Exploiting Functional Discourse Grammar to Enhance Complex Arabic Relation Extraction Using A Hybrid Semantic Knowledge Base - Machine Learning Approach

Taha Osman

Computer Science, Nottingham Trent University, taha.osman @ntu.ac.uk

Hussein Khalil

Computer Science, Misurata University, hussein.khalil@misuratau.edu.ly

Mohammed Miltan

Arabic Department, Faculty of Arts, Misurata University, mmlitan@gmail.com

Khaled Shaalan

Faculty of Engineering & Information Technology, The British University in Dubai, Khaled.shaalan@buid.ac.ae

Rowida Alfrjani

Computer Science, Nottingham Trent University, rowida.alfrjani@ntu.ac.uk

Relation extraction from unstructured Arabic text is especially challenging due to the Arabic language complex morphology and the variation in word semantics and lexical categories. The research documented in this paper presents a hybrid Semantic Knowledge base - Machine Learning (SKML) approach for extracting complex Arabic relations from unstructured Arabic documents; the proposed approach exploits the principles of Functional Discourse Grammar (FDG) to emphasise the semantic and pragmatic properties of the language and facilitate the identification of relation elements. At the initial phase, the novel FDG-SKML relation extraction approach deploys lexical-based mechanism that utilises a purposely built domain-specific Semantic Knowledge to encode the semantic association between the identified relations' elements. The evaluation of the initial stage evidenced improved accuracy for extracting most complex Arabic relations. The initial relation extraction mechanism was further extended by integrating its output into a Machine Learning classifier that facilitated extracting especially complex relations with significant disparity in the relation elements' presence, order, and correlation. Using Economics as the problem domain, experimental evaluation evidenced the high accuracy of our FDG-SKML approach in complex Arabic relation extraction task and demonstrated its further improvement upon integration with machine learning classifiers.

Additional Keywords and Phrases: Arabic Relation Extraction, Natural Language Processing, Semantic Web Base, Functional Discourse Grammar, Hybrid Knowledge-Based Machine Learning Classification

1 INTRODUCTION

The volume of published information on the Web is growing rapidly with the increase number of Internet's users. According to the Internet World Stats, the number of Internet's users have exceeded 4.5 Billion at the time of writing this paper[1]. As most of the published information is unstructured text, the need for systems that can automate the extraction of useful information from the unstructured documents is becoming ever more desirable, which contributed to the development of Information Extraction as a major research area in computational linguistics. Information Extraction has two essential tasks, Named-Entity Recognition, also known as Entity Extraction or Entity Identification, and Relation Extraction, which is based on recognising the semantic relation between named entities [Martinez-Rodriguez et al. 2020].

Relation Extraction is critical to the identification of the problem domain's key events where structured knowledge is extracted from unstructured raw text; therefore, it is considered as a key task to the majority of Information Extraction applications such as semantic search, question answering, knowledge harvesting, sentiment analysis and recommender systems [Konstantinova 2014]. Many developed systems have focused on extracting binary relations such as (is-a, part-of) for facts extraction purposes, whereas more recent research efforts have focused on extracting complex relation (n-ary relations) for event extraction purposes, which involves extracting two or more binary relations to determine the event. For example, the sentence below contains four relations about the main entity مؤشر توبيكس الياباني [Japan's Topix Index], which are: *hasCloseDay*[مؤشر توبيكس الياباني, الخميس], *decreasBy*[مؤشر توبيكس الياباني, 1.24%], *lossPoints*[مؤشر توبيكس الياباني, 62.43] and *hasLevel*[مؤشر توبيكس الياباني, 4.992.52]. This type of relation is complex relation (4-ary relation) in which four binary relations must be extracted to determine the event about the main entity مؤشر توبيكس الياباني [Japan's Topix Index]. In this study, the focus is on extracting complex semantic relations from Arabic unstructured documents relevant to the Economic problem domain.

أنهى مؤشر توبيكس الياباني تعاملاته اليوم الخميس ، بانخفاض واضح بنسبة 1.24٪ ، بخسائر 62.43 نقطة ، ليصل إلى مستوى 4992.52 نقطة.

Japan's Topix index ended its trading on Thursday, with a clear drop of 1.24%, with losses of 62.43 points, to reach the level of 4,992.52 points.

The advances in the research and development of Arabic Information Extraction systems are not as remarkable in comparison to European Information Extraction systems [Alruily et al. 2011; Khalil and Osman 2014], which is reflected in the scarcity of Arabic Natural Language Processing resources and tools; this has subsequently contributed to the deficit in the investment in Arabic textual analytics applications, such as sentiment analysis, expert (recommender) systems and cybercrime.

Arabic text poses many challenges that influence the performance of the Relation Extraction task such as the variations in the Arabic word representation including root, prefixes, suffixes and clitics. Moreover, the complexity of the Arabic sentence structure makes it difficult to engineer patterns that recognise the great variation in the presence, order, and correlation of the relation elements (i.e., the named entities and the relation's term 'trigger word') as the sentences often contain multiple relation trigger words [Zitouni 2014; Atwan and Mohd 2017].

Functional Discourse Grammar, in particular the semantic function is a set of rules and processes that govern the semantic of sentences in a given language regardless of the structure of sentence. Operating on the semantic level of the grammatical component, process's patterns are identified by a main *Predicate* 'trigger word' such as أنهى [ended], which specifies the state of operation (i.e., an event or sequence of events of a specific kind) that are described in the sentence. The *Predicate* has corresponding arguments that consist of an *Agent* such as مؤشر توبيكس الياباني [Japan's Topix Index], which can be used to determine the event, and *Patient* that can be simple entities such as rate, date, number or complex

[1] https://www.statista.com/statistics/617136/digital-population-worldwide/

sub-phrases that might represent another process, which express the result of the operation based on the meaning of the *Predicate*. Our hypothesis is that the adoption of Functional Discourse Grammar, can help to emphasise the semantic and pragmatic properties of the Arabic language [Attia and Somers 2008; Hengeveld and Mackenzie 2006] and thus facilitate the identification of the relation elements in the Arabic sentence. Moreover, we believe that, where possible, Arabic Natural Language Processing efforts can benefit from targeting a particular problem domain, where deep knowledge of the domain's key characteristics, such as domain's concepts and the likely relations (taxonomy) between them, can contribute to the understanding of unstructured text through syntactic and semantic linguistic processing.

Hence, we developed a novel hybrid Semantic Knowledge base – Machine Learning (SKML) relation extraction approach that exploits the principles of Functional Discourse Grammar (FDG) to extract Arabic complex relations from domain-specific unstructured text. Our proposed approach initially employs Functional Discourse Grammar to recognise the relations' elements in the Arabic sentence. In the first stage, the novel relation extraction approach deploys lexical-based mechanism that utilises a purposely built domain-specific Semantic Knowledge base for encoding the semantic association between the identified relations elements. Although the semantic knowledge base (lexical) approach proved successful in the accurate extraction of most complex Arabic relations, it was further extended by integrating its output into a Machine Learning classifier to facilitate extracting especially complex relations with significant disparity in their elements' presence, order, and correlation. We deployed genetic algorithms to optimise the feature selection for the machine learning training process, which resulted in further improvement in the accuracy of relation extraction.

This paper is organised as follows: section 2 addresses the common challenges on Arabic Relation Extraction. The related literature in relation extraction from unstructured Arabic texts is reviewed in section 3. Section 4 describes the semantic modelling of the problem domain underpinning the novel Semantic Knowledge base. Section 5 details the novel FDG-inspired Semantic Knowledge base relation extraction approach. The integration of the Semantic Knowledge base relation extraction with Machine Learning classification is described in section 6, which also details the experimentation results evidencing improved extraction coverage and accuracy. Section 7 presents our conclusions and plans for further work.

2 CHALLENGES ON ARABIC RELATION EXTRACTION

Arabic text poses many challenges that have influenced the development of language processing tools such as short vowels, absence of capital letters, complex morphology and considerable use of the discretisation (i.e., diacritic signs) [Farghaly and Shaalan 2009]. More critical to Relation Extraction is the morphological and semantic variations in the Arabic word representation, which includes root, prefixes, suffixes and clitics. The variation in the written Arabic word can be disambiguated by using Arabic discretisation that is used for the full representation of short vowels. Without discretisation, the semantic of some sentences can be interpreted incorrectly; for example, if discretisation is not applied, then the phrase "كتب الولد في المدرسة" may take the meaning: "the boy's books are in the school", "the boy wrote in the school", among other meanings. However, a large portion of the Arabic material published online lacks discretisation, which is bound to impact the accuracy of the Information Extraction techniques.

Moreover, Arabic sentence has multiple types which adds to the complexity of identifying comprehensive relation patterns. An Arabic sentence can be a verbal sentence (i.e., the sentence starts with a verb phrase: verb-subject-object or verb-object-subject); nominal sentence (i.e., the sentence starts with a noun phrase: subject-verb-object or subject-adjective without a verb, e.g., الشمس مشرقة [the sun is shining]); a sentence that is presented by only one word e.g., اعطيتمونيها [you gave it to me]; or can be a combination of the mentioned sentences [Attia and Somers 2008]. Moreover, according to the authors in [Boujelben et al. 2014a] "each sentence, either nominal or verbal, can be preceded by a conjunction (و /wa/and,

ثم /thomma/then), adverb (**عندما**/EndmA/when), negation particle (**لن**, **لا**, **لم** / ln, lA, lm/ not) or combination (**وعندما** /wEndmA/ and when)”.

The challenge in Arabic Relation Extraction is further exacerbated by the complexity of the Arabic sentence structure that makes it difficult to engineer patterns that recognise the great variation in the presence, order, and correlation of the relation elements, i.e. the named entities and the relation’s term (trigger word), as often sentences contain multiple relation trigger words. For example, the complex structure of the sentence below contains multiple relation trigger words that express different classes of relations for the main entity Index المؤشر العام لسوق دبي المالي [The general index of the Dubai Financial Market], which are: indexIncreaseBy [المؤشر العام لسوق دبي المالي,4.99%], hasCloseTime[المؤشر العام لسوق دبي المالي,29 مايو 2014], indexHasLevel [المؤشر العام لسوق دبي المالي,5.087.47 نقطة], indexWinPoints [المؤشر العام لسوق دبي المالي,241.69].

**إختتم المؤشر العام لسوق دبي المالي تعاملاته لجلسة اليوم الخميس الموافق 29 مايو 2014, على ارتفاع كبير بنسبة 4.99% ليصل إلى مستوى
5.087.47 نقطة، بمكاسب بلغت 241.69 نقطة**

The general index of the Dubai Financial Market (DFM) has finished its trading session on Thursday 29 May 2014, at a high of 4.99% to reach 5.087.47 points, with earns of 241.69 points.

This study focuses on addressing the challenge of extracting relations that are represented by one trigger word (e.g., ‘drop’ represents the relation ‘DecreasedBy’) or several trigger words (e.g., ‘increase’ and ‘inflation’ are complementary to represent the relation ‘hasIncreaseInflation’).

3 RELATED WORK

This section presents the related literature for Relation Extraction task with a focus on methods for extracting semantic relations from unstructured Arabic texts. Despite the limited volume of research efforts in extracting Arabic semantic relations in comparison to other languages, these efforts can be categorised into linguistic (rule-based or knowledge-based) methods, machine-learning methods and hybrid approaches that combine both methods in an attempt to improve the performance of Relation Extraction task.

3.1 Rule-based approaches

Rule-based approaches extract relations by using syntactic and semantic rules that are handcrafted based on linguistic and domain-specific information. Hence, these approaches are well suited to extract relations from domain-specific problems where detailed semantic knowledge can be extensively exploited in building the relations’ pattern recognition rules.

A straightforward rule-based approach has been applied in [Hamadou et al. 2010] to extract relations among Arabic named entities by using the 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 and organisation names. 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.

A study in [El-salam et al. 2016] has been conducted to extract the binary relation between two Arabic named entities from the Web using the semi-supervised techniques. Using initial seed relation instances as input, the suggested pattern-based system uses a generic search engine, Google^{TM}, to extract compatible candidate relations that are validated and selected in an iterative process. Four experiments were carried out to evaluate the extraction of 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.

An Arabic ontology learning system has been proposed in [Albukhitan and Helmy 2016] based on basic morpho-syntactic pattern recognition. The patterns were hand-crafted rules that were created based on two relation types. The first type is based on hierarchical conceptual relations for extracting taxonomical relations. The second type of relations is based on an entity-predicate method that depends on parsing sentences to capture the triple of subject, action and object to capture generic (non-taxonomic) relations. The preliminary evaluation demonstrated good results for precision, but the recall was low, which is anticipated as without the aid of domain-specific knowledge or computational intelligence it is difficult to achieve good recall results for the Arabic Relation Extraction task.

