Journal of Theoretical and Applied Information Technology
31st December 2014. Vol.70 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AUTOMATED ARABIC ANTONYM EXTRACTION USING A CORPUS ANALYSIS TOOL

[1]LULUH ALDHUBAYI, [2]MAHA ALYAHYA
[1,2]Information Technology Department, College of Computer and Information Sciences,
King Saud University
E-mail: [1]laldubaie@ksu.edu.sa , [2]malyahya@ksu.edu.sa

ABSTRACT

The automatic extraction of semantic relations between words from textual corpora is an extremely challenging task. The increasing need for language resources supporting Natural language processing (NLP) applications has encouraged the development of automated methods for the extraction of semantic relations between words. The use of corpus statistical and similarity distribution methods can help in the task of semantic relation extraction between pairs of words. In this paper, we present 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 uses LogDice and pattern co-occurrence to classify the extracted pairs into antonyms. Results of evaluation show that our approach is able to extract the antonym relations with a precision of 76%.

Keywords: *Antonym Extraction, Sketch Engine, Arabic Lexicon, Semantic Relation, Arabic NLP*

1. INTRODUCTION

One of the most challenging tasks in natural language processing (NLP) is extracting semantic relations. The task involves finding instances of predefined relations between pairs of entities. Determining the sematic relations between two words would greatly improve the accuracy of NLP applications. NLP applications, which are influenced by semantic relations, include word sense disambiguation, sentiment analysis, and discourse processing. However, current Arabic lexical resources are insufficient for Arabic language processing tasks due to their limited coverage. For instance, Arabic WordNet (AWN) covers only general concepts and needs to be extended to encompass more specific domains [1], [2]. Extracting semantic relations from text using a manual approach is labour intensive, expensive, and time consuming. Some authors have argued that an automatic approach might be helpful in extracting semantic relations and enriching lexical resources [3], [4], [5], [6], [7], [8], but automatic approaches do not involve straightforward procedures. The use of corpus statistical and similarity distribution methods are useful for extracting the semantic relations between pairs of words.

This paper is organized as follows: Section 2 presents related work in the area of semantic relation extraction. Section 3 provides details of the approach we adopt for semantic relation extraction and the tools used. Section 4 presents the experiment and the results obtained. Section 5 presents the evaluation of the results obtained. Section 6 concludes with a discussion, including future research directions.

2. RELATED WORK

Despite the importance of the Arabic language, few corpus-based semantic relation extraction studies have focused on Arabic. This is due to the limited number of resources serving the language and the scarcity of well-annotated corpora. Both the language and lack of tools make it difficult to construct an Arabic lexical corpus. According to [9], the language is complex in three aspects: morphology, syntax and semantics. The Arabic language has a large number of grammar rules, which give rise to challenges in modeling the language in a formal structure. In addition, the absence of diacritics in the written text creates ambiguity. Moreover, automatically distinguishing between proper names, acronyms, and abbreviations is difficult because capitalization is not used in Arabic [10].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Al-Saif and Markert [11] produced an Arabic Discourse Treebank—the LADTB—a news corpus in which all discourse connectives are identified and annotated with the discourse relations they convey, as well as with the two arguments they relate. This is a valuable addition to the Arabic corpus linguistics, and can be used for training and testing automated methods for relation extraction; however, the relation extraction process is manually performed by human annotators using a specially designed tool.

With regards to relation extraction methods in the literature, they can be classified into supervised machine learning methods [12], semi-supervised pattern-based, and bootstrapping approaches. In supervised machine learning methods, the problem is described as a binary classification task, and a classifier is trained using a set of negative and positive examples of specific semantic relations. On the other hand, semi-supervised and bootstrapping approaches only require a small set of seed instances or a few hand-crafted patterns for specific relations to start the extraction process.

Lexicon-based corpora or semantic knowledge bases like WordNet [13] and Cyc [14] require a great deal of effort to create. A lexicon-based corpus can be encoded manually, which yields well-encoded outcomes, but this requires extensive human effort. The WordNet Synset construction described in [4] is a semi-automatic approach that uses initial seeds and bootstrapping methods. However, this approach still requires human judgment to review or validate the extracted Synsets before adding them to the lexicon-based corpus.

