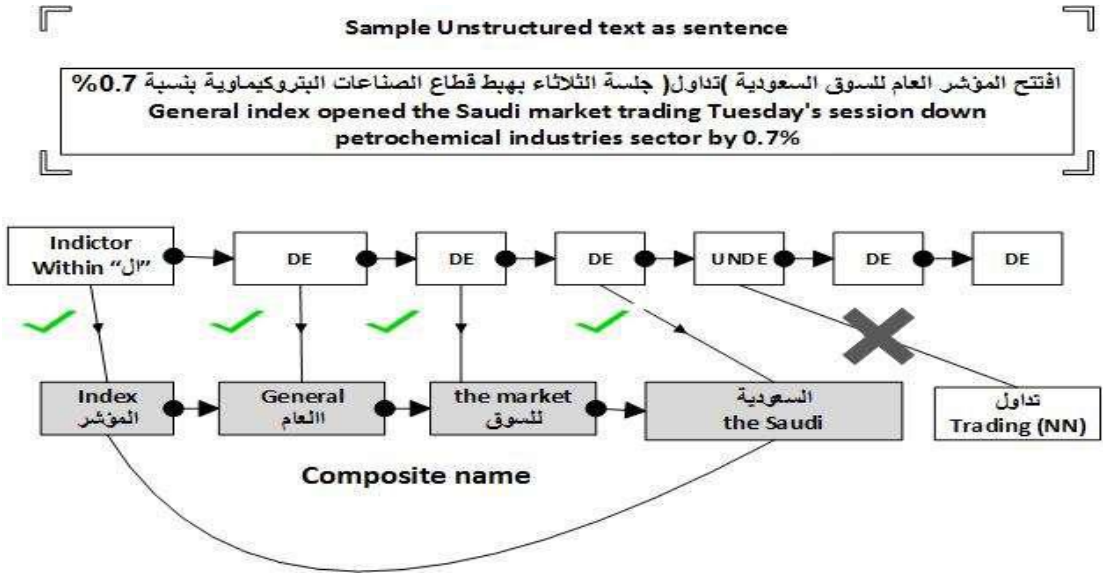


Figure 22: The first pattern for extracting Arabic composite names

- The second pattern is used to extract the composite names that are not attached to the definite article "ال" (the), such as "شركة" (Company) and "سهم" (Share). In this pattern, the word immediately following the indicator may be either DE or INDE. However,
- similar to the previous pattern rules, all the consequent words must be of type DE, as illustrated in Figure 24.

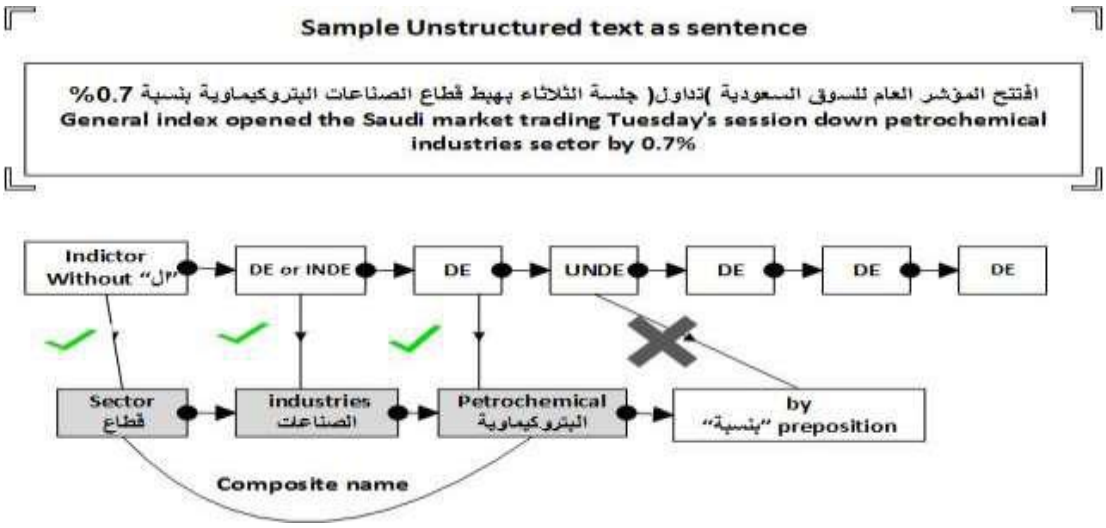


Figure 23: The second pattern for extracting Arabic composite names

The figure 25 shows the steps to extract the organisation's composite names by using genitive processing of an organisation name extraction pattern

```
Pseudocode detailing the implementation of the linguistic analysis for composite name extraction.

- read text
 Segment text into sentences
Split text into tokens (tokenisation)
for each line in the file:
 (for each token in the text )
 if (The PoS of the token =NNP or DTNN or DTNNS or DTJJ ) then
 The kind of token = Definite noun
 endif
 if (The PoS of token = NN or NNS or JJ) then
 if (the first letter in word = "ل" or "و") then
 if (the first letter in word = "ل") then The kind of token = INDEIN
 if (the first letter in word = "و" ) then The kind of token = INDECC
 else
 The kind of token = Indefinite noun
 endif
 endif
 endfor
//this stage, building the genitive patterns to use them to extract composite names
(for each token in the text )
 if Phrase matches GR1 or GR2 OR GR3 OR GR4 then
 the kind of Phrase = Definite noun
 endif
 end for

// extract
for each token , kind is Definite noun or Indefinite noun
 if P = P1 or P2 then
 P = the composite names
 end
 end for
Key,
GR1,GR2,GR3 and GR4 are kindS of the AGR Patterns
P1 and p2 are kind of the Linguistic patterns
the Phrase is sequences of tokens
```

Figure 24: Pseudocode detailing the implementation of the linguistic analysis for composite name extraction.

6.3. **Discussion of the results and error analysis**

At the time of compiling this study, we could not find any published research that evaluated NLP efforts at extracting Arabic named entities comprising composite names. Hence, an evaluation could not be compared against published efforts in the field. The ARB_ECON corpus was used to implement the experiments. The experiment aimed to evaluate the

performance of our algorithm in extracting the composite names dependent on number of words in the composite names and the AGR within the composite name. We observed that the increase in the number of composite words slightly affects the recognition accuracy, so, the names that contain two words demonstrated better performance in terms of precision (96%), as shown in Figure 4, compared to three words (93%), four words (92%) and five or more words (93%). Nevertheless, the recognition accuracy remains very high overall as illustrated in figure 26.

In a few cases, we noted that the word coming after the indicator was classified as indefinite or definite; meanwhile the original meaning of this word is not related to composite names. For example: "مليون سهم تقريباً بالجلسة الماضية" (the last session has almost a million shares). In this case, the word "تقريباً" (almost) is classified as an indefinite noun. The system decided that the "سهم تقريباً" is a composite name but this is unlikely to be correct, because the word "تقريباً" takes another meaning in this sentence.

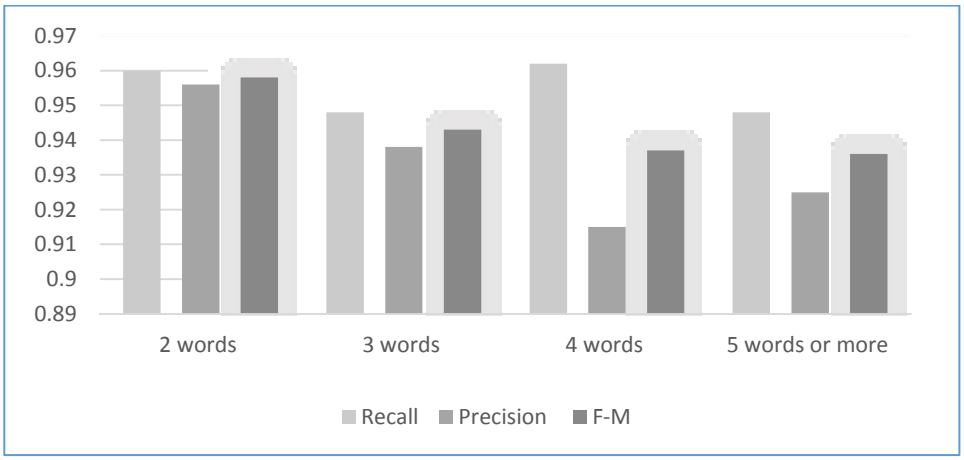


Figure 25: Impact of composite names' length on Precision

The second experiment compares the use of different AGR patterns on composite name extraction. The composite name can exist in different forms in the Arabic text, as shown in figure 27. In this experiment, we arranged our experiment into four categories of genitive rules, as mentioned above. During the second experiment, we observed that the names that used the first pattern demonstrated better performance obtaining precision (1.0%) compared to the second pattern (0.95%), third pattern (0.96%) and fourth pattern (0.94%). Furthermore, recall for most of the categories achieved very high results. Therefore, the genitive pattern method proved its consistency for the recognition of Arabic composite

names. According to our observation on both experiments, we found the last results were improved, with much better performance than the first experiment. Having said that, the process, which focused on composite names within the AGR patterns, was relatively consistent. Figure 28 shows the result of the second experiment.

As previously mentioned, there is a lack of resources for evaluation of Arabic language using the same corpora. Hence, we did not give precedence for comparison with other systems, similar to [80] the present approach to extract composite NE in Spanish text.

The results obtained were 92 % for precision and 91% for recall, which compares well with the result of our Arabic NER algorithm. And table 30 shows a sample of Arabic compose names.

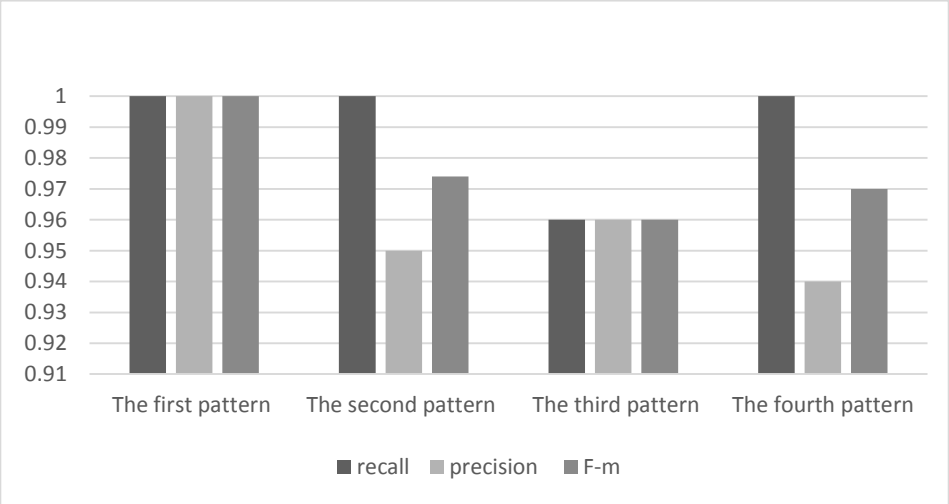


Figure 26: Recall,Precision, and F-meagure of AGR patterns

On other hand, table 29 shows the comparison between the first patterns and second patterns for extract Arabic composite names.

Table 29: comparison between the first pattern and second pattern

Patterns	Recall	Precision	F-measure
First pattern	0.97	0.94	0.95
Second Pattern	0.95	0.91	0.93

From the previous results, it can be shown that the results of our pipeline are satisfactory in the Recall term, which shows the Recall term has achieved the higher value rather than the Precision value in all experemnts. Because the EANER pipeline utilizes the economic

intrinsic indicators to determine the start word for the composite names, which caused to easily to repaginate the named entity.

Table 30: Samples of the Arabic composite names with AGR

Patterns	Arabic composite names example	English trans
First	سهم قاريونس لخدمات وصيانة وبرمجة الحاسوب	Garyounes services company, maintenance and computer programming
Second	شركة الاتقان لصيانة وتوريد الالة الكهربائية	Al Etgan company for the maintenance and supply of electric machine
Third	شركة العربي الافريقي الدولي لتداول الاوراق المالية	Arab African International Inc. Securities
Forth	شركة التصنيع وخدمات الطاقة	Industrialisation and Energy Services Company

It is noteworthy that the analysed results were affected by problems associated with the shortcomings of the POS tagger, as well as grammatical mistakes in the original text.

The Stanford tagger was used to find the POS tags of words in the text. It tagged some words as nouns, although they were verbs, along with other errors; for example, the word سهمis tagged as" س+هم". It is separated into the proclitic (س) and (هم) as a pronoun. In addition, sometimes for words, which were translated from foreign language words, such as "جي في سي" (JVC), the Stanford tagger identify theses as three words. Some syntactic analysis errors were caused by grammatical mistakes in the authored text, such as " شركة اسمنت ابيض" (White Cement Company). The rules of the Arabic language do not allow three words or more to be joined together to compose an indefinite Type (نكرة). Our approach cannot deal with these names, because the base of our algorithm does not allow three or more words to be joined together in an indefinite manner. This type of mistake rare happened in Arabic text.

6.4. **Summary**

In the previous two chapters, the Arabic NER problem was tackled by using the rule-based approach and gazetteer list approach. Many Arabic proper nouns have been extracted such as locations, numbers, dates, person names and organisations that were related to the

economic domain. In this chapter, we presented a novel approach for extracting composite names from documents authored in the Arabic language. Our approach relies on the classifying of pronouns into definite noun (الإسم المعرفة) and indefinite noun (الإسم النكرة). This is considered by us to be key to extracting composite names, which is extended to provide for definitive classification within phrases using complex genitive pattern recognition. Experimental evaluation was performed on financial documents with varied authoring styles to reveal good precision and recall results. It also confirmed that our error correction mechanism applied to the output of the POS tagging process results in noticeable improvement in the effectiveness of our composite names extraction approach. The chapter also highlights unresolved problems relating to the complex Arabic POS tagging process and to syntactic analysis errors stemming from common misuse of the Arabic language grammar.

Chapter 7

7. Arabic discourse grammar and Machine learning approach for relation extraction

7.1. **Introduction**

The volume of information published on the Web is growing quickly with the increase in number of Internet users. According to the Internet World Stats, the number of Internet users exceeded 3,631,124,813 at the time of write this thesis. As most of the published information is unstructured, i.e. written in natural language text, the need for systems that can automate the extraction of useful information from unstructured documents is becoming ever more desirable, which has contributed to the development of information extraction (IE) into a major research area. Nevertheless, the advance in the research and development of Arabic IE systems is not as remarkable in comparison to European languages[57], which is reflected in the scarcity of Arabic Natural Language Processing (NLP) resources and tools. Building on our original approach to composite Arabic named entity recognition[103], the research documented in this chapter proposes a novel, knowledge-based approach to relation extraction from unstructured Arabic text, which is based on the principles of the principles of Functional Discourse Grammar. We further improve the approach by integrating it with Machine Learning relation classification, resulting in a hybrid relation extraction algorithm that can handle especially complex Arabic sentence structures. The accuracy of our relation classification efforts is extensively evaluated by means of experimental evaluation.

7.2. **Related Work**

Despite the limited volume of research efforts in extracting Arabic semantic relations compared to European languages, these efforts can be similarly categorised into linguistic (rule-based) methods, statistical or machine-learning methods and hybrid approaches that combine both methods in an attempt to improve the relation extraction accuracy. It is worth noting that the focus of our research is on the extraction of generic (non-taxonomic) relations, hence in this review will not consider the works that are primarily concerned with extracting ontological (taxonomic) relations such as the efforts published by Zamil and Al-Radaideh in [104] and Bouaziz and et al, in [105].

Rule-based approach

Rule based approaches extract the relations by using syntactic and semantic rules that are hand-crafted based on part of speech and domain-specific information. Hence, these approaches are well suited to extract relations from specific problem domains where detailed semantic knowledge can be extensively exploited in building the relations' pattern recognition rules [106].

El-salam, Shimaa M. Abd, in [107] present an approach to extract the binary relation between two Arabic named entities from the Web by using the semi- supervised techniques. Using intitial seed relation instances as input, the suggested pattern-based system uses a generic search engine, Google™, to extract compatible candidate relations that are validated and selected in an iterative process. The authors reported that four experiments were carried out to evaluate their approach to extract four common relations on different domains. The success of the approach is dependent on the recall of the utilised search engine and the volume of seed relations.

Hamadou, Ben, Piton, and Héla in [108] used a rule-based approach to extract the relation between Arabic named entities by using NooJ Platform. The relation extraction linguistic patterns are based on basic grammar rules and the lexical composition of the problem domain, in particular the key concepts of person names and organization. It is difficult to assess the applicability of the suggested approach given the limited use of the Arabic grammar rules and the fact that only one relation type is evaluated.

Albukhitan and Helmy in [109], propose an Arabic ontology learning system based on basic lexico-syntactic pattern recognition. The patterns are hand-crafted and are either based on hierarchical conceptual relationships for extracting taxonomical relations, or on an entity-predicate method that depends on parsing sentences to capture the triple of subject, action and object to capture generic (non-taxonomic) relationships. The preliminary evaluation demonstrates good results for Precision, but the Recall is low, which is anticipated as without the aid of domain-specific knowledge or computational intelligence it is difficult to achieve good recall results for Arabic relation extraction.

Machine learning approach

Based on a set of features that can include syntactic, semantic and lexical features, Machine Learning approaches have been successfully deployed for relation extraction from unstructured text.

Al-Yahya, Maha, Luluh, and Sawsan in [110] present a relation extraction approach based on distant supervision machine learning. A seed ontology is used to to generate the training corpus, which is used in machine learning to extract antonyms from another corpus set. The new antonym is added to the original ontology after manual verification. The objective of the reported work is ontology enrichment rather than generic relation extraction, hence the involvement of human verification and the open research question of the optimum method for deciding the accuracy of the pattern matching score.

Mohamed, et al, in [111] present a distant supervision approach for extracting Arabic Relations. To counter the lack of annotated Arabic corpora, they source the DBpedia public linked dataset to build the training data. Their relation classifiers achieve 70% for the relation detection F-measure. However, the DBpedia dataset does not provide comprehensive coverage of relation types, particularly for Arabic, and might require to be complemented by manually trained data [49].

Hybrid approach

There are several definitions of what constitutes a hybrid approach, but in this work, we imply an approach that combines the advantages of domain knowledge in rule-based systems with the learning capabilities of computational intelligence algorithms. Despite their clear advantages and popularity in extracting relations from European languages[112]-[114], the adoption of this approach to extract relations from Arabic text are limited. One of the most significant contributions is that by I. Boujelben et al in [115], where they present an approach that utilises a hybrid approach where linguistic modules are used to improve the output of the machine learning relation classifiers. The training data combine automatically extracted systaltic and semantic features with the manually annotated relation indicator in each word. The output of the machine learning process is a set of relation extraction rules that are subjected to an optimisation process to select the highest quality rules. Targeting generic (non-specific) domains contributes to the complexity of this interesting approach, and

therefore they propose to deploy hand-crafted rules to deal with some of the ensuing challenges such as handling relation negation and moderating the role of POS tags in determining the relation indicators.

It can be concluded that hybrid approaches can offer significant improvement over the individual rule-based and machine learning methods, especially for domain-specific relation extraction where there is a clear advantage in initially exploiting domain knowledge in hand-crafting the relations pattern matching rules, which in turn can generate a richer and more accurate set of training data for the machine learning relation extraction classifiers. In this regard, rule-based methods need to exploit the sophisticated Arabic grammar rules in order to counter the complexity of the Arabic language, not only in morphology but also in the syntactic sentence structure. For instance, most reported rule-based systems employ basic Arabic syntax grammar rules to define the linguistic relation p[116] patterns (token order), primarily using the three basic features of the relation language: subject (S), object (O) and predicate (P) [104]

Traditional Transformational (Generative) Grammar (TG) attempts to structure natural language as an abstract set of generalised syntactic rules that are detached from the context of use, which can be very useful for recognising ordered patterns of binary relations. However, Arabic language sentences often contain complex (high order) relations where one subject has several predicates or several objects with varying order of the features in the sentence [117]. On the other hand, the Functional theories of grammar consider the functions of language and its elements to be key to the understanding of linguistic processes and structures[118], thus emphasising the semantic and pragmatic properties of a language. Hence, we believe that Functional grammar offers a more flexible abstraction for modelling the complex Arabic language relations, and adopted the principles of Functional Discourse Grammar [119], an advanced version of Functional grammar, as the basis for building a novel approach to relation extraction from Arabic natural text. As indicated in the motivation section, the problem domain of choice is the financial domain, and this case study particularly focuses on articles related to news about the stock market, which contain very complex sentences that is bound to challenge the proposed relation extraction approach.

7.3. **Arabic relation extraction based on Functional Discourse Grammar**

Functional Discourse Grammar (FDG) was adopted as the basis to build our relational extraction algorithms as it emphasises the semantic and pragmatic properties of the language, thus facilitating the identification of relation patterns in the Arabic language complex sentence structure that often contains complex relations where one subject has several predicates or several objects with varying order of the features in the sentence.

7.3.1. Overview of Functional Discourse Grammar

The Theory of FDG appeared in Amsterdam University, at the hands of the Dutch linguist (Simon C. Dik) and his colleagues in the 1970s[120]. It was first discussed in the context of processing the Arabic language in the beginnings of 1980s at the hands of the linguist Ahmed Moutaouakil, a professor in Mohammed V University, Rabat[121]. Since the formation of the Functional Grammar Theory, it has aimed to fulfil the linguistic theorisation of "Computational efficiency" in natural language understanding, and the proceeding models of the theory have considered this in formalisation and modelling.

The Functional Grammar theory witnessed successive models through which the theory had evolved, starting from the nucleus model reaching the proposed models in the last generation which has come to be known as (The Theory of Functional Discourse Grammar) developed by Kees Hengeveld and Lachlan Mackenzie [119]

The theory is established based on the idea that language has three primary functions, (1) pragmatic function [topic and focus, theme and tail], (2) semantic functions [agent, patient, recipient], (3) syntactic functions [subject, object].

This study is focused on the computation of the semantic functions of Arabic in the economic discourse (stock market). Figure 28 explains how the principles of functional discourse grammar we applied to the sentence structure of FDG relation extraction algorithm.

Operating on the semantic level of the grammatical component, process patterns are identified by a main predicate trigger word, which has corresponding arguments that consist of an agent and other complementary elements such simple entities (rate, date , number), or complex sub-phrase that might represent another process.

The next section presents the development of a novel relation extraction algorithm based on FDG.

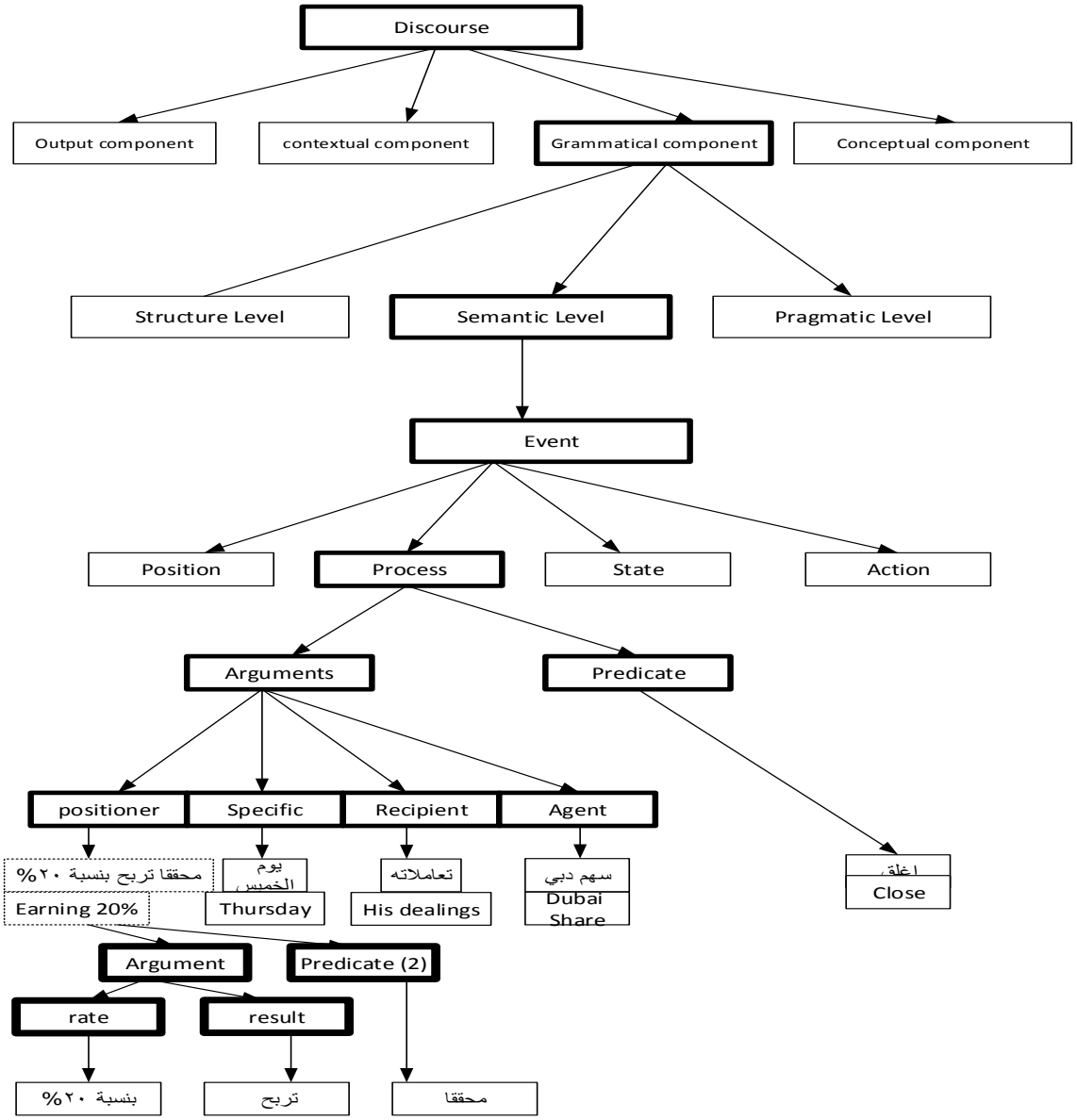


Figure 27: Function Discourse Grammar architecture

7.3.2. Relation Extraction algorithm based on Functional Discourse Grammar

As discussed in the literature survey, rule-based efforts in Arabic relation extraction mainly employed the syntax grammar that relied on the sequencing of the subject, verb, and object

in the sentence structure. In this research, we introduce a novel approach for semantic relation extraction that relies on the functional discourse grammar (FDG) rules. The algorithm is based on Functional Grammar (FG), in particular the semantic function grammar level. This type of grammar is a set of rules and processes that govern the semantic of sentences in a given language regardless the structure of sentence. FDG approach relies on two main terms: a predicate and an argument. The predicate term specifies the kind of state of operation (i.e. an event or sequence of events of a specified kind that has described in the sentence), and the argument terms express the result of the operation. The main argument term is the agent, which can be used to determine the event. The other argument terms are based on the meaning of predicate such as (Money, Number, date). The main target of the algorithm is to extract several types of relations based on the semantic structure of the relation that is defined in the economic Knowledge-base as defined in the domain analysis (knowledge modelling) stage, such as relation between Organisations and Number, relation between Location and Numbers, and relation between Organisations and Date. Figure 29 the mechanism of Function Discourse Grammar algorithm.

The semantic modelling of the problem domain maps naturally to the realm of Functional grammar. The semantic function of the Functional Grammar corresponds directly to the Semantic Web knowledge representation in RDF triples that are encoded as a set of subject, predicate and object nodes.