A rule-based system called ASRextrator has been introduced in [Mesmia et al. 2017] for annotating and extracting 18 types of Arabic semantic relations from a collection of Arabic Wikipedia corpus. The approach is based on performing Named-Entity Recognition to annotate named entities. After that, linguistic analysis was applied to identify segments of text that contain in their extremities annotated named entities. Thereafter, regular expressions rules were identified upon the extracted segments, which were utilised to establish a set of analysis transducers. The latter were regrouped in an analysis cascade and then their order was permuted until a specific one with few errors was found.

3.2 Machine learning approaches

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. The approach requires a training dataset in addition to a test dataset.

An Arabic Relation Extraction approach has been presented in [Al-Yahya et al. 2014] based on distant supervision machine learning. A seed ontology was used to generate the training corpus, which then was used by machine learning algorithm to extract antonyms from another corpus set. The new antonyms were added to the original ontology after manual verification. The objective of the reported work is ontology enrichment rather than generic Relation Extraction; hence, the human involvement in verifying the correctness of relation pattern recognition is required.

A distant supervision approach for extracting Arabic relations has been introduced in [Mohamed et al. 2015]. To counter the lack of the annotated Arabic corpora, the authors sourced the DBpedia public linked dataset to build the training data. Their relation classifier achieved 70% F-measure. However, the DBpedia dataset does not provide comprehensive coverage in terms of relation types, in particular, Arabic relation types, hence, their approach might require to be complemented by manually trained data [Aljamel et al. 2015].

An Arabic Relation Extraction approach that is based on the characteristics of the ArabicWikipedia articles has been presented in [Zakria et al. 2019]. According to the summary part of Wikipedia articles, the authors extracted sentences, which contain the relation between a principle entity (i.e. can be identified from the title of the page or from the first sentence in the page) and a secondary entity, then they extracted syntactic and lexical features from the extracted sentences in order to build a training dataset. DBpedia was used to identify the type of the Name Entities that were considered as one of the lexical features. The prepared training dataset was used to train a Naïve Bayes classifier for detecting the type of the extracted relation. The proposed approach achieved satisfactory results recording 0.89, 0.89, and 0.9 for F-measure, precision, and recall respectively.

3.3 Hybrid approaches

There are several definitions of what constitutes a hybrid approach, but in this work, we imply a hybrid approach that combines the advantages of utilising the 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 [Abacha and Zweigenbaum 2011; Gormley et al. 2015], the adoption of such hybrid approach to extract relations from Arabic text is still limited.

One of the most significant contributions is the work published by authors in [Boujelben et al. 2014b] where they have presented a hybrid approach that utilises the linguistic modules to improve the output of the machine learning relation classifiers. The training data contains a combination of automatically extracted systaltic and semantic features from the identified clauses (i.e. contains two named entities) and manually annotated words (i.e. relation indicators) that present the semantic relations between the identified named entities. The output of the machine learning process is a set of relation extraction rules that has been subjected to an optimisation process to select the highest quality rules. Targeting generic (non-specific) domains has contributed to the complexity of this interesting approach. Therefore, the authors have proposed to deploy hand-crafted rules to deal with some of the ensuing challenges during determining the relation indicators such as handling relation negation and moderating the role of part of speech tags.

The researchers in [Lahbib et al. 2013] propose a hybrid approach for extracting semantic relations from vocalised Arabic text, which combines linguistic knowledge with statistical similarity calculus rather than machine learning. The experimental evaluation has showed that the success rate of the extracted relations reached 75%.

Hybrid approaches are proposed to offer improvements over the rule-based and machine learning approaches, especially for domain-specific Relation Extraction where there is a clear advantage in initially exploiting the 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 [Abacha and Zweigenbaum 2011; Al-Zoghby et al. 2018]. In this regard, rule-based approaches for Arabic need to exploit sophisticated linguistic rules in order to counter for the complexity in the unstructured text, not only at the word morphological level but also at the syntactic sentence structure level. For instance, most reported rule-based systems have employed basic Arabic syntax grammar rules to define the linguistic relation patterns (i.e. token order), primarily using the three basic grammatical features of the syntactic/phrasal relation: subject, object and predicate [Al Zamil and Al-Radaideh 2014].

Traditional Transformational (Generative) Grammar attempts to structure natural language as an abstract set of generalised syntactic rules [Horrocks 2014] that are detached from the context of use, which can be very useful for recognising ordered patterns of binary relations in a sentence. However, Arabic language sentences often contain complex (high order) relations where one subject is represented by several predicates or several objects with varying order of the features in the sentence [Sarhan et al. 2016]. On the other hand, the Functional theories of grammar [Nichols 1984] consider the functions of a 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. This observation indicates that Functional grammar is able to offer a more flexible abstraction for modelling the complex relations of the Arabic language. We have therefore adopted the principles of Function Discourse Grammar [Hengeveld and Mackenzie 2006] an advanced version of Functional grammar, as the basis for building a novel hybrid approach to extract complex relations from Arabic natural text.

4 SEMANTIC MODELLING OF THE DOMAIN KNOWLEDGE

Domain knowledge refers to the specific and specialised knowledge about a particular field. In the information processing discipline, domain knowledge chiefly represents the key concepts and relations that describe the problem domain. In this study, we adopt Semantic Web technologies for the modelling and processing of the domain knowledge base. The Semantic Web is an extension of the current Web, where information is given a well-defined meaning, encouraging cooperation among human users and computers [Domingue et al. 2011]. The Semantic Web represents the technologies and methods that organise the knowledge in a formalised concept ontology, which provide efficient support for linking and sharing data between resources as well as allow computer machines to read, understand and retrieve the meaning of a specific semantic information on the Internet. Moreover, Semantic Web technologies can present the domain knowledge in a structured and consistent way, which facilitates the qualitative interpretation of domain specific contents in a way that people can understand. Semantic modelling is formalised using the Resource Description Framework (RDF) data model [Pan 2009] that lend well to relations modelling as it is based on a *semantic triple* that connects resources (Subject and Object) with object/data relation (*Predicate*).

This section describes the semantic modelling of the problem domain to capture its knowledge. The process starts with modelling the domain knowledge into a conceptual model that is translated into a formal semantic ontology; the ontology can then be populated with the domain's relevant key concepts and relations.

We adopted the Economic domain as the problem domain used in our research work. It provides a rich source of information due to its significance in our daily life and therefore represents an important decision-support use case. However, using text analytics to infer insight from Arabic Economic text can be challenging. At the named entity recognition level, the Arabic language orthographic and morphological complexity makes it challenging to mine the rich set of economic indicators such as the stock market, industry sources etc. [Kumar and Ravi 2016]. The indicators can also contain Latin words (such as the names of companies) that are written in Arabic letters (e.g. SMN are written as أس أم إن). Such Latin words can be classified by the POS tagger into three words (noun, verb or particles), which makes the automatic identification of Arabic composite names especially challenging. At the relation extraction level, the Arabic language is characterised by complex relations structure, which makes it challenging to generate patterns that recognise the variation in the presence, order, and correlation of the relation elements (i.e. the named entities and the relation's term 'trigger word') as the sentences often contain multiple relation trigger words [Atwan and Mohd 2017].

4.1 Capturing the domain knowledge in concept maps

Domain knowledge was captured using the knowledge (concept) map modelling [Coffey et al. 2006], an approach initially developed to help any individual or group to describe their ideas about any topic in a pictorial form. Similar to semantic ontologies, concept maps primarily model knowledge as concepts interlinked by relations and hence provide a human-centred interface to display the structure, content, and scope of an ontology.

With a view of creating a knowledge base that underpins the relation extraction activity from the use-case Economic domain, the concept map (see Figure 1) also models the interaction with the data source, i.e. the news domain. Hence, the modelled concept map covers knowledge about economic areas, such as country, economic indicators, stock market, share, products, currency, etc. as well as terms of the online news domain as document title, date, author, source, etc.

4.2 Translating the domain knowledge into a semantic ontology

In the Semantic Web, an ontology is used to represent the schema or taxonomy of the domain knowledge. Using the Protégé[2] tool, the Economic conceptual model was formalised into a machine-readable format and encoded as Economic semantic ontology (as shown in Figure 2), which represents the template box (T-box) of the Economic knowledge base. The classes in the formalised semantic ontology have been categorised into Superclasses such as Market Entities, Business Entities, Place, Economic Indicators, Natural Resources and Currency, and Subclasses such as Stock Market, Index, Share and. The object properties in the ontology define the relation between classes, for example, the object property *Locatedin* represents the relation between the classes Stock market and Country, also, the object property *DocumentBelongto*

[2] https://protege.stanford.edu/

represents the relation between the classes Market Entities and Documents. Furthermore, the data property represents the class attributes; for example, *openDate* and *pointNumber* are the main attributes for the class Index; also, *hasBirthdate* and *hasEmail* are attributes related to the class Person. The construction can be aided by the automatic extraction of the domain's key terms. The contribution in [Maynard et al. 2008] remains a key source to guide this process.

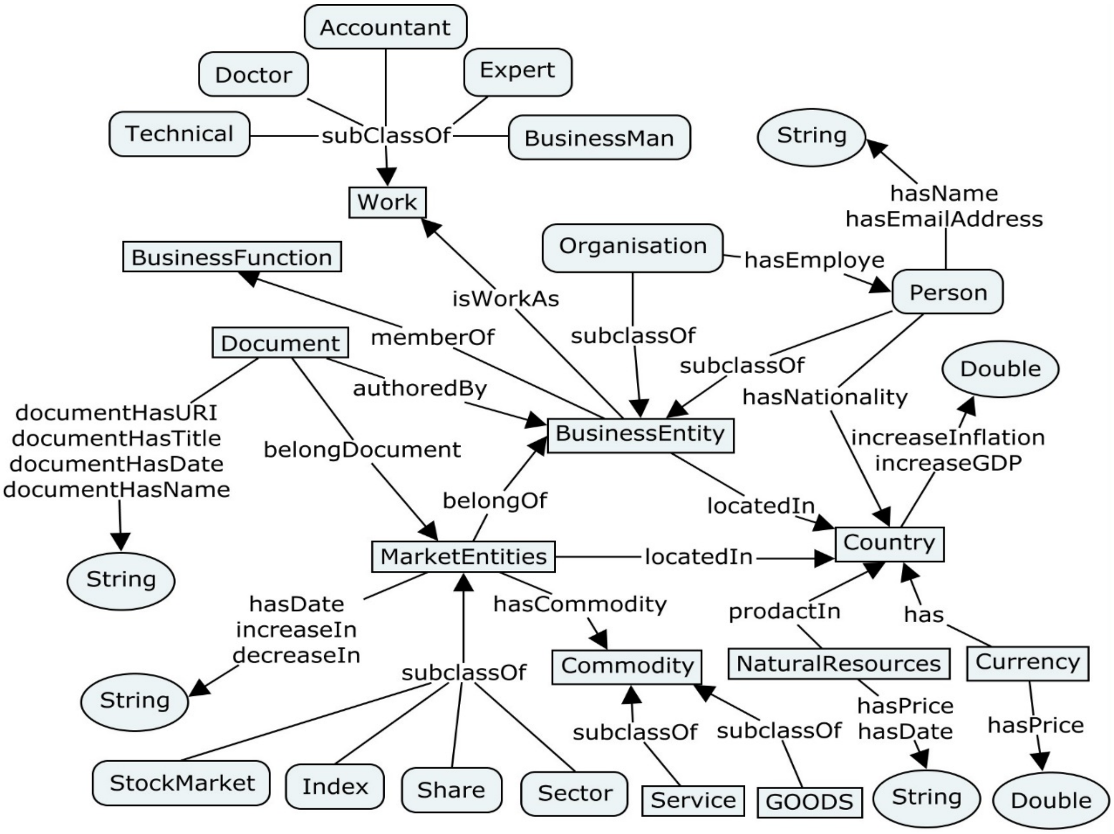


Figure 1: The concept map of the Economic Domain

To facilitate the utilisation of the ontology in the lexicon (rule-based) extraction of Arabic relations, the Arabic equivalent terms of the classes and relations were inserted into the ontology using the *rdf:label* annotation property, e.g. (السهم [*share*], المؤشر [*Index*], سوق الاوراق المالية [*Stock Market*], مدينة [*City*], الدولة [*Country*], المنتج [*Industry*], مؤشرات اقتصادية [*Economic Indicators*], etc.), and object property such as (ارتفع [*Increase*], انخفض [*Decrease*], **انتاج** [*Make*], ينتمي [*belong*]). The *rdf:label* annotation property was also used to add synonyms terms to the ontology's classes and object properties, which enhances the recall of relation extraction mechanism.