Pattern-based methods are the most common methods for relation extraction from text. 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 be hand-crafted or can be automatically generated. One of the earliest works on pattern based extraction methods is that of Hearst [15] on hyponyms. The method was based on using five manually identified lexico-syntactic patterns to extract the hyponym relation. Although this approach achieved good results, the process of manually hand-crafting patterns is time-consuming, and it is difficult to comprehend all possible patterns, especially when the domain or discourse of the text changes. There are two alternatives to identifying these patterns, using either bootstrapping and corpus tools, or by using machine learning algorithms to learn patterns from text and extract semantic relations. For many language-processing tasks including relation extraction, annotated (i.e., labeled) data is lacking and too expensive to create in large quantities, therefore bootstrapping techniques are desirable.

Espresso [16] is an example of using the bootstrapping approach for semantic relation extraction. Espresso uses an algorithm for extracting semantic relations using a bootstrapping algorithm to identify generic patterns automatically. Identified patterns are used to extract a range of semantic relations including meronymy and hyponymy. The bootstrapping starts with seed pairs and extracts all sentences in which these pairs co-occurred, and then generalizes the patterns. Espresso ranks patterns according to reliability measures that depend on precision and the number of antonyms discovered are given.

Similarly, the work in [17] presents an automatic pattern construction approach to extract 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. Based on the seed pairs, a corpus is analysed, patterns are constructed, and confidence values are computed for each pattern and used to extract new antonym/synonym pairs. Using seed terms to bootstrap a pattern search is also used in [5]; however, in their project, the patterns are generated manually. Another approach that uses seed pairs of antonyms to bootstrap a pattern is presented in [18]. That approach to extracting antonyms uses dependency patterns that are learned from a 450 million word treebank containing texts from Dutch newspapers. Using a set of seed pairs, patterns are identified and are used for finding new pairs of antonyms. A treebank is useful for generating dependency patterns expressing relations between words that occur far away from each other; this is more difficult for textual patterns.

A machine learning algorithm for pattern identification is presented in [19]. The algorithm classifies analogous (synonyms and antonyms and associations) word pairs, and can be used to solve multiple-choice analogy questions, synonym questions, 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.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A machine learning approach for hierarchical relation extraction is presented in [20]. The method is a SVM (Support Vector Machine) approach using features such as part of speech, entity subtype, entity class, entity role, semantic representation of sentence and WordNet synonym set.

For the Arabic language, similar approaches for relation extraction can be found in the literature. The work presented in [21] describes a method for extracting relations from text for the purpose of question-answering. The approach is based on Rhetorical Structure Theory (RST). The authors identified four rhetorical relations: cause, evidence, explanation, and purpose. Punctuation and cue phrases (patterns) are used to guide the relation extraction process. A similar approach using RST, but for the task of Arabic text summarization is presented in [22], where the authors identify cue phrases (patterns) for the rhetorical relation, and use these to generate summarized text.

Another similar study presented in [23] describes a method to detect casual relations that are expressed in Modern Standard Arabic. The approach is based on the development of patterns based on a set of syntactic features acquired by analyzing an Arabic corpus. Cue words and Part-of-Speech tags were used for extracting casual relations patterns.

Classified as a supervised machine learning method, rule mining from Arabic language text can be used for relations extraction as described in [24]. An Arabic language corpus is used to mine lexical (Part of Speech (PoS)), semantic (word category), and numerical (number of words) features. Features are learned from annotated samples and rules are generated for extracting sematic relations.

The work presented in [25] is an attempt to improve the semantic relations already existing in Arabic WordNet (AWN) [1]. The authors use a linguistic method based on morpho-lexcial patterns to extract semantic relations. Arabic Wikipedia articles are used, as they have a structure that can be used for pattern definition and semantic relation extraction. The method consists of two phases: morpho-lexical pattern recognition and semantic relation enrichment. In the first phase pairs of Synsets that are linked by semantic relations are extracted from AWN [1]. These extracted pairs are used to select Wikipedia articles, once selected sentences are tagged morphologically. Next, the morpho-lexical pattern is identified and used for extracting new relations.

Arabic Wikipedia has also been used to build ontologies and extract relations; for example, the work presented in [26] describes a methodology for identifying ontology instances. The Arabic version of Wikipedia is used as a knowledge source from which concepts and semantic relations are extracted. The algorithm is restricted to extract semantic relations between the article and the features it contains using the Wikipedia "Infoboxes" only.