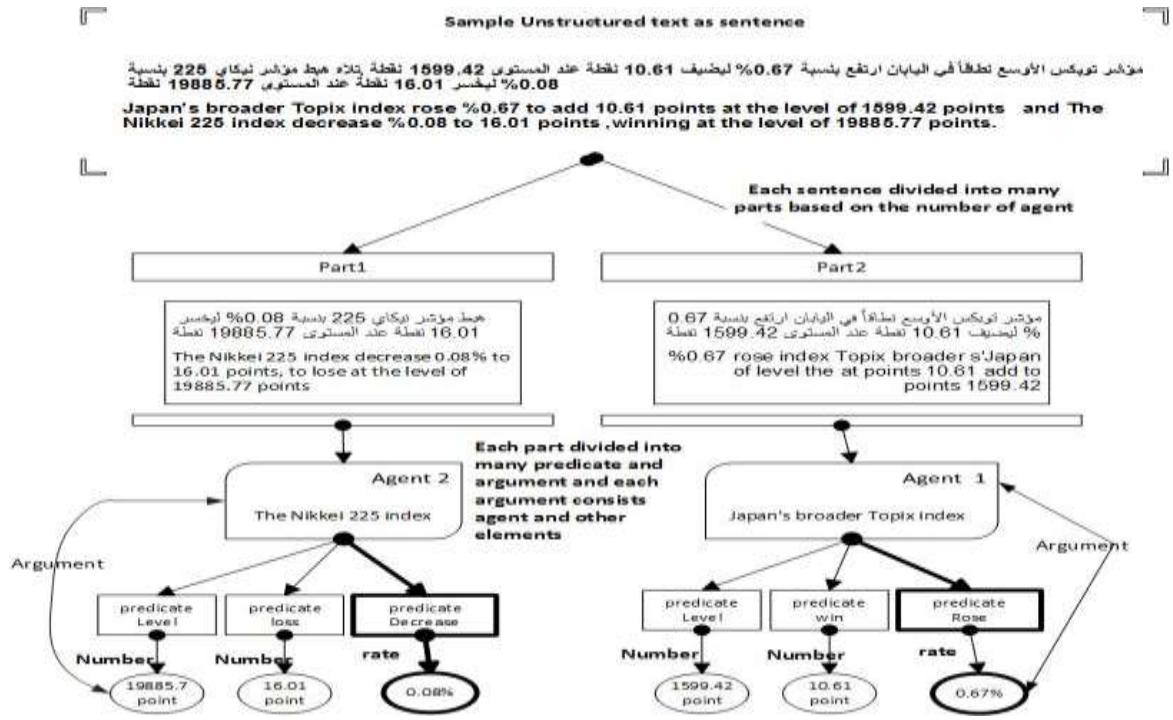


Figure 28: The mechanism of FDG algorithm

The main strategy of our approach is capturing the first proper nouns as agent (“مؤشر توبكس”, “Japan's broader Topix index”), and then find the first predicate (“هبط”,” Decrease”) that could be verb/noun/adjective. This predicate determines the probable type of the relation’s target (object) such as (“10.61 points”); this describes the result of the action in the sentence to complete the triple of the relation components. Moreover, the algorithm will be used to extract the relations in the sentence by matching them with the relations in the knowledge-base. Table 31 illustrates how FDG algorithm extracts different types of relations from the sentence listed below.

Table 31: Extracting different types of relations using the FDG algorithm

The general index of the Abu Dhabi's stock market has finished the session on Thursday by decrease 1.24% with a loss of 62.43 points to close at 4.992.52 points. أنهى **المؤشر العام لسوق أبوظبي** تعاملاته في جلسة يوم الخميس على هبوط واضح نسبته 1.24% بخسائر بلغت 62.43 نقطة، ليصل إلى مستوى 4.992.52 نقطة									
	1	2	3	4	5	6	7	8	9
Agent	المؤشر العام لسوق أبوظبي								
Predicate		أنهى		هبوط		بخسائر		ليصل	
argument			الخميس		1.24		62.43		4.992.52
Relation 1									
Relation 2									
Relation 3									
Relation 4									

The steps below detail function description of propose approach:

1) Identify the named entities in the text by using specific token called General Annotation Token (GAT), and add the terms features to each GAT token such as (token name, POS , kind, type and root). These features will be used to identify the type of the token in the sentence. GAT tokens are then appended to a Named Entity Array (NE_Array) according to their position in the sentence. Table 32 shows the classify the features to GAT tokens.

Table 32: The list of features of each GAT tokens

سهم الانوار القابضة منى باكبر الخسائر اليوم بنسبة 2.42% ليغلق على 0.322 ريال Al Anwar Share was the biggest loser today by 2.42% to closed 0.322 Real					
	entity name string	Features			annotation token
No		kind	Root	PoS	
01	سهم الانوار القابضة	Share		NNP	GAT
02	منى	predicate	منى	VERB	GAT
02	الخسائر	predicate	خسر	NNP	GAT
03	%2.42	Percentage		NUMBER	GAT
04	ليغلق	Predicate	غلق	VERB	GAT
05	ريال 0.322	Money		NUMBER	GAT

2) Find the first agent in the array (sentence), which is the first token of the proper noun (NNP) category. For example, "سهم الانوار القابضة", "Al Anwar Holding shares", and assign the name entity feature of the agent.

3) Starting from the first position in the array, process the GAT tokens in the array (sentence) sequentially as follows:

• If the entity name string feature of the first GAT token in the array (sentence) equals the entity name string feature of the agent then skips to the next element.

• If the type of the GAT token is predicate assign it as predicate and read the root feature of the token.

• If there are two sequential predicates in the sentence then ignore the first one and assign the second as the main predicate.

•If assigned the agent as the named entity and the predicate then the algorithm will assign the third part of relation based on the root of predicate for example:

- If the root if predicate is (rise"خفض", "رفع" decrease) the result term is rate value (percentage)

- If the root if predicate is (open"اغلق", "فتح" close) the result term is date value.

• After classifying the relation components (agent, predicate and result terms) then will matching these components with the triple of relations that retrieved from knowledge-base.
• If the triple of relation which extracted from unstructured text matching the triple of the relation in the knowledge-base, then the several features will be added to relation such as relationship subject, relationship object, relation name ,kind of relationship subject and kind of relationship object and then will use these feature to inject each relation has extracted into knowledge-base . And go to step 4 to extract next relation.
4) Choosing the next agent in the sentence, if the algorithm finds the new NE has category as NNP , and does not has relation with the first agent then will assign the new NE as the new agent and return to step 2.

The following examples illustrate the operation of FDG algorithm in extracting relations from different Arabic sentences forms based on the position and number of the agents, predicate and argument type.

Figure 30 illustrates the relation extraction process from a sentence that contains one agent and several predicates with different argument type.

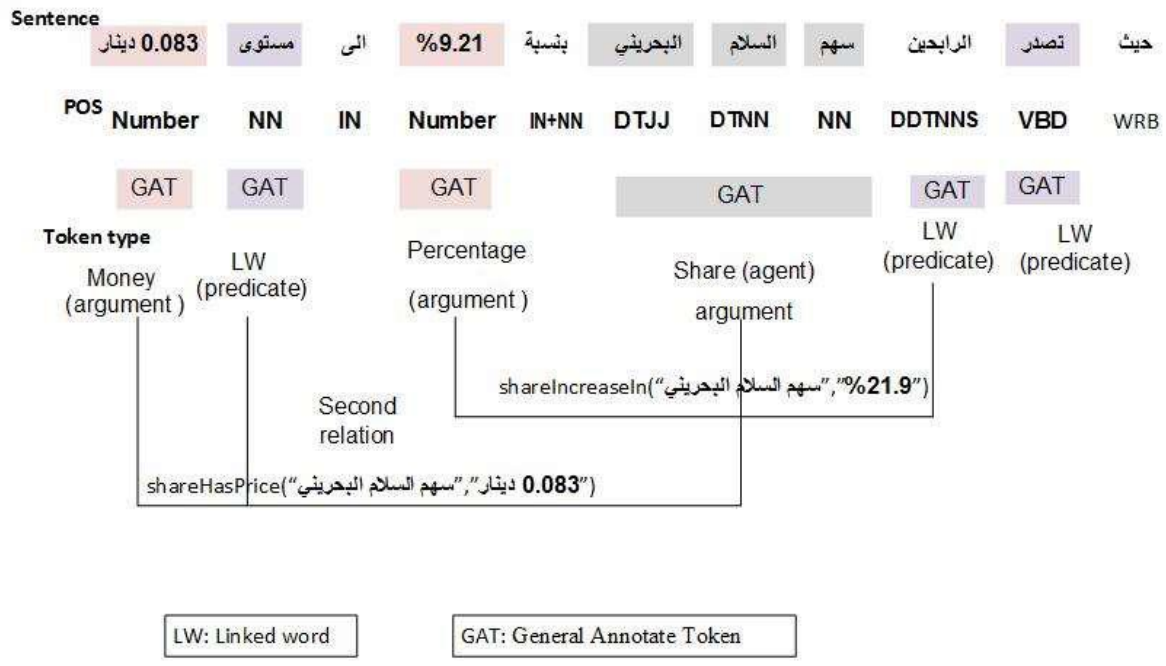


Figure 29*:* FDG approach for extracting the relation from sentence with one agent and several predicates of different types

The second example shows how the FDG algorithm processes a sentence containing two agents where each agent has two predicates and two argument types. Figure 31 illustrates the relation extraction process from a sentence that contains several agents and several predicates and goals.

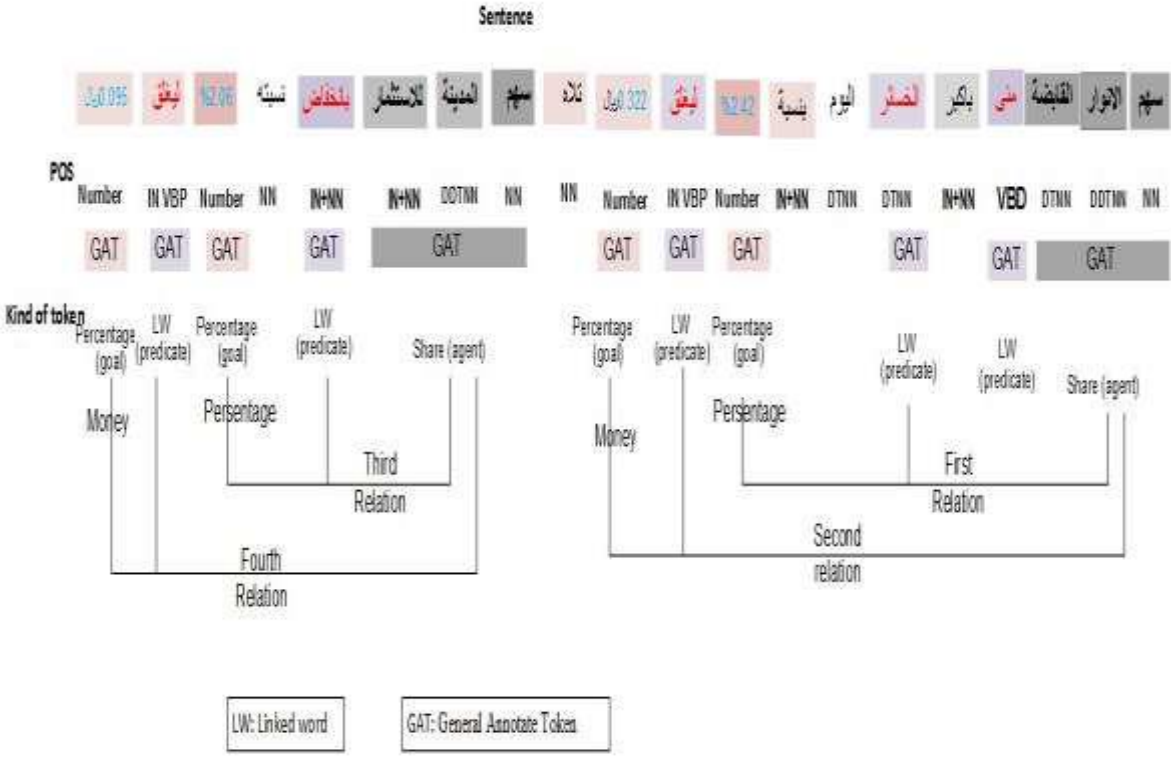


Figure 30: FDG algorithms extracts relations from a sentence with several agents and several predicates and goals

In the Figure 32 is the FDG algorithm for complex Arabic relation extraction expressed as pseudocode.

```
Read list of GAT AS NE_Array
For (each GNE in array)
{
      PNE= NEArray.PoS    (the PoS of the first NE)
      If PNE=NNP
      {
        Agent=NEArry[I].string  (the name of first NE as agent )
      }
  For (each GNE in array)
   {
      TNE= NEArray[J].PoS
    If (TNE=NNP)
      {
        Agent1=NEArray[J].string (the name of first NE as agent )
          If (Agent=Agent1)
           {
               Agent=Agent1
               T_agent=Agent.type  (the type of the GNE  such as Share )
           }
      }
      If (PNE=in the list of the Synonyms )  (as predicate )
       {
            Predicate =NEArray[J].string  (the name of first NE as predicate )
            TP = NEArray[J].Root  (save the root of predicate )// value of predicate
       }
    // choose the type of result based on the predicate features (TM= time , AC= number
          and       NNP = named entity )
     If (TNE=TM or TNE=AC or TNE=NNP)
        {
          Goal=NEArray[J] (the name of first NE as goal )
           TG = NEArray[J].type  (save the type of goal ) // the type of  goals

        }
      If( T_agent ,  TP and TG match the list of relation )
      {
        Assign the name of relation and inject to KB
       }
  }
}
```

Figure 31: Pseudocode detailing the implementation of the linguistic analysis for composite name extraction.

In order to illustrate the advantage of utilising semantic functions grammar in FDG relation extraction algorithm over traditional (syntactical) rule-based systems, we drive the example in Table 33 below. The function discourse grammar offers advantages over the traditional rule-base systems

Table 33: Example showing how the FDG algorithm reduces the number of instance classes

أما البلدان التي سجلت أدنى معدل للناتج المحلي الإجمالي فهي البحرين بنسبة 5.5٪ والأردن بنسبة 5.3٪. Countries with the lowest GDP ranking were Bahrain with 5.5% and Jordan with 5.3%.					
Instance relation					
Relation subject	Relation object	Trigger word	FDG rules	Synctactic rule-based systems	State of instance relation
البحرين (Bahrain)	5.5٪	أدنى lowest	Extracted	Extracted	True positive
الأردن (Jordan)	5.3٪.	أدنى lowest	Extracted	Extracted	True positive
البحرين (Bahrain)	5.3٪.	أدنى lowest	NO	Extracted	False positive
5.3٪.	الأردن	أدنى lowest	NO	Extracted	False positive
5.3٪.	البحرين	أدنى lowest	NO	Extracted	False positive
5.5٪	البحرين	أدنى lowest	NO	Extracted	False positive

7.3.3. Experimental Evaluation

A set of experiments have been conducted in order to evaluate the performance of the proposed algorithm. The experiments were conducted using the ARB_ECON corpus that was collected from different Arabic economic news websites. the details of this corpus is presented in Chapter 4 section 3. The specifications of the corpus is listed below in the Table 34.

Table 34:The specifications of the ARB_ECON corpus

Annotation name		N of Annotation name	Discretion
Document		1300	Number of document
Sentences		6055	
Token		189290	
NEs		24977	
	Location	3219	City /country
	Person	1619	
	Organisation	5214	Index/Share/sector/company Stock market
	Date	4106	Date /day/year
	Numbers	10819	Price/number of point/percentage/

The first set of experiments evaluates the performance of our algorithm when extracting relations from sentences with varying structural complexity, and the second set of experiments tests the algorithm's performance in extracting different relation types.

We divided the first set of experiments into three categories: the first is extracting the relation from the simple sentences, the second is extracting the relations from the complex sentences and the third is extracting the relation from more complex sentences.

The results of the first category experiment are shown in Table 35; it evaluates the extraction the relation from the simple structured sentences that has one agent and one predicate.

Table 35*:* The Recall, Precision and F–measure for the first experiment

Corpus	Recall	Precision	F-measure
ARB_ECON Corpus	0.84	1	0.91

The results of the second category experiment are shown in Table 36; it evaluates the relation extraction from the complex sentences that have one agent and more than one predicate.

Table 36: The Recall, Precision and F–measure for the second experiment

Corpse	Recall	Precision	F-measure
ARB_ECON Corpus	0.90	0.86	0.88

The results of the third category experiment are shown in Table 37; it evaluates the relation extraction from more complex sentences that have more than one agent, and where agents have more than one predicate.

Table 37: The Recall, Precision and F–measure for the third experiment

Corpse	Recall	Precision	F-measure
ARB_ECON Corpus	0.94	0.76	0.84

Figure 33 shows the results of the algorithm based on the experiment methods

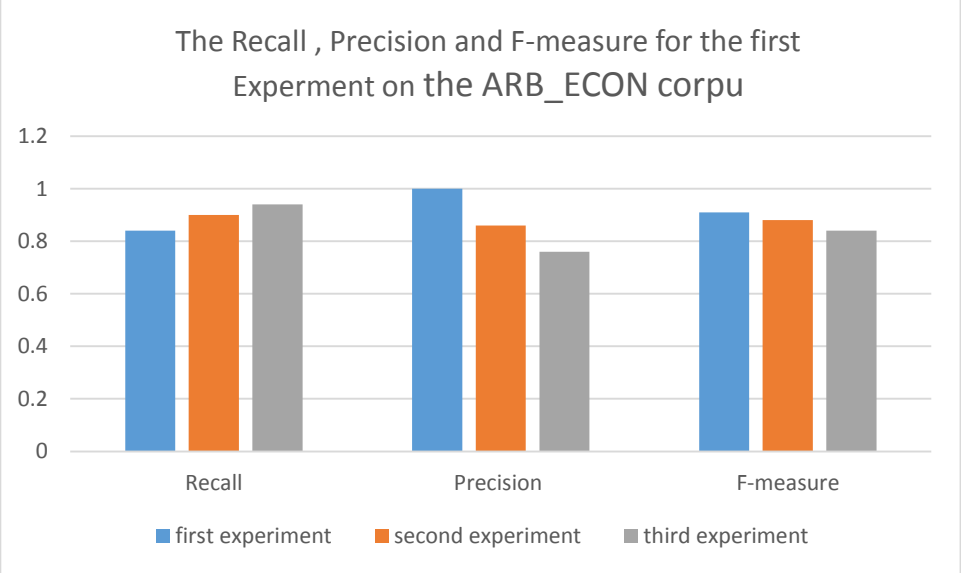


Figure 32: Graphical represented Precision, Recall and F-measure for testing FDG algorithm on the ARB_ECON corpus

Figure 33 collates the results of the first set experiments. In general, we can observe that the Precision of relation classification drops slightly with the increased sentence complexity, but overall it is evident that the algorithm achieves very good accuracy in relation extraction

from relatively complex Arabic non-structured text, scoring a commendable (0.84 F-measure) for the more complex sentence structure. It is curious, however, that the Recall exhibits almost opposite performance to Precision, where the lowest Recall (0.84) was registered against the relation classification of the simplest sentence. This can be explained by the fact that the simple sentence structure in the first experiment usually contains one subject, one object and one predicate (trigger word). In this type of sentence, if the algorithm fails to detect one of the elements of the relation, then the algorithm cannot extract this relation, thus reducing the overall Recall rate. On other hand, the second experiment and third experiments use complex structured sentences, which normally contain several subjects, predicates and objects. Therefore, taking into account that one of the main features of our FDG algorithm is that it can detect relation patterns in complex sentences, this will result in a higher Recall rate for sentences that are richer with relation components, similar to the sentences exemplified Table 38 below.

Table 38: An example explains why the recall is giving opposite result to precision

المقابل هبط سهم الخزف الاردنية بنسبة 4.63% تلاه سهم الاتحاد العربي خسر بنسبة 3.37% وسهم الاردنية الاماراتية ارتفع بنسبة 2.78% On the other hand, the Jordan Ceramic share was the biggest loser by 4.63%, followed by Arab Union that lost 3.37% and the United Arab Emirates share rose 2.78%					
No	Subject	Predicate	Object	State one	State two
01	سهم الخزف الاردنية Jordanian alkhuzf Share	هبط Decrease	%4.63	True positive	True positive
02	سهم الاتحاد العربي Alaitihad Alearabiu Share	خسر Decrease	%3.37	True positive	True positive
03	سهم الاردنية الاماراتية Jordanian Emirates Share	ارتفع Increase	2.78%	True positive	True negative
04	سهم الاتحاد العربي Alaitihad Alearabiu Share	خسر Decrease	2.78%		False positive

From the example above, if we consider that the algorithm can detect all the named entities and trigger words in the sentence, then the algorithm will extract all the possible relations, which are relations 1,2,3 in Table 36. But, if we consider that the algorithm failed to extract the trigger word (" ارتفع","Increase") then the algorithm will fail to extract the relation 3 in the Table 36 then the FDG algorithm will extract a new relation 4 as false positive relation. As a result, the probability of extracting the false positive relation in the complex structured sentence is more than the simple structured sentences. However, this complexity of the sentences will lead to increase in recall value and decrease the precision value.

In addition, the Recall is not an important factor in this situation because we are using the knowledge-base approach effectively which make it very high. It is because the classes of the relations are already defined in KB. This is different from the approaches that use the statistical techniques because it does not have the knowledge about the entities that can take part of the relation. However, the Recall is not useful in this case and F-measure is not either because the Recall destroys the assessment of the result. Therefore, we are going to rely on the precision in this work.

The results of the second set of experiments evaluating the performance of the FDG algorithm for extracting the different types of relations are illustrated in Table 39 and Figure 34.

Table 39: The System performance after applied the second experiment

Name of relation	Precision	Recall	F-measure
Org – PercentageOfIndicatorFluctuation	0.91	0.98	0.95
Org- DateOfEconomicActicvity	0.81	1.0	0.89
Org-FinancialValue	1.0	0.78	0.88
country – Industry	0.85	0.78	0.82
country – IncreaseGDP	0.58	0.91	0.71
country – DecreaseGDP	0.3	1.0	0.46
country – IncreaseInflation	0.38	0.80	0.52
country – DecreaseInflation	0.60	0.85	0.70

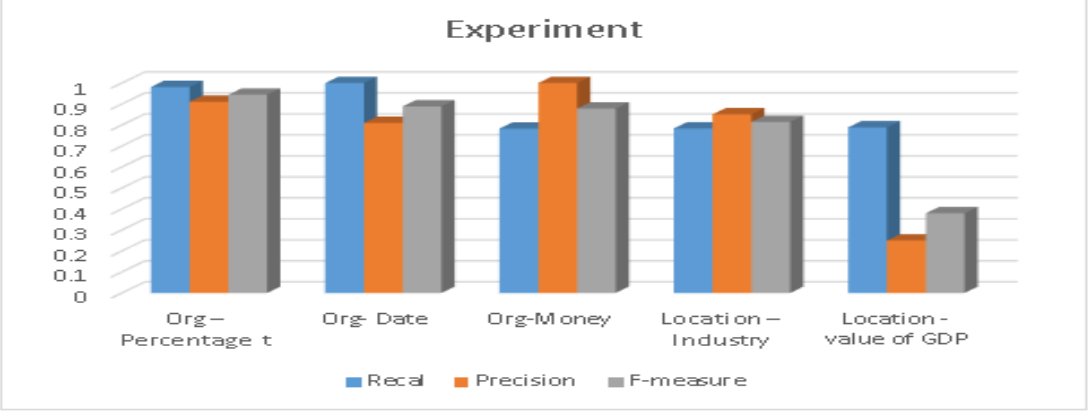


Figure 33*:* The Recall Precision and F–measure for the second method on the ARB_ECON corpus

The experimental result is quite satisfactory for most relations, scoring, for instance, (0.91) extraction precision for the relation between organisation and PercentageOfIndicatorFluctuation and (0.85) for the relation between Country and industry. However, there are some limitations that affected the accuracy of the relations extraction algorithm as detailed below.

7.3.4. Limitations of the algorithm implementation

The experimental evaluation revealed that the FDG algorithm's performance was impacted by errors in the document text and the particularly complex sentence structure as follows:

- **Missing or incorrect named entity**

Some of the ambiguous relations are caused by an incorrectly named entity or missing named entity in the sentence. As mentioned previously, the sentence sometimes contains several agents that represent the named entity. This agent is considered the main part of the triple for extracting the relation. In some cases, the named entity is undetected in the sentence which was caused by the missing or incorrectly named entity. As a consequence, the algorithm could not extract the relation in the example below:

%سهم النورس تراجع بنسبة 1.21% ليغلق على 0.652 ريال، سهم سيمبكورب هبط بنسبة 1.00
Al Nawras shares retreated 1.21% to close at 0.652 SAR. Sembcorp decrease1.00%

In the above sentence, there are two relations, the first one is shareDecreaseBy (i.e. سهم النورس,1.21) and the second is "shareDecreaseBy (i.e. ("سهم سيمبكورب", "Sembcorp"),1.00)". However, the named entity recognition process identified only one named entity and discarded the other entity, because the (سهم سيمبكورب) was translated from foreign language, which caused the Arabic named entity pipeline cannot recognise this name entity; and, then the FDG algorithm considered the entity "1.00 "is related to the first named entity (سهم "النورس"). Table 40 explains the absent or incorrectly named entity error.

Table 40 : An example showing the missing or incorrect named entity

No	Relation			State	
01	سهم النورس	shareDecreaseBy	1.21%	selected	correct
02	سهم سيمبكورب	shareDecreaseBy	1.00%	Not selected	
03	سهم النورس	shareDecreaseBy	1.00%	Selected	Not correct

- **Missing the word that describes the relation (trigger word)**

Some ambiguous relations are caused by the algorithm's failure to recognise the word that describes the relation (Triger word) in the sentence due to its absence in relation indicator words list in the knowledge-base. In the example below the predicate (trigger) word caused the algorithm to be unable to recognise the relations.

"حيث تصدر الرابحين سهم السلام البحريني بنسبة 9.21%الى مستوى 0.083 دينار"
"where the Al Bahrain AL SALAM share was the Leaders the winners by21.9 % to 083 .0 dinar"

In the sentence above, the root of the word ("الرابحين","winners") is not included in the knowledge-base as a synonym for the shareIncreseBy relation, which caused the algorithm to fail to recognise this relation shareIncreseBy (" **سهم السلام البحريني**","**9.21** ").

- **Nested Named Entity**

One of the problems that affected our results was the nested named entities. In some cases, the named entity contained other named entities which caused unexpected relations to be extracted. See Table 41.

Table 41: Nested Named entity problem

صعد المؤشر العام لبورصة قطر خلال جلسة اليوم بـ 231 نقطة ما يعادل 2.6% The general index of the Qatar Exchange rose during the session today 231 points equivalents to 2.6%				
Relation	First entity	Predicate	Rate	State of relation
Relation 1	المؤشر العام لبورصة قطر Index	صعد indexIncreaeBy	2.6% Number	Correct
Relation 2	قطر Country	صعد IncreaseGDP	2.6% Number	Incorrect
Relation 3	قطر Country	صعد IncreaseInflation	2.6% Number	Incorrect

As shown in the table 39 above, the algorithm has extracted three different relations. The first relation is between the Index and Numbers which describes the state of the index, and the second relation and the third relation between Country and Number which describes the state of the economy for the country (i.e. GDP, Inflation). In the sentence above, the target relation is the first relation, yet the algorithm extracted two more unexpected relations. Because the named entity "i.e. المؤشر العام لبورصة قطر, the general index of Qatar" contains another named entity "i.e. قطر, Qatar" this causes the algorithm to consider that there was a semantic relation between the Country and Percentage.

- **Failing to extract relations with especially complex sentence structure**

These complex sentence structures contain several relation trigger words within the same clause. This can lead to extract additional incorrect relations. An example of these relation is shown in Table 42 below.