Here, it is worth noting that the semantic modelling of the problem domain maps naturally to the paradigm of Functional Discourse Grammar as its semantic function [*Agent*, *Predicate*, *Patient*] corresponds directly to the Semantic Web knowledge representation in RDF triples that are encoded as a set of [subject, predicate, object] nodes. In addition, the resulting semantic Economic ontology is utilised to generate an association between each relation's *predicate* and the corresponding *Patients*. Each relation *predicate* term is represented by a particular *object property*, that in turn is associated with a *range* of corresponding relation *objects*.

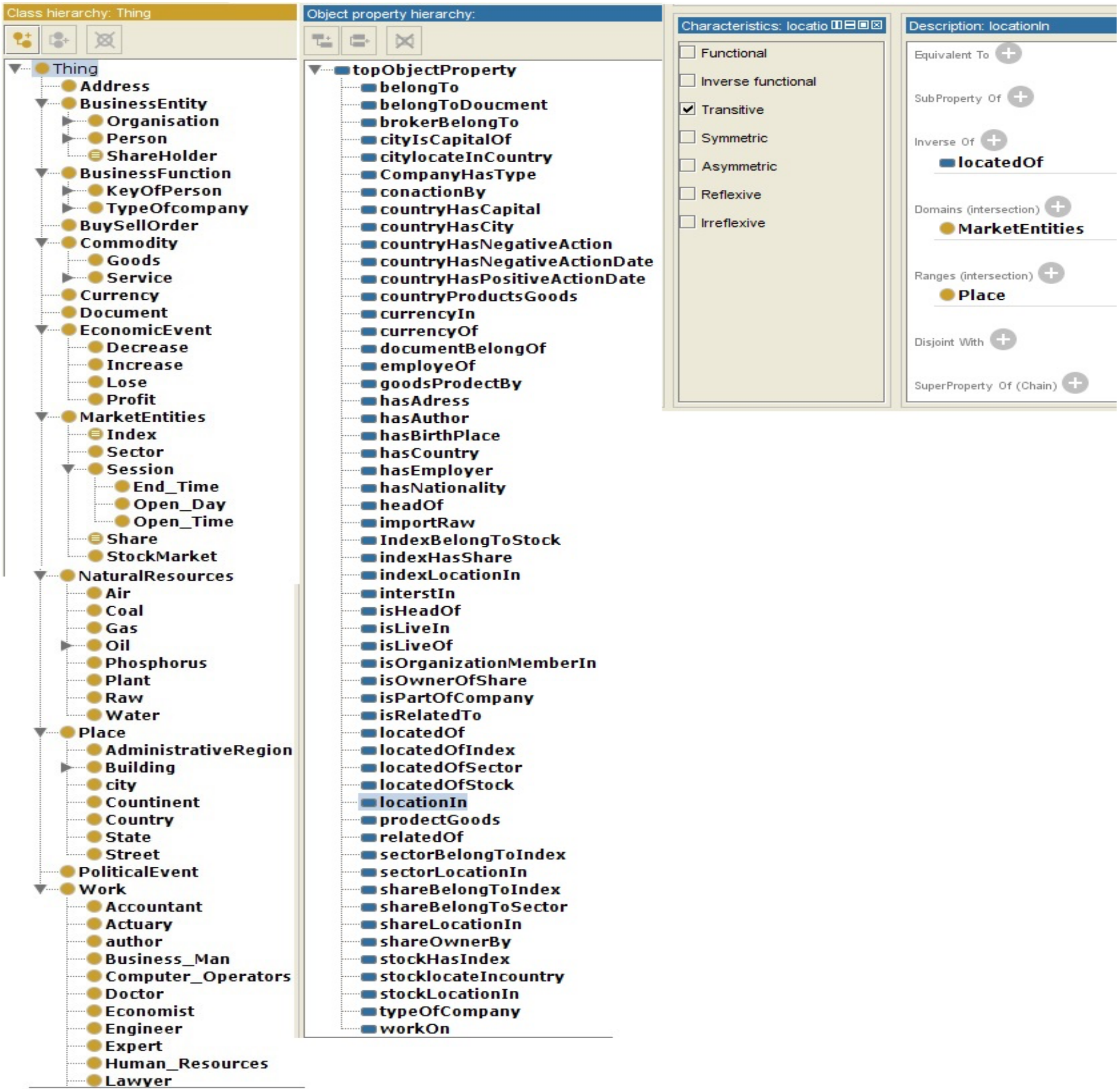


Figure 2: Screenshot presenting the class and object properties of the semantic Economic ontology

5 A NEW SEMANTIC KNOWLEDGE BASE APPROACH FOR COMPLEX ARABIC RELATION EXTRACTION BASED ON FUNCTIONAL DISCOURSE GRAMMAR (FDG-SK)

The proposed approach FDG-inspired Semantic Knowledge base (FDG-SK) extraction is the initial stage of our novel hybrid FDG-SKML relation extraction approach, which employs Functional Discourse Grammar (FDG) to recognise the relations' elements in the Arabic sentence and deploys lexical-based mechanism that utilises a purposely built domain-specific Semantic Knowledge base (SK) for encoding the semantic association between the identified relations elements to improve the performance of Relation Extraction from unstructured Arabic documents; in particular extracting complex semantic relations between several entities such as relations between Organisations and Number, relations between

Location and Numbers, and relations between Organisations and Date. Figure 3 illustrates the architecture of the proposed FDG-SK approach, which comprises the following main components: Linguistic Pre-processing, Arabic Named Entity Recognition, Semantic Arabic Relation Identification and Semantic Knowledge base Population.

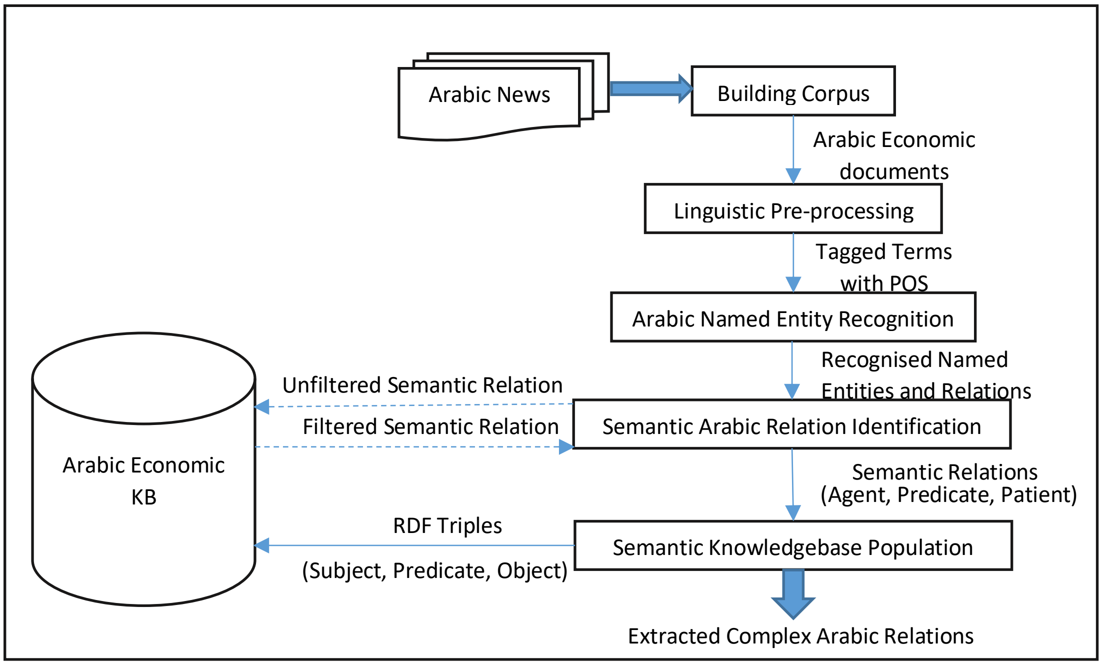

Arabic News
Building Corpus
Arabic Economic documents
Linguistic Pre-processing
Tagged Terms with POS
Arabic Named Entity Recognition
Recognised Named Entities and Relations
Arabic Economic KB
Unfiltered Semantic Relation
Filtered Semantic Relation
Semantic Arabic Relation Identification
Semantic Relations (Agent, Predicate, Patient)
RDF Triples
(Subject, Predicate, Object)
Semantic Knowledgebase Population
Extracted Complex Arabic Relations

Figure 3: System Architecture of the FDG-SK Approach

In general, the proposed FDG-SK approach processes unstructured Arabic Economic documents, which are collected from online Arabic News by removing the non-essential symbols, generating the POS tag for each token, and recognising the Arabic named entities as well as the Arabic relation terms (trigger words). Thereafter, the proposed system exploits Functional Discourse Grammar to identify semantic relations from the pre-processed documents, and then semantically classify the identified relations using the semantic Economic Knowledge base. Finally, the system populates the Economic Knowledge base with the semantic representation of the extracted Arabic relations, a process that is technically termed as RDF reification [Manola et al. 2004]. The next subsections explain in detail the role of each component.

5.1 Natural Language Processing

5.1.1Linguistic pre-processing

As indicated in the introduction section, this research focuses on extracting complex semantic relations from Arabic unstructured documents, which present Economic activities related to the Economic domain. In general, accessible Arabic corpora with sufficient number of annotated examples (named entities and relations) are very limited, particularly for domain-specific Natural Language Processing efforts. Hence, we resigned to build our own Arabic economic corpus from different Arabic economic news resources [3]. More than 1300 documents were collected from specialist websites that are listed in Table 1.

Table 1: Arabic Economic Corpus resources

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

Motivated by the challenge of extracting Arabic composite names, the linguistic pre-processing stage was developed and implemented as part of an earlier research work that used the genitive principles of Arabic grammar for composite Arabic NER [Khalil et al. 2020], where the list of POS tags was based on assigning the tag (e.g. noun, verb, adjective, etc.) to each word via using the Stanford POS tagger [Rabiee 2011]. A sample of POS tags for an Arabic sentence from the evaluation corpus is shown in Table 2. Our composite Arabic NER algorithm registered high precision of 95% for the recognition of the most complex composite names.

Table 2: Arabic part of speech tagging

جاء سهم الأسمنت الوطنية مرتفعا بنسبة 12.01%

National Cement share price has risen by the rate of 12.01%

Word	Tags	Description
جاء	VBD	Verb, Past Tense
سهم (Share)	NN	Noun, Singular
الاسمنت (Cement)	DTNN	Noun, Singular, definite
الوطنية (National)	DTNN	Noun, Singular, definite
مرتفعا (has risen)	RB	Adverb
بنسبة (by)	NN	Noun, Singular
12.01	CD	Numeral
%	SYM	Symbol

5.1.2Arabic Named Entity Recognition

Named-Entity Recognition is a fundamental task for Natural Language Processing as it is essential for extracting key terms that are related to a specific domain from unstructured text [Shaalan 2014]. Arabic Named-Entity Recognition is especially challenging compared to other languages such as English due to several factors including: 1) Arabic is morphologically rich and proper nouns lacks capitalisation. 2) Arabic is orthographic with discretisation, and is highly inflectional and

[3] https://olympuss.ntu.ac.uk/pages/cmp3osmantm/ArRE-EconDataset/

derivational [Khalil and Osman 2014]. For example, the preposition or conjunctive may attach to one word as a prefix to the nominal, such as لخدمات [for services] or وخدمات [and services]. 3) Due to lack of capitalisation researchers resort to use indicator patterns to help identify Named Entities such as الملك [the King] [Shaalan and Raza 2009] or شركة [Company]. Moreover, Arabic composite names can comprise different phrases, such as embedded place and/or owner name 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. We successfully negotiated these challenges in our earlier research published in [Khalil et al. 2020] where we present a novel approach that uniquely exploits semantically-structured domain knowledge and Arabic genitive grammar rules, in particular definite nouns (معرفة) and indefinite nouns (نكرة), to devise linguistic patterns that extract composite names of arbitrary length for composite name recognition; we adopted this approach for the recognition of the domain named entitles such as company name, stock market name, share name, etc.

We initially identify simple Named Entities using a combination of Gazetteer lists that were collected from domain-relevant semantically-structured knowledge bases such as Maknaz [Maknaz.org 2022] and DBpedia, then handcrafted pattern recognition rules, based on Arabic genitive grammar, are used to detect composite Named Entities such as شركة قاريونس لخدمات وصيانة وبرمجة الحاسوب [Garyounis Company for Computer Services, Maintenance and Programming].

