ONTOLOGICAL LEXICON ENRICHMENT: THE BADEA SYSTEM FOR SEMI-AUTOMATED EXTRACTION OF ANTONYMY RELATIONS FROM ARABIC LANGUAGE CORPORA

Maha Al-Yahya[1], Sawsan Al-Malak[2], Luluh Aldhubayi[3]

[1,3] Information Technology Department, King Saud University, P.O. Box 51178, Riyadh 11543, Saudi Arabia

[2] Computer Science Department, College of Computer & Information Sciences, King Saud University, P.O. Box 51178, Riyadh 11543, Saudi Arabia

Email: [1]malyahya@ksu.edu.sa, [2] 434203315@student.ksu.edu.sa, [3]laldubaie@ksu.edu.sa

ABSTRACT

Ontological lexicons are considered a rich source of knowledge for the development of various natural language processing tools and applications; however, they are expensive to build, maintain, and extend. In this paper, we present the Badea system for the semi-automated extraction of lexical relations, specifically antonyms using a pattern-based approach to support the task of ontological lexicon enrichment. The approach is based on an ontology of "seed" pairs of antonyms in the Arabic language; we identify patterns in which the pairs occur and then use the patterns identified to find new antonym pairs in an Arabic textual corpora. Experiments are conducted on Badea using texts from three Arabic textual corpuses: KSUCCA, KACSTAC, and CAC. The system is evaluated and the patterns' reliability and system performance is measured. The results from our experiments on the three Arabic corpora show that the pattern-based approach can be useful in the ontological enrichment task, as the evaluation of the system resulted in the ontology being updated with over 300 new antonym pairs, thereby enriching the lexicon and increasing its size by over 400%. Moreover, the results show important findings on the reliability of patterns in extracting antonyms for Arabic. The Badea system will facilitate the enrichment of ontological lexicons that can be very useful in any Arabic natural language processing system that requires semantic relation extraction.

Keywords: Antonym Extraction, Ontology, Arabic Lexicon, Semantic Relation, Arabic NLP

1.0 INTRODUCTION

A lexicon is defined as "the knowledge that a native speaker has about a language. This includes information about the form and meanings of words and phrases, lexical categorization, the appropriate usage of words and phrases, relationships between words and phrases, and categories of words and phrases" [1]. Lexicon is an essential element for natural language processing (NLP) applications. For some applications, such as machine translation, lexicon is a critical resource [2]. An important aspect of such lexicons that renders them effective, reusable, and sharable within the community is building them upon standards such as semantic Web standards in the form of ontologies. An ontological lexicon is a lexicon designed using an ontological model and developed as an ontology. Ontological lexicons play a vital role in NLP applications such as language analysis, semantic annotation, summarization, machine translation, sense disambiguation, generation of lexical-competence questions used in standard language tests, and other applications that rely on implicit information in the text.

Although ontological lexicons provide a rich source of knowledge for NLP applications, like other types of computational lexicons, they are expensive to build, maintain, and extend [2] [3] [4]. Moreover, the task of relation extraction is essential in any ontological lexicon development. Relation extraction focuses on the extraction of structured relations from unstructured sources such as Web documents or textual corpora [5]. Lexical relation extraction is a type of relation extraction that is concerned with lexical relations between words (lexemes) in a language. Examples of lexical relations include similarity-, synonymy-, and contrast-antonymy, homonymy, polysemy, hyponymy, and other relations [6]. The antonym relationship is used to express the binary contrast or opposition in meaning between two words. Extracting and identifying the antonym relationship in a text is important in various NLP applications that involve language understanding or language acquisition [7].

Existing approaches described in the literature for relation extraction include pattern-based methods [8], supervised methods [9] [10] [11], and bootstrapping methods using seeds and a large collection of corpora [12] [13] [14]. Within each method, various algorithms have been studied on the task of relation extraction, covering various relation types and various domains.

Most of the literature on relation extraction has been explored and studied very well for English and other European languages; however, little work has been done for the Arabic language. Major works in Arabic include the extraction of rhetorical relations [15] [16] and general semantic relations explicitly stated in a text [17], Wikipedia relations [18], synonyms [19], relations between named entities [20], verbs [21], spatial relations [22], grammatical relations [23], and ontological relations [24] [25]. The only study in our review that explicitly mentions antonymy as one of the extracted relations is [19], which uses morpho-lexical patterns applied on a set of Wikipedia articles as a corpus to enrich WordNet, but the results obtained for antonyms were 0%. Our study fills this gap in the field of automated antonym extraction for Arabic. Inspired by the work presented in [26] for the Dutch language, we adopt a similar approach for the Arabic language, antonym extraction using a seed ontology for bootstrapping and corpora to identify the patterns and extract new antonyms.

In this paper, we aim to answer the following question: Can pattern-based approaches to antonym extraction be useful for the enrichment of an Arabic ontological lexicon, and how reliable are these patterns? To answer this question, we present the Badea system, which implements a pattern-based method for the semi-automatic extraction of antonyms from Arabic language corpora using a seed ontology. The method uses an ontology of "seed" pairs of antonyms to facilitate the extraction of lexico-syntactic patterns in which the pairs occur. These patterns are then used to find new antonym pairs in a set of Arabic language corpora. We test the Badea system on three corpora: the King Saud University Corpus of Classical Arabic (KSUCCA) [27], the King Abdul Aziz City for Science and Technology Arabic Corpus (KACSTAC) [28], and the Corpus of Cotemporary Arabic (CAC) [29]. The antonyms extracted were subsequently evaluated, and the pattern reliability and performance of the system were measured. The correctly extracted antonyms were used to enrich the SemTree ontology [30] [31], an ontology-based lexicon for Arabic semantic relations.

The remainder of this paper is organized as follows: Section 2 presents a review of related work on the task of relation extraction in general and for Arabic in particular. Section 3 describes our method of antonym extraction, including materials, experiments, and results. Section 4 analyzes the results and discusses key findings. Finally, section 5 concludes this paper with a summary of the findings and recommendations for future work.

2.0 RELATED WORK

Relation extraction is defined as "the task of discovering semantic connections between entities. In text, this usually amounts to examining pairs of entities in a document and determining (from local language cues) whether a relation exists between them" [32]. Works reported in the literature on relation extraction uses numerous methods that, according to [33], can be divided into four main classes: knowledge-based methods, which usually rely on patterns and are thus sometimes called pattern-based methods, supervised methods, semi-supervised methods, and self-supervised (unsupervised) methods. This section reviews the existing works related to relation extractions in general and for the Arabic language.

Pattern-based methods are one of the earliest and most common approaches to the relation extraction task [8]. They rely on pattern-matching rules that are manually crafted. When patterns are manually crafted for a specific domain, it is usually called a knowledge-based method, as it depends on the knowledge within the domain. A pattern is a linguistic form or structure in which semantically related words occur in a sentence in a given language. Patterns for various semantic relations can either be handcrafted or automatically generated. One of the earliest works on pattern-based extraction methods is the method proposed for hyponyms [8]. The method is based on the use of five manually identified lexico-syntactic patterns to extract the hyponym relation. Although this approach achieves satisfactory results, the process of manually handcrafting patterns is time-consuming, and it is difficult to comprehend all possible patterns, especially when the domain or the discourse of the text changes. In general, pattern-based methods involve high labor costs in crafting the patterns, and the patterns might not be comprehensive. Two alternatives can be used to identify these patterns: bootstrapping with corpus tools and the use of machine learning algorithms to learn patterns from text and then extracting semantic relations.