Table 42: Challenge in extracting especially relation from especially complex structures

No	Arabic sentence	Entities		Trigger word of relation		Relation extracted	
		Entity 01	Entity 02	Trigger Word 1	Trigger Word 2	Relation1	Relation2
1	التضخم في السودان يتجاوز 30٪ Inflation in Sudan Increase 30%	السودان Sudan	30%	Increase	Inflation	True positive	
2	التضخم في السودان يتجاوز 30٪	السودان Sudan	30٪	Increase			False Positive

	Inflation in Sudan increase 30%						

The first sentence in Table 40 above represents the relation that describes the increase value of inflation for a specific country (i.e.السودان , Sudan), and this relation has two trigger words to describe the relation (i.e. Increase, Inflation). The strategy used to extract the semantic relation by using the FDG algorithm is based on one trigger word that refers to the relation between two entities as the predicate. In the sentences above, the correct relation between two entities is increaseInflation is based on two trigger words (trigger word1, trigger word 2). Therefore, to extract this type of relation we need to determine the two trigger words. The first trigger word refers to the state rate for the country such as increase and decrease, and the second trigger word refers to the type of state such as Inflation to describe the relation. For this reason, the FDG algorithm has exacted the relation 2 in table 48, but this relation incorrect because the FDG algorithm used on the first trigger word and ignored the second trigger word. The results that are shown in table 40 are limited and unexpected.

For the Economic problem domain, the GDP and Inflation events within the text are extremely important economic indicators and need to be adequately analysed if the textual analytics output is to be used in a decision support mechanism. Hence; as mention previous, the main propose of our framework is building the high-quality knowledge-base in order to uses in supporting decision making for recommender system. Therefore, we investigated the integrating of rule-based FDG algorithm and Machine Learning based techniques in a hybrid approach that addresses some of the limitations of the rule-based approach individual.

7.4. **Hybrid relation extraction approach**

The evaluation of our knowledge (rule) based FDG algorithm revealed its limitation in extracting relations from sentences with particularly complex sentence structure complex, such as sentences where the relations are described by more than one trigger word. Therefore, we investigated integrating the rule-based FDG algorithm and Machine Learning based relation classification in a hybrid approach in order to address the afore-mentioned limitations. The extracted relations by the rule-based FDG algorithm will be used to identify candidate relation instances in training datasets.

7.4.1. The proposed approach

The hybrid approach relies on the FDG algorithm to extract the relations based on the first trigger word, then Machine Learning classification is used to extract the multiple relations that might exist in the same clause. Figure 35 shows the sentences consists relation between Country and Percentage hasIncreaseInflation("مصر","29.6 "), and this relation has two trigger words that are used to predict the class relation: first is ("يقفز"," rises") and second is ("التضخم"," Inflation") .

التضخم في مصر يقفز إلى 29.6 في المئة

The Inflation in Egypt rises to 29.6 percent

Figure 34: Example shows the relation has two trigger words

In this approach, we focused on improving four types of relations between Country and Numbers: hasIncreasGDP, hasDecreaseGDP, hasIncreaseInflation and hasIncreaseInflation.

The hybrid approach proposed for relation extraction is illustrated in figure 37. The hybrid relation extraction approach consists the following four stages;

1 The main purpose of this stage is to identify the candidate relation instance based on the first trigger word.
2 Generate the feature set associated with the relation instances. Moreover, we used Genetic Algorithm (GA) to optimise the feature selection process.
3 Preform the mapping between the relation instance and relation class is applied to build the training datasets.
4 Building the ML classification model and testing the ML algorithm.

We used two supervised ML algorithm which are applied for relation extraction: Support Vector Machine (SVM) and k-Nearest Neighbours algorithm (KNN), and then the Genetic Algorithm-based (GA) method is used to obtain the best feature subset selection. Both algorithms have been utilised successfully in most NLP tasks including and Relation Extraction [112]. SVM is a linear or non-linear classifier, which is a mathematical function that can distinguish two different kinds of objects [44], [122]. The second algorithm applied is k-Nearest Neighbours algorithm (KNN). It is a non- parametric algorithm which is used for classification and regression. It is a simple algorithm showing accurate results with a small number of features [112], [123]. Figure 36 shows Architecture of Hybrid approach proposed.

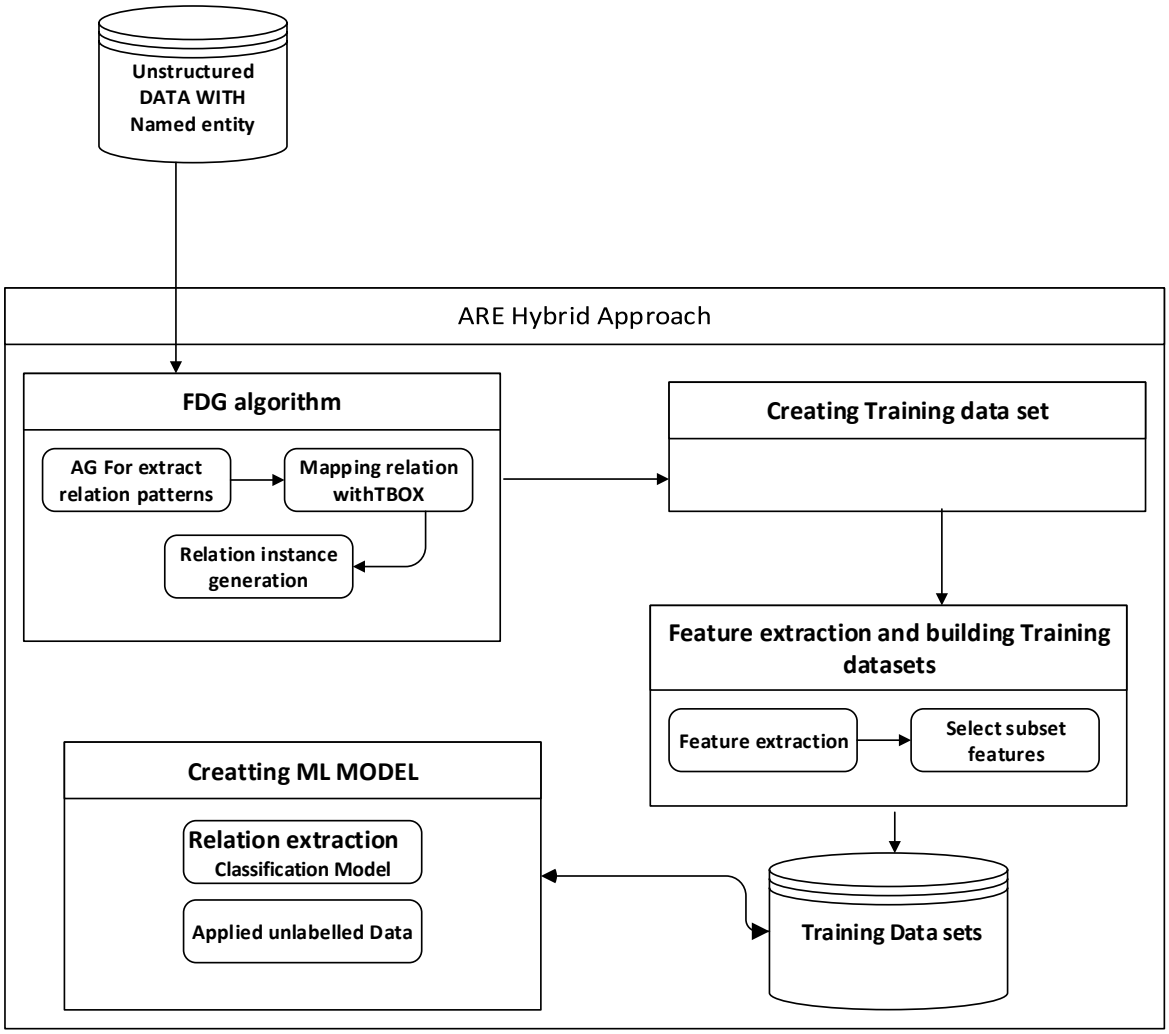


Figure 35: Architecture of Hybrid approach proposed

7.4.2. Identify candidate relation instances

Using the rich domain-specific relation taxonomy in the semantic knowledgebase, the FDG algorithm achieved excellent Recall in identify domain relevant relations. In the context of the hybrid approach, the FDG algorithm has been used to identify and annotate the relation instance for any entity pair in the document sentences. The relation instance that are extracted from the document have been annotated manually in order to determine the True Positive Relation and False Positive Relation. In Table 43, the output of FDG algorithm for extracting the relation between country named entity and number. This output has identified relations 1 &3 as True Positive relations and relation 2 as False Positive relation.

Table 43: An example shows the list of relation are extracted from document

وتليها مع نفس النسبة روسيا التي بدت تعاني بشكل اكبر واما ثالثاً فحلت الامارات العربية حيث من المتوقع ان تشهد هبوط من حيث الناتج المحلي الاجمالي بنسبة 1.5%. وتتصدرها دولة الفيليبين والتي من المتوقع ان يرتفع ناتجها الاجمالي الى نسبة تتجاوز 1%		
NO	Relation	Type relation
01	الامارات العربية حيث من المتوقع ان تشهد هبوط من حيث الناتج المحلي الاجمالي بنسبة 1.5%. Arab Emirates hasDecreasInflation 1.5 %	True Positive Relation
02	روسيا التي بدت تعاني بشكل اكبر واما ثالثاً فحلت الامارات العربية حيث من المتوقع ان تشهد هبوط من حيث الناتج المحلي الاجمالي بنسبة 1.5%. Russia hasDecreasInflation 1.5 .%	False Positive Relation
03	الفيليبين والتي من المتوقع ان يرتفع ناتجها الاجمالي الى نسبة تتجاوز .% Philippines hasDecreasInflation 1.5 .%	True Positive Relation

7.4.3. Building the training Datasets

The Arabic language suffers from the lack of resources and particularly resources for relation extraction task purposes. Some Arabic corpora are annotated, namely ACE multilingual training dataset, but it is not freely accessible. For these reasons, we decided to build our own datasets for training and testing the ML approaches. We have relied on FDG algorithm to extract relation instance for building the training datasets. This type of training dataset will be used with the hybrid approach. However, there are some of instances where we needed to build more comprehensive baseline. We need to complement FDG algorithm output by manually annotate relations. To achieve this, we used JAPE rules as the regular rule to annotate any two entities in the sentence in the unstructured document and which represent the entities of the target relation. These relations are assumed to be a class instance

in the training datasets, and these training datasets will use for training and testing other target relations. Table 44 shows the list of training datasets based on the number of documents and the relations.

Table 44*:* List of training datasets

No	Name of Dataset	No of Document	The relations name
1	Relation between Share and percentage	307	increaseBy decreaseBy
2	Relation between Index and percentage	121	increaseBy decreaseBy
3	Relation between country and GDP	165	increaseGDP decreaseGDP
4	Relation between country and Inflation	231	increaseInflation decreaseInflation
5	Relation between Sectored and Share	79	increaseBy decreaseBy belongTo
6	Relation between Share and Money	81	hasValue
7	Relation between Country and Date	172	hasGDPDate hasInflationDate
8	Relation between country and percentage for Hybrid approach	244	countryExportGoods countryImportGoods hasIncreaseInflation hasIncreaseInflation
9	Relation between Country and Industry	111	countryExportGoods countryImportGoods
		1511	

In the table 45, the main characteristic for each training datasets have explained such as type of data sets, number of true positive instances and true negative instances

Table 45: Illustrating the characteristic of training datasets

Dataset	Name of relation	Mentioned relation	
Share to percentage	IncreaseIn	True positive	329
	DecreaseIn	True positive	383
True positive instances	712	True negative Instance	1464
Country Inflation	increaseInflation	True positive	168
	decreaseInflation	True positive	108
True positive instances	276	True negative Instance	20
Index Date	hasOpenDay	True positive	60
	hasOpenDate	True positive	13
	hasCloseDay	True positive	21
	hasCloseDate	True positive	10
True positive instances	132	True negative Instance	30
Sector – Share	belongTo	True positive	329
True positive instances	329	True negative Instance	05
Country - Date	GDPDate	Right Instance	135
	InflationDate	Right Instance	87
True positive instances	222	True negative Instance	84
Country – number Hybrid approach	hasGDP	Right Instance	181
	hasInflation	Right Instance	137
True positive instances	318	True negative Instance	479

7.4.4. Feature extractions

The feature is an individual measurable property of the phenomenon being observed. Also, it is a crucial step for algorithms in effective pattern recognition, classification and regression. In this approach, the backbone of relation extraction is the sentence, and each sentence may contain many clauses and the clause sometimes contains two or more entities. In these studies, several learning features have been assigned to build the training dataset. These features are divided into three categories: lexical feature, semantic features and numeric features. Some of these features have been used in different studies [44], [115]. Namely features 13, 17, 19 as listed in Table 46 below, according to my knowledge this set of features not exploited in other relation classification works in Arabic documents. We believe that these features will assist to the ML model to classify the relation instance as true or false. In the lexical features, the window size is four words which was established heuristically. We have tried to use other number of words, but we have found it not effected.

Table 46: List of the features

No	Feature Category	Feature	Name of feature
1	Lexical feature	L_Ws_B_NEs	List of words between NEs
2		POS_w1_b1	category of the first word before the first NE
3		Str_w1_b1	string of the first word before the first NE
4		POS_w1_a2	category of the first word after the second NE
5		Str_w1_a2	The first word after the first NE
6		Str_w2_b1	The two words before the first NE
7		Str_w3_b1	Three words before the first NE
8		Str_w4_b1	The fourth words before the first NE
9		Str_w2_a2	The two words after the second NE
10		Str_w3_a2	The Three words after the second NE
11		Str_w4_a2	The fourth words after the second NE1
12		PoS_ws_B_NEs	PoS of words between NEs
13	numeric features	LengthOfRelation	The lent of relation
14	Semantic features	order	Direction of relation e.g.

15		Domain	The Kind of relation subject
16		Range	The Kind of relation object
17	numeric features	N_Of_FirstNe	How many times appears the class type of subject relation in the relation
18		Distance	the number of words between the NEs
19		N_Of_secondNE	How many times appears the class type of object relation in the relation

7.4.5. Establishing the ML algorithms' valuation parameters

Common evaluation methods have been used for the ML algorithms, such as the holdout test and K-fold cross validation. In K-fold cross validation the system grouped the documents into K partitions of equal size and each partition is used in turn as test set, with all remaining documents as the training set, while the holdout test is used to randomly select documents as the testing data with the other documents used as the training set. We have applied an experiment that is aimed at obtaining the best subset features based on the SVM and KNN algorithms and K value. The k-fold cross validation selected the training datasets in order to make sure all the documents have been used in training and testing datasets, and also used different values for k (5,10) in order to obtain the best performance for the k-fold cross validation with the K volume. This experiment has applied on two datasets relation; Organisation - Organisation and Place – Number. The table 47 below comparison between SVM and KNN modal based on the terms of the F1-measure.

Table 47: Comparison between SVM and KNN modal

Dataset relation	Number of Relation	Number of document	Type of class entity	K-	F-m SVM	F-m KNN
organisation organisation Sector – Share	1	79	Proper noun Proper noun	5	0.696	0.622
				10	0.670	0.640
Country and percentage Place - number	2	231	Proper noun Number	5	0.771	0.763
				10	0.705	0.687

this experiment aimed to select optimum machine learning classifiers between SVM and KNN. The experiment was implemented based on different criteria such as the numbers of relation instances, number of documents, type of class entity and the value of the k-fold

parameter. Table 46 shows more details about the result of this experiment. However, our analysis asserts that the SVM relation classifier shows more accurate results in terms of F-measures than KNN when applied to Arabic datasets. This result is in close agreement with results of [122], [124].

7.4.6. Features selection

In this research, we utilize Genetic Algorithm (GA) to find best subset features to build a robust learning model and subsequently the relation classification [125]. The GA algorithm is one of several algorithms which are used to select the best features from a large volume of features. The GA uses randomized search and optimisation techniques conducted by the principles of evolution and natural genetics [126], and has been widely successfully applied the feature selection process[125], [127], [128].

We have chosen the GA as it demonstrated to be more accurate, albeit computationally more expensive[129]. However, since feature selection in our algorithm is a one-off process, the computational overhead is irrelevant.

Several experiments have conducted to evaluate feature selection method and ML algorithms. The first experiment aimed to evaluate the SVM algorithm by selecting the best features for its training data set. We adopted GA for feature selection. However, GA has several parameters that should be selected to relevant to our target domain. These parameters are population size, a mutation rate and crossover rate. The values of these parameters have been chosen heuristically. Table 48 shows an example of these experiments to compare between two groups of these parameters. The parameters of the GA which has been used to select the best features for SVM algorithm are (uniform rate = 0.6/mutation rate = 0.001/pop size = 50). See Table 48.

Table 48: The performance of the SVM algorithm based on a set of groups parameter of GA

N	Uniform Rate = 0.6/mutation Rate = 0.001/ pop Size = 50				Uniform Rate = 0.5/mutation Rate = 0.015/ pop Size = 30	
	Dataset	S.No	Subset features	F-m	Subset features	F-m
01	Country Percentage	01	1011011101000110110	0.762	0000011000100000100	0.663
		02	1010010110110101010	0.761	1100010001011111001	0.761
		03	0110010000100010010	0.755	0110011101010110001	0.705
		04	1101011101010110111	0.775	0101001010001000001	0.771
02	Country	01	0100001101010101100	0.878	1100001111000101001	0.902

	Industry	02	1000111111011110000	0.877	0000001101010111011	0.885
		03	1000011111010110000	0.877	0010011101010111011	0.860
		04	0000001110100011011	0.879	0111011101010111011	0.873
03	Share Sector	01	1111010111100111000	0.994	0001100101110110000	0.994
		02	0010111011010110010	0.994	0110001100010000110	0.994
		03	0110111100110110000	0.994	0011111110110111010	0.994
		04	1100010100010110110	0.994	0111101110010001110	0.994
04	Country Date	01	0000111011010101000	0. 710	1011111011001111000	0.727
		02	0100110011010100000	0.745	1010111011010100100	0.749
		03	1011100111011010110	0.741	1011001111110000110	0.743
		04	1011101011010010000	0.751	1011101111011000000	0.751
05	Country Number	01	1011011101101010100	0.749	0100101001100101100	0.736
		02	0101111011101000010	0.741	1111011011110101111	0.728
		03	1100011001111000010	0.744	0101100100101000100	0.748
		04	1001011011110110111	0.734	0001100111101100100	0.745

This experiment aimed to test our datasets from two aspects; firstly, to choose the best subset of features for each training dataset by using GA which are then used to evaluate the SVM algorithm. Secondly, to evaluate the most participation frequency of features in the selected subsets. The feature selected from of the previous experiment has been used to conducted this experiment. We have chosen the highest F-measure value of each subset of features for both groups. We have selected the 10 highest subsets features based on the term the f-measure value from 50 samples. Table 49 shows the Accuracy frequency of the features.

Table 49:Accuracy frequency the participation of features

Features	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Accuracy frequency	6	5	4	8	4	3	7	8	6	7	5	4	4	4	6	6	3	2	4

The result shows that the Lexical feature has outperforms the other type of features categories include the features 4,7,8 and 10. Consequently, the focusing on to choose the words before the first named entity and the words after named entity as features, could increase the approach performance.

We believe that lexical features could be more important for extracting relations by annotating some specific phrases as the features that appears in the sentences and related to

the second trigger word such as (“معدل التضخم”,” Inflation rate”) and (“الناتج المحلي الاجمالي”,” Gross domestic product”). On other hand, the participation of the numeric features caused low performance.

The Second experiment has conducted in order to prove the effectiveness of the feature subsets compares with the all features, we applied the machine learning classifiers algorithm by using all features. This result of this experiment is compared with the results of the features subset result. Table 50 and figure 37 shows listed the result for the sub features and all features.

Table 50: Comparison between the sub of features and the total features

no	Relation Name	Subset features	All feature
1	Country Percentage	0.775	0.649
2	Country Industry	0.879	0.761
3	Share Sector	0.994	0.990
4	Country Date	0.752	0.641

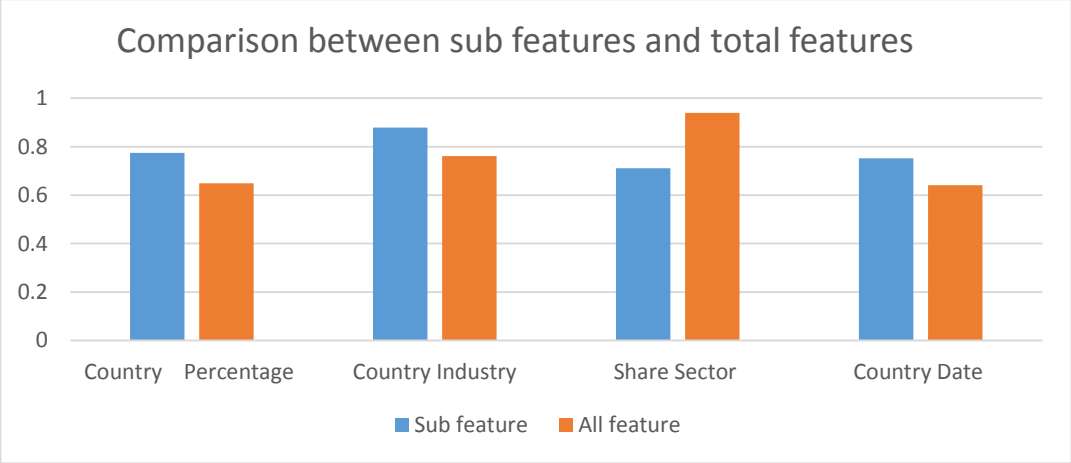


Figure 36:Evaluation the performance of ML classifiers based on sub features and total features

The result proved that the optimised set of features has improved the overall classification accuracy. It can also be observed that the SVM algorithm outperformed KNN and was hence adopted with the K=10 cross validation to build the relation extraction ML classification model.

7.4.7. Building relation classification models

In the previous sections, we created different training datasets and chose the best subset features for each training dataset in order to build the ML classification module. As highlighted previously, we decided to use the SVM algorithm to build the ML relation classification model. Each model will be evaluated by using unlabelled data to extract relations. Moreover, to obtain the optimum results by constructed training datasets, several experiments have been conducted that utilise the SVM algorithm to report on the appropriate implementations for successful information extraction for our domain. Each experiment has used about 30 documents to test the ML classification module which were not used in the training datasets. Two relation extraction ML models were created for the training datasets and testing the ML model.

7.4.8. Hybrid approach experimental evaluation and discussion

The experimental results of the hybrid approach are illustrated in Table 51 and Figure 38. Four relation classes have been tested to evaluate the hybrid approach, representing the relations that the rule-based FDG algorithm exhibited low accuracy in extracting. It is clear that the hybrid approach has significantly improved the classification accuracy for the aforementioned relations, taking the average F-measure from 0.6 to 0.77, but more significantly, improving the classification accuracy for one of the relation classes from 0.46 to 0.77.

Table 51:comparison (Precision, Recall, F-measure) between the hybrid approach and rule based approach

Relating name	Precision		Recall		F-measure	
	Hybrid	Rule-based	Hybrid	Rule-based	Hybrid	Rule-based
IncreaseGDP	0.73	0.58	0.875	0.91	0.80	0.71
decreaseGDP	0.75	0.3	1.00	1.00	0.77	0.46
IncreaseInflation	0.72	0.38	0.86	0.80	0.78	0.52
DecreaseInflation	0.67	0.60	0.8	0.85	0.73	0.70

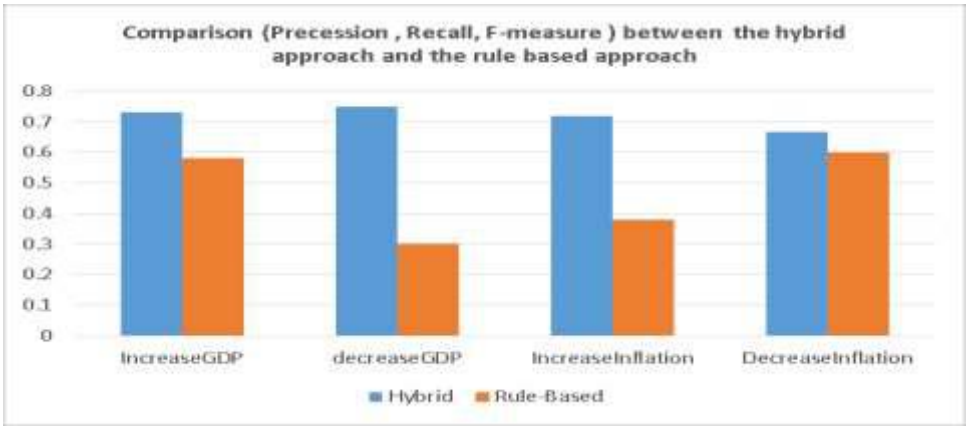


Figure 37: comparison (Precession, Recall, F-measure) between the hybrid approach and rule based approach

In this study, the problem of the complex relation extraction that relying on two triggers words was addressed by applying the hybrid approach. The experimental results show that the adopting of a hybrid approach has resulted in the highest performance.

Overall, the results indicate that the hybrid approach presents higher results than the rule-based approach indivisibly in terms of precision in extracting the relations which are based on two trigger words. However, the last two approaches have been found to be more accurate than the first in terms of recall. In this section, we will discuss three cases; the first one is trying to answer the question why the precision term in the hybrid approach was higher than rule-based approach? secondly, we are trying to answer the question why the recall was lower than rule-based approach? and finally, we are tried to answer the question why the hybrid approach recall was higher than precision?

In the first case, as mentioned above, the hybrid approach is a mix between the two techniques, with the rule-based approach being used to build the instance class relation. This is approach achieved with the highest accuracy in annotating relation based on the first trigger word (Increase, Decrease) that can cause an increase in the value of the recall. The ML approach was used to recognise the type of the class instance relation (GDP, Inflation) based on the second trigger word. Therefore, the precision term has higher. However, extracting the relation by the rule-based approach and the ML approach fails to identify this relation which causes an increase in the recall and a decrease in the precision. On the other hand, in some cases the rule-based approach cannot annotate the relation which means that

the other approach will not be able to annotate the relations and that causes a decrease in the two terms (recall, precision). In the ML approach, building the class instance relation is relying on extracting the named entity pairs from the sentence. And the features will be used by the relation classifier to decide if the relation is true or false. So, most of the time the ML approach performed with a high recall, in particular if the pair of NEs of the relation is clear in the sentence.

Consequently, our analysis asserts that the decrease in Recall in the hybrid approach is because the ML approach used the class instance relation that was extracted from the rule-based approach without using all the named entity pears in the sentence to build the class instance relation. The increase in Precision of the hybrid approach compared to the rule based approach because the problem is divided into two parts and by using a suitable approach, either FDG or ML approach to improve the results.

Regarding the final case, for different reasons the Recall increased more than the Precision. One of these reasons, is that sometimes the rule-based approach extracts the relation incorrectly which consequently affect the result of the ML stage, in addition to the different number of true positive instances for each class relation in the training dataset. In table 43, we can see that the number of has GDP class relations is more than the number of Inflation class relations which caused the increase in the precision value for GDP relation more than the inflation relation. See Table 39 and Figure 49.

7.5. **Summary**

This chapter was concerned with extracting the semantic relation between Arabic named entities. Several efforts have been reviewed. The related work showed that many statistical, clustering, semantic, syntactic and machine learning methods have been applied to extract the relation between Arabic entities. This related work showed that almost all of these efforts have used rule-based systems employing basic Arabic syntax grammar rules to define the linguistic relation. In this chapter, the FDG was adopted as the basis to build the relational extraction algorithms as it emphasises the semantic and pragmatic properties of the language, thus facilitating the identification of relation patterns in the Arabic language especially where there is a complex sentence structure that can often contains complex relations, where one subject has several predicates or several objects with varying order of

the features in the sentence. The results have argued that using Arabic semantic function grammar rules to extract semantic relations for a specific domain based on the domain knowledge is a useful task to obtain higher precision and that the semantic function grammar method will improve the state of the art for Arabic relation extraction. Also, our hybrid approach is more effective in addressing relations which have more trigger words.