For each identified relation (e.g. ارتفع [increased]), the knowledge base relation lexicon is further enriched by adding the synonyms of the identified relation (صعد [risen]) as well as using stemming to add the different forms of the relation word (e.g. ارتفعوا [increased, *"plural"*]), thus maximising the Recall of relation name recognition.

5.2 Arabic Relation Extraction

The proposed FDG-SK approach performs the Arabic relation extraction task in two main phases starting with identifying the semantic Arabic relations in the processed Arabic sentences, and then populating the semantic representation of the extracted complex relations into the modelled Economic knowledge base.

5.2.1Semantic Arabic Relation Identification

We have been successful in progressing knowledge-based research in information extraction and textual analytics. Our FDG-SK approach relies on the semantic modelling of the domain knowledge to build information extraction and textual analytics rules [Aljamel et al. 2018; Alfrjani et al. 2019]. Therefore, in the context of relation extraction, it was fitting to adopt the principles of Functional Discourse Grammar, in particular the semantic function, to emphasise the semantic and pragmatic properties of the Arabic language in order to identify the complex relation patterns in Arabic sentences.

The Theory of Functional Discourse Grammar have witnessed successive models in the context of processing the Arabic language since the beginnings of 1980s, which was developed in [Hengeveld and Mackenzie 2006]. As shown in Figure 4, the theory is established based on the idea that language usually has three primary functions grammars: pragmatic function [topic and focus, theme and tail], semantic functions [*Agent*, *Predicate*, *Patient*] and syntactic functions [subject, verb, object]. The semantic function grammar is a set of rules and processes that govern the semantic of sentences in a given language regardless of the structure of sentence. Operating on the semantic level of the grammatical component, process's patterns are identified by a main *Predicate* trigger word such as اغلق [closed at], which specifies the state of operation (i.e. an event or sequence of events of a specific kind) that are described in the sentence. The *Predicate* has corresponding arguments that consist of an *Agent* such as سهم دبي [Dubai shares], which can be used to determine the event, and *Patient* that can be simple entities such as rate, date, number or complex sub-phrases that might represent another process, which express the result of the operation based on the meaning of the *Predicate*.

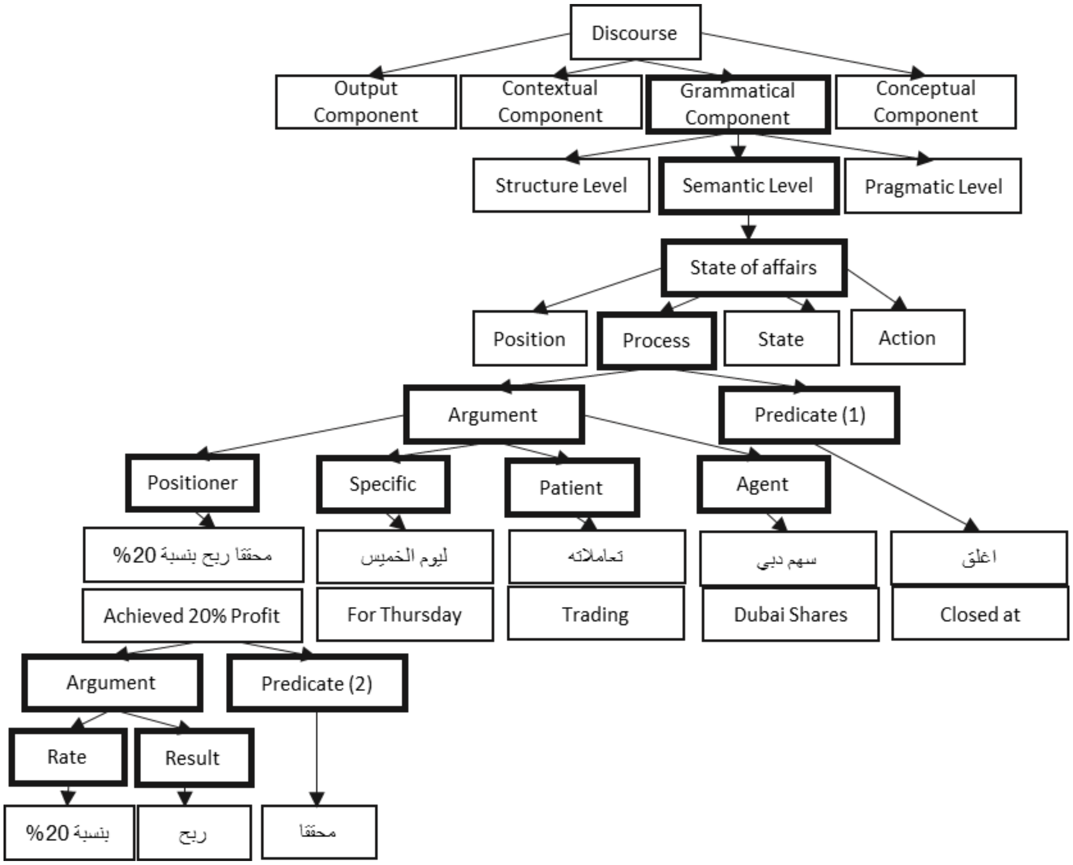


Figure 4: Example illustrating the representation of Function Discourse Grammar

As discussed in the literature survey, rule-based efforts in Arabic relation extraction have mainly employed syntactic functions that relied on the sequencing of the subject, verb, and object in the sentence structure. Hence, this study focuses on the computation of the semantic functions of Arabic in the economic discourse with particular focus on the stock market. Algorithm 1 presents the implementation of the new Semantic Arabic Relation Identification (SARI) algorithm for identifying semantic relation from different Arabic sentences forms based on the position and number of the *Agents*, *Predicate* and argument type, which can be illustrated as follows:

ALGORITHM 1: Semantic Arabic Relation Identification (SARI) Algorithm

```
Input: Pre-processed Documents D, Identified Named Entities NE, Identified trigger words TW, Semantic Economic
Knowledge base SEK
Predicate-Set= Generate-Predicate-Set (SEK) // list of predicate terms that have been retrieved from SEK
P= Count (Predicate-Set)
Do for i=1:D
  Sentences= Identify-Sentence (Document[i])
  S= Count (Sentences)
  Do for j=0:S
    Annotated-Tokens= CreateTokens (Sentence[j], NE, TW) // List of annotated named entities and trigger words
    T= Count (Annotated-Tokens)
    Do for a=0:T  // search for Agent
      If (Annotated-Token[a].KIND equals ‘Named Entity’ AND Annotated-Token[a].POS equals ‘NNP’)
        Agent = Annotated-Token[a]
      End If
    End For
    Do for a=0:T  // search for Patient
      If (Annotated-Token[a].KIND equals ‘Named Entity’ AND Annotated-Token[a].CLASS equals (M OR D OR P)
        // M= Money , D= Date, P= PERCENTAGE
        Patient = Annoated-Token[a]
      End If
    End For
    Do for a=0:T  // search for Predicates
      If (Annotated-Token[a].KIND equals ‘Relation Term’)
        Do for c=0:P // search through the predicate set
          If (Annotated-Token[a].ROOT equals Predicate-Set[c])
            Predicate = Annotated-Token[a]
            Predicate.SUBJECT = Predicate-Set[c].SUBJECT // identify the type of the relation subject to the predicate
            Predicate.OBJECT = Predicate-Set[c].OBJECT // identify the type of the relation object to the predicate
            If (Agent.CLASS equals Predicate.SUBJECT AND Patient.CLASS equals Predicate.OBJECT)
              Relation-Pattern= new Relation-Pattern (Agent, Predicate, Patient) // semantic association
            End If
          End If
        End For
      End If
    End For
  End for
End for
Output : Extracted Semantic Relations
```

i) The SARI algorithm takes as input a set of pre-processed Arabic Economic documents, which were linguistically processed using Natural Language Processing phase as mentioned in section 5.1.1 and resulted in tokenised terms with their relevant features (e.g. string, root, POS). In addition, the pre-processed documents contain the recognised Named Entities (e.g. country, index, share, money, percentage, date) and the recognised relation terms (e.g. closed, increased), which were recognised using Arabic Named-Entity Recognition approach as mentioned in section 5.1.2.

ii) For each pre-processed document, the algorithm identifies sentences and creates for each sentence a General Annotated Tokens (GAT) array that contains the recognised Named Entities and the recognised relation terms (trigger words) in order of appearance. Then, SARI algorithm tags each named entity in the GAT array as well as each trigger word with token’s features as shown in Table 3.

iii) The algorithm traverses through the GAT array elements until it finds an element of the feature:kind ‘Named Entity’ and feature:POS ‘NNP’ such as سهم الانوار القابضة [Al Anwar Share] and annotates it as an *Agent*.

iv) The algorithm searches for elements of the feature:kind ‘Named Entity’ and feature:class (Money, Percentage or Date) and annotates them as *Patient*.

v) The algorithm continues the linear search to find elements of the feature:kind ‘Relation Term’, which are then annotated as *Predicate* if the algorithm finds their feature:root within the problem domain *Predicate* set. The latter indicates a list of *Predicate* terms that have been retrieved from the modelled Economic knowledge base where each *Predicate* term represents a particular object property (i.e. relation). The object property is in turn associated with a particular type of relation subject (i.e. domain) and relation object (i.e. range). For example, the domain of the object property ShareIncreasedBy is ‘Share’ whereas the range is ‘Number’. Once the *Predicates* are identified, the algorithm associates them semantically with the type of their domain and range based on their matched object properties.

Table 3: GAT tokens classification

سهم الانوار القابضة منى بأكبر الخسائر اليوم بنسبة 2.42% ليغلق على 0.322 ريال Al Anwar Share registered the biggest loss today by 2.42% to close at 0.322 Riyal					
GAT	**Features**				
Array Index	**String**	**Kind**	**Class**	**Root**	**POS**
GAT[1]	سهم الانوار القابضة (Al Anwar Share)	Named Entity	Share	سهم	NNP
GAT[2]	منى (registered)	Relation term	-	منى	VBD
GAT[3]	الخسائر (loss)	Relation term	-	خسر	DDTNNS
GAT[4]	%2.42	Named Entity	Percentage	-	Number
GAT[5]	ليغلق (close)	Relation term	-	غلق	VBD
GAT[6]	0.322 ريال (Riyal)	Named Entity	Money	-	Number

vi) The algorithm uses the type of the semantically associated relation subject and relation object of the annotated *Predicate*s to determine whether the annotated *Patients* and *Agent* belong to them or not in order to complete the semantic function relation’s main parts [*Agent*, *Predicate*, *Patient*]. Here, it is worth noting that the semantic modelling of the problem domain maps naturally to the paradigm of Functional Discourse Grammar as its semantic function [*Agent*, *Predicate*, *Patient*] corresponds directly to the Semantic Web knowledge representation in RDF triples that are encoded as a set of [subject, predicate, object] nodes. For example, if the type relation subject and relation object of the *Predicate* انخفض [SharedecreasedBy] is Share and Percentage respectively, then the annotated *Agent* and *Patient* should be an element with feature:class ‘Share’ and ‘Percentage’ respectively. Whereas, if the type relation subject and relation object of the *Predicate* يفتح [hasOpenDate] is ‘Share’ and ‘Date’ respectively, then the annotated *Agent* and *Patient* should be an element with feature:class ‘Share’ and ‘Date’ respectively. The RDF semantic modelling of the domain provides for the use of the sophisticated SPARQL[4] query language to interrogate the knowledge base. In the example illustrated below, SPARQL’s ASK query was used to investigate each identified relation, which returns ‘True’ for the correct identified relation, otherwise, it returns ‘False’.

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
ASK  {
      Owl:ShareDecreasedBy rdfs:domain owl:Share .
      Owl:ShareDecreasedBy rdfs:range  xsd:Percentage .
}
The returned result: True
```

[4] https://www.w3.org/TR/rdf-sparql-query/

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
ASK {
   Owl: hasOpeneTime rdfs:domain owl:Share .
   Owl: hasOpenTime rdfs:range  xsd:Number .
}
The returned result: True
```

In order to illustrate the added advantage of utilising semantic knowledge base approach with Functional Discourse Grammar, we drive in Table 4 below an example demonstrating how the SARI algorithm avoids extracting incorrectly identified relations "false positives" in a complex sentence; this is achieved by evaluating the semantic association between the identified *Agents* / *Predicates* / *Patients* as shown in the first two rows. In contrast, deploying Functional Discourse Grammar without utilising the semantic knowledge can lead to extracting incorrect Arabic relations "False Positive" as illustrated in the last two rows of Table 4. Hence, relying on using both Functional Discourse Grammar and semantic knowledge base can enhance the performance of Arabic relation extraction.