Similar to the work described in [26], but using an Lexical Markup Framework (LMF) standardized dictionary instead of Wikipedia entries for ontology enrichment, the work presented in [27] uses a rules-based system which relies on lexico-syntactic patters for ontology elements extraction. The approach is based on manual analysis of the LMF dictionary and the definition of a set of rules to allow for the identification of ontology entities. These rules are then used on the LMF dictionary to extract concepts, relations, and triples for ontology enrichment.

Despite the existence of work in the area of semantic relation extraction, the coverage is still rather limited, and there is still need for enriching this important area of research. Our method for semantic relation extraction from Arabic language corpora is a pattern-based bootstrapping approach using the corpus analysis tool Sketch Engine. The work we describe in this paper focuses on antonym relation extraction.

3. METHODOLOGY

There are many lexical-semantic relations in a language, such as antonymy, synonymy, and hyponymy. This study focuses only on antonym relationships between nouns. An antonym is defined as the contradictory or opposite meaning between two words or lexical elements. The current study focuses on the general definition of an antonym and does not cover semantic dimensions of the relation. For instance, two Arabic words, such as (hot, cold), can have a distinct antonymic relation. Alternatively, the pairs (cold, chilly) can have the same meaning with different degrees of contradiction but cannot be considered as synonyms and used in a similar context. Thus, an automatic thesaurus would consider this type of semantic relation as a near-synonym without defining the depth of the relation. To tackle the problem of automatic definition of antonyms, this study uses

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the antonym patterns co-occurrence hypothesis and similarity distribution measurements to estimate the probability that an Arabic pair, (x, y), has an antonymic relationship.

Our method is classified as a pattern-based bootstrapping approach using the corpus analysis tool Sketch Engine [28] and a set of seed antonym pairs from the SemTree ontology [29], an ontology based lexicon for Arabic semantic relations, to obtain the most frequent patterns that co-occur with the seeds. These patterns are then used to extract new antonym pairs other than the initial seeds. The frequency of pattern occurrence in the corpus, in addition to the number of antonym pairs co-occurring with the pattern, provides a measure of reliability of the pattern to extract new antonym pairs from the corpus. A good pattern is an antonym pattern that is able to extract new antonym pairs. Running a good antonym pattern in a large corpus using the corpus query language (CQL) of the Sketch Engine is capable of extracting many new antonym pairs. For evaluating the quality of the newly extracted antonyms from the corpus, an Arabic native speaker reviewed the extracted antonym pairs before applying the bootstrapping method to the extracted antonymous pairs. Starting from initial seeds, bootstrapping involves many stages to obtain semantic patterns. Using the learned good patterns, new seeds are extracted. This process is repeated over many iterations to obtain new seeds.

3.1 Sketch Engine Tool

A robust corpus analysis tool applicable for an Arabic language corpus was essential to conduct the experiment. According to [30], the recently developed Sketch Engine lexicon tool provides corpora in different languages, including Arabic. In addition to featuring a robust and advanced query system, Sketch Engine is an efficient web-based tool for developing vast corpora. Sketch Engine offers various features, such as word sketch, collocations and a thesaurus [31]. It is used to build a detailed statistical profile of any word in a corpus, which enables lexicographers to understand the words or collocations, their behaviors, usages, as well as indicating the connotations they may carry. Sketch Engine is considered a good corpus analysis tool and has been shown to be reliable in conducting linguistics research, such as finding collocations and language patterns using statistical measurements [32],[33].

To date, Sketch Engine offers five different Arabic corpora, as shown in Table 1. The corpora have various sizes, text genera, and annotations. The corpus selected for the experiment should satisfy two main criteria: it should be comprehensive and have multi-text genera. A corpus with a large number of tokens and vocabulary would be expected to feature comprehensive language and semantic word diversity. Using such a corpus in this experiment is essential to yield good antonymous patterns and pairs.

Table 1. Description of Arabic corpora available on Sketch Engine.