Chapter 8

8. Constructing a semantic knowledge-base

8.1. **Introduction**

In the previous stage, we had focused on extracting the successful economic facts such as named entity and the relation between them from published information on the Web. This information will become more useful if we can organise this information in a semantic knowledge-base to facilitate the inference of a new information and to facilitate sophisticated, intelligent query from the knowledge-base for the benefit of the end user. Many efforts have used Semantic Web technologies for the improvement of natural language applications such as Adala, Asma, Nabil Tabbane, and Sami Tabbane. In [130] They presented a novel approach for automatic extracting of Semantic Web services that used NLP techniques to match a user request, expressed in natural language, with a Semantic Web service description.

Moreover, other researchers Albukhitan, Saeed, Ahmed Alnazer, and Tarek Helmy in [131] presented a Semantic Web service that supports semantic annotation of Arabic language documents.

We believe that the Semantic Web technologies are best suited to model this knowledge-base domain for the following reasons: domain knowledge is key in the knowledge-based information extraction tasks, and Semantic Web offers technologies to naturally model concepts and relations representing the domain knowledge. In addition, it offers reasoning capability allowing us to infer new facts from the modelled knowledge.

The formal representation of semantically tagged knowledge allows for sourcing information from Linked Open Data to further enrich the domain knowledge-base.

Hence, Semantic Web technologies can aid in the extracting information from domain specific resources and also contribute towards building advanced rules to query and infer interesting information from the resulting knowledge-base for the benefit of an intelligent exploration and recommended systems.

8.2. Utilizing the Semantic Web in Arabic IE

Arabic data on the Web are present in the unstructured form of Web pages. This results in an enormous amount of Arabic information that is meaningless and has no relationship. Therefore, an urgent quest to find solutions that allow the Arabic data on the Web be more understandable and meaningful for machines. Improving the Semantic Web is the solution to this issue. To take the benefits of Semantic Web technology, a Semantic model of Arabic data must be developed that is appropriate for Arabic data requirements and structure. The Semantic Model of Arabic data aims to develop an application that allows a machine to read and understand what it is presented instead of publishing data in human readable Arabic documents. That means allowing machines to understand the semantics of Arabic information on the Web, and extract new information and new relationships. Moreover, a structured representation of the information would improve their use in information extraction applications.

8.2.1. Overview of the Semantic Web

The term Semantic Web refers to W3C's vision of technologies that allow people to establish data stores on the Web, create vocabularies, and write rules for processing data empowered by technologies such as RDF, SPARQL, OWL and SKOS. The architecture of the Semantic Web is illustrated in Figure 39 below; it consists of several layers that make the Semantic Web model more acceptable. The Semantic Web model is a group of layers that are in a hierarchical form:

URI and Unicode layer is utilised to represent the resources and identify things and concepts. The UNICODE component is the standard international character set.

XML layer is a mark-up language utilised for data exchange. It is used to transport and save data in structured Web documents in the form of the user's vocabulary.

The RDF (Resource Description Framework) layer, is the HTML of Semantic Web and a basic data for writing simple statements about Web resources.

Web Ontology Language (OWL) layer include RDFs, it has semantics defined, and this semantics can be utilised for reasoning within ontologies and knowledge-bases described using these languages.

Simple Protocol and RDF Query Language (SPARQL) layer, is an SQL-like language and is used to query RDF Documents.

Rules Interchange Format (RIF) layer, is a language for representing rules on the Web and connecting several rule-based systems.

Logical layer is used to enhance the ontology language further and to allow for the writing of application-specific declarative knowledge. Proof layer concerns the description of inference results and data source.

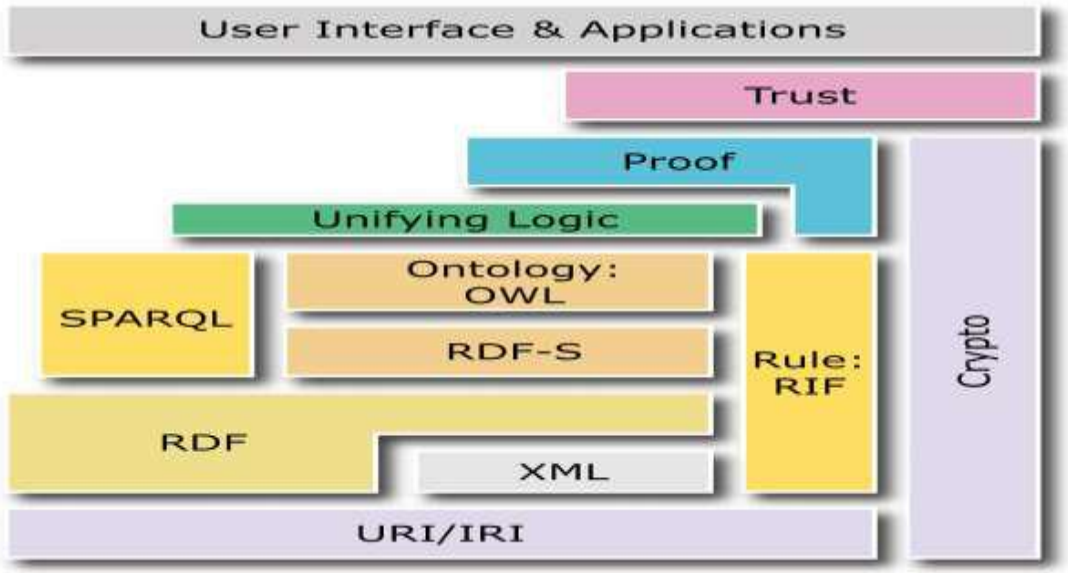


Figure 38:Semantic Web Architecture

8.2.2. Semantic Web tools

There are several tools available for the Semantic Web that can be employed to build the RDF triple, ontology, queries, conducting reasoners, and other services. The researchers can choose one or more tools they feel the most appropriate for reaching their goal. In this study, we have used the Protégé tool as the modelling tool and the Jena tool that allows insertion and extraction data from the knowledge-base.

- **Protégé tool**

  The protégé tool is one of the popular Semantic Web tools. It is a free open-source tool platform developed at Stanford Medical Informatics that is considered a suitable tool to build domain models and knowledge-based applications with ontologies. Stanford University's general Protégé 2000 ontology editor tool can be utilised to construct

ontologies, e.g., RDFS, OWL 1.0, and OWL 2.0, to access ontologies, to store ontologies, to visualise classes, properties, and relations. Also, Protégé can be utilised to implement reasoners [132].

- **Jena API**

Jena is an API that can be utilised to read, construct, process RDF documents, navigate through an RDF graph, querying of an RDF dataset via SPARQL and inference using OWL ontologies. Jena provides a collection of tools and Java libraries to assist in the development of Semantic Web and linked-data applications [133].

8.2.3. Arabic language supporting the Semantic Web

There is a lack of support to technology, tools and applications is the main challenges facing the improvement of the Arabic Semantic Web research. Al–Khalifa, Hend, and Areej Al–Wabil in [134], presented a paper to review Semantic Web tool, especially investigating support for the Arabic language. They demonstrated through a pilot test that the Arabic language has limited support for processing Arabic script in particular Semantic Web tools.

Beseiso, Ahmad, and Ismail in [135] investigated Arabic support in some existing Semantic Web technologies and established the ability to implement the Semantic Web with Arabic applications. The authors present various studies conducted in the Arabic information extraction area. Several tools, like Protégé, Jena, Sesame and KOAN were used in evaluating Arabic support. They studied several types of standard support level and the tool's functionality and the support level for the Arabic language. The evaluation found that Protégé and Jena were satisfactory in supporting the Arabic language with the Resource Description Framework (RDF), but limited support for the Arabic language with the OWL language. Also, while Jena had support for Arabic with OWL, it was limited with SPARQL query. On the other hand, concerning Arabic support, the results for the other tools were not satisfactory. However, the Arabic domain still needs further investigation to catch up with other language domains such as English and French.

After showing the lack of Semantic Web technology that supports the Arabic language, in the next section, we will present some attempted work to support Arabic Semantic Web research.

8.2.4. Existing work in Semantic Web based Arabic information extraction

In this work, we touch on different research communities evolving around the fields of Semantic Web applications. Many efforts will be presented related to Arabic Semantic Web applications in the domain of ontology construction and utilisation. Ontology is one of the most important tools in Semantic Web applications which offer defined and standardised shape of interoperable, machine understandable repositories [136]. These ontologies may then be classified as follows; the first one is the domain specific ontology which is concerned with the specific meaning of terms as they are inferred in that domain. Secondly is the upper ontology which is concerned with the general concepts that are related to a wide range of domains [137]. In the Arabic domain, there are numbers of efforts focused on building the ontologies based on a specific domain.

Maynard and Diana in[138], aimed to inquire into NLP techniques for ontology population, and how to verify that term recognition is helpful for many tasks of extracting information. Through a combination of rule-based learning and ML, TRUCKS is used to enhance traditional statistical techniques of term recognition.

This work tries to explain the relation between term recognition and information extraction and clarifies the difference between the methods used in each. In this work, the Balanced Distance Metric (BDM) was designed to evaluate ontology-based information extraction, which uses the similarity between the key and response instances in an ontology to determine the correctness of the extraction. The approach presents a dependable technique for using NLP techniques for extracting terms and ontology population.

Ben Saleh and Alkhalifa in [38], produced the first Arabic annotation tool for annotation of Arabic news on the Web. The tool was built to discover Arabic named entities based on the location ontology.

Another work on Arabic Web annotation has been proposed by Albukhita, Alnazer and Helmy in [131], which aims to create service that supports the semantic annotation of Arabic language text. This work consists of four components: ontology pre-processing, document NLP analysis, entities extraction and the relation extraction component. The application uses some documents related to three types of domain: health, food and nutrition.

Alromima et al. in [139], proposed the semantic-based retrieval approach for Arabic text. In this method, the Vector Space Model (VSM), was used to build the search engine index, and the Web Ontology Language (OWL) has been utilised to construct and implement the Arabic place nouns domain ontology from the Arabic corpus. The approach consists of two phases: the first phase is offline and uses the VSM model to create and maintain the index of the Arabic information retrieval approach and create and implement the Arabic domain body from the Arabic body. The second one is an online phase which concerns the user query. The corpus of the Holy Quran scripts was used as an experiment in the approach. The authors report that the results of the proposed approach in terms of precision and recall were better than term-based method.

H. Alfeel in [140], attempted a mapping of the Arabic info box in Wikipedia, which aimed to use the Wikipedia data set to improve the Arabic DBpedia chapter by increasing the mapping of properties and templates in the Arabic chapter. The authors reported that this work contributed by adding 52 mappings from Wikipedia templates to DBpedia classes.

H. Al-Feel et al in [141], presented a new approach that aimed to support Arabic Question and Answering (QA). This approach proposed a method to translate the Arabic questions into triples and retrieve the answer by matching the RDF data against the triples which are utilised to create the SPARQL query. Linguistic and semantic features have been used to solve ambiguity when mapping the words with the ontology content.

8.3. **Motivation**

Most of the efforts discussed have employed Semantic Web technologies to build sematic applications through different methods. Some of them have focused on constructing the ontology and using it to improve Semantic Web processing [142]-[146]. On the other hand, some researchers have been concerned with links between NLP and the Semantic Web to build Semantic Web applications [38], [131]. As seen earlier, almost all Arabic studies have focused on Semantic Web based applications based on the ontology, and there is limited research which looks at links between the Semantic Web and unstructured text on the Web documents. [131], [147] reported that the reason for the limited research in this area is due to the lack of resources and tools and they also recommended a focus on research in NLP to improve Semantic Web applications. For this reason, we need to present a high-quality framework that exploits the Semantic Web in Arabic information extraction from the Web

to bridge the gap between traditional information extraction tools that use NLP techniques and the semantic representation of information. Analysing an enormous amount of data on the Web to extract the useful information that will lead to improving the semantic intelligent explorations and close the gap between the Web documents and Semantic Web is still an issue in Arabic domain. Moreover, most of the studies have ignored addressing the economic domain in the Arabic language which is considered to hold some of the most valuable data on the Web. This work makes the following contributions to the body of work for Arabic language domains:

- A comprehensive framework is presented to analyse the unstructured Arabic documents on the Web and convert it into structured forms, and then use this information to improve the semantic recommender system.
- Arabic grammar rules have been used to improve NLP tasks that lead to improved Semantic Web technologies results.
- This study methodology is based on the economic domain.

8.4. Bridge the gap between Natural Language processing and the Semantic Web

Ontology population from the text is a significant task that integrates NLP, Knowledge Representation, and Semantic Web techniques to extract asserted knowledge from texts according to specific ontologies. Most of the information on the Web exists as unstructured text and ontology population plays an essential role in bridging the gap between structured and unstructured data, thus helping realise the vision of Semantic Web where contents are equally consumable by humans and machines [148].

Recently, the domain specific ontologies have become the integrated components of some Semantic Web technologies [149], and also the populating ontology of with domain related data has become an important task in the Semantic Web application. To perform these tasks manually needs more effort and is cost-intensive. Therefore, NLP technologies can play a major role in the extraction of NEs and the relations between them in the specific domain to automate the process of ontology population. The automatic extraction of the relations and the NEs are the critical steps in several information extraction applications which use ontologies, such as automatic indexing, terminology mining, knowledge discovery and so on. In this work, we have used the NLP technology to make the data on the Web machine

readable by annotating the documents on the Web with the semantic tags to represent the information in a more structured manner. Several Arabic named entities and relation between then have extracted from the Arabic text. This information is represented by using Semantic Web technologies (ontology). Figure 40 illustrates how the NLP stage can help to automate the population of the semantic knowledge-base with the information extraction from natural text. However, in the same token, as we mention before the semantic analysis of the domain knowledge can also help the NLP process to benefit from structured data available on the Web such as DBpedia data set to enrich the gazetteers list using to recognise the NEs.

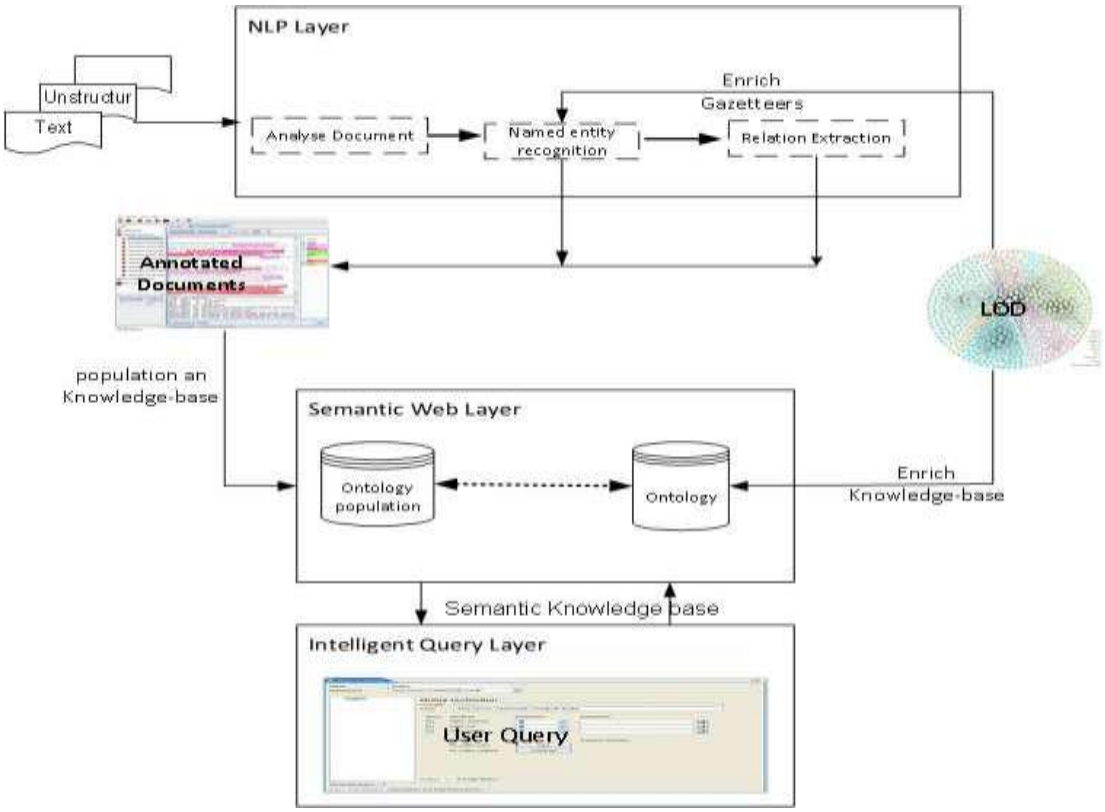


Figure 39: Architecture illustrate how bridge the gap between the unstructured data and structured data.

8.5. **Populating semantic knowledge-base with domain relevant information**

An important step in this phase is the proper linking of the entities and relations extraction in the text to the existing instances in the knowledge-base. In this section, we will present several ways have used to represent the information that extracted from the textual text in the knowledge-base.

8.5.1. Constructing Knowledge-base

As explained in chapter 4, the ontology plays an increasingly significant role in knowledge management and is utilised as a standard knowledge representation for the Semantic Web. In the Semantic Web, the ontology has used to represent the schema or taxonomy of the domain knowledge. The ontology is the structural component of conceptual relationships. In the initial stage, we have analysed the domain knowledge to build the ontology of our domain knowledge by translated the events in the domain knowledge map. The ontology consists the title of classes and the relations between classes that will be used to tag the extracting NEs and relationship from the text semantically.

8.5.2. Populating semantic knowledge-base

In NLP, a relation usually indicates linking between entities in the text. There are many types of relations such as semantic relations, grammatical relations and co-reference relations [140]. In this study, the focus is on the extraction of the semantic relation between several entities from unstructured documents related to the economic domain; specifically, the binary relation and complex relation indicating the economic activity, in preparation for injecting the relation into the semantic knowledge-base.

The semantic relation triples (S, P, O) that were extracted from unstructured data will be converted into RDF to be inserted into the semantic knowledge-base. The semantic knowledge-base will be populated with the relations triples extracted from the source unstructured text as well as domain relevant data extracted from LOD such as the information about the country, e.g., the population the name of a country, name of a capital city of a country).

8.5.2.1. Frame-based ontology population from Arabic economic news text

In the first stage, the NLP techniques have used to annotate the named entities and extract the relation between named entities from unstructured text, and then we needed to convert contents of the text to RDF representations. In Figure 42, the text-RDF pipeline illustrates integrating the IE task and Semantic Web technology to map the RDF triple with semantic relation extraction triples.

Our pipeline is designed to engineer the ontology mappings by matching the main three elements in the semantic relation triple with existing triples in the ontology. The subject, predicate and object are the main terms of the relation triple which are extracted from the unstructured documents.

In some cases, these relations are the binary relations that are represented where each instance of the relation links an entity to another entity or value, while other types of relations link the entity to more than just one entity or value. This kind of relation is called n-ary relations. The processes that are performed to populate the entities, facts and events extracted from the financial news are explained by the text- RDF pipeline in Figure 41.

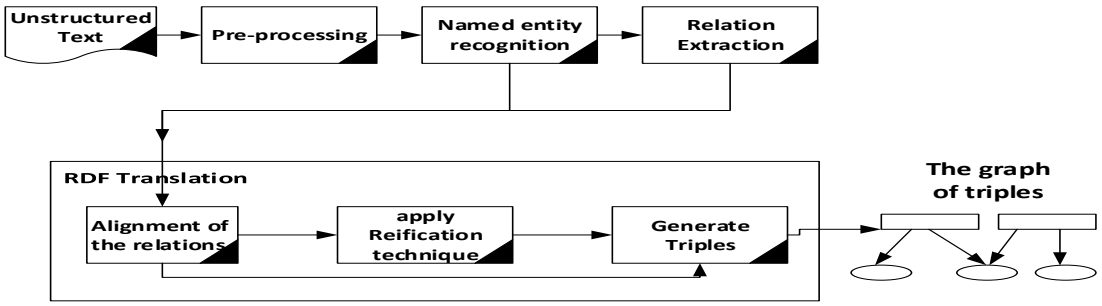
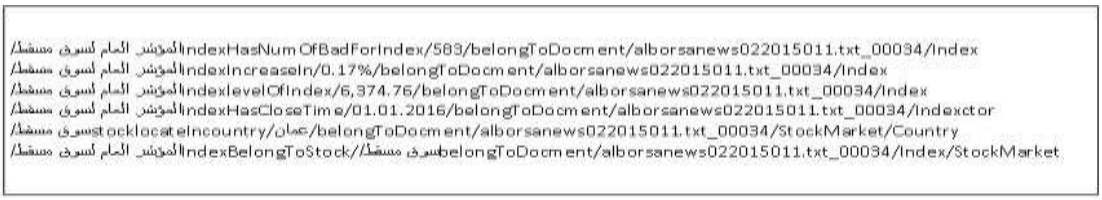


Figure 40: The text-RDF pipeline architecture for populating data into KB.

As seen in Figure 42, the first stage is the NLP stage, which aims to extract the NEs and the relation between the NEs. At this stage, we have used the rules based approach by employing Arabic grammar rules to extract several Arabic named entities such as the composite Arabic names. Also, we have utilised the Function Discourse Grammar to extract the complex Arabic relation extraction. Moreover, then the lists all the entities and triple relations which were extracted from textual documents into the list. This list is an intermediate list which contains all the relations in the documents to be easily processed. The second stage is semantic annotation stage, where the NEs and corresponding relations are tagged with the appropriate class or relation in the ontology, resulting in an RDF resource(triple). The Jena framework has been used to write the RDF triples into semantic knowledge-base by using ontology. In some cases, the reification RDF technique was used to build RDF triples that should connect an individual to more than just one individual or value. Figure 42 shows the RDF graph can be stored in the subject, predicate and object of the relation model, the source

node of each relation is placed on the subject, the edge label in the predicate and the target node in the object.



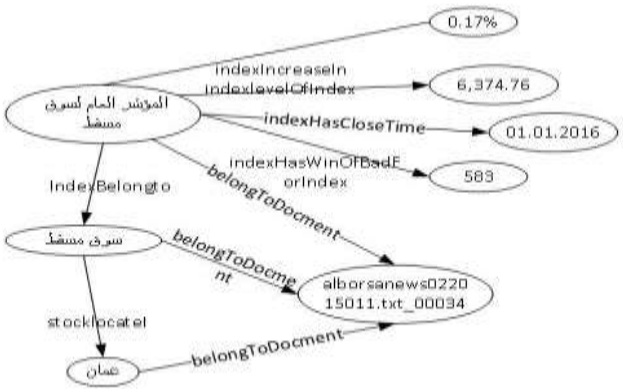


Figure 41: Example of representing a data structure

- **Representing the Binary Relations**

The relationships that describe a relation between two entities in the same domain, without any additional information is called the binary relations. Almost all of the relations that are extracted from the unstructured text are binary relations. To represent these types of relations in the Semantic Web knowledge-base, we need to convert the semantic relation triples to RDF triples. In this case, the Jena API has been used to represent these types of relations as RDF triples. Figure 43 shows how the Jena API applied this task.

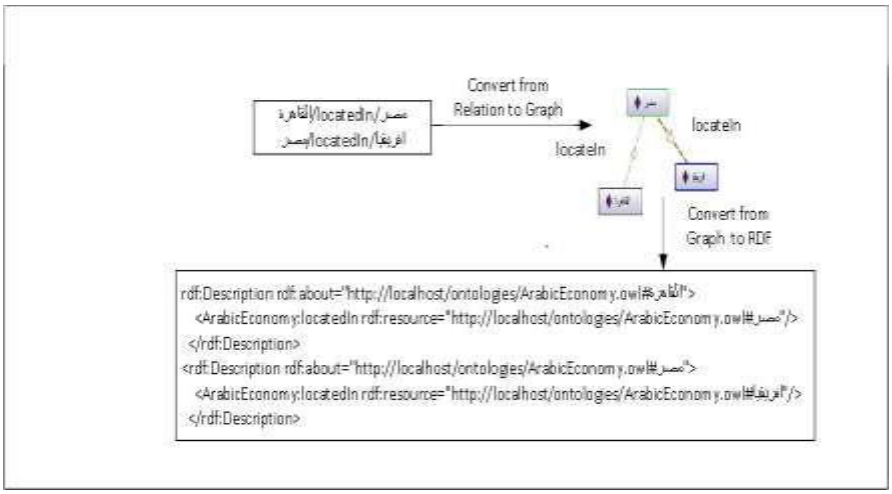


Figure 42: An example shows how the binary relations represented

As seen in Figure 43, the relations are extracted from unstructured text such as city ("مصر" (Egypt), "القاهرة" (Cairo)) and have been represented into RDF triples and injected into the knowledge-base.

- **Representing the n-ary relations**

The n-ary relation is the binary relation that involves more than two arguments [150]. As known, the limitations of OWL are that it represents the binary relations between classes. Sometimes, the extracted relationship links an individual with more than one other individual or value. So, we need a more natural and convenient way to represent the n-ary relations. As shown in Figure 44, the sentence consists two relations that represent the relation between the individual as country and two objects: the first one is a value as a percentage of increased inflation – increaseInflation(مصر,26.9%), and the second is a date that represents the time of the event,– dateOfInflation(مصر,2016). This type of relation is 2-ary relation and is called the n-ary relation. To represent the n-ary relations as an RDF triple, we need specific techniques because the n-ary relation cannot be simply be split into binary relations.

سجل معدل التضخم السنوي ارتفاعا في مصر 26.9% في يناير/ كانون الثاني 2017 أعلى مستوى خلال نحو 30 عاماً

The annual inflation rate rose to 26.9% in Egypt in January 2017 the highest level in nearly 30 years

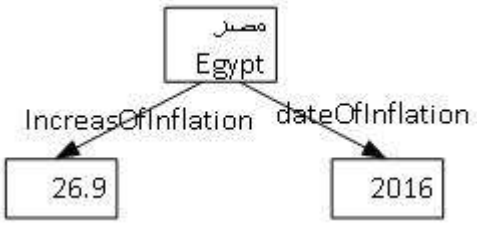


Figure 43: The n-ary relation example

In this work, the Reification approach will be used for the representation of the n-ary relations. The Reification approach that works to introduce a new triple for a statement and then uses three new triples with subject, predicate and object to describe the original statement. The reification technique that applied by Jena has used to add additional information about the triple [151]. We have chosen the Reification for the following reason: many studies widely use the reification approach to tackle the problem of representing the complex relations [152], [153]. Moreover, it works as well as with the Jena API [25], and it is sufficient enough for this work.

- **Reification approach to representing the complex relations**

RDF reification vocabulary has been used to represent n-ary relations into RDF triples. The RDF reification vocabulary represents the relations as the statement and individuals which are instances of the statement. The statement consists of a subject, predicate and object triple and the reification technique has been used to add the additional information about the triple [151]. In the next example, will illustrate how the reification techniques used to represent the n-ary relation. Figure 45 shows the sentence consists different relations.