Table 4: Example illustrating the differences between using semantic function against syntactic function

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%.**

Agent	***Predicate***	***Patient***	**Extracted**	**State**
البحرين (Bahrain)	أدنى lowest	٪5.5	Yes	True positive
الأردن (Jordan)	أدنى lowest	٪5.3	Yes	True positive
٪5.3	أدنى lowest	الأردن	No	False positive
٪5.5	أدنى lowest	البحرين	No	False positive

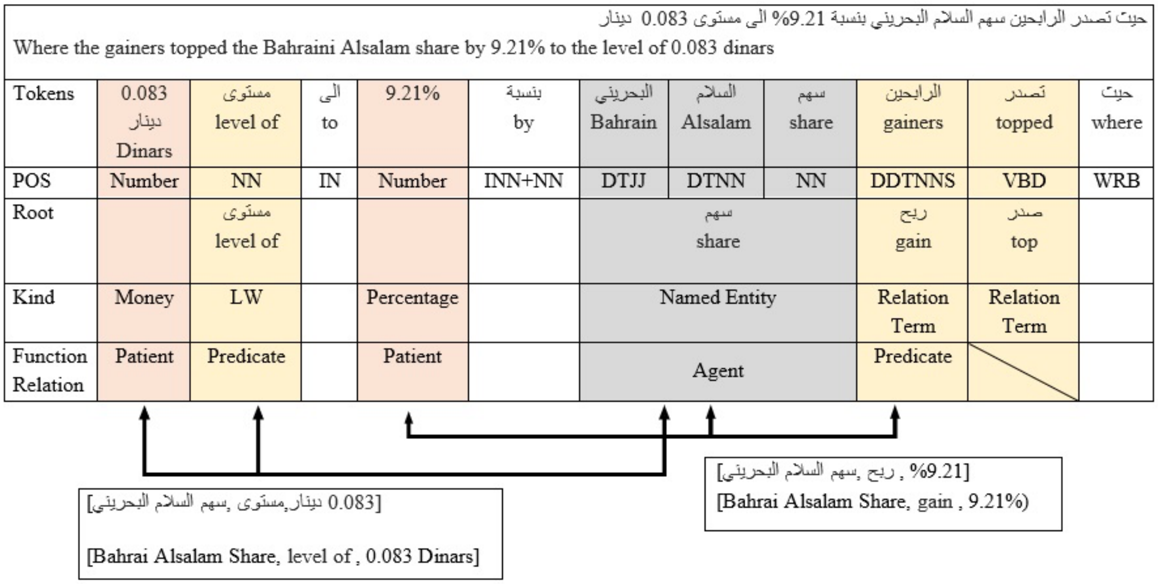
GDP: Gross Domestic Product

vii) The algorithm continues the linear search in the GAT array for an element with feature:kind 'Named Entity' and feature:POS 'NNP' that can be recognised as a new *Agent* if it has no relation to the first annotated *Agent* (i.e. identifying another semantic function relation in the entire processed sentence). In this case, the algorithm identifies the relevant *Predicate* and *Patient* for the new *Agent* via re-performing the above steps iv-vii; otherwise, a new sentence is processed.

Figure 5 illustrates the operation of our SARI algorithm in identifying semantic relation patterns from a complex sentence that contains one *Agent*, several *Predicates* and *Patients*. Based on SARI algorithm, سهم السلام البحريني [Bahrai Alsalam Share] is the only *Agent* in the sentence, hence, it will be associated with the two *Predicates* (الرابحين [gainers], مستوى [level of]) and results in two relation patterns. In the next stage, the algorithm associates the *Patients* (Percentage 9.21%, Money 0.083Dinars) with the created relation patterns based on the type of the semantically associated range of the identified *Predicates*.

5.2.2Semantic Knowledge base Population: Relation Representation using RDF reification

As mentioned in the introductory section, this research focuses on extracting complex relations, also known as n-ary relations. The focus is on extracting event-based relations as illustrated in Table 5, where multiple relations related to the main entity (the *Agent*) المؤشر العام لسوق دبي المالي [the general index of the Dubai Financial Market] are identified. The relation indexIncreaseBy (the general index of the Dubai Financial Market, 4.99%) is the main relation, whereas the other relations are complementary to the main relation; this type of relation is called the 5-ary relation.



حيث تصدر الرابحين سهم السلام البحريني بنسبة 9.21% الى مستوى 0.083 دينار

Where the gainers topped the Bahraini Alsalam share by 9.21% to the level of 0.083 dinars

Tokens	0.083 دينار Dinars	مستوى level of	الى to	9.21%	بنسبة by	البحريني Bahrain	السلام Alsalam	سهم share	الرابحين gainers	تصدر topped	حيث where
POS	Number	NN	IN	Number	INN+NN	DTJJ	DTNN	NN	DDTNNS	VBD	WRB
Root		مستوى level of				سهم share			ربح gain	صدر top	
Kind	Money	LW		Percentage		Named Entity			Relation Term	Relation Term	
Function Relation	Patient	Predicate		Patient		Agent			Predicate		

[0.083 دينار,مستوى ,سهم السلام البحريني]

[Bahrai Alsalam Share, level of , 0.083 Dinars]

[9.21% , ربح ,سهم السلام البحريني]

[Bahrai Alsalam Share, gain , 9.21%)

Figure 5: Example for identifying the relation in a sentence with one agent and several predicates of different types

Table 5: The list of binary relations in the sentence

إختتم **المؤشر العام لسوق دبي المالي** تعاملاته لجلسة اليوم الخميس الموافق 29 مايو 2014 على ارتفاع كبير بنسبة 4.99% ليصل إلى مستوى 5.087.47 نقطة، بمكاسب بلغت 241.69 نقطة

The general index of the Dubai Financial Market has finished its trading session on Thursday 29 May 2014, at a high of 4.99% to reach 5.087.47 points, with earns of 241.69 points.

Subject	Predicate	Object	Explain
المالي دبي لسوق العام المؤشر Dubai Financial Market Index	indexIncreaseBy	4.99%	This relation describes the state of index
المالي دبي لسوق العام المؤشر Dubai Financial Market Index	hasCloseTime	29 مايو 2014	This relation describes the time the index has closed
المالي دبي لسوق العام المؤشر Dubai Financial Market Index	indexHasLevel	5.087.47 نقطة [point]	This relation describes the level of index based on the number of points
المالي دبي لسوق العام المؤشر Dubai Financial Market Index	indexWinPoints	241.69 نقطة [point]	This relation describes the number of points the index has won
المالي دبي لسوق العام المؤشر Dubai Financial Market Index	belongToDocument	NEWS991.txt	This relation describes the index belongs to a document

N-ary relations cannot be represented in the Semantic Knowledge base by simply splitting them into binary relations because significant information may be lost. For instance, for the relation indexIncreaseBy (the general index of the Dubai Financial Market, 4.99%), information about the date of this action and how many points the index loss or win in this action can be lost during the query stage. Hence, RDF reification vocabulary has been used to represent the complex relations as RDF triples. The RDF reification vocabulary represents the relations as the statement and individuals that 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 [Noy et al. 2006]. Figure 6 illustrates how the reification techniques have been used to represent the complex relation of the exemplified sentence in Table 5.

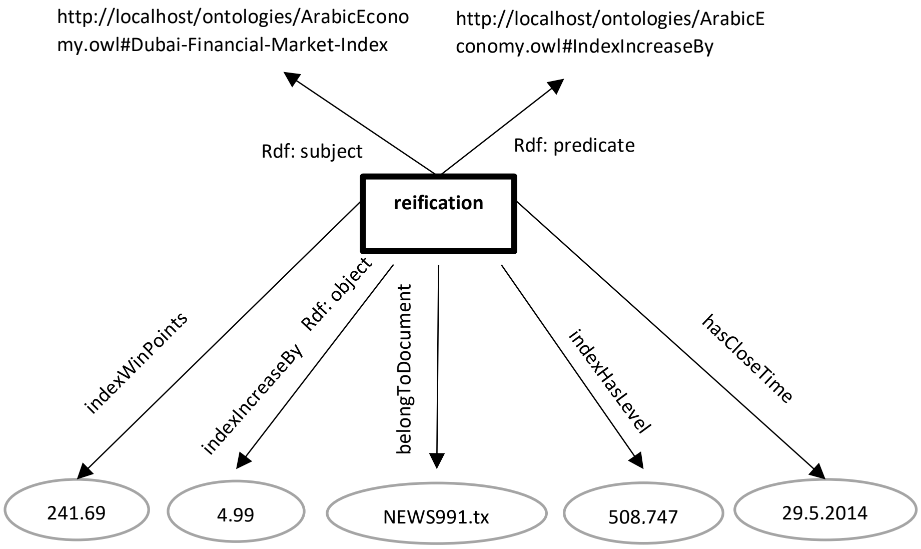


Figure 6: An example illustrates the reification technique represent the n-ary relations

5.3 Experimental Evaluation of the FDG-SK approach

This section evaluates the performance of the proposed SK-FDG in extracting Arabic relations. The specification of the corpus used in the experimental evaluation (ARB-ECON) is detailed in Table 6 below; a total of 1300 documents containing 6055 sentences and 189290 words were collected from different Arabic economic online news resources5 as described in section 5.1.1. Arabic Economic relations were manually identified within the collected documents, which represented the baseline for evaluating the performance of the proposed FDG-SK approach. Equations 1, 2 and 3 were used to compute the Precision, Recall and F-measure of the obtained results respectively.

The experimental evaluation assesses the performance of the proposed FDG-SK approach in extracting Arabic sentences with varying structure complexity and for different types of relations.

Table 6: The specifications of Arabic Economic dataset

Annotated Entity	Number
Document	1300
Sentences	6055
Words	189290
All Named Entities	24977
Named Entity: **Location** Include: City /country	3219
Named Entity: **Organisation** Include: Index/Share/Sector/Company/Stock Market	5214
Named Entity: **Date** Include: Date/Day/Year	4106

[5] https://olympuss.ntu.ac.uk/pages/cmp3osmantm/ArRE-EconDataset/

Annotated Entity	Number
Named Entity: **Numbers**	10819
Include: Price/Number of point/Percentage	
Named Entity: **Person**	1619

$$\textbf{Precision} = \frac{|\{\text{relevant relations}\} \cap \{\text{retrieved relations}\}|}{|\{\text{retrieved relations}\}|} \qquad (1)$$

$$\textbf{Recall} = \frac{|\{\text{relevant relations}\} \cap \{\text{retrieved relations}\}|}{|\{\text{relevant relations}\}|} \qquad (2)$$

$$\textbf{F-measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (3)$$

5.3.1Evaluating the performance of the FDG-SK approach for extracting Arabic Economic relations from sentences with different structure complexity

For this evaluation, we conducted three experimental evaluations: Experiment 1 evaluates the extracted Arabic Economic relations from simple structured sentences where each one *Agent* has one *Predicate*, Experiment 2 evaluates the extracted Arabic Economic relations from complex sentences where one *Agent* has more than one *Predicate*; and Experiment 3 evaluates the extracted Arabic Economic relations from more complex sentences that have more than one *Agent* and each *Agent* has more than one *Predicate*.

Table 7 collates the results of the above conducted experiments. In general, we can observe that the precision of the extracted Arabic Economic relations drops slightly with the increase of the complexity of the sentence structure. However, the overall results have indicated that the FDG-SK approach achieved very good accuracy in extracting Arabic Economic relations from relatively complex Arabic unstructured texts, scoring a commendable 0.84 f-measure for the more complex sentence structure, which is comparable to the recorded f-measure for extracting economic relations from English documents in [Aljamel et al. 2015], where the Relation Extraction task was performed using distantly-supervised machine learning.

Table 7: Evaluating the performance of the FDG-SK approach for extracting Arabic Economic relations from sentences with different structure complexity

Experiments	Recall	Precision	F-measure
EXP- 1	0.84	1	0.91
EXP- 2	0.90	0.86	0.88
EXP- 3	0.94	0.76	0.84

It is curious, however, that the recall's impact by increasing the sentence complexity is almost opposite to the precision as the lowest recall is 0.84, which was registered against the extracted Arabic Economic relations from sentences with simple structure. This can be explained by the fact that the simple sentence structure in the first experiment contains one *Agent*, one *Patient* and one *Predicate*. For this type of the sentence, if the FDG-SK approach fails to detect one of the semantic function relation elements [*Agent*, *Predicate*, *Patient*], then it cannot extract this relation, thus resulting in reducing the overall recall rate. On other hand, Experiment 2 and Experiment 3 use complex structured sentences, which normally contain several *Agents*, *Predicates* and *Patients*. Therefore, considering that one of the main features of our FDG-

SK approach is that it can detect relation patterns in complex sentences, this results in a higher recall rate for sentences that are richer with relation components, similar to the sentences exemplified in Table 8.