Patterns can be automatically generated using supervised methods for relation extraction. In these methods, examples of relations are labeled in a training document set where relations are tagged. The model can learn and predict the relations in new data sets by using machine learning techniques. These systems can be easily adapted to any other domain as long as training data are available. Supervised methods are either based on the extracted features [9] or use kernel methods [10] [11]. In feature-based methods, features capable of expressing the learning examples accurately are extracted (feature extraction) [4]. In kernel-based methods, a kernel function that is most appropriate for a specific relation is developed and used to measure the similarity between two entities. In these methods, instead of a feature extraction process, a kernel function that is effective in calculating the similarity of two entities is generated [34].

A study that employs a supervised approach is presented in Turney [35]. It uses a machine learning algorithm for pattern identification that classifies analogous (synonyms, antonyms, and associations) word pairs and can be used to solve multiple-choice analogy, synonym, and synonym-antonym questions. The algorithm is based on a standard supervised machine learning approach, with feature vectors based on the frequencies of patterns in a large corpus. Supervised learning algorithms depend on the availability of labeled data for the target relation types that must be extracted. They are effective for limited and similar documents; however, for varied and large-scale documents, supervised methods are limited [36].

Approaches such as bootstrapping can be helpful, as they require less labeled data and can handle varied document and corpora. Bootstrapping is classified as a semi-supervised or weakly supervised [5] approach. With bootstrapping, patterns and instances are learned simply by iterations starting with a small number of seeds.

Espresso [12] is an example of the use of the bootstrapping approach for semantic relation extraction. Espresso uses an algorithm to extract semantic relations and a bootstrapping algorithm to identify generic patterns automatically. The patterns identified are used to extract a range of semantic relations, including meronymy and hyponymy. The bootstrapping starts with seed pairs, followed by the extraction of all sentences these pairs co-occur in and then generalization of the patterns. Espresso ranks patterns according to ~~a~~ reliability measures that depends on precision and the number of relations discovered. Wang *et al*. [37] describes an automatic pattern construction approach for extracting verb synonyms and antonyms from an English newspaper corpus. Instead of relying on a single pattern, multiple patterns are used to extract results and maximize recall. The approach is based on seed antonyms and synonyms extracted from WordNet. The corpus is analyzed and patterns are constructed on the basis of the seed pairs. Confidence values are then computed for each pattern and used to extract new antonym/synonym pairs.

Another approach that uses seed pairs of antonyms to bootstrap a pattern is described in Lobanova *et al*. [14] [26]. Their approach extracts antonyms using dependency patterns learned from a 450 million-word treebank containing texts from Dutch newspapers. Using a set of seed pairs, patterns are identified and used to find new pairs of antonyms. A treebank is useful for generating dependency patterns expressing relations between words that occur far away from each other, an activity that is more difficult with textual patterns. Similarly, another study based on bootstrapping is the work described in Mohammad *et al*. [13], which uses seed terms to bootstrap a pattern search; however, in their proposed method, the patterns are generated manually.

The bootstrapping method is strongly dependent on the corpus and seed quality; therefore, the quality of the seeds has a huge impact on the quality of the patterns identified [38]. The bootstrapping method has several advantages, including low labor costs, high reliability, and easy implementation. Therefore, it is a widely used method in the area of relation extraction. However, one of the main problems of bootstrapping is "error propagation" [5]. Error propagation refers to the fact that errors in initial stages could generate more errors in later stages and affect extraction accuracy.

A relevant area of the literature rich in relation extraction is that of ontology learning and ontology enrichment. Ontology learning is the process of building an ontology and-ontology enrichment is the process of "extending an existing ontology with additional concepts and semantic relations" [39]. Ruiz-Casado *et al*. [40] [41] described an automatic approach for identifying lexical patterns representing semantic relationships between concepts, with the new patterns being used to extend the ontology with new relations. The approach uses Wikipedia as a source to generate patterns for various relations such as hyperonymy, hyponymy, holonymy, and meronymy. It processes the Wikipedia definitions and determines the sense of each entry by mapping it to the WordNet synset. Wikipedia hyperlinks in the definition are used to extract the pattern for that relation using WordNet relations, and the patterns identified are used to discover new relations other than those in WordNet. Similarly, the LexO framework proposed by Wandmacher *et al*. [42] uses resources such as Wiktionary and

WordNet to extract semantic relations and build the ontology. The extraction is based on the use of Wiktionary, WordNet, and a corpus to generate a hypothesis base from which confidence in each hypothesis is computed. An ontology is created from this hypothesis base by interpreting certain lexical-semantic relations as ontological statements.

A supervised approach for relation extraction for ontology enrichment is presented in Wang *et al*. [43]. The method uses support vector machine (SVM) and features such as parts of speech, entity subtype, entity class, entity role, semantic representation of sentences, and a WordNet synonym set. Specia and Motta [44] reported a hybrid approach for relation extraction that semantically annotates raw text with semantic relations. The approach uses a domain ontology and linguistics tools comprising lexical databases, a lemmatizer, a syntactic parser, a part-of-speech tagger, a named entity recognition system, and pattern matching. The system analyses raw text for linguistic triples (syntactic relationships), with identification of the relations, relying on the knowledge available in the domain ontology and a lexical database and on pattern-based classification and sense disambiguation models.

Wang *et al*. [45] presented a method for synonym extraction based on multiple approaches: two are rule-based and one is a machine learning approach. They use a machine readable dictionary as the corpus. Their results compare their lexicon-based method to corpus-based methods and show that it performs well, despite its computational simplicity. This supports the notion that pattern-based approaches achieve comparable performance to other state-of-the-art relation extraction methods.

2.1 Relation Extraction for Arabic

Although relation extraction has received great attention in English and some other European languages, less attention has been given to relation extraction for Arabic. In this section, we review key studies in the area of relation extraction from Arabic language text.

Similar to approaches for relation extraction in other languages, Arabic language relation patterns can be manually created or automatically extracted. Sadek *et al*. [15] reported on a method based on Rhetorical Structure Theory (RST) for extracting relations from Arabic news website text for the purpose of answering a question. They identify four rhetorical relations, cause, evidence, explanation, and the purpose and use punctuation and cue phrases to guide the relation extraction process. A similar approach using RST for Arabic text summarization from Arabic news websites is described by Ibrahim and Elghazaly [46]. In their approach, cue phrases for the rhetorical relation are identified and used to generate summarized text. A similar approach, presented by Sadek [16], detects causal relations expressed in Modern Standard Arabic. The approach is based on the development of patterns from a set of syntactic features acquired by analyzing an Arabic corpus (Contemporary Arabic Corpus – science domain) and uses cue words and part-of-speech tags to extract patterns for casual relations. They report a recall value of 75%, with a precision of 77%.

Rule mining from Arabic language text is another approach for relation extraction [20]. In this approach, an Arabic language corpus is used to mine lexical (POS), semantic (word category), and numerical (number of words) features. Features are learned from annotated samples (Arabic corpus), and rules are automatically generated for extracting semantic relations. They report a recall value of 53.56%, with a precision of 70% and an F-measure of 60.65%.

Boudabous *et al*. [19] attempted to improve the semantic relations already existing in Arabic WordNet (AWN) [47]. They use a linguistic method based on morpho-lexical patterns to extract semantic relations. Arabic Wikipedia articles are used, as they have a defined structure that can be used for pattern definition and semantic relation extraction. The method comprises two phases: morpho-lexical pattern recognition and semantic relation enrichment. In the first phase, pairs of synsets linked by semantic relations are extracted from AWN. These extracted pairs are then used to select Wikipedia articles after selected sentences are tagged morphologically. Next, the morpho-lexical pattern is identified and used to extract new relations. The results report a precision value of 39%; however, for the antonym relations, they are not satisfactory (0%).