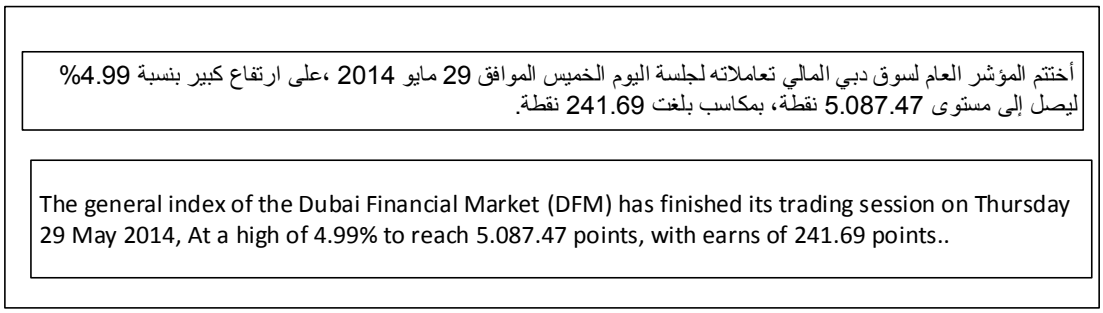


Figure 44: An example shows the sentence consists different types of relations

In the example above, the sentence consists of different relations; each relation has represented as the binary relations. Table 52 shows listed several binary relations in this sentence.

Table 52: The list of binary relations in the sentence.

No	Subject	Predicate	Object	Explain
1	المؤشر العام لسوق دبي المالي Dubai Financial Market Index	indexIncreaseBy	4.99%	This relation describes the state of index
2	المؤشر العام لسوق دبي المالي Dubai Financial Market Index	hasCloseTime	29 مايو 2014	This relation describes the time the index has closed
3	المؤشر العام لسوق دبي المالي Dubai Financial Market Index	indexHasLevel	5.087.47 نقطة (point)	This relation describes the level of index based on the number of points
4	المؤشر العام لسوق دبي المالي Dubai Financial Market Index	indexWinPoints	241.69 نقطة (point)	This relation describes the number of points the index has won
5	المؤشر العام لسوق دبي المالي Dubai Financial Market Index	belongToDocument	NEWSNEWS 991.txt	This relation describes the index belongs to a document

As mention before, a common way to represent n-ary relations is to break down them directly into binary relations between two entities. However, in doing so, significant information may be lost. For instance, "indexIncreaseBy" we cannot sure to which the date of this action and how many points the index loss or win in this action.

In the example above, the sentence consists of different relations that are related to the main entity in the sentence ("المالي دبي لسوق العام المؤشر","the general index of the Dubai Financial Market"). One of these relations is the main relation, and the other relations are complementary to the main relation.

Suppose that we want to express the previous example in figure 36 as produced in natural language tasks.

The previous example shows that the relation "indexIncreaseBy" is the main relation and other relations have been used to add the additional information about the main relation; this type of relation is called the 5-ary relation. The main idea behind utilising the reification techniques is that it overcomes many difficulties that happen at the query stage. Figure 46 shows the representation of the 5-any relation by the reification technique and figure 47 shows the output of applying the reification technique.

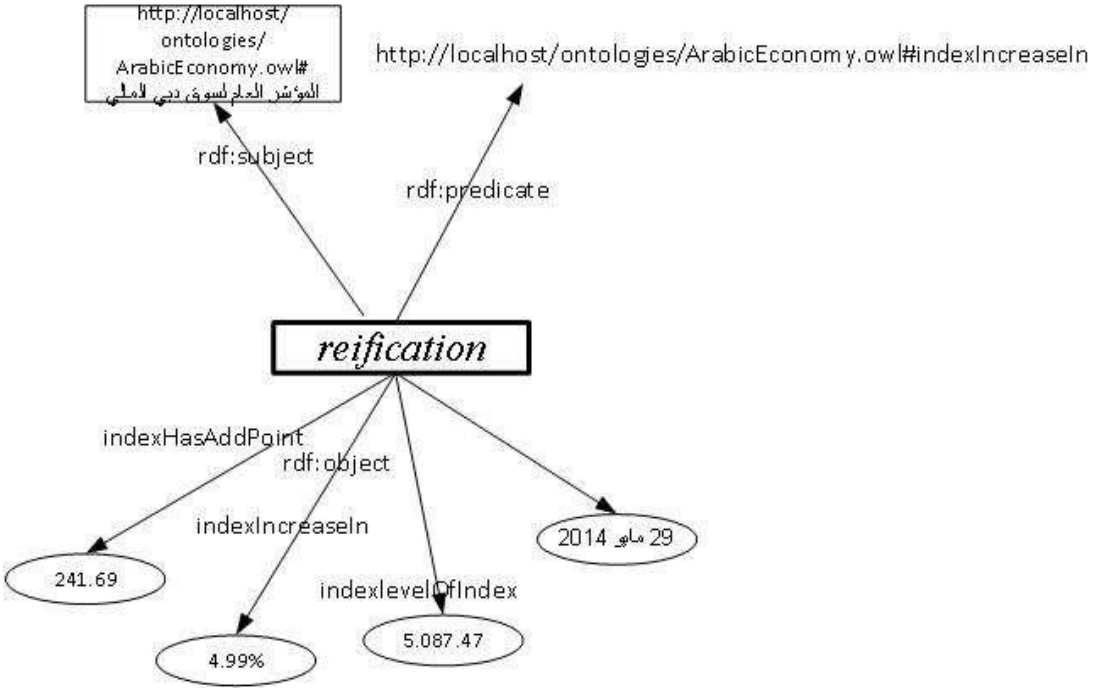
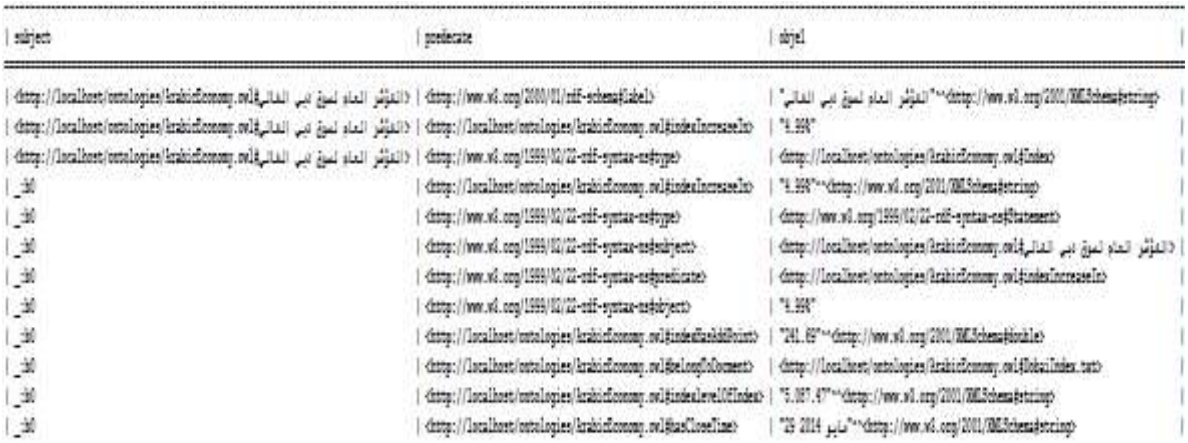


Figure 45: An example illustrates the reification technique represent the n-ary relations



subject	predicate	object
<http://localhost/ontologies/arabicEconomy.owl#المؤشر العام لسوق دبي المالي>	<http://www.w3.org/2000/01/rdf-schema#label>	"المؤشر العام لسوق دبي المالي"^^<http://www.w3.org/2001/XMLSchema#string>
<http://localhost/ontologies/arabicEconomy.owl#المؤشر العام لسوق دبي المالي>	<http://localhost/ontologies/arabicEconomy.owl#indexIncreaseIn>	"4.998"
<http://localhost/ontologies/arabicEconomy.owl#المؤشر العام لسوق دبي المالي>	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http://localhost/ontologies/arabicEconomy.owl#Index>
_:b0	<http://localhost/ontologies/arabicEconomy.owl#indexIncreaseIn>	"4.998"^^<http://www.w3.org/2001/XMLSchema#string>
_:b0	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http://www.w3.org/1999/02/22-rdf-syntax-ns#Statement>
_:b0	<http://www.w3.org/1999/02/22-rdf-syntax-ns#subject>	<http://localhost/ontologies/arabicEconomy.owl#المؤشر العام لسوق دبي المالي>
_:b0	<http://www.w3.org/1999/02/22-rdf-syntax-ns#predicate>	<http://localhost/ontologies/arabicEconomy.owl#indexIncreaseIn>
_:b0	<http://www.w3.org/1999/02/22-rdf-syntax-ns#object>	"4.998"
_:b0	<http://localhost/ontologies/arabicEconomy.owl#indexRankAtStart>	"241.89"^^<http://www.w3.org/2001/XMLSchema#double>
_:b0	<http://localhost/ontologies/arabicEconomy.owl#belongToDocument>	<http://localhost/ontologies/arabicEconomy.owl#GlobalIndex.txt>
_:b0	<http://localhost/ontologies/arabicEconomy.owl#indexLevelOfIndex>	"5,187.47"^^<http://www.w3.org/2001/XMLSchema#string>
_:b0	<http://localhost/ontologies/arabicEconomy.owl#hasCloseTime>	"28 2014 مارس"^^<http://www.w3.org/2001/XMLSchema#string>

Figure 46: Representing the n-ary relation in the knowledge-base

8.6. **Using linked open data to enhance Arabic information extraction system**

The term LOD refers to a set of pieces of information that are linked together on the Web. It refers to data published on the Web so that it is machine-readable data [98]. The linked data relies on the documents which are represented in RDF format. There are several datasets saved as linked data such as FOAF and DBpedia datasets.

DBpedia data set has become one of the most popular datasets in LOD. This information has been extracted from Wikipedia and converted into structured data, to make this information accessible on the Web. The DBpedia knowledge-base provides localised versions in 125 languages, which describe 38.3 million things for all these versions and the 23.8 million of things that are also available in the English version of DBpedia.

The DBpedia dataset contains a set of features, one of which is a label feature, which contains 38 million labels in the different languages [154]. At the time of writing this thesis, the DBpedia has around 19 localised DBpedia chapters. There are still several languages which do not get much attention. Unfortunately, the Arabic language remains one of these languages which still needs more effort to take positions in the DBpedia dataset. There are several studies that worked to overcome the challenges which faced the Arabic language in Dbpedia [140], [155], [156]. There are a few types of Arabic research which use the DBpedia to improve information extraction [141]. In the context of this work, enriching ontology utilises of Linked Data sources would provide benefits. The DBpedia data set has been utilised to enrich the Arabic economic knowledge-base by using the piece of information of

the main entity, and we try to extract the data that related to the main entity from DBpedia. As the Arabic language, does not publish and link large semantic data in DBpedia, especially for the economic domain, therefore, most information extracted from DBpedia is based on the English chapters by using the label features.

The SPARQL query engine techniques [157] have been used to retrieval the numbers of RDF triples from DBpedia based on the entities mentioned in the unstructured text. For instance, the country entity has been used to retrieve several pieces of information as the RDF triples that are related with the specific country and this information is not mentioned in the unstructured text. The example below (Figure 48) shows how the pipeline retrieving the information from DBpedia.

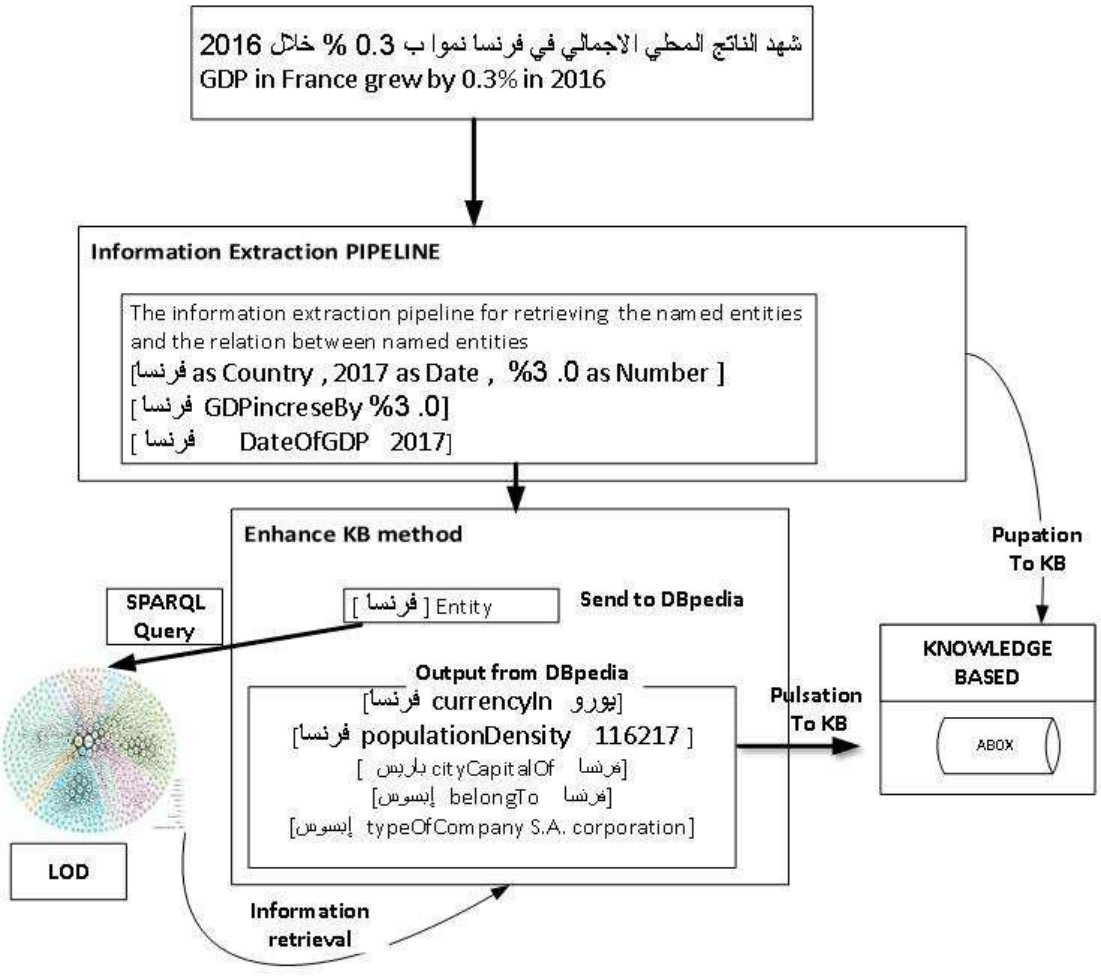


Figure 47: The architecture pipeline for extraction RDF triples from LOD

Figure 49 illustrates the several steps that are used to extract the information from LOD. In a first step, the Information extraction task is applied to extract the useful information from

the unstructured Arabic text such as NEs and relation between them and represent this information as the RDF triples into the Knowledge-base. In a second step, some of the named entities such as country and Company will be utilised as parameters to build an SPARQL query to extract more information that related for these entities. For example, the SPARQL query will use to model the equation (*can retrieve all the companies that belong to the France country including the type and home page of the companies?*). Figure 49 shows SPARQL query that uses to retrieve all the companies belong France country.

```
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
 PREFIX foaf:  <http://xmlns.com/foaf/0.1/>
SELECT  ?countyNampe  ?companyName ?companyType
?companyHomePgae
where {
?company  dbpedia-owl:Company.
?company  rdfs:label ?companyName.
?company  dbo:type ?type.
? ?type rdfs:label ?companyType.
 ?company foaf:homepage ?companyHomePgae.
?company dbo:locationCountry ?z.
?z  rdfs:label ?CountryName.
FILTER (lang(?companyName) = \"ar\" )
FILTER (lang(?companyType) = \"ar\" )
FILTER (lang(?companyHomePgae) = \"en\" )
FILTER (?countyNampe = "فرنسا")
 }
```

Figure 48: The SPARQL query for retrieving the information from LOD

In a third step, the SPARQL query is executed on DBpedia to retrieve the list of RDF triples that are related to the specific country such the companies belong to the country. In a final

step, the Jena API is used to write the RDF triple into the knowledge-base. Figure 50 shows the information that is extracted from LOD and populated into Knowledge-base.

Figure 49: A screen shot representing the information extracted from LOD in the Knowledge-base

In other examples, the SPARQL query is used to model the equation (e.g. *can we retrieve the information about France country including currency unit and capital city?*). Figure 50 shows how SPARQL query was used to retrieve the information related to a specific country (France) such as the name of the capital city and currency unit.

8.7. **Intelligent interrogation of the semantic knowledge-base**

Semantic data, which covers structured data, is becoming widely available on the Web and semantic data exploration is becoming an important activity in a range of application domains, such as government organisations, education, life science, cultural heritage, and media. Recently, several explorations of semantic data methods have made proposals, including novel interfaces and interaction, means; for example, semantic data browsers, ontology/content visualisation environments and semantic wikis. Data exploration refers to the process of efficient and effective knowledge from data even if we are uncertain as to what we are looking for [158]. This process normally involves developing reasoning techniques for inferring new facts and sophisticated query methods for interrogating the knowledge-base.

8.7.1. Overview of inference on the Semantic Web

An OWL based ontology plays an important role in the Semantic Web. It can be utilised to describe the intended use of resources and can use powerful DL reasoning tools. Inference on the Semantic Web can be characterised by extracting new relationships. On the Semantic Web, data is represented as a set of relationships between resources. Inference means that automatic processes can infer new RDF triples based on that data and based on some additional information in the form of vocabulary, such as a set of rules.

There are two types of reasoning inference: ontology reasoner and rule based reasoner. The ontology reasoning is that support OWL ontology language and can be used as a plug-in for either protégé. There are several types of ontology reasoner such as Pellet, FaCT++, HerMiT, ELK [159] . To process the ontologies that are described by OWL-DL a reasoner is used, and it has several services to process the ontologies. One of them is to test the hierarchy of the classes in ontology to complete the inferring ontology class hierarchy [160], while another service is consistency checking. The class hierarchy automatically computed by the reasoner is named the inferred hierarchy. The Protégé tool that is used in building OWL ontologies has several kinds of resonators. Pellet reasoning is one of the kinds of reasoning that is offered by Protégé.

Also, the second type of reasoners is Rule Based Reasoner. Jena can provide several reasoners inside the application based on user requests and requirements. Jena essentially supports RDFS and OWL reasoning but also has the support of a generic reasoner. The generic reasoning is a rule based reasoner that supports user defined rules, Forward chaining, hybrid execution strategies and tabled backward-chaining are supported [133]. The Jena reasoner builds a new RDF model consisting of asserted and derived tuples. The Jena inference subsystem allows a range of reasoners for deriving additional facts including Transitive reasoner which implements transitive and reflexive properties, RDFS rules reasoner containing RDFS entailments, OWL reasoner, DAML reasoner and the Generic rule reasoner for supporting user-defined rules [161]. With Jena Generic Rules, these types of rules are defined by the user, and it is widely used which allows users to implement them according to their needs and requirements.

8.7.2. Developing the economic knowledge-base inference engine

The economic domain is one of the important domain because it comprises a lot of entities and events. Using Semantic Web technologies to model the economic domain knowledge and represent the information into ontology is the useful way. However, in some cases, this information need advanced techniques to extract the hidden knowledge from them such as inference engine (reasoning) to support the decision making for intelligent application such as recommender system. The inference engine uses to obtain answers or replies to queries from a knowledge-base. There are several processes can apply by the reasoning such as:

8.7.2.1. Advanced use of object properties

The object property characteristics allow for enriching the information by using a different type of property characteristic. One of these property characteristics is Transitive Properties. The property relates individual Ind1 to individual Ind2, and also individual Ind2 to individual Ind3; then we can infer that individual Ind1 is related to individual Ind3 via property P. Figure 52 shows an example of a Transitive Property the relation locatedIn is representing the relation between the City with County, and Country with continent. In this example, if the individual ("طهران"," Tehran") as City is located in ("ایران"," Iran") as Country, and ("ایران"," Iran") as Country is located in ("اسیا"," Asia") as Continent, then after Appling the reasoner can infer that ("طهران"," Tehran") is located in ("اسیا"," Asia"). In our ontology, we have three classes (City, Country and continent). These classes have three individuals (Tehran, Iran and Asia) respectively. The triples of these individuals in our knowledge-base will be as below:

Ourl:Tehran rdf:type ourl:City

Ourl:Iran rdf:type ourl:Country

Ourl:Asia rdf:type ourl:Continent

The relations between these individuals by using locatedIn object property are :

Ourl:Tehran ourl:locatedin ourl:Iran

ourl:Iran ourl:locatedin ourl:Asia

Because the characteristic of locatedIn object property is transitive then after Appling the reasoner it will generate a new triple which is

Ourl:Tehran ourl:locatedin ourl:Asia

Figure 51 can be subdivided into two main parts: The first part on the left side (1) shows the Tehran City is located in Iran Country and Iran is located in Asia. The second part on the right side (2) shows the reasoning generated a new RDF triple which is Tehran located in Asia.

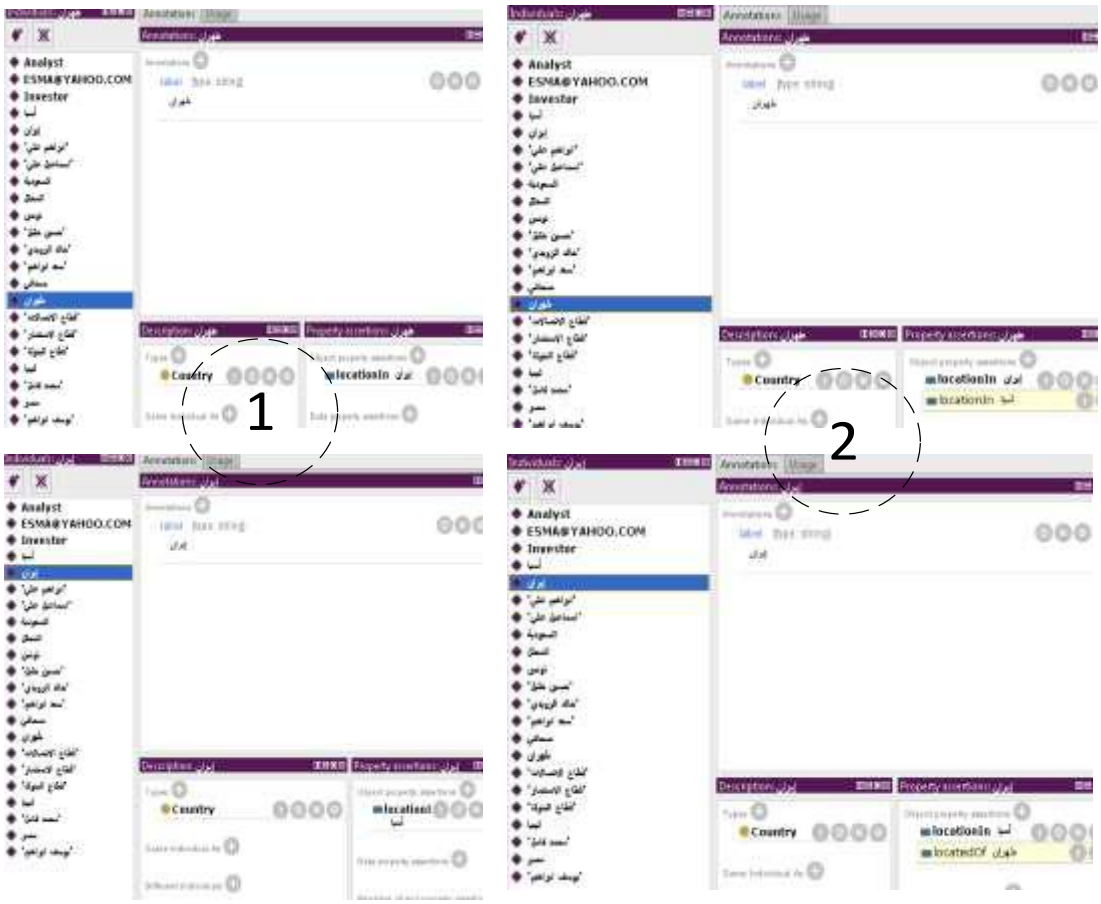


Figure 50: The reasoning applies the transitive properties characteristics in the knowledge-base

8.7.2.2. Automated Classification using the necessary and sufficient condition

The important task of a reasoner is to automatically compute the class hierarchy of the ontologies built by using the OWL-DL. Therefore, the reasoner will be used to compute and maintain multiple inheritance in the ontology, and keep the ontology in a maintainable and logically correct state. Figure 53, shows the reasoner has classified each person as an employee such as Investor, Journalist and Analyst as the instance in the targeted user class. For example in our ontology we have three classes (Investor, Journalist and Analyst) and these classes have three individuals (علي_اسماعيل, خالد_الرويدي , سعد_ابراهيم) respectively. The triples of these individuals in the knowledge-base will be as below:

Ourl: علي_اسماعيل rdf:type ourl: Journalist

Ourl: الرويدي_خالد rdf:type ourl: Analyst

Ourl: ابراهيم_سعد rdf:type ourl: Investor

Ourl: Journalist rdf:type ourl: Journalist

Ourl: Analyst rdf:type ourl: Analyst

Ourl: Investor rdf:type ourl: Investor

Ourl: ابراهيم_سعد ourl:hasEmployer ourl: Investor

Ourl: علي_اسماعيل ourl:hasEmployer ourl: Journalist

Ourl: الرويدي_خالد ourl:hasEmployer ourl: Analyst

If we add the following axioms to the ontology :

user □ hasEmployer some Journalist

user □ hasEmployer some Journalist

user □ hasEmployer some Analyst

after Appling the reasoner it will classify all the persons that have the work as (Investor, Journalist and Analyst) as the target user for the system. Then the reasoner will generate the following triples:

Ourl: علي_اسماعيل rdf:type ourl: User

Ourl: الرويدي_خالد rdf:type ourl: User

Ourl: ابراهيم_سعد rdf:type ourl: User

Figure 52, can be subdivided into two main parts: The first part on the left side (1) shows the several persons has different job. The second part on the right side (2) shows the reasoning classify all the persons that have the work as (Investor, Journalist and Analyst) as the target user for the system.

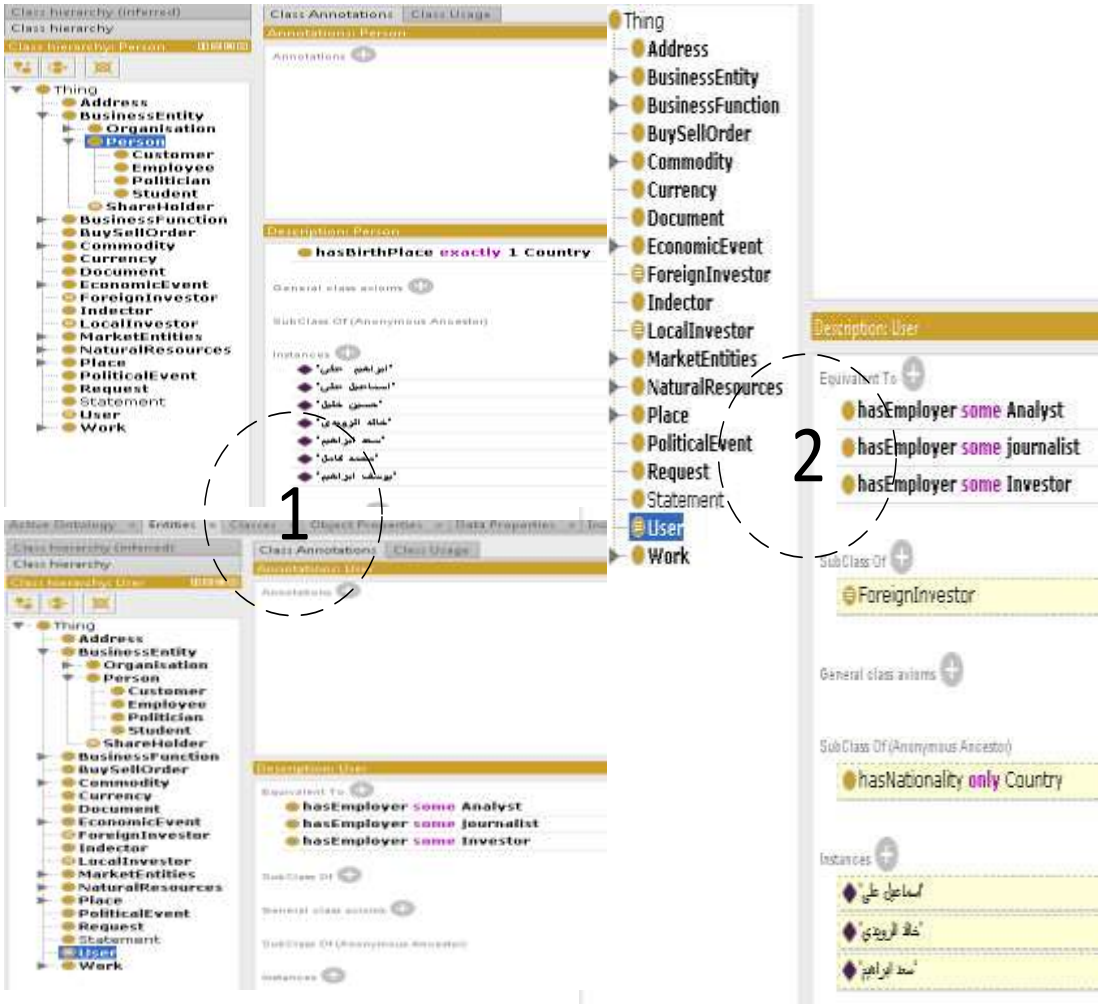


Figure 51: Example how the reasoning automatically compute the class hierarchy in the knowledge-base

8.7.2.3. Explicit reasoner rules

For instance, our proposed approach for a Semantic Web recommender system is to employ the Jena reasoning approach, which involves user defined rules. In this study, the Jena rules have been used to create different rules that are relevant to the economic domain. These rules describe the relations between several economic indicators that allow the system to predict different events which help the user to take decisions regarding their field of interest. Table 53 shows some economic indicators and Table 54 shows the relation rules between the indicators which should assist in the financial decision-making process.