Corpus	Tokens	Description
Arabic Web Corpus	174,239,600	170-million-word Arabic Web Corpus, Arabic Wikipedia, Corpus of Contemporary Arabic, and specialized Arabic corpora for news, computer science, and legal texts
KSUCCA King Saud University Corpus of Classical Arabic	59,693,146	Classical Arabic text
arTenTen12	6,637,387,738	Modern Standard Arabic
OPUS2 Arabic	406,527,277	OPUS: Open source Parallel Corpus in many languages. (Arabic and 39 languages).
Quran annotated corpus [Unvoweled Arabic] [Voweled Arabic] [Unvoweled Latin] [Voweled Latin]	128,243	Four Quranic corpora in different four scripts.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.2 Arabic Corpora

The vast Arabic corpus arTenTen [34] is the largest Arabic corpus available on Sketch Engine. arTenTen is made up of 5.8 billion tokens and 177,011,938 sentences. The 10-billion-word corpus is tagged only with sentence, paragraph and document tags. However, according to [34], syntax tags, in common with PoS tags, are under-represented, and only 150 million tokens are tagged with 30 features (e.g. PoS tags and gender). The corpus was gathered from the web using the SpiderLing tool; thus, the corpus is a combination of contemporary and classical text types.

The King Saud University Corpus of Classical Arabic (KSUCCA) [35] is another corpus available on Sketch Engine that is tagged with many linguistic features. Moreover, the text type is more classical and clustered into many categories, such as religion, linguistics, and science. However, the corpus contains only 50 million tokens. In addition, selecting antonym patterns is restricted to those with the highest frequency, which indicates the pattern's popularity in the language. A larger corpus, regardless of the genera and annotation availability, would be more useful in extracting new antonyms. Consequently, we selected arTenTen, which is the largest Arabic corpus available in Sketch Engine.

3.3 Sketch Engine Association Scores

Sketch Engine uses two types of statistics. The first type is based on grammatical relations. Grammatical relations, such as those that appear in sketch grammar (e.g. subject, object, adjective-of and construct-state), in the corpus are used to measure the association score of two words based on the triples $\|w_1, R, w_2\|$. The statistics available in Sketch Engine include the mutual information score, association score, dice and LogDice. LogDice [36] is a popular measurement of semantic similarity between collocation candidates. The score is independent of the size of the corpus and considered stable. Thus, this score is able to confirm whether a relation exists between two words,

$$LogDice = 14 + Log_2 \frac{2frequency\ XY}{frequency\ X + frequency\ Y}.$$

3.4 Procedure

Pattern learning: Our method uses a corpus-based distribution method to collect the most frequent antonym pairs in the Arabic corpus arTenTen. These can be used to retrieve the most frequent patterns surrounding the pairs [6]. This approach helps to reduce odd patterns by filtering the frequency and number of patterns in co-occurring antonym pairs. The final result with this approach is a set of learned antonym patterns associated with frequencies and LogDice scores.

Antonym pairs extraction: This process uses learned patterns to extract new antonym pairs automatically. It uses the CQL in Sketch Engine to run the selected patterns and find the most frequent pairs in arTenTen. CQL provides a robust language to restrict the retrieved pairs by excluding numbers, identical pairs (x, x) and prepositions. We also used PoS tags to extract noun-noun pairs.

The algorithm is summarized in the following steps:

1. Starting with a set of 57 seed antonym pairs (noun-noun) and running the CQL on arTenTen corpus, we obtained the ten most frequent pairs (Table 2). We argue that obtaining the most frequent pairs can identify good antonym patterns. We define a good pattern as "*a frequent pattern that co-occurs with many different antonym-pair initial seeds*".
2. The LogDice association scores of the initial seeds are computed to define a threshold of antonym pair co-occurrences. Table 2 shows the ten most frequent pairs in the corpus and those with LogDice scores above 7.0.
3. Pattern extraction is executed by defining the CQL expression as,

   [word] "1st Antonym" [word]{1,3} "2nd Antonym" within <\s>

   The expression shows the antonym pairs with a distance of (1, 2 or 3) words between the pairs in the sentence boundaries. This distance was selected after many experiments. We concluded that a distance with less than one word or more than three words would not yield good antonym patterns. In addition, placing one word before the first antonym helped to acquire good patterns.
4. Pattern learning is implemented by selecting only good patterns. As pointed out, good patterns are frequent patterns that co-occurred with many different antonym pairs (the initial seeds). The pattern learning step aims to avoid

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

idiomatic antonym patterns that have a high frequency but only co-occur with one antonym pair, such as (ظهر الحق و زهق الباطل). After many experiments, we selected a minimum pattern frequency of 100 and a minimum association of two antonym pairs. The learned pattern must satisfy these two requirements. For instance, the pattern (from X to Y) has a frequency of 21,040 and co-occurs with seven antonym pairs (شرق ، جنوب-شمال-كثير ،شر-خير ،خاص-عام ،موت-حياة ،غرب-الباطل-الحق ،قليل). However, another candidate (A matter of X or Y, قضية س أو ص) has a frequency above 100 but only co-occurs with one antonym pairs (life-death, موت-حياة). In the latter case, the number of antonyms co-occurring might change in later iterations.