Arabic Wikipedia has also been used to build ontologies and extract relations. For example, Al-Rajebah *et al.* [48] [18] presented a methodology for identifying ontology instances in which the Arabic version of Wikipedia is used as a knowledge source from which concepts and semantic relations are extracted. The algorithm extracts the semantic relations between the article and the features it contains using Wikipedia “Infoboxes.” The approach enriched the ontology with a total number of 760,000 triples.

Amar *et al*. [17] presented a similar method that uses a Lexical Markup Framework (LMF) standardized dictionary instead of the Wikipedia entries for ontology enrichment. The method uses a rule-based system that relies on lexico-syntactic patterns for ontology element extraction. The approach is based on manual analysis of the LMF dictionary for astronomy and the definition of a set of rules to allow for the identification of ontology entities. These rules are then used for the LMF dictionary to extract concepts, relations, and triples for ontology enrichment.

For the same purpose of building an ontology, Lahbib *et al*. [49] reported on a method for relation extraction from vocalized Arabic corpora, specifically the "hadith" corpora. The method uses a hybrid approach with statistical and linguistic methods; thus, syntactic dependencies are used to infer semantic relations. Further, it uses morphological and syntactic analysis to identify semantic relations associated with word usage and not lexical relations. The relations extracted are mainly situational relations, and the authors report a success rate between 50% and 65%.

Imam *et al*. [25] described an approach for Arabic summarization using a domain ontology and extraction of ontology relations that are explicitly stated in the text: taxonomic and semantics. Their approach is evaluated against other summarization systems and against human experts. It scores precision and recall values that are comparable to a human's (precision of 53% against 56%, and recall of 47% against 52%).

Saad *et al*. [21] used a pattern-based approach for extracting verb patterns. Their method first generates a list of all possible Arabic verb patterns and then reduces the list based on Arabic morphological rules. The reported results are 96% accurate. Alnairi *et al*. [22] described a method for spatial relation extraction using a rule-based approach. These rules and patterns are extracted from a spatial relation annotation corpus in addition to other technologies. Their approach focuses on three specific types of spatial relations. Their system performance results in a recall of 81.70%, a precision of 91.1%, and an F-measure of 86.08%. Hammadi *et al*. [23] described a method for grammatical relation extraction in Arabic using a rule-based approach. Creators of the work report an F-measure of 83.60%. Their approach aims to identify the object, subject, and predicate in simple and complex Arabic sentence structures. Nasri *et al*. [51] described a method for the semantic analysis of Arabic text. Their approach first builds an Arabic ontology exploiting some existing linguistics resources and then uses a combination of syntactic parsing tools along with the ontology to extract semantics from the Arabic text. Their approach is still under testing; therefore, the authors do not report any performance results.

Our previous work reported in [52] presents a pattern-based bootstrapping approach using Arabic language corpora and a corpus analysis tool (Sketch Engine) to extract the semantic relations (antonyms) between word pairs. The algorithm is run on the arTenTen corpus [53] and uses LogDice and pattern co-occurrence to classify the extracted pairs into antonyms. The approach utilizes the Sketch Engine corpus query tool and the metrics generated by the engine; no system was designed or implemented for the task. Similarly, our current work presented in this study uses the pattern-based approach with a seed set of antonyms, but the method is implemented in a Web-based system called Badea. Moreover, in this work, we test the approach on three other Arabic corpora.

Despite the existence of work in the area of relation extraction for Arabic, the coverage is still rather limited, and work is still required to enrich this important area of research. According to our review of the literature for Arabic language relation extraction, the studies cover rhetorical relations [15] [16] and general semantic relations explicitly stated in a text [17], Wikipedia relations [18], synonyms [19], relations between named entities [20], verbs [21], spatial relations [22], grammatical relations [23], and ontological relations [24] [25]. The only study in our review that explicitly mentions antonymy as one of the relations extracted is [19], which uses morpho-lexical patterns applied to a set of Wikipedia articles as a corpus to enrich WordNet, but the results obtained for antonym relations were 0%. Our earlier work performed antonym extraction for Arabic, but no system was developed for the task. It was based on queries and measurements generated by the corpus analysis tool. However, this study uses the patterns generated in [52] for the extraction of antonyms using the Badea system.

Our study fills this gap in the field of automated antonym extraction for Arabic and explores the usefulness of pattern-based approaches to antonym extraction for the enrichment of an Arabic ontological lexicon and investigates the reliability of these patterns. Inspired by the work in [26] for the Dutch language, we adopt a similar approach for the Arabic language, using seed pairs of antonyms and patterns for antonym extraction from a set of corpora. These are designed and developed in a system called Badea. Can pattern-based approaches yield satisfactory results when applied to different corpora?

3.0 MATERIAL & PROPOSED METHOD

Patterns

Antonym pairs from the SemTree ontology [30] [31] were used as seeds to extract initial patterns [52]. The SemTree ontology is an ontological lexicon for the Arabic language. It is an extended version of the SemQ ontology [54] [55], which was designed to provide a semantic representation for (noun) antonym pairs found in the Holy Quran. SemTree utilizes the top-level classes found in SemQ and provides properties for two semantic relations: synonymy and antonymy. It contains a total of 110 Arabic synonym pairs and 70 Arabic antonym pairs. Fig. 1 shows the SemTree ontology classes and relations, as well as a sample of the individuals.

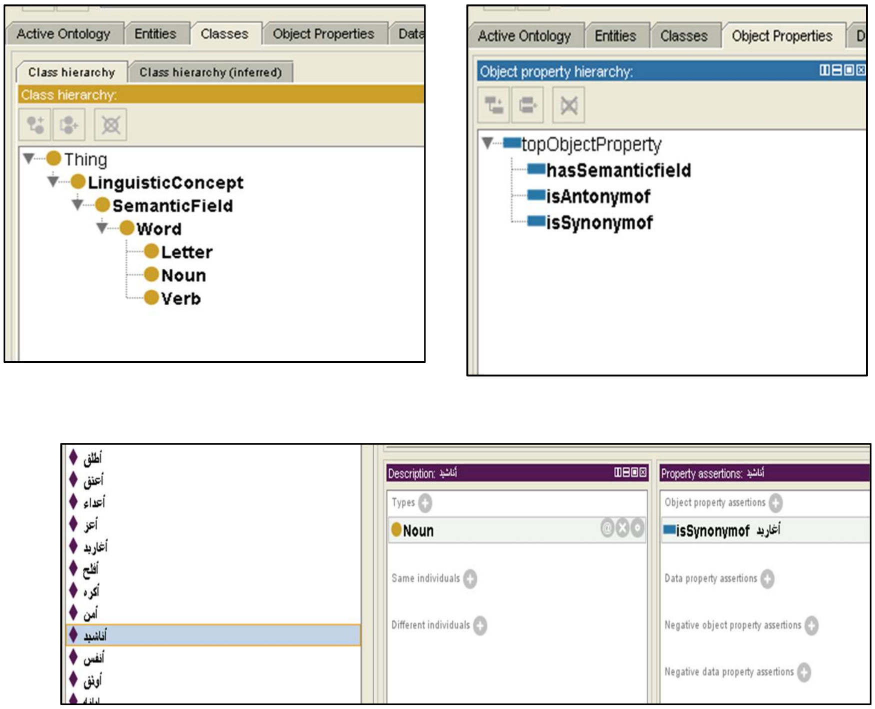

Fig. 1. SemTree ontology classes, relations, and individuals

In our earlier study [52], this set of seed antonyms from the SemTree ontology was used to extract lexico-syntactic patterns using a corpus-based analysis tool, Sketch Engine [56]. Starting with the seed ontology of antonym pairs, using the arTenTen corpus [53], we were able to identify and examine the most frequent antonym pairs (frequent pairs) from the seeds and record their frequencies using Sketch Engine. This was done by querying the corpus for the co-occurrence of the antonym pairs in the sentence boundaries in the corpus. 912 different lexico-syntactic patterns for antonyms were identified. Table 1 shows the most frequent antonym patterns for the antonym pair "حياة" (life) and "موت" (death) in the arTenTen corpus. Fig. 2 shows the relation between the pattern extraction process (earlier study [52]) and our current antonym extraction system, Badea.