The rules have been built based on several kinds of economic indicators that describe the relation between them. Four indicators have been chosen to build the rules, Interest Rate, Inflation, Gross Domestic Product and Exchange rates. Economic indicators measure the

current state of the economy. They can measure specific sectors of an economy, such as the housing or retail sector or they can give measurements of an economy as a whole, such as GDP or unemployment.

Below are some of the important economic indicators that we use to exemplify building reasoning rules for the economic knowledge-base.

Inflation is “the rate at which the general level of prices for goods and services is rising and, consequently, the purchasing power of a currency is falling. Central banks attempt to limit inflation and avoid deflation, to keep the economy running smoothly” [37].

The interest rate is “the amount charged, expressed as a percentage of principal, by a lender to a borrower for the use of assets” [37].

Gross Domestic Product (GDP) is “a monetary measure of the market value of all final goods and services produced in a period (quarterly or yearly)” [162].

Exchange rates are “among the top factors that distinguish the health of a country's economy. Also, known as a forex rate, the foreign exchange rate is the value of the currency of one nation about another nation's currency” [163].

Investors and traders usually keep a close eye on the state of these indicators which give them clues about the state of the economy.

Table 53: List of the economic indicators.

No	Feature
1	The GDP Increase
2	The GDP Decrease
3	The Inflation Increase
4	The Inflation Decrease
5	Interest rate Increase
6	Interest rate Decrease
7	Exchange rate Increase
8	Exchange rate Decrease
9	The Index Increase
10	The Index Decrease

Table 54: The relation between different economic indicators.

No	Relation Between Rules	Local investor	Foreign investor
1	Interest rate Increase+ GDP Decrease	Not good	Not good
2	Inflation Increase+ GDP Increase	Good	Good
3	Exchange rate Increase+ Inflation Increase	Good	Good
4	Interest rate Increase + Financial market index decrease	Not good	Not good
5	Financial market index decrease+ GDP Decrease	Not good	Not good
6	Exchange decrease +Financial market index decrease	Good	Good
7	Exchange decrease+ GDP Decrease	Not good	Not good
8	Interest rate Increase + Exchange decrease	Not good	Good
9	Interest rate Decrease+ Exchange rate Decrease	Good	Not good

We can see that there is a relation between the indicators that affect the economic state. In the following examples, several Jena rules have been developed based on the aforementioned economic indicators to infer a new knowledge by exploiting the existing information in the knowledge-base.

As described in [164], [165] the drop in the interest rate indicator encourages people to spend more money and subsequently consuming spindly, which could lead to a rise in inflation. The opposite holds true if the interest rate indicator is increased then the result is that the economy slows and inflation decreases[164]. For the use case scenario, two rules have been written: the first rule has been written based on the state of the interest rates indicator, and the effect on the economic growth for the country; if the interest rates indicator has decreased then the economy state could lead to a rise in inflation indicator. From the previous scenario, the rule has created as shown in Figure 53.

```
First Rule
[r1:
(?R1 rdf:type rdf:Statement)(?R1 rdf:subject ?S1) (?S1 rdfs:label ?t)
(?R1 pre: has interestrateIndector ?Indecator)(? Indecator rdfs:label ?D)
(?R2 pre:hasValue ?VIRate)(?R2 pre:hasValue “decrease”)
(?R2 pre: interestratesDate ?date)
-->
(?S1 pre:economicExpectIncrease "GDP ")
(?S1 pre:dateeconomicExpectIncrease?date)
]
// Second Rule
[r1:
(?R1 rdf:type rdf:Statement)(?R1 rdf:subject ?S1) (?S1 rdfs:label ?t)
(?R1 pre: has interestrateIndector ?Indecator)(? Indecator rdfs:label ?D) (?R2 pre:hasValue ?VIRate)  (?R2
pre:hasValue “increase”)
 (?R2 pre: interestratesDate ?date)
-->
(?S1 pre:economicExpectDecrease "GDP ")
(?S1 pre:dateeconomicExpectDecrease ?date)
]
```

Figure 52: The first rules to generation a new knowledge-based on economic indicators

A second rule has been written based on the output of the previous rules to give the investor a recommendation about investment in a specific country. In the next rules, the reasoner will apply the rules which will help the user to decide concerning this country as either a good

country for investment or not based on the state of GDP and Inflation indicators, see figure 54.

```
[r1:
(?R1 rdf:type rdf:Statement)(?R1 rdf:subject ?S1)(?S1 rdfs:label ?t)
(?R1 pre: has inflationIndector ?Indecator)(? Indecator rdfs:label ?D)(?R2 pre:hasValue
?VIRate)(?R2 pre:hasState "decrease")(?R2 pre: interestratesDate ?date)(?S1
pre:economicExpectDecrease "GDP " )(?S1 pre:dateeconomicExpectDecrease ?date)
(?user per:requst Advacerabout ?S1)
->
(?advisor pre:hasRequst from ?user)
(?user pre:gotAdviser " the country not good for investment")
]
```

Figure 53: The second rules to generating a new knowledge about the state of country

8.8. **Advanced query mechanism for structured data exploration**

As previously highlighted, the information that is extracted from the unstructured text and the information that is inferred to by using the reasoner is presented into the knowledge-base as RDF triples. SPARQL is standard query language for RDF data retrieval, which is, a semantic query language for databases, able to retrieve and manipulate data stored in RDF format [157]. The SPARQL query language provides a set of operators, such as a conjunction, optional patterns, union, aggregate amongst others characteristics to achieve further complex functionality over the stored RDF data. The structure of SPARQL query language is shown in Figure 55.

PREFIX (Namespace Prefixes) e.ge. PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT (Result Set) **E.g. SELECT** ?name ?address
FROM (Data Sate) **E.g. From** <http://localhost/ontologies/ArabicEconomy.owl#>
WHERE (Query Triple Pattern) **E.g. WHERE {**?user Ourl:hasJob ?Work .**}**
ORDER BY . DISTINCT ect (Query Triple Pattern) **E.g. ORDER BY ?**name

Figure 54: The structure of the SPARQL query language

In this study, the SPARQL query has been used to explore and query the information that is stored in the semantic knowledge-base. Some SPARQL queries have been created to support the user query for the Arabic Financial Knowledge-base Recommender (FKBR) system. The following examples explain different kinds of SPARQL queries that show the intelligent exploration whereby with very complex queries can retrieval a lot of useful data from the knowledge-base.

Example 01: The query in Figure 56 SPARQL query example that models the question "*can retrieve all the users that have the same details such as has same nationality and interested in investing in the same country and same industry sector*?". We can write a query which finds those users and explores their information: The output of the SPARQL query is that the list of all users that have same details.

```
PREFIX Ourl: <http://localhost/ontologies/ArabicEconomy.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT  ?L2 ?workName ?interestName  ?NationalityName
WHERE {
?user rdf:type Ourl:User.
?user rdfs:label ?L2 .
?user  Ourl:hasJob ?Work .
?Work rdfs:label ?workName .
?user  Ourl:interstIn ?interest .
?interest rdfs:label ?interestName .
?user  Ourl:hasNationality ?nationality .
?nationality rdfs:label ?NationalityName .
FILTER (??NationalityName = ?x1)
FILTER (?interestName = ?x2)
FILTER (?workName= ?x3)
}
```

Figure 55: The SPARQL query to retrieval the list of users which have same details.

In the second example, SPARQL query that models the question “Can we retrieve all the state of the Shares that belong to a specific Index for specific dates?”. We can write a query

that finds information about all the shares such as the price of the share and state of the share (increase or decrease) that belong to specific Index in a specific day. The output of the SPARQ query is that the list of all the shares belongs to the specific Index and specific day. Figure 57 shows a query that will retrieve information about a specific Index.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX Ourl: <http://localhost/ontologies/ArabicEconomy.owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
select distinct ?ShareName  ? shareName increase ?PRICE
?decrease
where { ?R rdf:type rdf:Statement .
?R rdf:subject ?S .
?S rdf:type Ourl:Share .
?S rdfs:label ?shareName .
?S Ourl:belongTo ?index .
?index rdf:type Ourl:Index .
?index rdfs:label ?indexName .
OPTIONAL{ ?R Ourl:shareHasPrice ?PRICE}.
OPTIONAL{ ?R Ourl:shareIncreaseIn ?increase}.
OPTIONAL{ ?R Ourl:shareDecreaseIn ?decrease}.
?R Ourl:belongToDocment ?doc .
?doc rdfs:label ?DocName .
?doc Ourl:documentHasDate ?Date .
FILTER (?indexName = ?x1)
FILTER (?Date = ?x2)
}
```

Figure 56: A SPARQL query that will retrieve information about all the shares belong to a specific index.

8.9. **Exploiting the knowledge-base in financial recommendation**

This section exemplifies how the resultant economic semantic knowledge-base can be meaningfully exploited in financial recommendation use-case scenario.

Recommender systems are information filtering systems that help users in finding the information that is related to their field of interest by implicitly or explicitly gathering and measuring the performance from other users [166]. Moreover, J. Bobadilla et al. have defined Recommender Systems as, "systems that provide consumers with personalised recommendations of goods or services and thus help consumers find relevant goods or services in the world of" information overload" [167]. With the rapid increase of data in diverse fields on the internet, recommender systems have been utilised in several domains.

J. Bobadilla et al., in [168] have classified the recommender system into three categories; collaborative filtering (CF), content-based filtering (CB), and hybrid filtering. The Knowledge-Based Recommendation System (KBRS) is a different type of recommender system in that it uses an alternative technique to produce a recommendation because it generates recommendations through the domain

knowledge. S. Bouraga et al. in [169], have presented a survey that has discussed the key ideas in the improvement of knowledge-based recommendation systems. Recently, economic recommendation systems have become one of the common applications to provide the user with advice about a specific query on different topics such as choosing the best goods and advising the user to choose an area for investments or which share is the best share in the stock market, based on different indicators. In the Arabic domain, there is still limited research on recommender systems, in particular in the economic domain. Next, an economic framework for an intelligent recommender system prototype will be presented.

8.9.1. Financial Knowledge-Based Recommender System framework (FKBR)

A simple FKBR prototype was built to demonstrate how the Arabic semantic knowledge-base can be intelligently explored. The prototype will allow us to review all the stages of the framework and to ensure that the framework can answer the research questions for this research, as well as providing an excellent opportunity to evaluate all the framework phases by testing the failure processes and effects analysis. The FKBR system consists of four main components:1) Information extraction component 2) Knowledge explicit Modelling, 3) Reasoner mechanism, 4) Interface component. Figure 58 illustrates the processes of the FKBR system scenario.

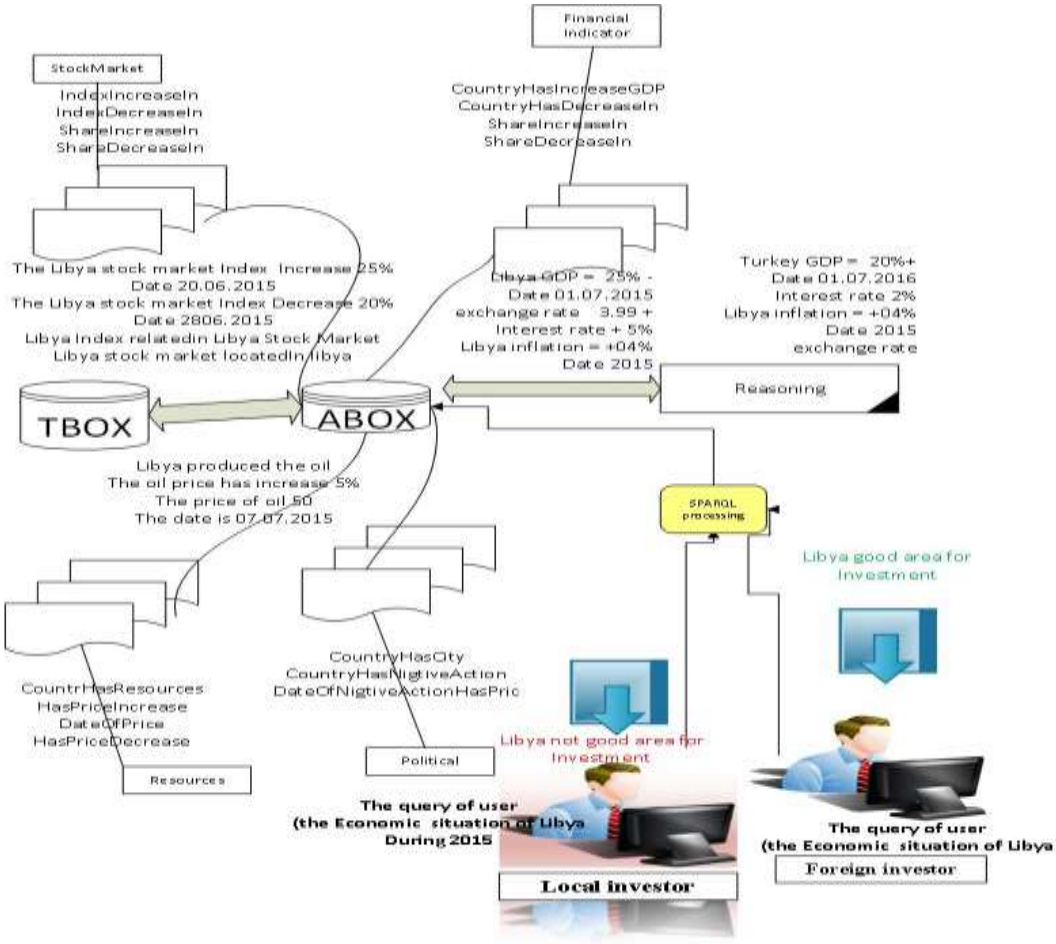


Figure 57: FKBR system scenario.

In the economic domain, the recommendation system plays the important role in assisting investors in decision making process. In this study, we have presented the SKBR system which able to constantly update the knowledge-base according to the consumer queries by using the IE techniques. By FKBR system, we demonstrate that IE techniques with the Semantic Web technologies to model the domain knowledge may complement each other more deeply. This will provide an opportunity to deliver integrated systems that will provide adequate and interesting results to users. In figure 59, the initial stage is the IE component that concerns to collect the data from several Arabic news on the Web and extract the useful information. The second stage is injecting the extracted information into knowledge-base by applying the Knowledge explicit modelling component that tries to represent the information as the RDF triples in the knowledge-base. The third component is reasoner mechanism component that aims to use the inference engine to model the user request to generate a new RDF triples, which will use to build the decision for recommendation system.

The final component is interface component that uses by the users to request and response their queries.

8.9.2. Implementation of the FKBR system prototype

We have implemented a prototype of the traditional recommender system model. The scenario applied in this prototype aims to use the existing information in the knowledge-base including the information about several countries such as economic indicators, stock market information, resources information and user information to infer a new knowledge about the state of these countries. The information has been collected from different Arabic online news sites. This information was gathered during the period between January 2017 and April 2017. In this prototype, we focused on implementing some rules and queries to evaluate the FKBR system.

To continue with our scenario, some of the information in the knowledge-base is about the countries of Qatar and Egypt. This information concerns the economic indicators for Qatar and Egypt in the year 2017 such as the GDP, inflation, change rate and interest rate.

Also, we have several users for our application. One of those users has an Egyptian nationality, and he is an investor. This information will be used with other information about the user such as the name of the user, nationality and type of user to recommend to the user if this country is suitable for investment as a local or foreign investor.

The following will explain each step in the scenario. In a first step, the user could ask this question "how is the state of investment in Qatar Country?". The system will create several triples consist of the information about the user request query, such as request number, date of the request, the name of the user and the country. The system will insert several triples in the knowledge-base. Figure 59 shows the green legend represents the information about the user request that is inserted in the knowledge-base.

In the second step, the information extraction component plays an important role by updating information of the knowledge-base. In this stage, the system automatically will contact the Web to collect many documents from Arabic financial news on the Web. This document will include analysis by NLP tasks to extract the useful information and then inject this information into the knowledge-base.

In the third step, the system will apply a reasoning with rules using Apache Jena to infer more information related to the user request. The inference engines have two main methods of reasoning forward and backward-chaining. Rules represent as a form of LHS => RHS. The left-hand side of the rule, LHS that is called the antecedent, body, and the right-hand side of the rule, RHS, is called the consequence, head[170]. In this stage, the reasoning that was involved in the query both forward-chaining that looks for rules where the LHS is true and add RHS facts to the knowledge-base, therefore it is data-driven. We have built the rules based on two types of the economic indicators (interesting rate and change rate) and other information available in the knowledge-base such as information about the target user and the information about the requesting.

As the result of this rule, the FKBR system could be expecting country's economic situation based on the previously mentioned indicators (interesting rate and change rate) and investors' type (local, foreign).

As mention in the figure 61; in the knowledge-base, there is information about the Qatar country such as:

- **The change rate indicator is decrease by 1.5% in 2017**
- **The interest rate indicator is increase by 5.0% in 2017**

And we have information about the investor such as:

- **The investor has nationality Egyptian**
- **The investor request information about Qatar in 2017**

The rules have built based on the following economic rules:

First rule

If the (interest rate indicator) is increased and (change rate indicator) is decreased and (investors' type) is local

then

The decision is this country could be not safe for investment

Second rule

If the (interest rate indicator) is increased and (change rate indicator) is decreased and (investors' type) is foreign

then

The decision is this country could be safe for investment

After executing these rules, the system will generate new triples that represent the recommendation advice about the user request such as:

"Qatar is safe contrary for investment because the interest rate indicator is increased and change rate indicator is decreased in 2017 "and also will classify the user as the foreign investor. Figure 60 shows the blue legend represents the information that is inferred by using theses rules.

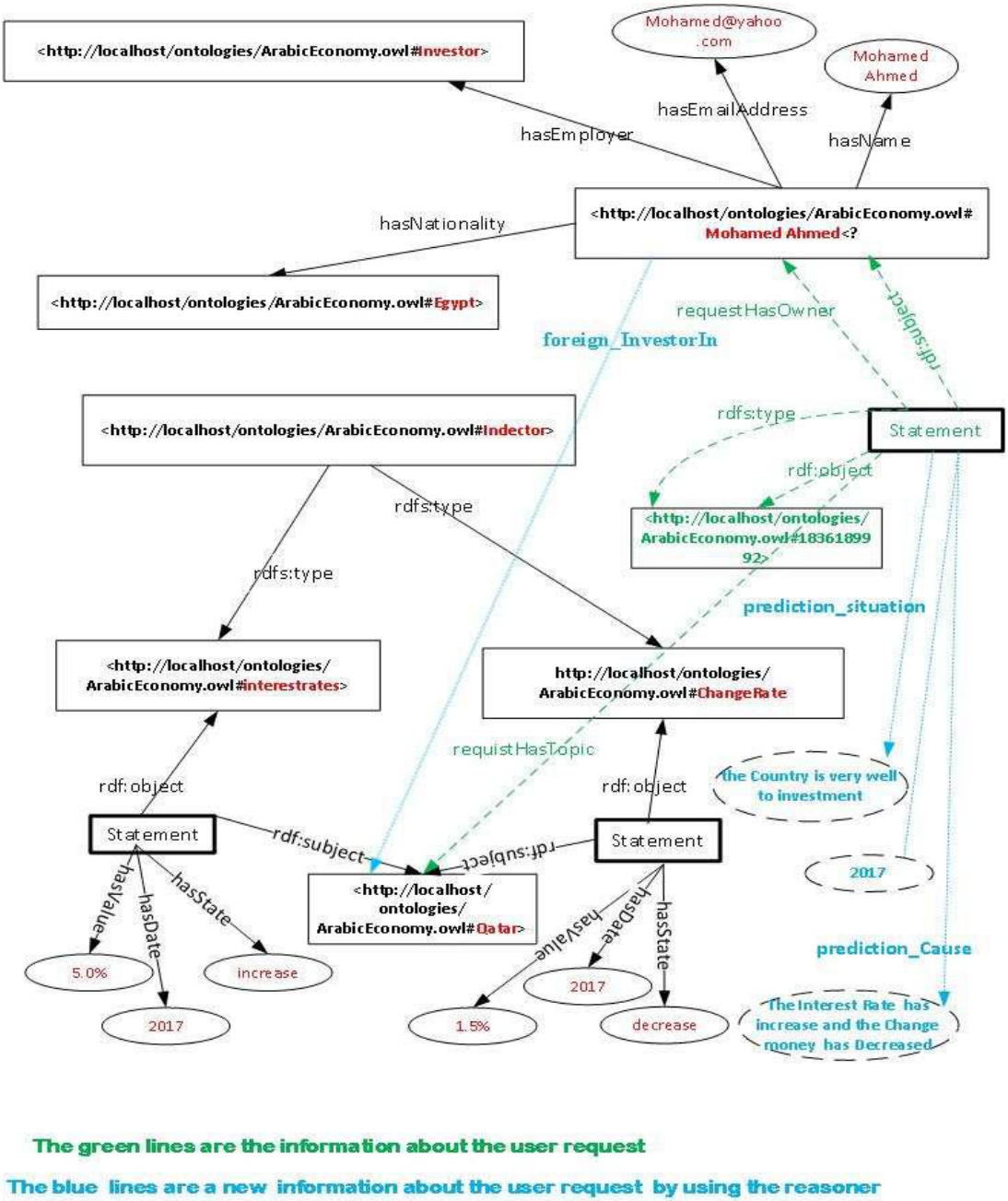


Figure 58: The reasoning result to infer information about the state of a specific country based on the type of investors.

In the final step, the SPARQL query approach is used to retrieve the result. The question of the user could be modelled into SPARQL query. The response of this query is to recommend advice to the user about this country which will show the recommender's decision that helps the user to decide to invest in this country or not. Also, the query will present more information related to the country such as the currency, name of cubital and the list of the company. Figure 60 shows the SPARQL query for retrieving the recommender decision that could be supported the investor to take their decision. Figure 61 shows the result of SPARQL query when the user requests Qatar country.