Table 8: Example illustrating high recall rate compared 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%

The value of the **Jordanian Alkhuzf share** decreased by 4.63% followed by **Alitihad Alarabi share** that decreased by 3.37%, while **the Jordanian Emirates share** increased by 2.78%

Relation	Agent	Predicate	Patient	State of the extracted relation
1	سهم الخزف الاردنية Jordanian Alkhuzf share	هبط Decrease	%4.63	True positive
2	سهم الاتحاد العربي Alaitihad Alarabi share	خسر Decrease	%3.37	True positive
3	سهم الاردنية الاماراتية Jordanian Emirates share	ارتفع Increase	2.78%	True negative
4	سهم الخزف الاردنية Jordanian alkhuzf share	خسر Decrease	%3.37	False positive

From the example in Table 8, if we consider that the FDG-SK approach can detect all the Agents and Predicates in the sentence, then it will extract all the possible relations 1-3. However, if the FDG-SK approach fails to extract the *Predicate* word ارتفع [increase] that is associated with relation 3, then it will consequently extract a new relation 4, which is a false positive relation. As a result, the probability of extracting the false positive relations from the complex structured sentences is more than from the simple structured sentences. However, this complexity of the sentences will lead to increase the recall value and to decrease the precision value. In general, Recall is not critical in the assessment of the FDG-SK approach as the modelled Arabic Economic knowledge base is used effectively to semantically classify the identified relation pattern, in contrast, statistical techniques do not conceptualise the knowledge about the entities that can take part of the relation extraction task.

5.3.2Evaluating the performance of the FDG-SK approach for extracting different types of relations

The fourth experiment (Experiment 4) evaluates the performance of the FDG-SK approach in extracting different relation types as presented in Table 9. Overall, the experimental results showed quite satisfactory relation extraction performance for most relation types, scoring, for instance, 0.91 precision for the relation type (organisation-percentage), and 0.85 precision for the relation type (location-industry). However, it can be noticed that there is performance degradation in the extraction of the last four relation types. This is attributed to the FDG-SK approach's difficulty in dealing with some especially complex relations, for instance, when there are several *Predicates* in the same clause. We have discussed this limitation as well as other limitations that affected the accuracy of the extracted Arabic Economic relations in the following section.

Table 9: Evaluating the performance of the FDG-SK approach for extracting different types of relations

Relation Type	Precision	Recall	F-measure
Org – Percentage	0.91	0.98	0.95
Org – Date	0.81	1	0.89
Org – FinancialValue	1	0.78	0.88
country – Industry	0.85	0.78	0.82
country – IncreaseGDP	0.58	0.91	0.71

Relation Type	Precision	Recall	F-measure
country – DecreaseGDP	0.3	1	0.46
country – IncreaseInflation	0.38	0.80	0.52
country – DecreaseInflation	0.60	0.85	0.70

5.3.3 Limitations of the FDG-SK approach

The experimental evaluation revealed that there are some limitations in the output of the Relation Extraction mechanism of the proposed FDG-SK approach that affected the performance of Arabic Relation Extraction task. We attribute these limitations to the following contributing factors:

i. Failing to detect the *agent* due to missing or incorrect named entity

In some cases, the FDG-SK approach failed to detect the *Agent* term, the main part of the semantic function triple [*Agent*, *Predicate*, *Patient*], because of a missed or incorrectly named entity, which can lead to a relation extraction error as some sentences can contain several *Agents*.

سهم النورس تراجع بنسبة 1.21% ليغلق على 0.652 ريال، سهم سيمبكورب هبط بنسبة 1.00

Al-Nawras share retreated 1.21% to close at 0.652 SAR. SembCorp decrease 1.00%

The sentence in the above example contains two relations, the first one is: shareDecreaseBy (سهم النورس [Al-Nawras], 1.21), and the second one is: shareDecreaseBy (سهم سيمبكورب [SembCorp], 1.00). However, Named-Entity Recognition process identified only the first named entity and discarded the second one because سيمبكورب [SembCorp] was transliterated from foreign language and does not follow the Arabic words definition; therefore, it was not recognised by the Arabic named entity pipeline. Consequently, the FDG-SK approach erroneously considered the numeric entity “1.00” relating to the first named entity سهم النورس [Al-Nawras].

ii. Missing the word describing the relation (*Predicate* trigger word)

Ambiguity in relation extraction can occur when the FDG-SK approach fails to recognise the *Predicate* trigger word that describes the relation in the sentence due to its absence from the modelled Arabic Economic knowledge base’s repository (i.e. absent of relation term). In the example below, the root of the word الرابحين [winners] is not included in our knowledge base as a synonym for the relation term shareIncreseBy, which caused the FDG-SK approach to fail to recognise this relation shareIncreseBy (سهم السلام البحريني [Bahraini Al-Salam share], 9.21).

حيث تصدر الرابحين سهم السلام البحريني بنسبة 9.21 ٪ الى مستوى 0.083 دينار

Where the Bahraini Al-Salam share was top of the winners by 9.21% to reach 0.083 Dinar

iii. Nested Named Entity

One of the problems that affected our results was the nested named entities. In some cases, named entity can contain other named entities, which can result in extracting unexpected relations. For example, in the provided sentence in Table 10, the FDG-SK approach extracted three different relations [Relation1, Relation2]. Relation1 represents the relation between the named entity Index and the named entity Percentage (i.e. rate value describes the state of the index). Whereas, both Relation2 present the relation between the named entity Country and the named entity Percentage (i.e. Inflation describing the state of the economy for the country). However, the correct extracted relation is Relation1, whereas Relation2 is an incorrectly defined relation. The is due to the fact that the named entity Index المؤشر العام لبورصة قطر [the general index of

Qatar] embeds the named entity Country قطر [Qatar], which caused the FDG-SK approach to consider that there was a semantic relation between the Country and Percentage.

Table 10: 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 equivalent to 2.6%				
Relation	**Agent**	**Predicate**	**Patient**	**State of relation**
Relation 1	المؤشر العام لبورصة قطر Index	صعد IndexIncreaeBy	2.6% Percentage	Correct
Relation 2	قطر Country	صعد IncreaseInflation	2.6% Percentage	Incorrect

iv. Failing to extract relations with especially complex sentence structure

Complex sentences can contain several relation triggers words (*Predicates*) within the same clause, which can result in extracting irrelevant relations. Examples of such relations are shown in Table 11.

Table 11: Challenge in extracting especially relation from especially complex structures

Arabic sentence	**Entities**		**Trigger word of relation**	
	Entity 1	Entity 2	Predicate 1	Predicate 2
التضخم في السودان يتجاوز 30٪ Inflation in Sudan exceed 30%	السودان Sudan	30%	Inflation	Increase
نمو الناتج المحلي الإجمالي في السودان 20% خلال أربع سنوات GDP growth in Sudan registered 20% in four years	السودان Sudan	20%	GDP	Increase

The sentences in Table 11 represent the relations that describe the *increase* in the value of *Inflation/GDP* for a specific *country* - السودان [Sudan], and these relations have two trigger words (*Predicates*) that describe the relation (*Increase, Inflation*) and (*Increase*, *GDP*). The mechanism used to extract a semantic relation by the FDG-SK approach is based on that one trigger word is used to represents a particular object property (i.e. relation) that refers to the relation between two entities, and therefore the algorithm will fail to extract the relation representing the weighed *increase* of *Inflation*/*GDP* for a specific country.

The undetected relations can be important for the deployment of the Information Extraction function; for instance, for the chosen Economic problem domain, the GDP and Inflation events represent major economic indicators and need to be adequately analysed if the textual analytics output is to be used in a decision support mechanism. Therefore, we investigated the integration of Machine Learning with the FDG-SK approach in a bid to address some of its discussed limitations.

6 INTEGRATING MACHINE LEARNING WITH THE FDG-SK APPROACH

The evaluation of our FDG-SK approach revealed some limitations in extracting relations from sentences with particularly complex structures and high variability in expressing the relation components, such as sentences where the relations are described by more than one *Predicate* (trigger word). For example, FDG-SK could successfully extract complex relations that are represented by one *Predicate* (e.g. the *Predicate* 'drop' represents the relation 'DecreasedBy') whereas it did not perform well for extracting complex relation that are represented by more than one *Predicate* (e.g. the *Predicate* 'increase' and the *Predicate* 'inflation' are complementary to represent the relation 'hasIncreaseInflation'). Therefore, we investigated integrating the FDG-SK approach with Machine Learning in a hybrid approach (FDG-SKML) in order to address the afore-mentioned limitations.

Figure 8 illustrates the architecture of hybrid FDG-SKML approach that comprises the following main components: Feature Extraction, Feature Selection and Building Machine Learning Classifier. In general, the hybrid approach relies on utilising the complex semantic relations extracted by the FDG-SK approach as candidate relations instances, then dataset features are generated from the candidate relations. The proposed hybrid approach also uses genetic algorithms optimisation to select the optimum feature subset in order to boost the performance of the system. The optimised features are used for training a machine learning classifier that facilitates extracting complex relations with significant disparity in the relation elements' presence, order, and correlation.

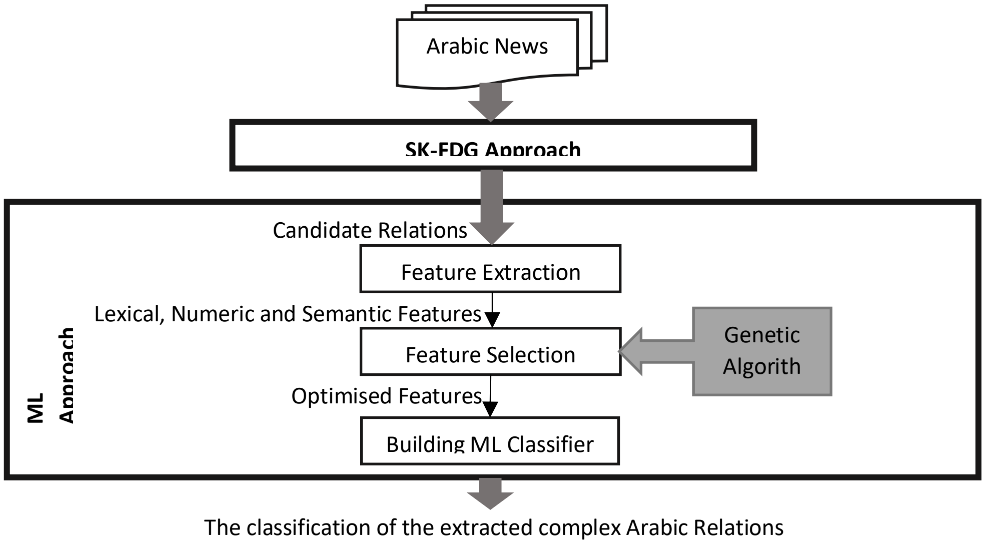


Figure 7: System Architecture of Integrating the FDG-SK approach with Machine Learning approach

6.1 Feature Extraction

A feature is an individual measurable property of the phenomenon being observed and its extraction is considered a crucial step for algorithms in effective pattern recognition, classification and regression. Moreover, the backbone of Relation Extraction is the sentence, and each sentence may contain many clauses and the clause sometimes contains two or more entities. Hence, the hybrid FDG-SKML approach extracts three categories of features from sentences that contain the candidate relation instances, which are lexical features, semantic features and numeric features as listed in Table 12. The listed features are commonly used in Relation Extraction [Boujelben et al. 2014b], but to our knowledge, the use of features 13, 17, 19 is unique in the study of Arabic relation classification. For the lexical features, the window size of 'four words' was established heuristically. Regarding the semantic features, semantic information about the candidate relations such as the domain of property and the range of the property were used.