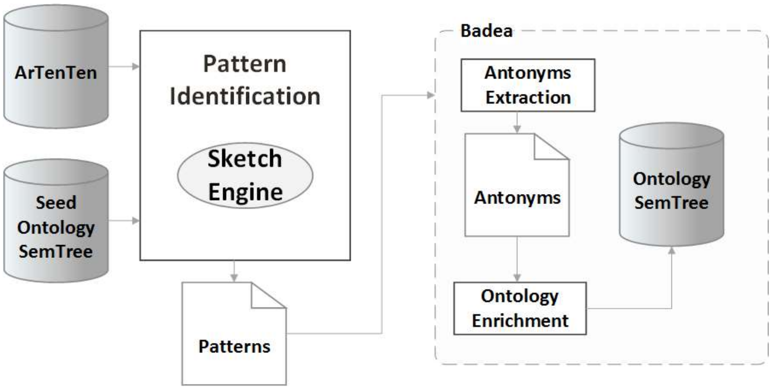

Fig. 2. Relationship between pattern extraction [52] and Badea

Table 1. Most Frequent Antonym Patterns for the Antonym pair "**حياة**" (life) and "**موت**" (death)

مسألة حياة أو موت	بين الموت و الحياة
من الموت إلى الحياة	من الحياة إلى الموت
قضية حياة أو موت	لا موت ولا حياة
معركة حياة أو موت	بمثابة حياة أو موت
بين الحياة و الموت	مباراة حياة أو موت
مسألة حياة او موت	مسالة حياة او موت
من الموت الى الحياة	عن الحياة بعد الموت
حياة أو موت	صراع حياة أو موت
على الموت توهب لك الحياة	وجود حياة بعد الموت
إلى الحياة بعد الموت	في الحياة وبعد الموت
يحبون الموت كما تحبون الحياة	ومن الموت إلى الحياة
هناك حياة بعد الموت	حب الحياة وكراهية الموت
معركة حياة او موت	ليت الموت أعدمني الحياة
قضية حياة او موت	حرب حياة أو موت

Corpora:
Three corpora were used in the experiments:

1. The King Saud University Corpus of Classical Arabic (KSUCCA), a recently developed corpus for classical Arabic [27]. It is freely available online as raw text and contains 50 million tokens. It is clustered into six domains: religion, literature, linguistics, science, biography, and sociology.
2. The Corpus of Contemporary Arabic (CCA) [29], which contains around one million words collected from contemporary Arabic websites and online articles.
3. The King Abdulaziz City for Science and Technology Arabic Corpus (KACSTAC) [28], which contains a large number of Arabic text genres, such as books, newspapers, magazines, journals, and online articles, all of which combine both classical and contemporary Arabic contents.

Badea System:
Our approach to antonym extraction was implemented in the Badea system. The general architecture of the system is depicted in Fig. 3. The system uses a pattern-based approach for relation extraction. Regular expressions are generated from the *lexico-syntactic patterns* and are used to extract new antonyms in the corpus.

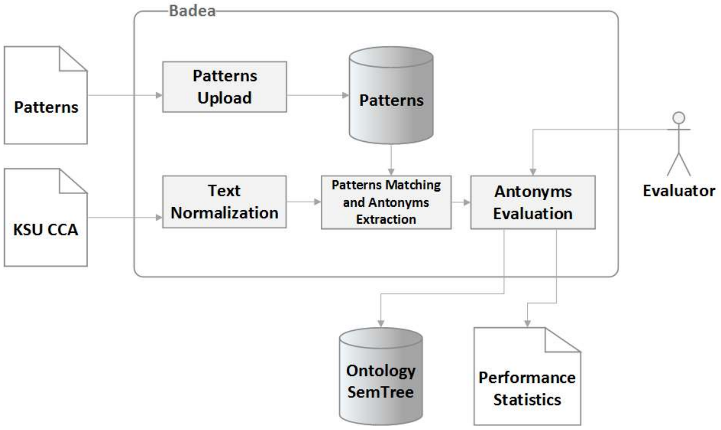

Fig. 3. Architecture of the Badea System and its major Components

The system first obtains the *lexico-syntactic patterns* and then converts these patterns to regular expressions that are used to find antonyms in the given corpus set through "string matching." Corpus preprocessing is required to clean the text, as the text may contain Arabic diacritics (Tashkeel), punctuation marks, and other unwanted symbols. Pattern matching is then carried out. Whenever a string match to one of the lexico-syntactic patterns is found, two relevant words are extracted from the string and stored as antonyms in the system. After all antonyms from the corpus have been extracted, a human evaluator judge the correctness of each antonym pair. Correct antonym pairs are added to the SemTree ontology. This process involves checking to determine whether the antonym pair already exists in the ontology or not. If it is new, the system creates individuals from the pair and adds the following triple to the SemTree ontology: *<word1> <isAntonymOf> <word2>*.

Three measures are computed for evaluation, the reliability of the lexico-syntactic pattern, the precision of the system, and system performance (ontology size after extraction). Pattern reliability [17] is a frequency threshold computed based on the number of times a pattern is able to extract a correct antonym pair. The reliability of each pattern is calculated as follows:

$$P = \frac{C_a}{C_o}$$

where C_a is the total number of correct antonyms that the pattern was able to extract and C_o is the total number of extracted antonyms. System precision is computed to evaluate how successful it is in extracting antonyms compared to human judgment. System precision is evaluated with a cumulative precision score: the ratio of correct antonyms to the total extracted. System performance is the measure of the increase in ontology size.

4.0 EXPERIMENTS RESULTS AND DISCUSSION

4.1 Experiment 1

The 912 *lexico-syntactic patterns* identified in our previous work [52] are used to extract new antonym pairs from the KSUCCA. The system was tested on a sample set from each of the six categories of the KSUCCA. The texts selected for the experiment ranged from 4,000 to 80,000 words. The results obtained for each category and sub-category are displayed in Tables 2 and 3, respectively. The system computes a pattern reliability measure for all patterns. Table 4 shows a subset of the reliable patterns for antonym extraction in the sample set of the KSUCCA, with reliability measuring 20% and higher. The reported precision of this experiment was very low. However, the system performance was very impressive.