```
PREFIX rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX Ourl: <http://localhost/ontologies/ArabicEconomy.owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
select distinct ?L1 ?Owner ?Topic ?decision ?dateOfDecision  ?causeOfDecision ?TYPE ?cityName
?companyName ?currency
 where {
?R rdf:type rdf:Statement.
?R rdf:subject ?S.
?S rdfs:label ?L1.
?S Ourl:hasRequest ?request.
?R Ourl:requestHasOwner ?userName.
?userName  rdfs:label ?name.
?R Ourl:requistHasTopic ?nameCountry.
?nameCountry  rdfs:label ?country.
OPTIONAL{ ?R Ourl:typeOfInvestor ?TYPE.  FILTER (?TYPE = ?tt1).FILTER (?Topic != ?x2)}
OPTIONAL{ ?R1 Ourl:typeOfInvestor ?TYPE .FILTER (?TYPE = ?tt2).FILTER (?Topic = ?x2)}
OPTIONAL{ ?nameCountry Ourl:hasCurrency ?currency)}
OPTIONAL{ ?nameCountry Ourl:countryHasCity ?city.?city Ourl:rdfs:label ?cityName  }
OPTIONAL{ ?nameCountry Ourl:CountryHasCompany ?company.?company Ourl:rdfs:label
?companyName  }
?R Ourl:prediction_situation ?decision.
?R Ourl:prediction_Date ?dateOfDecision .
?R Ourl:prediction_Cause ?causeOfDecision .
FILTER (?Owner = ?x1)
}
```

Figure 59: The SPARQL query for retrieving the recommended decision

Figure 60: A screen shot for the result recommender system

8.9.3. Discussion of results

The main purpose of the evaluation a prototype framework is that to get rapid feedback on the usability of prototypes. In general, the FKBR system gives a good result by giving a better explanation of its recommendation and advice based on inference process.

However; in some cases, the FKBR system responses invalid the query results when applied some rules and queries. We can conclude there are two types of gaps in the knowledge-base representation. The first is unstructured document data gap that is related limited information in the Arabic documents and incorrect information in the Knowledge-base. Moreover, the second is bootstrapping data gap that is a related lack of Arabic structured data.

This is attributed to three factors: the limited information in the Arabic documents, Lack of structured data and the incorrect information in the Knowledge-base.

- **unstructured document data gap**

This type of gap consists the following types:

- **The limited information in the Arabic documents**

One of the important factors that have greatly affected to perform the FKBR system is the limited information in the Arabic documents. There is some information missing in the documents that affected the performance of the FKBR system. For example, the information related to the share, which help the investors to find out the best times of day to buy and sell a share. For example, if the user requests the information about the share of the specific company then the system will apply the relevant rules and queries to response the user

request. In some cases, there is data gap between the information requested by the user and the available data in the knowledge-base; as a result, this data gap will invalid the query results. We can overcome the problem of data gap by enriching the knowledge-base by other information sources. However, the availability of this information is limited especially in the Arabic language.

On the other hand, in some cases the documents contain information about the shares, but, this information is missing some details. For example, in the figure 62 screenshot from investing news shows the missing data, the investing news (https://sa.investing.com/news/) one of the important Arabic Website that contains daily information about the economic domain. However, sometimes these documents may lose important information such as in the figure 63 the price of the share mentioned as a number without the currency unit. The currency unit is important measurement to recognise several names entities such as price. The absent this measurement dues difficult to recognise the type of number.

الأسهم في كندا تغلق مرتفعة في نهاية تداولات يوم الجمعة، حيث صحبت – Investing.com
المؤشرات للأعلى، وقد شجلت مكاسب في قطاعات المواد، الطاقة والصناعات.

عند نهاية التداولات في تورونتو، ستاندرد آند بورز تي إس إكس أغلق على ارتفاع عند 0.65%.

من بين الأسهم القيادية اليوم في ستاندرد آند بورز تي إس إكس برز سهم فورتونا سيلفر ماينز إنك (TO:FVI)، الذي ارتفعت قيمته 7.64% أو 0.48 نقطة وبلغ سعره 6.76 عند الإغلاق. في المقابل، سهم مجموعة جران تيارا انرجي (TO:GTE) واصل ارتفاعه عند 6.30% أو 0.170 نقطة وأغلق عند سعر 2.870، في حين سهم Yamana Gold Inc (TO:YRI) زاد 6.01% أو 0.20 نقطة بسعر 3.53 عند نهاية

Figure 61: A screen shot from investing news shows the missing data

Also, in some cases, extracting the relation between the entities is a difficult task because of that the pair of entity does not appear in one sentence as shown in figure 64 the pair of entity could be in two separate sentences. In this study, the sentence is the baseline to extract the relation between entities. Each entity pair for a targeted relation that is mentioned in a sentence in unstructured data is identified and annotated as the relation. For example, in figure 63, to extraction the relation between the Index and the Share we need to find a link word or predicate to connect these entities. During the analysis, our data, we have noted this type of relation is difficult to extract directly because each entity appears in separate sentence such as ("العام البحرين بورصة مؤشر ","Bahrain Stock Exchange "(has appeared in the first sentence and ("الوطني البحرين بنك سهم ","National Bank of Bahrain shares" (appeared in the second sentence.


أغلق مؤشر بورصة البحرين العام (BB) خلال جلسة التداول اليوم
الأربعاء على تراجع طفيف بنسبة - 0,02 % ليصل إلى مستوى
1.168.80 نقطة. و بلغ حجم التداول في بورصة البحرين ما يقرب من
414.758 ألف سهم، فيما بلغت قيمة التداول 62.388 دينار بحريني.
وشهد مؤشر البحرين الإسلامي تراجعاً بنسبة - 0,74 % ليصل إلى
مستوى 754.30 نقطة. بينما ارتفع مؤشر استيراد بنسبة 0.18 %
ليصل إلى مستوى 1.247.78 نقطة.
وارتفعت أسهم بنك البحرين الوطني(NBB) بنسبة 3.73 %, كما
ارتفعت أسهم شركة زين البحرين (ZAINBH) بنسبة 1.20 %. فيما
تراجعت أسهم مصرف السلام - البحرين (SALAM) بنسبة 6.32 %,
وتراجعت أسهم البنك الأهلي المتحد (AUB) بنسبة - 0.76 %,
وتراجعت أيضاً أسهم شركة البحرين للاتصالات السلكية واللاسلكية
(BATELCO) بنسبة 0.65 %.
وارتفع مؤشر قطاع البنوك التجارية 0.90 نقطة، بينما تراجع مؤشر
قطاع الخدمات 3.22 نقطة، واستقرت قطاعات الاستثمار، التأمين،
الفنادق والسياحة والصناعة دون تغير.

Figure 62: A screen shot from random news shows the two entities each entity in separate sentence

Therefore, we have applied some reasoning rules to extract this type of relation to improve the result. The main idea of the rule is that if the index and the share appeared in the same document, then the reasoning will infer a new relation between the index and the share (Share belongto Index). This type of rule is suitable with the document that mentioned one Index. The Figure 64 shows the screen shortcut for this rule and the result.

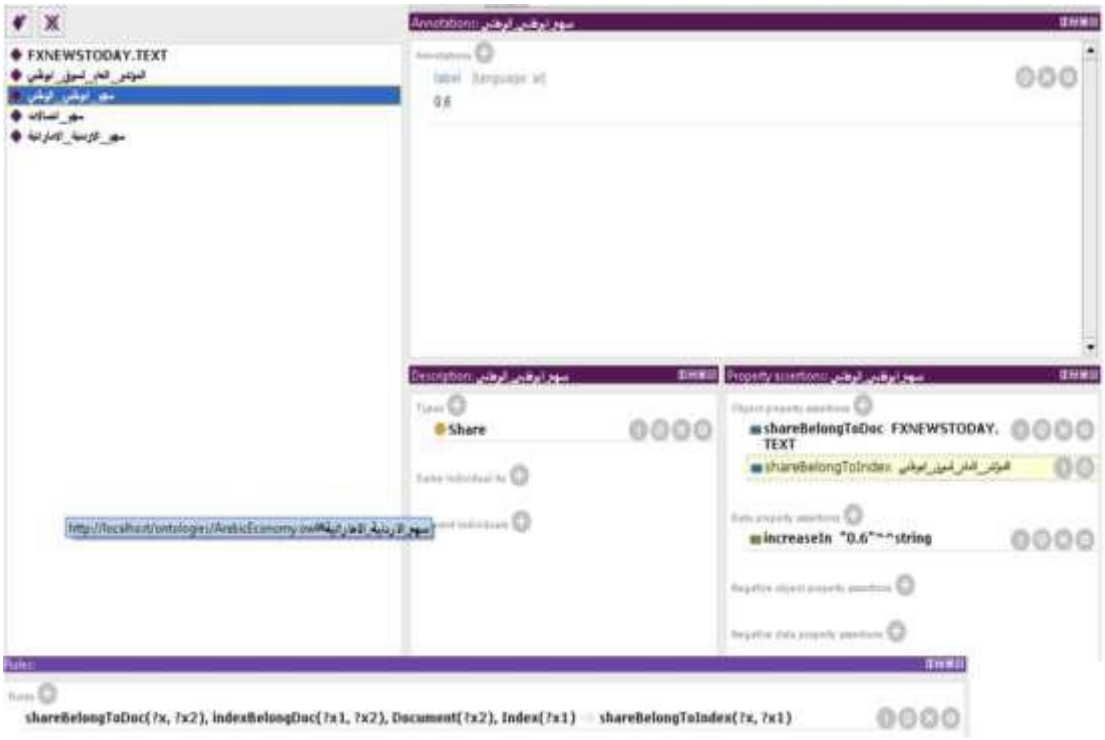


Figure 63: A screen shot for the rule and the result to extract the relation between Index and Share when the documents consist of one Index.

However, in some cases, this rule unsuitable for extracting this relation because the document contains more than one Index see figure 65. So, in future, we need more investigation how to solve this problem because it is important to apply some rules and queries.

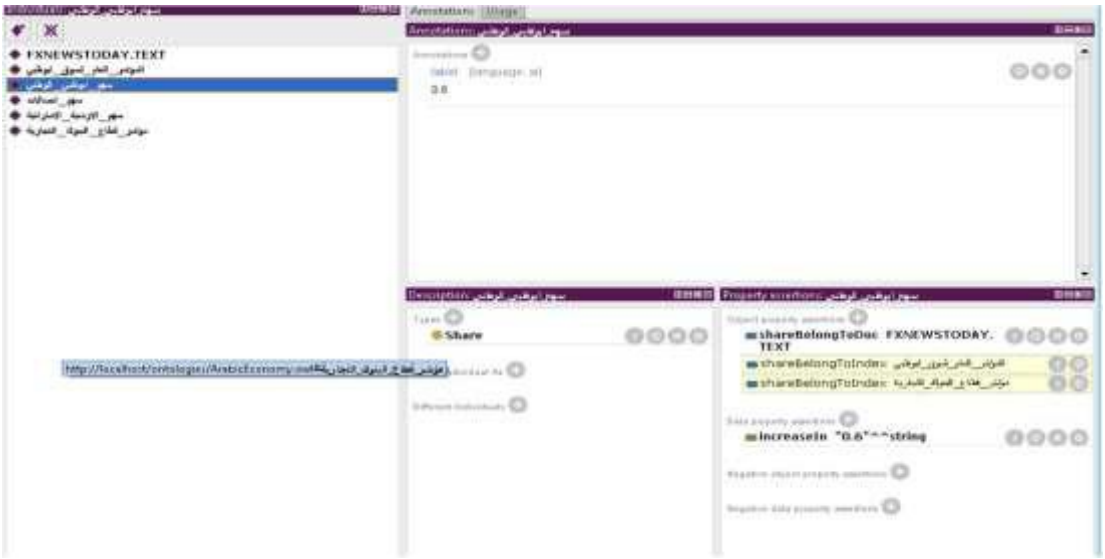


Figure 64: A screen shot the rule and the result to extract the relation between Index and Share when the documents consist more one Index.

- **The incorrect information in the Knowledge-base**

Knowledge Base population is a process of extracting facts about entities and relation between the entities from a large amount data and utilised it to augment a knowledge-base (KB). These processes involved two tasks NLP technique and Semantic Web technology task. The NLP task is an important task that concerns to convert the unstructured data into structured information and populate this information in the KB.

In this study, NLP technique involved the NE recognition and relation extraction plays an important role in the development and building the KB. There are several methods have been applied to extracted and improve the NLP tasks as mention in the previous chapters. Experiments for this method have shown very accurate results. Many times, we have noted that the FKBR system retrieval incorrect information from the KB, which is caused to retrieve incorrect answers for some queries. Figure 66 shows the inaccurate information that was extracted by the NLP stage and populate to the KB. This type of errors originated from the incorrect annotated of some Arabic semantic relation. The incorrect extraction for the semantic relation often led to the generation of incorrect information by the FKBR system. A deep analysis shows that this error is correct logically because there is entity pair appeared in the same sentence for the target relation that is relation between Country and City. However, semantically incorrect because of the ("الرياض","Riyadh") city is not a city in the ("فرنسا ","France") country. It is noteworthy that the errors results happened through problems associated with the NLP tasks that are mentioned in the discussion of the result in the previous chapters.

Figure 65: A screen shot for Incorrect Result Due to the NLP task

- Bootstrapping data gap

This type of the data gap related to the lack of availability Arabic information in the structured public information.

- **Lack of Arabic structured data**

There are several efforts have used the Linked Open Data to populate and enrich an ontology like DBpedia and Freebase datasets because they consist much information. It is very useful get more information about a specific domain such as economic, sport and political. In economic domain, DBpedia consists much information about different topics in an economic domain such as information about places, organisations and financial information. The availability of Arabic language entities in DBpedia is still limited that affects the enrichment of Arabic knowledge-base from DBpedia. For example, as mention in the previous section; in some cases, when Appling the SPARQL query that models the question "what is the stock indexes names in Arabic". The SPARQ query could not retrieve any stock index name in the Arabic language.

8.10. **Summary**

In this chapter, we were concerned with using the Semantic Web to model the domain knowledge to represent the information that is extracted from the unstructured data into a reusable format (RDF triples) which are saved in the knowledge-base ontology. This knowledge-base can be used for several purposes which covers a part of the economic and finance domain such as information about stock markets and information about the economic indicators of countries. We proposed the financial knowledge-base recommender system to exploit the resultant knowledge-base to produce a recommended investment decision and intelligently explore the knowledge-base by end users in terms of our motivation scenario.

This knowledge-base recommender system utilized Semantic Web technologies such as reasoning and SPARQL queries to achieve our proposed motivation scenario. The reasoning mechanism was used to infer new knowledge from existing information in the knowledge-base such as the economic indicators of countries that will support the users in the stock market investment decision-making process. Moreover, the SPARQL query has been used to express queries across the ontology to be used by users to intelligently explore the knowledge-base.

Chapter 9

9. Conclusions and Future Work

9.1. Overview

Although there is an extensive amount of Arabic information that can be found on the Web and other resources, most of this information is unstructured data. The Arabic language is a language of considerable interest to the NLP community mainly due to its economic and political importance. In this thesis, we presented a comprehensive framework that will intelligently extract the information from an Arabic unstructured data on the web by employing a Semantic Web technique to improve the intelligent exploration of unstructured documents written in Arabic.

This framework consists of several phases, with each one developed to improve information extraction challenges in the Arabic domain. In the initial phase, we performed a comprehensive analysis of the problem domain to capture the knowledge that would eventually allow for the construction of a comprehensive semantic knowledge-base for the target domain. The captured knowledge was modelled into a knowledge map containing information about the chief domain concepts and the interrelations connecting them, as well as the lexica published in public sources. The modelled knowledge was formalised into a machine-readable format and encoded as a semantic ontology. The resulting ontology maintains high coverage of lexical terms as well as semantic representation of concepts; therefore, the ontology resulted from this research would be useful for Arabic NLP applications beyond the scope of this work, especially for applications focusing on the economic or financial domains.

In second phase, we have developed and implemented the rule-based NER pipeline to extract and classify NEs from Arabic economic documents. A set of syntactical rules and patterns has been built, whilst considering features, such as prefixes and suffixes of words, morphological and POS information, information about the surrounding words and their tags. Also, predefined economic and general indicator lists were utilised, such as gazetteer

lists, and an Arabic named entity annotation corpus from the economic domain. The Arabic NEs pipeline has achieved an overall improvement for extracting the Arabic named entity, with the exception that when dealing with the organisation's names for the composite name because the initial approach is not enough to recognise the composite names. This is because the main challenge was to predict the boundaries of the NE, particularly with long and composite named entities.

Therefore, advanced Arabic grammar rules are used to recognise the Arabic composite names. These advanced Arabic grammar rules aim to classify the Arabic words as the definiteness and indefiniteness nouns and employ the genitive rules to build different patterns in order to recognise Arabic composite names. The evaluation produced high results and would be helpful in solving ambiguity problems in the research's domain. Also, the current research results are meaningful because the rules of the pattern in our approach are quite logical and simple. Furthermore, the semantic analysis of the domain knowledge can be used to help the NLP process to benefit from structured data available on the Web (DBpedia data set) to enrich the gazetteers lists used to extract the Arabic NEs.

In the third phase, a novel, knowledge-based approach to relation extraction from unstructured Arabic text was built. It is based on the principles of Functional Discourse Grammar. This type of grammar is a set of rules and processes that govern the semantic aspects of sentences in a given language, regardless of the structure of the sentence. FDG relies on two main terms: a predicate and an argument. We have enumerated several problems related to the Arabic relation extraction task, in addition to the problems related to the complex structure of the Arabic sentences. The FDG algorithm registered satisfactory results in extracting Arabic relations from unstructured texts, but exhibited some limitations when dealing with especially complex Arabic sentence structures, as in sentences that contained relations that have more than one trigger word. We addressed this problem by integrating the rule-based FGD approach with Machine Learning relation classification, resulting in a hybrid algorithm that improved the precision of our relation classification approach.

The fourth stage of the research focused on the population of the semantically tagged data into the economic domain's semantic knowledge-base. This knowledge-base covers part of the economic and financial domain, such as information about stock market and information about the economic indicators of countries. We proposed the financial knowledge-base

recommender system to exploit the resultant knowledge-base to produce a recommended investment decision and intelligently explore the knowledge-base by end users in terms of our motivation scenario.

This knowledge-base recommender system utilised the Semantic Web technologies, such as reasoning and SPARQL queries to achieve the proposed motivation scenario. The reasoning mechanism using the specific rules is used to infer a new knowledge from existing information in the such as the economic indicators of countries, which will support the users in stock market investment decision-making processes. Moreover, SPARQL query has been used to express queries across the ontology to be used by users to intelligently explore the knowledge-base.

In the final stage, we have presented a user scenario to evaluate the performance of the framework and to identify the most important issue that affects the performance of the framework.

In general, the framework gives acceptable results by providing a better explanation of its recommendation and advice based on the inference process.

During this evaluation, we observed that there are several challenges faced when attempting to improve the creation of an Arabic financial intelligent application that relies on unstructured data on the Web, such as the complexity of Arabic sentences structured in terms of composite names and the difficulties of building grammar based rules because of the variability in the way the relations can be expressed in Arabic sentences.

We have concluded that the research and development in the area of Arabic Information Extraction is significantly underdeveloped, which puts the exploitation of the huge information resource on the Web at a serious disadvantage in the era of globalisation, Internet culture, information technology and the knowledge economy. It is hoped that this effort will contribute towards redressing the situation as the presented framework showcases a comprehensive methodology for exploiting domain-specific information published on the Web starting from analysis of the target domain knowledge to the final stage of organising the extracted knowledge in a semantic knowledge-base and building intelligent rules that encode the requirements of the beneficiary group in the new knowledge inference engine.

9.2. **Thesis contributions**

In my opinion, the thesis has addressed the aim of this work as expressed in the title; i.e. building "a Hybrid NLP-Semantic Knowledge Base for the Intelligent Exploration of Arabic Documents" has comprehensively been addressed by the design and implementation of the Arabic information extraction (AIE) framework, as detailed in the previous chapters. The research and development effort in this thesis aimed to respond to the significant research problem posed at the start of this thesis.

Question 1 How can Arabic grammar rules be exploited to improve recognition of Arabic composite names?

Initially, this research investigated the effective handling of Arabic grammar rules to extract Arabic words, such as names of companies, indices, shares, ports, organisations, and countries which are related to our case study within the Arabic financial domain. At this stage, a major challenge in composite name consideration was to predict the boundaries of the names and the length of the Arabic NE.

In this study, we presented a novel approach for extracting composite names from documents authored in the Arabic language. Our approach relies on the Arabic grammar rules for classification of pronouns into Definite Nouns "الاسم المعرفة" and Indefinite Nouns " الاسم النكرة", and extends them to provide classification within phrases using complex genitive pattern recognition. Experimental evaluation was carried out on financial documents with varied authoring styles revealed good precision and recall results. In this study, from the results of the assessment, it was found that the accuracy of the retrieval performance for extracting the composite names from unstructured documents based on different experiments was more than 96%. On the other hand, similar approaches addressing composite name problems in the Arabic language were not found. Based on this, the proposed approach was not compared against any other method. However, using the Arabic grammar rule improved the extraction performance of the named entity recognition task with regards to accuracy. The results presented in this thesis could be a useful reference for other researchers when comparing their performance.

Question 2 Can the ontology-based approach be used to improve the state of art of Arabic relation extraction?

In this study, we used a knowledge-based approach for extracting the semantic relation out of Arabic documents on the web. A new approach was presented for extracting semantic

relations between high-level (concepts in the ontology) and low level (instances in the text) aimed at detecting and classifying semantic relations between entities according to predefined concepts and relations in the ontology.

The main objective of our work is to extract several types of relations based on the existing relationships in the ontology from the sentence and predict the trigger word based on several synonyms that describe a semantic relation in the ontology. Each synonym represents the root of several Arabic words. This method has helped to overcome the limitations of identifying the target relation that we aim to extract from the text. In addition, it improves the capability of detecting the trigger word that expresses the relation between the entities, regardless of the position of that word in the sentences and its form. We observed that this mechanism performs much better for extracting the semantic relation on clearer and simpler sentences, especially after increasing the number of synonyms for the relations in the ontology.

Question 3 Can exploitation of the structure of Arabic grammar be useful in improving the current state of the art relation extraction?

Relation extraction is a useful task for many NLP systems. In contrast to the critical attempts concerning English or other European languages, the efforts of extracting the relations in Arabic language domains are relatively limited. This limitation might be due to several challenges that face extracting the Arabic relation from the Arabic text, such as the trigger word of relation that describes the relation. This trigger word could take a variety of forms and positions in the sentence and sometimes more trigger words describe the relationship. Also, there is another limitation to extract the relation in the Arabic sentences, which is a complex of relations in the sentence. Many researchers have used the syntax grammar to extract the relations which deals with the relationship between words in the sentence by assigning the subject, object, and predicate for each relation.

However, Arabic language sentences often contain complex (high order) relations where one subject has several predicates or several objects with varying order of the features in the sentence. On the other hand, the functional theories of grammar consider the functions of language and its elements to be key to the understanding of linguistic processes and structures; thus emphasising the semantic and pragmatic properties of a language. Hence, we believe that functional grammar offers a more flexible abstraction for modelling the complex Arabic language relations. In so doing, we adopted the principles of Functional

Discourse Grammar, an advanced version of Functional grammar, as the basis for building a novel approach to relation extraction from Arabic natural texts. As indicated in the motivation section, the problem domain of choice is the financial domain, and this case study particularly focuses on articles related to news about the stock market, which contain very complex sentences that is bound to challenge the proposed relation extraction approach.

Question 4 If a knowledge-base approach is adopted for IE, can classification technique that are based on ML play a positive role in the IE process?

The rule-based approach proved effective in IE, but some limitations that were observed when dealing with the overly complex sentence structures were managed by integrating it with Machine Learning relation classification.

In this study, the rule-based approach has been used to extract the semantic relations from the Arabic text by employing the semantic discourse grammar rule. This method provides a better result in different complex relations. However, during a critical analysis of the results, we have noted that the results of some relations still needed improvement, especially the relations that based more than one trigger word to extract the relation. Meanwhile, we acknowledge that the rule based approach does not readily satisfy the extraction of this type of relation.

Therefore, we decided to create a new approach that will improve the relation extraction task. The Hybrid approach between ML approach and rule based approach is used to improve the results of these relations.

Overall, the results indicate that the hybrid approach presents higher results than the rule-based approach regarding the F-measure (0.80, 0.74, 0.71) in recognising the relations which are based on two trigger words.

Question 5 Given the complex structure of the Arabic language, how do the linguistic features impact on the relation classification in the Arabic sentence structure?

We have comprehensively investigated the linguistic influence of the information extraction process when attempting to build training data sets for ML driven relation extraction algorithm. Our analysis showed that the lexical feature has produced a better result than other types of feature categories, including the features which focus on choosing words before the first named entity and the words after named entity. These linguistic features could increase the approach's performance because of the characteristic of Arabic linguistic grammar in the

Arabic sentences. To help with the selection of the most optimal feature that can improve the accuracy of ML relation classification, we deployed an optimisation algorithm; specifically the Genetic Algorithm, which in fact proved that the optimisation of the feature selection did improve overall accuracy.

Question 6 Can the adoption of Semantic Web standards for modelling the domain knowledge improve the information extraction processes and the subsequent intelligently exploration of the extracted information?

Yes, it improves the NLP because we can enrich the gazetteer list by using the structured datasets published in the Linked Open Data cloud. Hence, it was straight-forward to import the relevant key concepts and relations from Wikipedia because we are using the same formalisation for our semantic modelling of our economic knowledge-base. Also, it can improve intelligent exploration processes because it offers advanced integration techniques and advanced reasoning approaches.

We presented a hybrid NLP-Semantic framework that uses fundamental NLP techniques and proprietary grammar rules to semantically tag Arabic texts for a specific domain (Economy was our use case), which was followed by knowledge inference in an appropriately modelled semantic knowledge-base to discover new facts in the mined text by building rules that allowed intelligent exploration of the eventual knowledge-base.

Question 7 How can the gap between the human language and formalised knowledge be bridged to improve intelligent exploration of unstructured documents?

In this work, we presented a framework for integrating domain-specific knowledge with NLP technique that was proved to be effective in solving most classification problems to improve the intelligent exploration of online Arabic unstructured data on the Web. In this framework, we first used an NLP technique to address the Arabic document on the Web to extract the useful information and to populate the information into a knowledge-base. Secondly, we have used the Semantic Web technologies to model the economic domain knowledge in order to build a knowledge-based recommender systems. We have focused on improving the performance of the NLP tool to improve the intelligent exploration by employing the Arabic grammar rules to extract useful information from Arabic unstructured data on the Web. During the evaluation of the framework, we noted that improving the intelligent exploration relies on two issues: First, the quality of the information extracted from the unstructured data relies on improving the NLP tools, which was evident from the

results obtained from extracting the composite names along with relation extraction between the entities; the second is the quality of advanced rules that use this information to infer a new knowledge. These rules describe the relation between information in the knowledge-base. We have noted from our discussion that the core task that effectively bridged the gap between NLP techniques and Semantic Web technologies and improved the intelligent exploration in Arabic system has enhanced NLP tools in terms of overcoming the challenges that are related to Arabic NE recognition and relation extraction, such as composite names and complex relations.

9.3. **Further work**

Based on the aforementioned limitations, some outstanding research issues are thus highlighted for this research effort to leave lasting impacts and go further.

- **Improve the FDG algorithm by including Co-reference resolution**

The relation extraction mechanism of the FDG algorithm is designed to extract relations between entity pairs in the same sentence. However, in the Arabic language context, the same named entity could be mentioned in different sentences in the same document to provide more information about that named entity. It is proposed that the FDG algorithm is improved to process the whole document to extract the relations between all named entities in the documents. However, this improvement requires processing the whole document to track the matched named entities. This process is called Co-reference resolution or Ortho-matcher process. To the best of the researcher's knowledge, no reported work has been published so far to develop a tool in NLP to perform Co-reference resolution in Arabic.