Table 12: List of the features

No	Feature Category	Feature	Description
01	Lexical feature	L_Ws_B_NEs	List of words between Named Entities
02	Lexical feature	POS_w1_b1	POS of the first word before the first Named Entity
03	Lexical feature	Str_w1_b1	String of the first word before the first Named Entity
04	Lexical feature	POS_w1_a2	POS of the first word after the second Named Entity
05	Lexical feature	Str_w1_a2	String of the first word after the first Named Entity
06	Lexical feature	Str_w2_b1	String of the two words before the first Named Entity
07	Lexical feature	Str_w3_b1	String of the three words before the first Named Entity
08	Lexical feature	Str_w4_b1	String of the four words before the first Named Entity
09	Lexical feature	Str_w2_a2	String of the two words after the second Named Entity
10	Lexical feature	Str_w3_a2	String of the Three words after the second Named Entity
11	Lexical feature	Str_w4_a2	String of the four words after the second Named Entity
12	Lexical feature	POS_ws_B_NEs	POS of words between Named Entities
13	Numeric features	LengthOfRelation	The length of the relation (number of words)
14	Numeric features	Order	Relation direction
15	Semantic features	Domain	The subject of the relation
16	Semantic features	Range	The object of the relation
17	Numeric features	N_Of_FirstNE	The order of the first Named Entity in the relation
18	Numeric features	Distance	Number of words between the Named Entities
19	Numeric features	N_Of_secondNE	The order of the second Named Entity in the relation

6.2 Feature Selection

The numerous features listed in Table 12 above represent distinctive characteristics of the training dataset and hence should be subjected to a selection process to remove undesirable features and arrive at a feature subset that will reduce the dimensionality of the training data and improve the effectiveness of the classification process [Chandrashekar and Sahin 2014]. As the types of the extracted features are not closely related, that makes the utilisation of manual feature selection techniques ineffective. Hence, the hybrid FDG-SKML approach utilises Genetic Algorithms to automate the process of selecting the optimum subset of features that can boost the accuracy of the Relation Extraction models [Hasanuzzaman et al. 2010].

Genetic Algorithms use randomized search and optimisation techniques based on the principles of evolution and natural genetics [Golberg 1989] and have been widely successfully applied the feature selection process [Anbarasi et al. 2010; Oliveira et al. 2010]. Although Genetic Algorithms are more computationally expensive compared to filter-based approaches [Xue et al. 2015], they generally yield more accurate selection results [Alromima et al. 2016]. Moreover, since feature selection in the proposed hybrid FDG-SKML approach is a one-off process, the computational overhead is of little significance.

6.3 Building the Machine Learning Classifier

The optimum set of features selected by the Genetic Algorithm is used to train machine learning algorithms that are widely adopted in Relation classification task; specifically, Support Vector Machine (SVM) and K-Nearest Neighbours algorithm (KNN) in order to classify whether the extracted candidate relation is true or represents a false positive.

SVM is a supervised machine learning algorithm that has been successfully deployed for numerous classification tasks including information extraction [Li et al. 2004] KNN is a non-parametric algorithm that is used for classification and is commonly used in information retrieval regression. It is a simple algorithm showing accurate results with a small number of features [Manning et al. 2008].

6.4 Experimental Evaluation of the Hybrid Approach

This section evaluates the proposed hybrid FDG-SKML approach, which is envisaged to enhance the performance of Arabic Relation Extraction task; in particular to classify the extracted complex relations that have several *Predicates* or several *Patients* into true and false relations. There is lack of Arabic language resources for evaluating relation extraction research. Some Arabic corpora are annotated, namely ACE multilingual training dataset, but it is not freely accessible. Hence, to evaluate our approach, the extracted candidate relation instances were manually annotated with positive class (i.e. correct relation) and negative class (i.e. incorrect relation). Moreover, we created some hand-crafted rules using the Java Annotation Pattern Engine [Thakker et al. 2009] to aid in identifying complex relations that the FDG-SK approach is likely to fail to extract as discussed in section 5.3.3; these relations were also added to the candidate relations set.

The extracted relation instances and their labelled classes were divided into four training datasets and one test dataset as shown in Table 13. The positive and negative classes will be used as baseline classes to evaluate the correctness of the extracted relations. Equations 4, 5 and 6 were used to compute the precision, recall and f-measure of the obtained results respectively.

Table 13: Dataset Specifications

Datasets		Relation's Entities	Type of relation	Positive Relations	Negative Relations	Total
Training Datasets	Dataset 1	Share-Number	IncreaseIn DecreaseIn	712	1464	2176
	Dataset 2	Country-Industry	IncreaseInflation DecreaseInflation	276	20	296
	Dataset 3	Sector-Share	BelongTo	329	05	334
	Dataset 4	Country-Date	GDPDate InflationDate	222	84	306
Testing Dataset	Dataset 5	Country-Number	IncreaseGDP decreaseGDP IncreaseInflation DecreaseInflation	318	479	797

$$\textbf{Precision} = \frac{True\ Positive\ Relations}{True\ Positive\ Relations + False\ Positive\ Relations} \quad (4)$$

$$\textbf{Recall} = \frac{True\ Positive\ Relations}{True\ Positive\ Relations + False\ Negative\ Relations} \quad (5)$$

$$\textbf{F-measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

The lexical, semantic and numeric features (total 19 features) were extracted from the training and the testing datasets as described in section 6.2. Next the Genetic Algorithm was applied to select the optimum feature set to train the relation extraction machine learning classifier. Finally, the machine learning classifiers were tested using the optimised testing dataset.

The two most common methods for evaluating machine learning algorithms are the holdout test and K-fold cross validation. In terms of the K-fold cross validation, the documents are grouped into K partitions of equal size, then each partition is used in turn as a test set while the other remaining partitions are used as a training set. With regard to the holdout test, it is based on randomly selecting documents to be a test set, whereas the other remaining documents to be a training set.

Table 14 presents the obtained results from the SVM (parameters: c = 0.7 and tau = 0.4, where c indicates to the cost associated with allowing training errors and tau indicates to setting the value of uneven margins) and KNN (parameters: k = 5, where k indicates to the number of the nearest neighbor instances) classifiers on two of the training datasets (Dataset 1 and Dataset 3) based on two K-folds (K=5 and K=10). The results assert that the SVM classifier outperforms KNN classifier for classifying whether the extracted Arabic relations are positive relations or negative relations, which is consistent with the findings in [Hmeidi et al. 2008]. Therefore, SVM classifier (c= 0.7 and tau = 0.4) was adopted for the rest of our experiments. In addition, the K-Fold cross validation method with K-fold=10 was adopted too as it was empirically found to perform best in evaluating Machine Learning algorithms [Witten et al. 2016].

Table 14: F-measure of the SVM and KNN relation classifiers

	K-fold=5		**K-fold=10**	
Training Dataset	**SVM**	**KNN**	**SVM**	**KNN**
Dataset 1	0.705	0.687	0.771	0.763
Dataset 3	0.696	0.622	0.670	0.640

Genetic Algorithms have several parameters that should be tuned to best fit a particular optimisation problem. These parameters are: population size, mutation rate and crossover rate. For instance, a relatively small population might not provide a sufficient sample size for the search space in order to reach an optimum solution, and a large population needs more evaluations per generation and hence slower convergence rate. In this study, the values of these parameters were chosen heuristically by means of experimentation on SVM classifier as shown in Table 15. The best configuration of GA's parameters was established as: uniform rate of 0.6, mutation rate of 0.001 and population size of 50.

Table 15: The performance of the SVM algorithm based on different configuration of parameters

Datasets	Dataset / Configuration	Uniform Rate = 0.6/mutation Rate = 0.001/ population Size = 50			Uniform Rate = 0.5/mutation Rate = 0.015/ population Size = 30	
		No	**Subset features**	**F-m**	**Subset features**	**F-m**
Training	Dataset 1	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
	Dataset 2	01	0100001101010101100	0.878	1100001111000101001	0.902
		02	1000111111011110000	0.877	0000001101010111011	0.885
		03	1000011111010110000	0.877	0010011101010111011	0.860
		04	0000001110100011011	0.879	0111011101010111011	0.873
	Dataset 3	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
	Dataset 4	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
Testing	Dataset 5	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

In addition to selecting the best subset of features by GA, the obtained results of the selected features also helped to establish the most influential feature category in the relation extraction process. This was achieved by frequent analysis of the most commonly selected features across all the datasets. In Table 16, the results showed that the lexical features (4,7,8,10) are the most frequently selected by GA. Therefore, it can be concluded that focusing on the words around the first named entity (sliding window) can improve the performance of classifying the extracted relations task. This finding can be exploited to improve the relation extraction process by annotating some specific phrases such as the features that appear in the sentences and are related to the second trigger word such as معدل التضخم [Inflation rate] and الناتج المحلي الاجمالي [Gross domestic product].

Table 16: Accuracy frequency the participation of features

	Lexical												**Numeric**	**Semantic**			**Numeric**		
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

Table 17 presents the obtained results of evaluating the effectiveness of using the selected features (i.e. optimised training datasets) against using all the features (i.e. non-optimised training datasets), which evidenced that the accuracy of the classified extracted relations was improved across all the training datasets.

Table 17: Evaluating the impact of feature selection on the performance of machine learning classifier

Training Datasets	**Selected Subset Features**	**All Features**
Dataset 1	0.775	0.649
Dataset 2	0.879	0.761
Dataset 3	0.994	0.990
Dataset 4	0.752	0.641

Finally, our hybrid FDG-SKML approach was evaluated for extracting complex relations described in section 5.3 that the FDG-SK approach struggled to extract (such as relations based on one *Agent* linked to several *Predicates* and/or several *Patients*. In this experiment, Dataset 5 was used to evaluate both approaches, which covers four types of relations: hasIncreasGDP (country, GDP), hasDecreaseGDP (country, GDP), hasIncreaseInflation (country, Inflation) and hasIncreaseInflation (country, Inflation). As shown in Table 18, the obtained results evidence that our hybrid FDG-SKML approach has significantly improved the classification accuracy of the aforementioned relations, improving the average f-measure of the FDG-SK approach from 0.6 to 0.77, but more significantly, improving the classification accuracy of the relation type hasDecreaseGDP (country, GDP) from 0.46 to 0.77. The obtained results also clearly indicate that the SVM implementation of our hybrid FDG-SKML approach clearly outperforms the KNN classifier implementation with the SVM recording better f-measure for all relation types.

Overall, the results indicate that the hybrid FDG-SKML approach outperforms the baseline FDG-SK approach in terms of the overall f-measure and precision, but there was no noted improvement for the recall results. This can be explained by the fact that the hybrid FDG-SKML approach uses the FDG-SK approach to extract from a sentence (e.g. التضخم في السودان يتجاوز 30٪ [Inflation in Sudan increased 30%]) the candidate relation based on the first *Predicate* such as Increase(السودان [Sudan],30%), and then deploys machine learning to recognise additional relation instances based on the second *Predicate* such as Inflation(Sudan, 30%) to complement the identification of the relation increaseInflation (السودان [Sudan],30%). Therefore, the 'relevant' relations space for the hybrid FDG-SKML approach is large, which can result, in some instances, in a greater number of falsely identified relations and a lower Recall rate.

Table 18: Comparison between the hybrid FDG-SKML approach (using SVM and KNN) and the FDG-SK approach

Measurement		**Precision**			**Recall**			**F-measure**		
Approach		FDG-SKML		FDG-SK	FDG-SKML		FDG-SK	FDG-SKML		FDG-SK
ML classifier		SVM	KNN		SVM	KNN		SVM	KNN	
Relation name	IncreaseGDP	0.73	0.66	0.58	0.87	0.79	0.91	0.80	0.71	0.71
	DecreaseGDP	0.75	0.62	0.3	1.00	0.79	1.00	0.77	0.68	0.46
	IncreaseInflation	0.72	0.53	0.38	0.86	0.56	0.80	0.78	0.54	0.52
	DecreaseInflation	0.67	0.37	0.60	0.8	0.71	0.85	0.73	0.45	0.70

It is worth noting that at the time of compiling this paper, to the best of our knowledge, there was no published research with public datasets and results that evaluated an NLP effort in extracting complex Arabic event-based relations from a specific problem domain, which is the main thesis of our proposed research. All the relevant published research focused on extracting binary Arabic relations of the 'is-a, kind-of', similar to the study in [Zakria et al. 2019]. Closer to the objectives of our research is the work in [Subburathinam et al. 2019], which extracts Arabic event relations using cross-lingual structure to train the relation event extractor from source language annotations and applying it to the target language. However, their approach does not directly process the Arabic text to extract event relations. Therefore, our experimental evaluation could not be directly compared against published works in the field.