Table 2. Results of Antonym Extraction for all Categories

Category	Antonyms extracted	Precision score
Religion	321	1.09%
Science	146	0.71%

Literature	104	0.67%
Linguistics	66	0.66%
Biography	52	0.33%
Sociology	21	0.91%
Total	733	0.81%

Table 3. Antonym Extraction Statistics for Sub-categories

Category	**Sub-category**	**Name of original text**	**Antonyms extracted**
Religion	Quran	القرآن الكريم	105
	Hadith	الآثار لمحمد بن الحسن	30
		الأدب لابن أبي شيبة	
		الأمر بالمعروف والنهي عن المنكر	
		الزهد لأبي داود السجستاني	
	Exegesis of Quran	تفسير القرآن من الجامع لابن وهب	16
		تفسير الثوري	
		الجزء فيه تفسير القرآن ليحيى بن يمان ونافع بن أبي نعيم القارئ ومسلم بن خالد الزنجي وعطاء الخراساني برواية أبي جعفر الترمذي	
	Qur'anic Studies	أحكام القرآن للجهضمي	62
		أخلاق أهل القرآن	
		غريب القرآن المسمى بنزهة القلوب	
	Hadith Studies	الإلزامات والتتبع للدارقطني	18
		المحدث الفاصل بين الراوي والواعي	
		رسالة أبي داود إلى أهل مكة وغيرهم في وصف سننه	
		مجموعة رسائل في علوم الحديث	
	Belief	أصول السنة، ومعه رياض الجنة بتخريج أصول السنة	39
		الإبانة عن أصول الديانة	
		كتاب الأيمان "ومعالمه، وسننه، واستكماله، ودرجاته"	
		شرح السنة معتقد إسماعيل بن يحيى المزني	
		كتاب الأصنام	
		تخريج العقيدة الطحاوية	
	Jurisprudence	جزء في مسائل عن أبي عبد الله أحمد بن حنبل	18

Category	Sub-category	Name of original text	Antonyms extracted
		مسائل أحمد بن حنبل رواية ابنه عبد الله	
	Principles of Jurisprudence	الرسالة للشافعي أبو عبدالله محمد بن إدريس المطلبي القرشي المكي	33
Literature	Grammar	الجمل في النحو	33
		اللامات	
		اللمع في العربية	
		علل التثنية	
		رسالة منازل الحروف	
		معاني الحروف	
	Language	إصطلاح المنطق لابن السكيت أبو يوسف يعقوب بن إسحاق	13
	Lexicons	الأزمنة وتلبية الجاهلية	29
		الزاهر في غريب ألفاظ الشافعي	
	Proverbs	الأمثال	29
		الأمثال المولدة	
Linguistics	Poetry	مقامات بديع الزمان الهمذاني	27
		نقد الشعر	
		معلقة طرفة بن العبد	
		أخبار أبي تمام	
		القوافي	
	Novels	كليلة ودمنة	15
	Eloquence	الألفاظ (الكتابة والتعبير)	24
		الإبل	
		الأصمعيات اختيار الأصمعي	
		الأمالي في آثار الصحابة للحافظ الصنعاني	
		الآمل والمأمول	
Science	History	الديباج	23
		الردة مع نبذة من فتوح العراق وذكر المثنى بن حارثة الشيباني	
	Geography	المسالك والممالك للاصطخري	23
		الكتاب العزيزي أو المسالك والممالك	
		التبصرة بالتجارة في وصف ما يستظرف في البلدان من الأمتعة الرفيعة والأعلاق النفيسة والجواهر الثمينة	
	Medicine	) مختصر في الطب) العلاج بالأغذية والأعشاب في بلاد المغرب	9
	Physics	الجماهر في معرفة الجواهر	9
	Astronomy	زيج الصابئ	78
	Philosophy	رسائل فلسفية	2
	Politics	رسالة ضمن «مجموع في السياسة«	5
	Miscellaneous	مفاتيح العلوم	20
Biography	Muhammad PBUH	مختصر الشمائل المحمدية	46
		الشمائل المحمدية	
	Other Biographies	من اسمه عمرو من الشعراء	6
		أحوال الرجال	
		أخبار أبي حفص عمر بن عبد العزيز رحمه الله وسيرته	

Category	Sub-category	Name of original text	Antonyms extracted
		أخبار المكيين من كتاب التاريخ الكبير لابن أبي خيثمة	
		أخبار النحويين البصريين	
		أخبار الوافدين من الرجال من أهل البصرة والكوفة على معاوية بن أبي سفيان	
Sociology	Ethics and Morals	آداب الشافعي ومناقبه	5
		آداب النفوس	
		الحث على طلب العلم والاجتهاد في جمعه	
		رسالة المسترشدين	
		الأعضاء والنفس	
	Genealogy	حذف من نسب قريش	16
		جمهرة نسب قريش وأخبارها	
		مختلف القبائل ومؤتلفها	

Table 4. Subset of Patterns with Reliability Measuring ≥ 20%

Pattern	Antonyms extracted	Reliability measure
الأمر بس والنهي عن ص	13	100%
وحب س وكراهية ص	1	100%
أنها س أو ص	1	100%
اليوم س أو ص	1	100%
في س أم في ص	1	100%
بعد س أو ص	48	72.7%
يحب س ويكره ص	2	66.7%
س كان أو ص	24	58.5%
فى س وفى ص	6	54.5%
في س لا في ص	2	50.0%
في س ولا في ص	13	39.0%
عن س وعن ص	16	38.0%
يعرف س من ص	6	37.5%
في س وبعد ص	1	33.3%
نحو س أو ص	1	33.3%
أن س أفضل من ص	1	33.3%
تخرج س من ص	7	31.8%
يخرج س من ص	11	30.6%
وضع س أو ص	1	25%
إما س وإما ص	9	23.7%
أحدهما س والآخر ص	4	23.5%
أمر س أو ص	1	20%
حق س بعد ص	1	20%
حد س أو ص	1	20%

4.2 Experiment 2

Although our first experiment resulted in very high level of system performance, as the ontology was updated with over 300 new antonym pairs, thereby enriching the lexicon with a 400% increase in the size of the lexicon, the precision of the system was extremely low. This was caused by a large number of patterns with very low scores, patterns that extracted many incorrect antonym pairs. Only 4.82% of the patterns performed well in extracting correct antonym pairs. The previous experiment resulted in a large number of extracted antonyms with a precision of only 0.81%.

The 912 lexico-syntactic patterns used in experiment 1 were analyzed and evaluated, and the reliability measure of each pattern was analyzed. The measure for the patterns varied, and some patterns had no score at all. Our analysis of the measures showed that patterns with a measure of 20% or higher provided an acceptable level of performance in extracting antonyms.

Our second experiment limited the extraction of antonyms to patterns with a measure of 20% or higher; therefore, only 44 lexico-syntactic patterns were used in the second experiment. Moreover, to improve system precision, we tested the system using three different corpora and compared the results.

Since the KSUCCA includes text only in classical Arabic, in this second experiment, we used two other corpora to cover different variations of the Arabic language, modern and contemporary Arabic. We used the Corpus of Contemporary Arabic (CCA) [29] and the King Abdulaziz City for Science and Technology Arabic Corpus (KACSTAC) [28]. We conducted the second experiment with the 44 lexicon-syntactic patterns on the complete KSUCCA and on subsets of the other two.

This resulted in a total of 1,789 antonyms for the same corpus with an improved precision of 27.61%. The precision on the three corpora was significantly better at 28.53% precision, and the total number of correct antonyms is larger, at 746 pairs. Details of the antonyms extracted and the precision and system performance for the different corpora in both experiments are shown in Table 5, and details of the pattern reliability scores are shown in Table 6.