- **Language texts Semantic Web query language expansion**

Applying a free-text query is an added-value to the benefits of this research work because it will enhance the user's interactivity with the framework and extend its accessibility to casual users rather than the domain experts. The challenge with this area of research lies in converting the naturel language query for the user to the standards for RDF querying language, such as SPARQL extension. A possible area of further research would be to utilise our Arabic Named Entity Recognition (ANER) pipeline FDG relation extraction to extract the key words and their interrelations from the user request query and expanding them with

matching concepts and object relations from the ontology, and then converting the original query to one or several standard RDF queries (SPARQL) query that can be directly fired against the Arabic knowledge-base.

- **Investigating the representation and application of fuzzy rules**

The need to deal with fuzzy information in Semantic Web languages is rising in importance, and, thus, calls for a standard way to represent such information. We believe that in our Arabic knowledge-base of this research, there are several fuzzy ground facts; for example, the reason for the level of GDP rate of a specific country to be considered as high or low. Thus, it is important to further investigate the representation of these fuzzy ground facts and the application of fuzzy rules on the resultant semantic knowledgebase for more explicit reasoning and to improve queries results.

- **Building Arabic language semantic resources**

As mentioned in this thesis, the Arabic language suffer from scarcity of structured Arabic information. For this reason, it is hoped that this framework is considered as a starting point to build an academic centre of Arabic language resources as a linked knowledge-base repository, which allows researchers to contribute domain-specific conceptual models in the form of standardised ontologies to encourage their reuse and dynamic population with current data.

References

[1] S. Sarawagi, "Information extraction," *Foundations and Trends in Databases,* vol. 1, *(3),* pp. 261-377, 2008.

[2] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," *ACM Transactions on Asian Language Information Processing (TALIP),* vol. 8, *(4),* pp. 14, 2009.

[3] P. Castells *et al*, "Semantic web technologies for economic and financial information management," in *The Semantic Web: Research and Applications*Anonymous Springer, 2004, pp. 473-487.

[4] S. Mesfar, "Named entity recognition for arabic using syntactic grammars," in *Natural Language Processing and Information Systems*Anonymous Springer, 2007, pp. 305-316.

[5] M. Hijjawi and Y. Elsheikh, "Arabic language challenges in text based conversational agents compared to the English language," *International Journal of Computer Science and Information Technology (IJCSIT),* vol. 7, *(5),* pp. 1-13, 2015.

[6] J. Maloney and M. Niv, "TAGARAB: A fast, accurate arabic name recognizer using high-precision morphological analysis," in *Proceedings of the Workshop on Computational Approaches to Semitic Languages,* 1998, pp. 8-15.

[7] M. K. Saad and W. Ashour, "OSAC: Open source arabic corpora," in *6th International Symposium on Electrical and Electronics Engineering and Computer Science, Cyprus,* 2010, pp. 118-123.

[8] A. Alsaad and M. Abbod, "Arabic text root extraction via morphological analysis and linguistic constraints," in *Computer Modelling and Simulation (UKSim), 2014 UKSim-AMSS 16th International Conference On,* 2014, pp. 125-130.

[9] R. Al-Shalabi and M. Evens, "A computational morphology system for arabic," in *Proceedings of the Workshop on Computational Approaches to Semitic Languages,* 1998, pp. 66-72.

[10] M. Gridach and N. Chenfour, "Developing a New Approach for Arabic Morphological Analysis and Generation," *arXiv Preprint arXiv:1101.5494,* 2011.

[11] K. Toutanova *et al*, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1,* 2003, pp. 173-180.

[12] stanford.edu, "The Stanford Natural Language Processing Group," 2016. [Online]. Available: https://nlp.stanford.edu/software/tagger.shtml. [Accessed: 07-Dec-2017].

[13] G. Kanaan, R. Al-Shalabi and M. Sawalha, "Improving Arabic information retrieval systems using part of speech tagging," *Information Technology Journal,* vol. 4, *(1),* pp. 32-37, 2005.

[14] H. Shouhani Rabiee, "Arabic Language Analysis Toolkit." 2011.

[15] S. Khoja, "APT: Arabic part-of-speech tagger," in *Proceedings of the Student Workshop at NAACL,* 2001, pp. 20-25.

[16] J. H. Yousif and T. M. T. Sembok, "Arabic part-of-speech tagger based support vectors machines," in *2008 International Symposium on Information Technology,* 2008, pp. 1-7.

[17] K. Shaalan and H. Raza, "Person name entity recognition for arabic," in *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources,* 2007, pp. 17-24.

[18] M. Oudah and K. F. Shaalan, "A pipeline arabic named entity recognition using a hybrid approach." in *COLING,* 2012, pp. 2159-2176.

[19] S. Zaidi, M. Laskri and A. Abdelali, "Arabic collocations extraction using gate," in *Machine and Web Intelligence (ICMWI), 2010 International Conference On,* 2010, pp. 473-475.

[20] A. Elsebai, F. Meziane and F. Z. Belkredim, "A rule based persons names Arabic extraction system," *Communications of the IBIMA,* vol. 11, *(6),* pp. 53-59, 2009.

[21] L. Han, T. Finin and A. Joshi, "UMBC CSEE Technical Report TR-CS-11-08 GoRelations: Towards an Intuitive Query System for RDF Data," 2012.

[22] P. N. Mendes, D. Weissenborn and C. Hokamp, "DBpedia Spotlight at the MSM2013 Challenge," *Making Sense of Microposts (# MSM2013),* 2013.

[23] H. Al-Feel, "A Step towards the Arabic DBpedia," *International Journal of Computer Applications,* vol. 80, *(3),* pp. 27-33, 2013.

[24] M. Radzimski *et al*, "FLORA–Publishing unstructured financial information in the linked open data cloud," in *International Workshop on Finance and Economics on the Semantic Web (FEOSW 2012),* 2012, pp. 27-28.

[25] T. Declerck *et al*, "Ontology-based multilingual access to financial reports for sharing business knowledge across europe," in *Internal Financial Control Assessment Applying Multilingual Ontology Framework*Anonymous 2010, .

[26] V. Chenthamarakshan, P. M. Desphande, R. Krishnapuram, R. Varadarajan, and K. Stolze, "WYSIWYE: An Algebra for Expressing Spatial and Textual Rules for Visual Information Extraction," Jun. 2015 [Online]. Available: http://arxiv.org/abs/1506.08454. [Accessed: 07-Dec-2017].

[27] A. Yates, *Information Extraction from the Web: Techniques and Applications,* 2007.

[28] C. Chang *et al*, "A survey of web information extraction systems," *IEEE Trans. Knowled. Data Eng.,* vol. 18, *(10),* pp. 1411-1428, 2006.

[29] M. Banko *et al*, "Open information extraction from the web." in *IJCAI,* 2007, pp. 2670-2676.

[30] O. Etzioni *et al*, "Web-scale information extraction in knowitall:(Preliminary results)," in *Proceedings of the 13th International Conference on World Wide Web,* 2004, pp. 100-110.

[31] M. Tadić, "XLike: Cross-lingual knowledge extraction," in *Machine Translation Summit XIV,* 2013, .

[32] J. Aguilar, P. Valdiviezo-Díaz and G. Riofrio, "A general framework for intelligent recommender systems," *Applied Computing and Informatics,* 2016.

[33] sakhr.com, "Sakhr Software - Sakhr." [Online]. Available: http://www.sakhr.com/index.php/en/. [Accessed: 07-Dec-2017].

[34] sakhr.com, "Sakhr Software - Knowledge Management." [Online]. Available: http://www.sakhr.com/index.php/en/solutions/knowledge-management. [Accessed: 07-Dec-2017]

[35] maknaz.org, "MAKNAZ." [Online]. Available: http://www.maknaz.org/. [Accessed: 07-Dec-2017].

[36] T. Helmy and A. Daud, "Intelligent agent for information extraction from arabic text without machine translation," in *Proceedings of the 1st International Workshop on Cross-Cultural and Cross-Lingual Aspects of the Semantic Web,* 2010, .

[37] D. M. Daoud, "Building an arabic application employing information extraction technology," in *Proceedings of the Second International Conference on Information Technology (ICIT05), Amman, Jordan,* 2005, pp. 1-9.

[38] L. M. B. Saleh and H. S. Al-Khalifa, "AraTation: An arabic semantic annotation tool," in *Proceedings of the 11th International Conference on Information Integration and Web-Based Applications & Services,* 2009, pp. 447-451.

[39] M. Alruily and M. Alghamdi, "Extracting information of future events from arabic newspapers: An overview," in *Semantic Computing (ICSC), 2015 IEEE International Conference On,* 2015, pp. 444-447.

[40] K. Shaalan and H. Raza, "Person name entity recognition for arabic," in *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources,* 2007, pp. 17-24.

[41] K. Nebhi, "A rule-based relation extraction system using DBpedia and syntactic parsing," in *Proceedings of the 2013th International Conference on NLP & DBpedia-Volume 1064,* 2013, pp. 74-79.

[42] W. Li *et al*, "A novel feature-based approach to chinese entity relation extraction," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers,* 2008, pp. 89-92.

[43] T. Arts *et al*, "arTenTen: Arabic corpus and word sketches," *Journal of King Saud University-Computer and Information Sciences,* vol. 26, *(4),* pp. 357-371, 2014.

[44] I. Boujelben, S. Jamoussi and A. B. Hamadou, "RelANE: Discovering relations between arabic named entities," in *Text, Speech and Dialogue,* 2014, pp. 233-239.

[45] rss.com, "RSS Reader & Podcast Hosting." [Online]. Available: https://rss.com/. [Accessed: 07-Dec-2017].

[46] R. Tairas, M. Mernik and J. Gray, "Using ontologies in the domain analysis of domain-specific languages," in *International Conference on Model Driven Engineering Languages and Systems,* 2008, pp. 332-342.

[47] D. Jones, T. Bench-Capon and P. Visser, "Methodologies for ontology development," 1998.

[48] M. Uschold and M. King, *Towards a Methodology for Building Ontologies.* Citeseer, 1995.

[49] C. P. Menzel and R. J. Mayer, "IDEF5 ontology description capture method: Concept paper," 1990.

[50] N. F. Noy and D. L. McGuinness, "Ontology development 101: A guide to creating your first ontology," 2001.

[51] S. Boyce and C. Pahl, "Developing domain ontologies for course content," *Educational Technology & Society,* vol. 10, *(3),* pp. 275-288, 2007.

[52] H. G. L. Kerschberg, "A Knowledge-Based Approach to Generating Target System Specifications from a Domain Model," .

[53] M. Horridge *et al*, "A Practical Guide To Building OWL Ontologies Using The Protégé-OWL Plugin and CO-ODE Tools Edition 1.0," *University of Manchester,* 2004.

[54] edmcouncil.org, "EDM COUNCIL: Financial Industry Business Ontology™." [Online]. Available: https://www.edmcouncil.org/financialbusiness. [Accessed: 07-Dec-2017].

[55] w3.org, "vCard Ontology - for describing People and Organizations." [Online]. Available: https://www.w3.org/TR/vcard-rdf/. [Accessed: 07-Dec-2017]

[56] A. Simón, L. Ceccaroni and A. Rosete, "Generation of OWL ontologies from concept maps in shallow domains," in *Conference of the Spanish Association for Artificial Intelligence,* 2007, pp. 259-267.

[57] H. Khalil and T. Osman, "Challenges in information retrieval from unstructured arabic data," in *Proceedings of the 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation,* 2014, pp. 456-461.

[58] J. Heflin, "An Introduction to the OWL Web Ontology Language," *Lehigh University.National Science Foundation (NSF),* 2007.

[59] V. Graudina and J. Grundspenkis, "Concept map generation from OWL ontologies," in *Proceedings of the Third International Conference on Concept Mapping, Tallinn, Estonia and Helsinki, Finland,* 2008, pp. 263-270.

[60] R. Grishman and B. Sundheim, "Message understanding conference-6: A brief history." in *Coling,* 1996, pp. 466-471.

[61] A. Borthwick, *A Maximum Entropy Approach to Named Entity Recognition,* 1999.

[62] W. Karaa and T. Slimani, "A New Approach for Arabic Named Entity Recognition." *International Arab Journal of Information Technology (IAJIT),* vol. 14, *(3),* 2017.

[63] M. OUDAH and K. SHAALAN, "NERA 2.0: Improving coverage and performance of rule-based named entity recognition for Arabic< a href="# afn1">*</a," *Natural Language Engineering,* pp. 1-32, 2016.

[64] M. S. Al Tayyar, "Arabic information retrieval system based on morphological analysis (AIRSMA): a comparative study of word, stem, root and morpho-semantic methods." 2000.

[65] H. Cunningham *et al*, *Developing Language Processing Components with Gate Version 5:(A User Guide).* University of Sheffield, 2009.

[66] S. Green and C. D. Manning, "Better arabic parsing: Baselines, evaluations, and analysis," in *Proceedings of the 23rd International Conference on Computational Linguistics,* 2010, pp. 394-402.

[67] H. Shouhani Rabiee, "Arabic Language Analysis Toolkit." 2011.

[68] M. Silberztein, T. Váradi and M. Tadic, "Open source multi-platform NooJ for NLP." in *COLING (Demos),* 2012, pp. 401-408.

[69] M. Mourchid, I. Blanchete and A. Mouloudi, "STANDARD ARABIC VERBS INFLECTIONS USING NOOJ PLATFORM," .

[70] A. B. Hamadou, O. Piton and H. Fehri, "Recognition and translation Arabic-French of Named Entities: case of the Sport places," *arXiv Preprint arXiv:1002.0481,* 2010.

[71] M. Boudchiche *et al*, "Alkhalil morpho sys 2: A robust arabic morpho-syntactic analyzer," *Journal of King Saud University-Computer and Information Sciences,* vol. 29, *(2),* pp. 141-146, 2017.

[72] A. Feriel and K. M. Khireddine, "AUTOMATIC EXTRACTION OF SPATIO-TEMPORAL INFORMATION FROM ARABIC TEXT DOCUMENTS," .

[73] S. Zaidi, M. Laskri and A. Abdelali, "Arabic collocations extraction using gate," in *Machine and Web Intelligence (ICMWI), 2010 International Conference On,* 2010, pp. 473-475.

[74] A. Alfaries *et al*, "A rule-based annotation system to extract tajweed rules from quran," in *Advances in Information Technology for the Holy Quran and its Sciences (32519), 2013 Taibah University International Conference On,* 2013, pp. 281-286.

[75] O. Zayed, S. El-Beltagy and O. Haggag, "A novel approach for detecting arabic persons' names using limited resources," in *Complementary Proceedings of 14th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing,* 2013, .

[76] M. Hasanuzzaman, S. Saha and A. Ekbal, "Feature subset selection using genetic algorithm for named entity recognition," in *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation, PACLIC,* 2011, pp. 153-162.

[77] R. Alfred *et al*, "Malay named entity recognition based on rule-based approach," *International Journal of Machine Learning and Computing,* vol. 4, *(3),* pp. 300, 2014.

[78] O. Zayed, S. El-Beltagy and O. Haggag, "A novel approach for detecting arabic persons' names using limited resources," in *Complementary Proceedings of 14th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing,* 2013, .

[79] H. Traboulsi, "Arabic named entity extraction: A local grammar-based approach." in *IMCSIT,* 2009, pp. 139-143.

[80] S. N. Galicia-Haro, A. Gelbukh and I. A. Bolshakov, "Recognition of named entities in spanish texts," in *MICAI 2004: Advances in Artificial Intelligence*Anonymous Springer, 2004, pp. 420-429.

[81] M. Aboaoga and M. J. Ab Aziz, "Arabic person names recognition by using a rule based approach," *Journal of Computer Science,* vol. 9, *(7),* pp. 922, 2013.

[82] W. Zaghouani, "RENAR: A rule-based Arabic named entity recognition system," *ACM Transactions on Asian Language Information Processing (TALIP),* vol. 11, *(1),* pp. 2, 2012.

[83] M. H. Btoush, A. Alarabeyyat and I. Olab, "Rule Based Approach for Arabic Part of Speech Tagging and Name Entity Recognition," *International Journal of Advanced Computer Science & Applications,* vol. 1, *(7),* pp. 331-335, 2016.

[84] K. Shaalan and H. Raza, "NERA: Named entity recognition for Arabic," *J. Am. Soc. Inf. Sci. Technol.,* vol. 60, *(8),* pp. 1652-1663, 2009.

[85] D. Campos, J. L. Oliveira and S. Matos, *Biomedical Named Entity Recognition: A Survey of Machine-Learning Tools.* INTECH Open Access Publisher, 2012.

[86] S. AbdelRahman *et al*, "Integrated machine learning techniques for Arabic named entity recognition," *IJCSI,* vol. 7, pp. 27-36, 2010.

[87] N. F. Mohammed and N. Omar, "Arabic named entity recognition using artificial neural network," *Journal of Computer Science,* vol. 8, *(8),* pp. 1285, 2012.

[88] R. E. Salah and L. Q. binti Zakaria, "A Comparative Review of Machine Learning for Arabic Named Entity Recognition," *International Journal on Advanced Science, Engineering and Information Technology,* vol. 7, *(2),* pp. 511-518, 2017.

[89] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes,* vol. 30, *(1),* pp. 3-26, 2007.

[90] M. A. Meselhi *et al*, "A novel hybrid approach to arabic named entity recognition," in *China Workshop on Machine Translation,* 2014, pp. 93-103.

[91] M. Oudah and K. F. Shaalan, "A pipeline arabic named entity recognition using a hybrid approach." in *COLING,* 2012, pp. 2159-2176.

[92] O. H. Zayed, S. R. El-Beltagy and O. Haggag, "A Novel Approach for Detecting Arabic Persons' Names using Limited Resources." *Research in Computing Science,* vol. 70, pp. 81-93, 2013.

[93] M. Oudah and K. Shaalan, "Studying the impact of language-independent and language-specific features on hybrid Arabic Person name recognition," *Language Resources and Evaluation,* pp. 1-28, 2016.

[94] M. Aboaoga and M. J. Ab Aziz, "Arabic person names recognition by using a rule based approach," *Journal of Computer Science,* vol. 9, *(7),* pp. 922, 2013.

[95] S. Alanazi, B. Sharp and C. Stanier, "A Named Entity Recognition System Applied to Arabic Text in the Medical Domain," *International Journal of Computer Science Issues (IJCSI),* vol. 12, *(3),* pp. 109, 2015.

[96] K. F. Shaalan, "Arabic GramCheck: A grammar checker for Arabic," *Software: Practice and Experience,* vol. 35, *(7),* pp. 643-665, 2005.

[97] R. Al-Zaidy *et al*, "Mining criminal networks from unstructured text documents," *Digital Investigation,* vol. 8, *(3),* pp. 147-160, 2012.

[98] C. Bizer, T. Heath and T. Berners-Lee, "Linked data-the story so far," *Semantic Services, Interoperability and Web Applications: Emerging Concepts,* pp. 205-227, 2009.

[99] J. Makhoul *et al*, "Performance measures for information extraction," in *Proceedings of DARPA Broadcast News Workshop,* 1999, pp. 249-252.

[100] S. Alanazi, *A Named Entity Recognition System Applied to Arabic Text in the Medical Domain,* 2017.

[101] K. Shaalan and H. Raza, "Arabic named entity recognition from diverse text types," in *Advances in Natural Language Processing*Anonymous Springer, 2008, pp. 440-451.

[102] (). *the definiteness and indefiniteness of words.Available online at http://www.learnarabiconline.com/definiteness.shtml,Accessed on 15th January 2015*.

[103] Hussein Khalil ,Taha Osman , Paul Bowden , Mohammed Miltan, "Extracting arabic composite name susing a knowledge driven approach," in ***17th International Conference on Intelligent Text Processing and Computational Linguistic****,* April 3–9, 2016 • Konya, Turkey, 2016, .

[104] M. G. Al Zamil and Q. Al-Radaideh, "Automatic extraction of ontological relations from Arabic text," *Journal of King Saud University-Computer and Information Sciences,* vol. 26, *(4),* pp. 462-472, 2014.

[105] I. B. Mezghanni and F. Gargouri, "Deriving ontological semantic relations between Arabic compound nouns concepts," *Journal of King Saud University-Computer and Information Sciences,* 2017.

[106] K. Shaalan, "Rule-based approach in Arabic natural language processing," *The International Journal on Information and Communication Technologies (IJICT),* vol. 3, *(3),* pp. 11-19, 2010.

[107] S. M. A. El-salam *et al*, "Extracting Arabic Relations from the Web," *arXiv Preprint arXiv:1603.02488,* 2016.

[108] A. B. Hamadou, O. Piton and H. Fehri, "Multilingual Extraction of functional relations between Arabic Named Entities using NooJ platform," 2010.

[109] S. Albukhitan and T. Helmy, "Arabic ontology learning from un-structured text," in *Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference On,* 2016, pp. 492-496.

[110] M. Al-Yahya, L. Aldhubayi and S. Al-Malak, "A pattern-based approach to semantic relation extraction using a seed ontology," in *Semantic Computing (ICSC), 2014 IEEE International Conference On,* 2014, pp. 96-99.

[111] R. Mohamed, N. M. El-Makky and K. Nagi, "ArabRelat: Arabic Relation Extraction using Distant Supervision," .

[112] A. Aljamel, T. Osman and G. Acampora, "Domain-specific relation extraction: Using distant supervision machine learning," in *Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference On,* 2015, pp. 92-103.

[113] A. B. Abacha and P. Zweigenbaum, "A hybrid approach for the extraction of semantic relations from MEDLINE abstracts," in *International Conference on Intelligent Text Processing and Computational Linguistics,* 2011, pp. 139-150.

[114] M. R. Gormley, M. Yu and M. Dredze, "Improved relation extraction with feature-rich compositional embedding models," *arXiv Preprint arXiv:1505.02419,* 2015.

[115] I. Boujelben, S. Jamoussi and A. B. Hamadou, "A hybrid method for extracting relations between Arabic named entities," *Journal of King Saud University-Computer and Information Sciences,* vol. 26, *(4),* pp. 425-440, 2014.

[116] G. Horrocks, *Generative Grammar.* Routledge, 2014.

[117] I. Sarhan, Y. El-Sonbaty and M. A. El-Nasr, "Arabic Relation Extraction: A Survey," *International Journal of Computer and Information Technology,* vol. 5, *(5),* 2016.

[118] J. Nichols, "Functional theories of grammar," *Annu. Rev. Anthropol.,* vol. 13, *(1),* pp. 97-117, 1984.

[119] K. Hengeveld and J. L. Mackenzie, "Functional discourse grammar," *Encyclopedia of Language and Linguistics,* vol. 4, pp. 668-676, 2006.

[120] J. H. Connolly, *Constituent Order in Functional Grammar: Synchronic and Diachronic Perspectives.* Walter de Gruyter, 1991*(14).*

[121] A. Moutaouakil, "Exceptive Constructions: From the Arabic Grammatical Tradition to Functional Discourse Grammar," .

[122] A. Aljamel, T. Osman and G. Acampora, "Domain-specific relation extraction: Using distant supervision machine learning," in *Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference On,* 2015, pp. 92-103.

[123] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician,* vol. 46, *(3),* pp. 175-185, 1992.

[124] I. Hmeidi, B. Hawashin and E. El-Qawasmeh, "Performance of KNN and SVM classifiers on full word Arabic articles," *Advanced Engineering Informatics,* vol. 22, *(1),* pp. 106-111, 2008.

[125] M. Hasanuzzaman, S. Saha and A. Ekbal, "Feature subset selection using genetic algorithm for named entity recognition," in *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation, PACLIC,* 2011, pp. 153-162.

[126] D. E. Golberg, "Genetic algorithms in search, optimization, and machine learning," *Addion Wesley,* vol. 1989, pp. 102, 1989.

[127] M. Anbarasi, E. Anupriya and N. Iyengar, "Enhanced prediction of heart disease with feature subset selection using genetic algorithm," *International Journal of Engineering Science and Technology,* vol. 2, *(10),* pp. 5370-5376, 2010.

[128] A. L. Oliveira *et al*, "GA-based method for feature selection and parameters optimization for machine learning regression applied to software effort estimation," *Information and Software Technology,* vol. 52, *(11),* pp. 1155-1166, 2010.

[129] H. Chouaib *et al*, "Feature selection combining genetic algorithm and adaboost classifiers," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference On,* 2008, pp. 1-4.

[130] A. Adala, N. Tabbane and S. Tabbane, "A framework for automatic web service discovery based on semantics and NLP techniques," *Advances in Multimedia,* vol. 2011, pp. 1, 2011.

[131] S. Albukhitan, A. Alnazer and T. Helmy, "Semantic Annotation of Arabic Web Resources Using Semantic Web Services," *Procedia Computer Science,* vol. 83, pp. 504-511, 2016.

[132] H. Knublauch *et al*, "The protégé OWL plugin: An open development environment for semantic web applications," in *International Semantic Web Conference,* 2004, pp. 229-243.

[133] A. Jena, "Reasoners and rule engines: Jena inference support," *The Apache Software Foundation,* 2013.

[134] H. Al-Khalifa and A. Al-Wabil, "The arabic language and the semantic web: Challenges and opportunities," in *The 1st Int. Symposium on Computer and Arabic Language,* 2007, .

[135] M. Beseiso, A. R. Ahmad and R. Ismail, "A survey of Arabic language support in semantic web," .

[136] A. M. Al-Zoghby, A. S. E. Ahmed and T. T. Hamza, "Arabic Semantic Web Applications–A Survey," *Journal of Emerging Technologies in Web Intelligence,* vol. 5, *(1),* pp. 52-69, 2013.

[137] M. Jarrar, "Building a formal arabic ontology," in *Proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks,* 2011, .

[138] D. Maynard, Y. Li and W. Peters, "Nlp techniques for term extraction and ontology population," in *Proceeding of the 2008 Conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge,* 2008, pp. 107-127.

[139] W. Alromima *et al*, "Ontology-based Query Expansion for Arabic Text Retrieval," *International Journal of Advanced Computer Science & Applications,* vol. 1, *(7),* pp. 223-230, 2016.

[140] H. Alfeel, "The roadmap for the arabic chapter of DBpedia," in *Wseas; 14th International Conference on Telecommunications and Informatics (TELEINFO'15),* 2015, .

[141] I. AlAgha and A. Abu-Taha, "AR2SPARQL: an arabic natural language interface for the semantic web," *International Journal of Computer Applications,* vol. 125, *(6),* 2015.

[142] A. Sayed and A. Al Muqrishi, "IBRI-CASONTO: Ontology-based semantic search engine," *Egyptian Informatics Journal,* 2017.

[143] S. Zaidi, M. T. Laskri and K. Bechkoum, "A cross-language information retrieval based on an arabic ontology in the legal domain," in *Proceedings of the International Conference on Signal-Image Technology and Internet-Based Systems (SITIS'05),* 2005, pp. 86-91.

[144] F. Z. Belkredim, A. El-Sebai and U. H. B. Bouali, "An ontology based formalism for the arabic language using verbs and their derivatives," *Communications of the IBIMA,* vol. 11, *(5),* pp. 44-52, 2009.

[145] H. S. Al-Khalifa *et al*, "SemQ: A proposed framework for representing semantic opposition in the holy quran using semantic web technologies," in *2009 International Conference on the Current Trends in Information Technology (CTIT),* 2009, pp. 1-4.

[146] S. A. Elzeiny and J. M. Alja'am, "A Web-Based Intelligent Tutoring System to Facilitate the Children's Understanding of Animals' Stories Through Keywords Extraction and Multimedia Elements," *International Journal of Computing & Information Sciences,* vol. 12, *(1),* pp. 11, 2016.

[147] A. Al-Nazer, S. Albukhitan and T. Helmy, "Cross-Domain Semantic Web Model for Understanding Multilingual Natural Language Queries: English/Arabic Health/Food Domain Use Case," *Procedia Computer Science,* vol. 83, pp. 607-614, 2016.

[148] F. Corcoglioniti, *Frame-Based Ontology Population from Text: Models, Systems, and Applications,* 2016.

[149] S. Dastgheib, A. Mesbah and K. Kochut, "mOntage: Building domain ontologies from linked open data," in *Semantic Computing (ICSC), 2013 IEEE Seventh International Conference On,* 2013, pp. 70-77.

[150] D. Zhou, D. Zhong and Y. He, "Biomedical relation extraction: from binary to complex," *Comput. Math. Methods Med.,* vol. 2014, pp. 298473, 2014.

[151] N. Noy *et al*, "Defining n-ary relations on the semantic web," *W3C Working Group Note,* vol. 12, *(4),* 2006.

[152] S. Ramanujam *et al*, "A relational wrapper for RDF reification," in *IFIP International Conference on Trust Management,* 2009, pp. 196-214.

[153] J. Hoffart *et al*, "YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia," *Artif. Intell.,* vol. 194, pp. 28-61, 2013.

[154] (08.03.2017). ***DBpedia*** [DBpedia]. Available: http://wiki.dbpedia.org/.

[155] A. S. Ismail, H. Al-Feel and H. M. Mokhtar, "Introducing a new arabic endpoint for DBpedia internationalization project," in *Proceedings of the 20th International Database Engineering & Applications Symposium,* 2016, pp. 284-289.

[156] H. Al-Feel, "A Step towards the Arabic DBpedia," *International Journal of Computer Applications,* vol. 80, *(3),* pp. 27-33, 2013.

[157] E. Prud and A. Seaborne, "SPARQL query language for RDF," 2006.

[158] S. Idreos, O. Papaemmanouil and S. Chaudhuri, "Overview of data exploration techniques," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data,* 2015, pp. 277-281.

[159] G. Meditskos and N. Bassiliades, "Rule-based OWL ontology reasoning systems: Implementations, strengths, and weaknesses," in *Handbook of Research on Emerging Rule-Based Languages and Technologies: Open Solutions and Approaches*Anonymous IGI Global, 2009, pp. 124-148.

[160] B. Glimm *et al*, "HermiT: an OWL 2 reasoner," *Journal of Automated Reasoning,* vol. 53, *(3),* pp. 245-269, 2014.

[161] A. Ameen, K. U. R. Khan and B. P. Rani, "Reasoning in Semantic Web Using Jena," *Computer Engineering and Intelligent Systems,* vol. 5, *(4),* pp. 39-47, 2014.

[162] (). *Gross domestic product*.

[163] (). *What economic indicators are most used when forecasting an exchange rate?*.

[164] (). *Relationship between inflation and interest rate, http://www.investopedia.com/ask/answers/12/inflation-interest-rate-relationship.asp, acessed 28.04.2017*.

[165] F. Alvarez, R. E. Lucas and W. E. Weber, "Interest rates and inflation," *Am. Econ. Rev.,* vol. 91, *(2),* pp. 219-225, 2001.

[166] S. Fraihat and Q. Shambour, "A Framework of Semantic Recommender System for e-Learning," *Journal of Software,* vol. 10, *(3),* pp. 317-330, 2015.

[167] P. Resnick and H. R. Varian, "Recommender systems," *Commun ACM,* vol. 40, *(3),* pp. 56-58, 1997.

[168] J. Bobadilla *et al*, "Recommender systems survey," *Knowledge-Based Syst.,* vol. 46, pp. 109-132, 2013.

[169] S. Bouraga *et al*, "Knowledge-based recommendation systems: a survey," *International Journal of Intelligent Information Technologies (IJIIT),* vol. 10, *(2),* pp. 1-19, 2014.

[170] J. Moskal and C. J. Matheus, "Detection of suspicious activity using different rule engines-comparison of BaseVISor, jena and jess rule engines." in *RuleML,* 2008, pp. 73-80.