7 CONCLUSION

The market size of Text Analytics applications is predicted to reach $30.7 billion by 2030 [MarketWatch. 2022]. However, the market share of Arabic text analytics remains insignificant despite the fact that Arabic is the fourth most used language in the Internet with over 168 million users at the time of writing this document. This is mainly attributed to the well-documented challenges in processing the Arabic natural language in terms of complex morphology, heavy use of discretisation and variation in word semantics, and complex sentence structure. However, the research reported in this paper focused on challenges that are associated with relation extraction from unstructured Arabic text. The Arabic language sentences often contain complex (high order) relations with great variation in the presence, order, and correlation of the relations' subjects, predicates and objects. Hence, in this work we introduced a novel hybrid Semantic Knowledge Base-Machine Learning approach (FDG-SKML) that exploits Functional Discourse Grammar to emphasise the semantic and pragmatic properties of the Arabic language in order to identify the relation patterns in Arabic sentences that are complicated and often contain complex relations. In addition, our novel FDG-SKML relation extraction approach benefits from the advantage of using a domain-specific knowledge base that encoded the semantic association between the relations' *Agents*, *Predicates*, and *Patients* (subject, *Predicates*, and objects), which in turn proved instrumental in identifying the domain-relevant relations in the unstructured text. Moreover, the proposed approach utilises Machine Learning to enhance the performance of Arabic relation extraction task, in particular for determining whether the extracted complex relation that has one *Agent* and several *Predicates* or several *Patients* represent true domain-relevant relation or a false negative.

The experimental evaluation revealed that using the novel FDG-SK relation extraction approach for relation extraction at the initial stage of our hybrid FDG-SKML approach registered satisfactory results in extracting Arabic relations from unstructured texts, but exhibited some limitations when extracting relations from sentences with particularly complex structures and high variability in expressing the relation components, such as sentences where the relations are described by more than one *Predicate* (trigger word). This shortcoming was addressed by integrating the semantic FDG-SK approach with Machine Learning classification, which significantly improved the relation extraction task from complex Arabic sentence structure with significant disparity in the relation elements' presence, order, and correlation. Moreover, we deployed feature selection optimisation to reduce the dimensionality of the search space, which resulted in further improvement in the accuracy of relation extraction.

We claim that the developed relation extraction methodology is applicable to any semantically-rich domain where the relation triggers can be consistently identified, which is critical for building the pattern-recognition based on the Functional Discourse Grammar principles.

The relation extraction mechanism of the FDG-SKML 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. Our future plans include investigating Arabic co-reference resolution to allow the FDG-SKML algorithm to process the whole document to extract the relations between all named entities in the document.

As has been reported in numerous publications, the lack of accessible Arabic language resources is a main hindrance to the advancement of Arabic Natural Language Processing and Information Retrieval research and development. We therefore believe that there is a pressing need for creating an Arabic language processing on-line consortium that collates gazetteers, taxonomies, ontologies, domain-specific corpora etc. Hence, our future plans also involve using our experience in knowledge base systems to investigate the use of Semantic Web technologies for creating an intelligent knowledge base for Arabic language resources.

REFERENCES

Abacha, A.B. & Zweigenbaum, P. 2011. February. A hybrid approach for the extraction of semantic relations from MEDLINE abstracts. In International conference on intelligent text processing and computational linguistics (pp. 139-150). Springer,Berlin, Heidelberg.

Albukhitan, S. & Helmy, T. 2016, October. Arabic ontology learning from un-structured text. In 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI) (pp. 492-496). IEEE.

Alfrjani, R., Osman, T. & Cosma, G. 2019. A hybrid semantic knowledgebase-machine learning approach for opinion mining. Data & Knowledge Engineering, 121, pp.88-108.

Aljamel, A., Osman, T. & Acampora, G. 2015, November. Domain-specific relation extraction: Using distant supervision machine learning. In 2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K) (Vol. 1, pp. 92-103). IEEE.

Aljamel, A., Osman, T., Acampora, G., Vitiello, A. & Zhang, Z. 2018. Smart information retrieval: Domain knowledge centric optimization approach. IEEE Access, 7, pp.4167-4183.

Alromima, W., Moawad, I.F., Elgohary, R. & Aref, M. 2016. Ontology-based query expansion for Arabic text retrieval. Int. J. Adv. Comput. Sci. Appl, 7(8), pp.223-230.

Alruily, M., Ayesh, A. & Zedan, H. 2011. Arabic Language in the Context of Information Extraction Task.

Al-Yahya, M., Aldhubayi, L. & Al-Malak, S. 2014, June. A pattern-based approach to semantic relation extraction using a seed ontology. In 2014 IEEE International Conference on Semantic Computing (pp. 96-99). IEEE.

Al Zamil, M.G. & Al-Radaideh, Q. 2014. Automatic extraction of ontological relations from Arabic text. Journal of King Saud University-Computer and Information Sciences, 26(4), pp.462-472.

Al-Zoghby, A.M., Elshiwi, A. & Atwan, A. 2018. Semantic relations extraction and ontology learning from Arabic texts—a survey. In Intelligent Natural Language Processing: Trends and Applications (pp. 199-225). Springer, Cham.

Anbarasi, M., Anupriya, E. & Iyengar, N.C.S.N. 2010. Enhanced prediction of heart disease with feature subset selection using genetic algorithm. International Journal of Engineering Science and Technology, 2(10), pp.5370-5376.

Attia, M. & Somers, H. 2008. Handling Arabic morphological and syntactic ambiguity within the LFG framework with a view to machine translation (Vol. 279). Manchester: University of Manchester.

Atwan, J. and Mohd, M., 2017. Arabic Query Expansion: A Review. Asian Journal of Information Technology, 16(10), pp.754-770.

Boujelben, I., Jamoussi, S. & Hamadou, A.B. 2014a. A hybrid method for extracting relations between Arabic named entities. Journal of King Saud University-Computer and Information Sciences, 26(4), pp.425-440.

Boujelben, I., Jamoussi, S. & Hamadou, A.B. 2014b, September. Relane: discovering relations between Arabic named entities. In International Conference on Text, Speech, and Dialogue (pp. 233-239). Springer, Cham.

Chandrashekar, G. & Sahin, F. 2014. A survey on feature selection methods. Computers & Electrical Engineering, 40(1), pp.16-28.

Coffey, J.W., Hoffman, R. & Ca˜nas, A. 2006. Concept map-based knowledge modeling: perspectives from information and knowledge visualization. Information Visualization, 5(3), pp.192-201.

Domingue, J., Fensel, D. & Hendler, J.A. eds. 2011. Handbook of semantic web technologies. Springer Science & Business Media.
El-salam, S.M.A., El Houby, E.M., Al Sammak, A.K. & El-Shishtawy, T.A. 2016. Extracting Arabic relations from the web. arXiv preprint arXiv:1603.02488.

Farghaly, A. & Shaalan, K. 2009. Arabic natural language processing: Challenges and solutions. ACM Transactions on Asian Language Information Processing (TALIP), 8(4), pp.1-22.

Golberg, D.E., 1989. Genetic algorithms in search, optimization, and machine learning. Addion Wesley. Reading.

Gormley, M.R., Yu, M. & Dredze, M. 2015. Improved relation extraction with feature-rich compositional embedding models. arXiv preprint arXiv:1505.02419.

Hamadou, A.B., Piton, O. & Fehri, H. 2010, May. Multilingual Extraction of functional relations between Arabic Named Entities using NooJ platform.

Hasanuzzaman, M., Saha, S. & Ekbal, A. 2010, November. Feature subset selection using genetic algorithm for named entity recognition. In Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation (pp. 153-162).

Hengeveld, K. & Mackenzie, J.L., 2006. Functional discourse grammar. Encyclopedia of language and linguistics, 4, pp.668-676.

Hmeidi, I., Hawashin, B. & El-Qawasmeh, E. 2008. Performance of KNN and SVM classifiers on full word Arabic articles. Advanced Engineering Informatics, 22(1), pp.106-111.

Horrocks, G. 2014. Generative grammar. Routledge.

Khalil, H. & Osman, T., 2014, March. Challenges in information retrieval from unstructured arabic data. In 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation (pp. 456-461). IEEE.

Khalil, H., Osman, T. & Miltan, M. 2020. Extracting Arabic Composite Names Using Genitive Principles of Arabic Grammar. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 19(4), pp.1-16.

Konstantinova, N. 2014, April. Review of relation extraction methods: What is new out there?. In International Conference on Analysis of Images, Social Networks and Texts (pp. 15-28). Springer, Cham.

Kumar, B.S. & Ravi, V. 2016. A survey of the applications of text mining in financial domain. Knowledge-Based Systems, 114, pp.128-147.

Lahbib,W., Bounhas, I., Elayeb, B., Evrard, F. & Slimani, Y. 2013. A hybrid approach for Arabic semantic relation extraction.

Li, Y., Bontcheva, K. & Cunningham, H. 2004, September. SVM based learning system for information extraction. In International Workshop on Deterministic and Statistical Methods in Machine Learning (pp. 319-339). Springer, Berlin, Heidelberg.

Maknaz.org. 2022. Maknaz-Expanded Arabic Thesaurus. [online] Available at: http://www.maknaz.org/ [Accessed 11 April 2022].

Manola, F., Miller, E. & McBride, B. 2004. RDF primer. W3C recommendation, 10(1-107), p.6.

MarketWatch. 2022. Text Analytics Market : Key Facts, Dynamics, Segments and Forecast Predictions Presented 2022-2030 (CAGR) of 18%. [online] Available at: <https://www.marketwatch.com/press-release/text-analytics-market-key-facts-dynamics-segments-and-forecast-predictions-presented-2022-2030-cagr-of-18-2022-03-24#:~:text=The%20global%20text%20analytics%20market%20size%20is%20forecast%20to%20reach,period%20from%202022%20to%202030.> [Accessed 11 April 2022].

Martinez-Rodriguez, J.L., Hogan, A. and Lopez-Arevalo, I., 2020. Information extraction meets the semantic web: a survey. Semantic Web, (Preprint), pp.1-81.

Maynard, D., Li, Y. & Peters, W. 2008. NLP Techniques for Term Extraction and Ontology Population. pp. 107-127

Mesmia, F.B., Zid, F., Haddar, K. & Maurel, D. 2017. ASRextractor: a tool extracting semantic relations between Arabic named entities. Procedia Computer Science, 117, pp.55-62.

Mohamed, R., El-Makky, N.M. and Nagi, K., 2015, November. ArabRelat: Arabic Relation Extraction using Distant Supervision. In KEOD (pp. 410-417).

Nichols, J. 1984. Functional theories of grammar. Annual review of Anthropology, 13(1), pp.97-117.

Noy, N., Rector, A., Hayes, P. & Welty, C. 2006. Defining n-ary relations on the semantic web. W3C working group note, 12(4).

Oliveira, A.L., Braga, P.L., Lima, R.M. & Corn´elio, M.L. 2010. GA-based method for feature selection and parameters optimization for machine learning regression applied to software effort estimation. information and Software Technology, 52(11), pp.1155-1166.

Pan, J.Z. 2009. Resource description framework. In Handbook on ontologies (pp. 71-90). Springer, Berlin, Heidelberg.

Rabiee, H.S. 2011, September. Adapting standard open-source resources to tagging a morphologically rich language: a case study with Arabic. In Proceedings of the Second Student Research Workshop associated with RANLP 2011 (pp. 127-132).

Sarhan, I., El-Sonbaty, Y. & Abou El-Nasr, M. 2016. Arabic relation extraction: A survey. International Journal of Computer and Information Technology, 5(5).

Manning, C.D., Schütze, H. & Raghavan, P. 2008. Introduction to information retrieval. Cambridge university press.

Shaalan, K. 2014. A survey of arabic named entity recognition and classification. Computational Linguistics, 40(2), pp.469-510.

Shaalan, K. & Raza, H. 2009. NERA: Named entity recognition for Arabic. Journal of the American Society for Information Science and Technology, 60(8), pp.1652-1663.

Subburathinam, A., Lu, D., Ji, H., May, J., Chang, S.F., Sil, A. & Voss, C. 2019. November. Cross-lingual structure transfer for relation and event extraction. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 313-325).

Thakker, D., Osman, T. & Lakin, P., 2009. Gate jape grammar tutorial. Nottingham Trent University, UK, Phil Lakin, UK, Version, 1.

Witten, I.H., Frank, E., Hall, M.A. & Pal, C.J. 2016. The WEKA workbench. online appendix for "Data Mining: Practical machine learning tools and techniques". In Morgan Kaufmann.

Xue, B., Zhang, M. & Browne, W.N. 2015. A comprehensive comparison on evolutionary feature selection approaches to classification. International Journal of Computational Intelligence and Applications, 14(02), p.1550008.

Zakria, G., Farouk, M., Fathy, K. & Makar, M.N. 2019. Relation Extraction from ArabicWikipedia. Indian Journal of Science and Technology, 12, p.46.

Zitouni, I. ed., 2014. Natural language processing of semitic languages (pp. 299-334). Berlin: Springer