5. We extracted new candidate pairs by querying the arTenTen corpus for the learned antonym patterns using the CQL in Sketch Engine. To find a word (noun) occurring in the pattern, the CQL replaces the pairs with empty wildcards, as shown below,

"from" 1:[word=""] "to" 2:[word=""] & 1.word != 2.word

Furthermore, more selective results such as only noun-pairs, can be retrieved using the tagged arTenTen corpus, as follows,

"from" 1:[tag="noun"] "to" 2:[tag="noun"] & 1.word != 2.word

However, not all tokens are tagged with PoS. In fact, only 50 million tokens are tagged in arTenTen on Sketch Engine. The first CQL expression retrieves pairs without specifying the PoS tag, resulting in the generation of thousands of odd pairs. In contrast, the latter expression retrieves only pairs in (noun, noun) form and, thus, generates fewer odd pairs. An odd pair can be defined as a pair that has number, prepositions or identical pairs (x, x).

Running the CQL in the arTenTen corpus retrieved hundreds of pairs. To reduce the scale and limit the processing time, we used the multi-level frequency distribution tool in Sketch Engine to select only the five most frequent pairs in each pattern.

6. As we reduced the scale to only five learned patterns, the search results generated 25 pairs. We then implemented the *Pairs Classification Stage* described below.

7. Finally, we applied the bootstrapping method to extend the extracted pairs and patterns. This involved using the extracted antonym pairs in each iteration as new seeds and repeating the process.

Table 2. Antonym seeds (i.e., the ten most frequent antonyms in the arTenTen corpus).

Rank	Seeds pairs X-Y	Seeds pairs X-Y (English)	Frequency (x,y)	LogDice
1	عام – خاص	Private Public	16,782	8.124
2	قليل – كثير	Many Few	13,244	8.386
3	فوق – تحت	Up Down	13,765	7.426
4	حياة – موت	Death Life	8,625	8.201
5	جنوب – شمال	North South	4,854	8.267
6	ارتفاع – انخفاض	Decrease Increase	2,564	7.537
7	باطل – حق	Truth Delusion	2,534	10.059
8	شر – خير	Good Bad	2,521	8.521
9	غرب – شرق	East West	2,337	9.007
10	علم – جهل	Illiteracy Knowledge	2,266	7.589

Pairs classification stage:

The classification process proceeded through three stages as follows:

Stage *1. Antonym classification using the LogDice association score:* To classify the 25 discovered pairs as antonyms or non-antonyms, we used the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

LogDice association score and set up a threshold. We used Sketch Engine to compute the LogDice score using its collocation tool. We defined a threshold of 7.0 as a minimum LogDice score. Thus, any pairs with a LogDice value below the threshold of 7.0 were discarded. We consider that a candidate pair with a LogDice score above 7.0 might be an antonym pair. The threshold was selected after pilot testing and comparing the scores of the initial noun seeds of both antonyms and synonyms. We concluded that a threshold value of 7.0 is reasonable because each of the 25 pairs exceeded the proposed threshold. However, some synonym pairs have a LogDice value of 7.0 or greater. This is expected, as the LogDice value computes the probability of candidate pairs X and Y appearing together in a similar context (collocation).

Stage 2. Antonym patterns co-occurrence hypothesis: Using only LogDice scores is not sufficient to classify antonym pairs. The LogDice score is used as a filter to reduce the probability of non-antonym pairs. To increase the precision of the classification, we added another score: the co-occurrence of antonym patterns. To do so, we used our initial five learned antonym patterns, replacing the wildcards with candidate pairs (X, Y). If two or more antonym patterns retrieved the candidate (X, Y) successfully, then this was considered a good indication of an antonym pair. For example, the candidate (البداية و النهاية) pair has a LogDice score of 8.6 and co-occurred with the five antonym patterns: (من البداية إلى النهاية، من البداية أو النهاية، في البداية أو النهاية، بين البداية و النهاية، لا البداية ولا النهاية). In contrast, a synonym pair, such as (اجتماع و اتفاق), has a LogDice score greater than 7.0, but never occurs with the five antonym patterns. Consequently, we used the combined scores of the LogDice and antonym patterns co-occurrence to classify the extracted pairs. We compiled a list of 25 extracted pairs with LogDice scores and co-occurrence of antonym patterns. A native Arabic speaker was asked to judge the accuracy of the classification. The judge considered five of 25 as co-hyponym pairs and the rest as antonym pairs.