Table 5. Antonyms Extracted and Precision for Different Corpuses

Corpus	*Total Number of Words*	*Patterns*	*Antonyms Extracted*	*Correct Antonyms*	*Precision Score*	*System Performance*
KSUCCA experiment 1	1,819,351	913	90,822	733	0.80%	400%
KSUCCA experiment 2	1,819,351	44	1,789	494	27.61%	423%
KACSTAC experiment 2	74,921	44	384	106	27.60%	
CCA experiment 2	83,843	44	441	146	33.11%	

Table 6. Subset of the Patterns Arranged by Reliability

Patterns	*Unique Antonyms Extracted*	*Unique Correct Antonyms*	*Pattern Reliability*
الأمر بس والنهي عن ص	13	13	100%
في س أم في ص	3	3	100%
اليوم س أو ص	1	1	100%
يحب س ويكره ص	1	1	100%
نحو س أو ص	3	3	100%
بعد س أو ص	66	52	79%
س كان أو ص	41	31	76%
في س ولا في ص	32	23	71.9%
فى س وفى ص	11	7	64%
في س لا في ص	5	3	60%
تخرج س من ص	25	12	48%
أحدهما س والآخر ص	17	8	47%
يعرف س من ص	16	7	44%
عن س وعن ص	42	18	43%
من س او ص	29	10	35%
حق س بعد ص	3	1	33%
إما س وإما ص	37	12	32%
لا س ولا ص	737	190	26%
حد س أو ص	4	1	25%

Patterns	*Unique Antonyms Extracted*	*Unique Correct Antonyms*	*Pattern Reliability*
س خير من ص	115	24	21%

4.3 Discussion

The results obtained from our experiments show that the pattern-based method is able to extract correct antonyms, which, in the first experiment, resulted in a system precision of 0.81% and a system performance of 400%. The latter measure indicates the increase in the size of the ontology resulting from the first experiment, which is a very encouraging result. In the second experiment, we were able to increase the system precision by selecting the best patterns based on their score from the first experiment. The results clearly indicate that both the first and second experiment were able to increase the lexicon size by 400% and 423%, respectively.

Regarding pattern reliability, our first experiment shows that only a few patterns (4.82%) were effective in identifying antonym relationships. This can be explained by that fact that some of these patterns are very commonly used in the Arabic language and do not necessarily discriminate antonymous pairs from other related pairs of words. For example, the pattern "ال للأمين ال الممثل," extracted using the seed antonyms "العام" and "الخاص," is not an antonymous pattern (the two words are usually considered antonyms, but in this specific context and sentence, they are not). Moreover, regarding the pattern reliability measure, Table 6 shows that some patterns were 100% reliable; however, when these patterns were revisited and the antonym pairs discovered for each pattern reviewed, only one unique pair was found for each pattern. The patterns with reliability measures in the range of 50–60% were better at discovering more antonym pairs. Therefore, it is important to consider the number of resulting unique antonym pairs when computing the pattern reliability measure.

An interesting finding resulting from this study is the fact that, even though some incorrect antonym pairs extracted were not actual antonyms, a semantic relation did exist between them, including synonymy and hyponymy. For example, there is a semantic relation between the words "سجود" and "ركوع," "صلاة" and "وضوء," "الدين" and "الوصية," "الكواكب" and "الشمس," "متوسط" and "متدرج," and "فرسا" and "برذون." This finding indicates that our method can aid in the discovery of other semantic relations in Arabic that are not necessarily antonyms, even when the original patterns are identified from antonym pairs.

In addition, it is important to note from our first experiment that, in the pattern identification process used in the experiment, we based our selection of patterns on two main criteria: the frequency of the pair and the frequency of the pattern. The results show that frequent patterns may not always yield correct antonyms; some less frequent patterns may also be good patterns.

The results of the second experiment, comparing the system performance on the three corpora, suggest that patterns extracted from the arTenTen corpus are more applicable to contemporary Arabic than to classical Arabic and yielded better results when applied to the contemporary Arabic corpora. An interesting finding from this is that the performance of patterns depends on the variation of the language used to extract the patterns.

Comparing the results with our previous reported results, we can see that our system performed as well as a corpus analysis tool (Sketch Engine). This may result from the fact that the patterns extracted were used to extract antonyms from the same corpus. However, in the current study, the patterns were generated from the arTenTen corpus and the antonyms were extracted from three different corpuses.

As emphasized in the first experiment, patterns with good scores do not necessarily extract more antonym pairs than others. It is important to distinguish patterns that performed well in our experiment from those that did not. Details of the pattern scores are shown in Tables 4 and 6. The results indicate that most of the patterns with perfect scores are specific to a certain context and that patterns with lower scores tend to be more general and are capable of extracting a larger collection of antonym pairs.

The pattern scoring we used is somewhat biased toward the patterns that were accurate. That is, when a pattern extracts a correct yet small number of antonyms, its score is 100%. Such patterns are usually very specific to the context, extracting the same antonym pair many times. More general patterns are able to extract larger numbers of antonym pairs, requiring the score to give bigger weight to its ability to extract antonyms rather than the accuracy. The latter leads to the use of a weighted pattern score that considers the number of correctly extracted antonyms in the scoring process by simply multiplying by the number of unique antonym pairs that the pattern correctly extracted. The resulting scores will reflect the performance of patterns better than their accuracy.

Although some improvements were reported in our second experiment, it is important to have a benchmark or a gold standard to evaluate our system against other systems for antonym relation extraction. An annotated corpus of antonyms would be a valuable addition to the research community in this field to enable a comparison of various approaches.

5.0 CONCLUSION AND FUTURE WORK

In the current paper, we presented a pattern-based method for the semi-automatic extraction of antonyms from Arabic language corpora using a seed ontology to support the semi-automated construction and enrichment of an Arabic ontological lexicon. The method uses an ontology of "seed" pairs of antonyms to facilitate the extraction of lexico-syntactic patterns in which the pairs occur. These patterns are then used to find new antonym pairs in a set of Arabic language corpora. The Badea system was developed to test the approach on three different Arabic corpora: the King Saud University Corpus of Classical Arabic (KSUCCA) [27], the KACSTAC corpora [28], and the CAC corpora [29]. The antonyms extracted are subsequently evaluated; the pattern reliability, precision, and performance of the system were measured. The correctly extracted antonyms were used to enrich the SemTree ontology [30] [31], an ontology-based lexicon for Arabic semantic relations.

The evaluation results from our experiments indicated that the pattern-based approaches implemented in the Badea system performed well on antonym extraction and the ontology enrichment task, with an increase in ontology size by 400% in the first experiment and 423% in the second experiment. However, regarding the precision of the system, the results were not as good as the ontology enrichment.

The system improved in the second experiment when patterns were selectively chosen and the corpora were varied to include not only classical Arabic but modern and contemporary Arabic texts as well. However, the improvement was not as good as that reported in our earlier work [52]. This may be a result of the fact that the patterns extracted were used to extract antonyms from the same corpus. However, in the current study, the patterns were generated from the arTenTen corpus and the antonyms were extracted from three different corpora.

The results from our experiments indicate that pattern filtering and using a corpus set with Contemporary and Modern Arabic enhanced system performance. Moreover, our results indicate that the pattern scoring technique was not sufficient. We introduced a modification to the pattern scoring technique that better reflects pattern reliability by incorporating the number of unique antonyms extracted into the score.

The shortcomings and limitations will direct our future work on the pattern-based method for antonym extraction. A human expert was involved in the evaluation of the correctness of the antonyms in our study. This is a time-consuming activity requiring considerable effort and significant human involvement in the whole process; therefore, it is advised that the number of resulting antonyms should be filtered, as the number was extremely high in our experiment, and that automatic methods of evaluation should be adopted. A gold standard such as Arabic WordNet can be used to compare the results obtained instead of a human expert. An important development that can aid in the evaluation is to deploy Badea as a Web-based application and adopt a crowdsourcing method, in which users can submit Arabic language texts for antonym extraction and participate in the evaluation of the right antonyms.

The results from computing the pattern reliability measure highlight interesting questions for further research in this area: Can pattern reliability measure(s) be predicted and computed accurately? Other than the correctness of the antonyms extracted, what factors influence the effectiveness of a pattern in eliciting a semantic relation between words?

Two important aspects of the design and implementation of the Badea system need to be highlighted. First, the Badea system is designed as a generic system so that it can be used to extract any type of semantic relation, given a set of patterns. Second, the SemTree ontology that it uses is based on Web standards, which means that, as a language resource, it can be shareable and reusable in many different Arabic NLP applications. As the next step, we intend to make it available for the community, accessible publicly via Badea APIs.

REFERENCES

[1] "LinguaLinks Library 5.0 Plus," *SIL International*. [Online]. Available: http://www.sil.org/resources/publications/entry/40892. [Accessed: 20-Dec-2015].

[2] R. Cole, Ed., *Survey of the State of the Art in Human Language Technology*. New York, NY, USA: Cambridge University Press, 1997.

[3] M. Sammer and S. Soderland, *Building a Sense-Distinguished Multilingual Lexicon from Monolingual Corpora and Bilingual Lexicons*. 2007.

[4] C. Huang, N. Calzolari, A. Gangemi, A. Lenci, A. Oltramari, and L. Prevot, *Ontology and the Lexicon: A Natural Language Processing Perspective*, 1st ed. Cambridge University Press, 2010.

[5] N. Konstantinova, "Review of Relation Extraction Methods: What Is New Out There?," in *Analysis of Images, Social Networks and Texts*, D. I. Ignatov, M. Y. Khachay, A. Panchenko, N. Konstantinova, and R. E. Yavorsky, Eds. Springer International Publishing, 2014, pp. 15–28.

[6] M. L. Murphy, *Semantic Relations and the Lexicon: Antonymy, Synonymy and other Paradigms*. Cambridge University Press, 2003.

[7] E. S. Gjergo and S. Delija, "The Role and Function of the Antonyms in Language," *Mediterranean Journal of Social Sciences*, vol. 5, no. 16, p. 703, Sep. 2014.

[8] R.G. Raj and S. Abdul-Kareem. "A Pattern Based Approach for The Derivation Of Base Forms Of Verbs From Participles And Tenses For Flexible NLP". *Malaysian Journal of Computer Science*, Vol. 24(2): Jun. 2011. pp 63-72.

[9] D.-T. Vo and C.-Y. Ock, "Extraction of Semantic Relation Based on Feature Vector from Wikipedia," in *Proceedings of the 12th Pacific Rim International Conference on Trends in Artificial Intelligence*, Berlin, Heidelberg, 2012, pp. 814–819.

[10] S.-P. Choi, S. Lee, H. Jung, and S.-K. Song, "An Intensive Case Study on Kernel-based Relation Extraction," *Multimedia Tools Appl.*, vol. 71, no. 2, pp. 741–767, Jul. 2014.

[11] H. Jung, S.-P. Choi, S. Lee, and S.-K. Song, "Survey on Kernel-Based Relation Extraction," in *Theory and Applications for Advanced Text Mining*, S. Sakurai, Ed. InTech, 2012.

[12] P. Pantel and M. Pennacchiotti, "Espresso: leveraging generic patterns for automatically harvesting semantic relations," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA, 2006, pp. 113–120.

[13] S. Mohammad, B. Dorr, and G. Hirst, "Computing Word-pair Antonymy," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA, 2008, pp. 982–991.

[14] A. Lobanova, G. Bouma, E. Tjong, and K. Sang, "Using a Treebank for Finding Opposites," presented at the TLT9, Tartu, Estonia, 2010, pp. 139–150.

[15] J. Sadek, F. Chakkour, and F. Meziane, "Arabic Rhetorical Relations Extraction for Answering 'Why' and 'How to' Questions," in *Proceedings of the 17th International Conference on Applications of Natural Language Processing and Information Systems*, Berlin, Heidelberg, 2012, pp. 385–390.

[16] J. Sadek, "Automatic Detection of Arabic Causal Relations," in *Natural Language Processing and Information Systems*, E. Métais, F. Meziane, M. Saraee, V. Sugumaran, and S. Vadera, Eds. Springer Berlin Heidelberg, 2013, pp. 400–403.

[17] F. B. B. Amar, B. Gargouri, and A. B. Hamadou, "Domain Ontology Enrichment Based on the Semantic Component of LMF-Standardized Dictionaries," in *Knowledge Science, Engineering and Management*, M. Wang, Ed. Springer Berlin Heidelberg, 2013, pp. 404–419.

[18] N. I. Al-Rajebah and H. S. Al-Khalifa, "Extracting Ontologies from Arabic Wikipedia: A Linguistic Approach," *Arab J Sci Eng*, vol. 39, no. 4, pp. 2749–2771, Sep. 2013.

[19] M. M. Boudabous, N. C. Kammoun, N. Khedher, L. H. Belguith, and F. Sadat, "Arabic WordNet semantic relations enrichment through morpho-lexical patterns," in *Communications, Signal Processing, and their Applications (ICCSPA), 2013 1st International Conference on*, 2013, pp. 1–6.

[20] I. Boujelben, S. Jamoussi, and A. B. Hamadou, "Enhancing Machine Learning Results for Semantic Relation Extraction," in *Natural Language Processing and Information Systems*, E. Métais, F. Meziane, M. Saraee, V. Sugumaran, and S. Vadera, Eds. Springer Berlin Heidelberg, 2013, pp. 337–342.

[21] E. M. Saad, M. H. Awadalla, and A. Alajmi, "Arabic verb pattern extraction," in *2010 10th International Conference on Information Sciences Signal Processing and their Applications (ISSPA)*, 2010, pp. 642–645.

[22] F. A. Alnairia, N. Omar, and M. Albared, "Extraction of Spatial Relation in Arabic Text Using Rule-Based Approach.," *International Journal of Advancements in Computing Technology*, vol. 4, no. 15, 2012.

[23] O. I. Hammadi and M. J. A. Aziz, "Grammatical Relation Extraction in Arabic Language," *Journal of Computer Science*, vol. 8, no. 6, pp. 891–898, 2012.

[24] M. G. H. Al Zamil and Q. Al-Radaideh, "Automatic extraction of ontological relations from Arabic text," *Journal of King Saud University - Computer and Information Sciences*, vol. 26, no. 4, pp. 462–472, Dec. 2014.

[25] I. Imam, N. Nounou, A. Hamouda, and H. Khalek Abdul, "An Ontology-based Summarization System for Arabic Documents (OSSAD)," *International Journal of Computer Applications*, vol. 74, no. 17, pp. 38–43, 2013.

[26] Lobanova, A., van der Kleij, T, and J. Spenader, "Defining antonymy: a corpus-based study of opposites by lexico-syntactic patterns," *International Journal of Lexicography*, vol. 23, pp. 19–53, 2010.

[27] M. Alrabiah, A. Al-Salman, and E. Atwell, "The design and construction of the 50 million words KSUCCA King Saud University Corpus of Classical Arabic," in *Workshop on Arabic Corpus Linguistics*, Lancaster University, UK., 2013.

[28] A. Al-Thubaity, M. Khan, M. Al-Mazrua, and M. Al-Mousa, "New Language Resources for Arabic: Corpus Containing More Than Two Million Words and a Corpus Processing Tool," in *2013 International Conference on Asian Language Processing (IALP)*, 2013, pp. 67–70.

[29] L. Al-Sulaiti and E. S. Atwell, "The design of a corpus of contemporary Arabic," *International Journal of Corpus Linguistics*, vol. 11, no. 2, pp. 135–171, May 2006.

[30] M. Al-Yahya, M. Al-Shaman, N. Al-Otaiby, W. Al-Sultan, A. Al-Zahrani, and M. Al-Dalbahie, "SemTree ontology for enriching Arabic text with lexical semantic annotations," in *2015 IEEE International Conference on Semantic Computing (ICSC)*, 2015, pp. 167–168.

[31] M. Al-Yahya, M. Al-Shaman, N. Al-Otaiby, W. Al-Sultan, A. Al-Zahrani, and M. Al-Dalbahie, "Ontology-Based Semantic Annotation of Arabic Language Text.," *International Journal of Modern Education & Computer Science*, vol. 7, no. 7, 2015.

[32] A. Culotta, A. McCallum, and J. Betz, "Integrating probabilistic extraction models and data mining to discover relations and patterns in text," in *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, 2006, pp. 296–303.

[33] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld, “Open Information Extraction from the Web,” *Commun. ACM*, vol. 51, no. 12, pp. 68–74, Dec. 2008.

[34] C. Zhang, W. Xu, S. Gao, and J. Guo, “A bottom-up kernel of pattern learning for relation extraction,” in *2014 9th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2014, pp. 609–613.

[35] P. D. Turney, “A Uniform Approach to Analogies, Synonyms, Antonyms, and Associations,” in *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, Stroudsburg, PA, USA, 2008, pp. 905–912.

[36] F. Mesquita, “Clustering Techniques for Open Relation Extraction,” in *Proceedings of the on SIGMOD/PODS 2012 PhD Symposium*, New York, NY, USA, 2012, pp. 27–32.

[37] W. Wang, C. Thomas, A. Sheth, and V. Chan, “Pattern-based synonym and antonym extraction,” in *Proceedings of the 48th Annual Southeast Regional Conference*, New York, NY, USA, 2010, pp. 64:1–64:4.

[38] C. Zhang, W. Xu, Z. Ma, S. Gao, Q. Li, and J. Guo, “Construction of semantic bootstrapping models for relation extraction,” *Knowledge-Based Systems, vol*. 83 issue C , 2015.

[39] G. Petasis, V. Karkaletsis, G. Paliouras, A. Krithara, and E. Zavitsanos, “Ontology population and enrichment: State of the art,” in *Knowledge-driven multimedia information extraction and ontology evolution*, 2011, pp. 134–166.

[40] M. Ruiz-casado, E. Alfonseca, and P. Castells, “Automatic extraction of semantic relationships for wordnet by means of pattern learning from wikipedia,” in *In NLDB*, 2005, pp. 67–79.

[41] M. Ruiz-casado, E. Alfonseca, and P. Castells, “Automatising the Learning of Lexical Patterns: an Application to the Enrichment of WordNet by Extracting Semantic Relationships from Wikipedia,” *Journal of Data and Knowledge Engineering*, vol. 61, pp. 484–499, 2007.

[42] T. Wandmacher, E. Ovchinnikova, U. Krumnack, and H. Dittmann, “Extraction, evaluation and integration of lexical-semantic relations for the automated construction of a lexical ontology,” in *Proceedings of the Third Australasian Workshop on Advances in Ontologies - Volume 85*, Darlinghurst, Australia, Australia, 2007, pp. 61–69.

[43] T. Wang, Y. Li, K. Bontcheva, H. Cunningham, and J. Wang, “Automatic Extraction of Hierarchical Relations from Text,” in *Proceedings of the 3rd European Conference on The Semantic Web: Research and Applications*, Berlin, Heidelberg, 2006, pp. 215–229.

[44] L. Specia and E. Motta, “A hybrid approach for extracting semantic relations from texts,” in *IN. PROCEEDINGS OF THE 2 ND WORKSHOP ON ONTOLOGY LEARNING AND POPULATION*, 2006, pp. 57–64.

[45] T. WANG and G. HIRST, “Exploring patterns in dictionary definitions for synonym extraction,” *Natural Language Engineering*, vol. 18, no. 3, pp. 313–342, Jul. 2012.

[46] A. Ibrahim and T. Elghazaly, “Arabic text summarization using Rhetorical Structure Theory,” in *2012 8th International Conference on Informatics and Systems (INFOS)*, 2012, p. NLP–34–NLP–38.

[47] W. Black, S. Elkateb, H. Rodriguez, M. Alkhalifa, P. Vossen, A. Pease, and C. Fellbaum, “Introducing the Arabic wordnet project,” in *Proceedings of the 3rd International WordNet Conference (GWC-06)*, 2006, pp. 295–299.

[48] N. I. Al-Rajebah, H. S. Al-Khalifa, and A. M. S. Al-Salman, “Exploiting Arabic Wikipedia for automatic ontology generation: A proposed approach,” in *2011 International Conference on Semantic Technology and Information Retrieval (STAIR)*, 2011, pp. 70–76.

[49] W. Lahbib, I. Bounhas, B. Elayeb, F. Evrard, and Y. Slimani, “A Hybrid Approach for Arabic Semantic Relation Extraction,” in *The Twenty-Sixth International FLAIRS Conference*, 2013.

[50] N. G. Ali and N. Omar, “Arabic keyphrases extraction using a hybrid of statistical and machine learning methods,” in *Information Technology and Multimedia (ICIMU), 2014 International Conference on*, 2014, pp. 281–286.

[51] M. Nasri, L. Abouenour, A. Kabbaj, and K. Bouzoubaa, “Toward a semantic analyzer for Arabic language,” in *22nd IBIMA*, 2013.

[52] ALdhubayi, Luluh and Al-yahya, Maha, “Automated Arabic Antonym Extraction Using A Corpus Analysis Tool,” *Journal of Theoretical & Applied Information Technology*, vol. 70, no. 3, p. 422, Dec. 2014.

[53] Y. Belinkov, N. Habash, A. Kilgarriff, N. Ordan, R. Roth, and V. Suchomel, “arTenTen: a new, vast corpus for Arabic,” in *WACL’2 Second Workshop on Arabic Corpus Linguistics*, 2013.

[54] H. S. Al-Khalifa, M. M. Al-Yahya, A. Bahanshal, and I. Al-Odah, “SemQ: A proposed framework for representing semantic opposition in the Holy Quran using Semantic Web technologies,” in *2009 International Conference on the Current Trends in Information Technology (CTIT)*, Dubai, United Arab Emirates, 2009, pp. 1–4.

[55] M. Al-Yahya, H. Al_Khalifa, A. Bahanshal, I. Al-Odah, and N. Al-Helwah, “An Ontological Model for Representing Semantic Lexicons: An Application on Time Nouns in the Holy Quran,” *The Arabian Journal for Science and Engineering (AJSE)*, vol. 35, no. 2C, pp. 21–35, 2010.

[56] A. Kilgarriff, P. Rychly, P. Smrz, and D. Tugwell, “The Sketch Engine,” in *Proceedings of EURALEX*, 2004.