4 EXPERIMENTAL RESULTS

4.1 Extracted Patterns

After running the initial ten noun-noun antonym seeds in the arTenTen corpus, we extracted 359 patterns. The frequency of patterns in the corpus ranged from 5 to 4763 occurrences. To maintain high performance, we used those patterns with frequency above 100 (n=137 patterns). We found that patterns with frequency less than 100 tended to be idiomatic expressions and were not useful in extracting new antonym pairs. Moreover, patterns that co-occurred with different antonym seeds tended to extract more new pairs. Thus, we used only a set of five good patterns with frequencies greater than 100 that co-occurred with at least three initial antonyms pairs (Table 3).

Table 3. Learned antonym patterns.

#	Patterns		Frequency in arTenTen	Antonyms pairs (seeds)
1	من س الى ص	From X to Y	21,040	7 pairs
2	من س أو ص	From X or Y	3,351	3 pairs
3	بين س و ص	Between X and Y	2,720	5 pairs
4	في س أو ص	In X or Y	1,534	3 pairs
5	لا س ولا ص	Neither X nor Y	480	3 pairs

4.2 Extracted Antonym Pairs

We ran the first pattern 'من س إلى ص - from X to Y and extracted a test set of 3,300 pairs from the arTenTen corpus. Sketch Engine provides an ordered list of the extracted pairs, together with their frequencies. The ordered list makes it easier to sort the most frequent pairs that co-occur with each pattern. To reduce the processing time, we applied our algorithm to the top 25 extracted pairs and evaluated their semantic relations. Table 4 shows the 25 extracted pairs manually tagged by the native Arabic speaker. The algorithm classification showed a precision of 76%. However, some pairs needed to be filtered, such as common words (أخرى another). Numbers and repeated words were filtered using the tagged corpus and CQL expressions, as mentioned in section 3. Table 5 shows that applying the antonymous patterns co-occurrence hypothesis to the 25 extracted pairs yielded reasonable results.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Most of the 25 extracted pairs co-occurred at least twice with the antonymous patterns. Therefore, the application of the antonymous patterns co-occurrence hypothesis can aid in filtering synonym pairs from the list. However, as the context of many co-hyponyms might be similar to that of antonymous pairs, this hypothesis alone is not sufficient to extract pairs.

Table 4. Association scores and human classification of the 25 extracted pairs.

#	Pairs X-Y (Arabic)	Pairs X-Y (English)	Association Scores			Human Classification
			T-score	MI	LogDice	
1	المحيط-الخليج	Gulf Ocean	140.082	9.806	9.523	Co-Hyponym
2	الظلمات - النور	Lightness Blackness	113.157	12.837	10.04	Antonym
3	البحر - النهر	River Sea	92.531	9.158	8.365	Antonym
4	العدم- الوجود	Presence Absence	86.565	10.961	9.026	Antonym
5	السماء -الأرض	Earth Sky	248.938	8.471	9.637	Antonym
6	الألف - الياء	A Z	73.984	12.207	9.957	Co-Hyponym
7	اليمين - اليسار	Left Right	157.68	12	10.937	Antonym
8	المهد - اللحد	Death Birth	59.287	16.689	11.68	Antonym
9	الأعلى- الأسفل	Lower Upper	91.229	11.149	8.981	Antonym
10	البداية- النهاية	End Begin	111.663	8.216	8.602	Antonym
11	قريب-بعيد	Far Close	228.458	11	11.002	Antonym
12	قول-فعل	Doing Saying	100.339	5.958	7.336	Antonym
13	ذكر-أنثى	Female Male	110.287	9.584	7.967	Antonym
14	ليل-نهار	Day Night	231.721	14.326	12.695	Antonym
15	الداخل-الخارج	Outside Inside	221.287	10.333	10.618	Antonym
16	الكتاب-السنة	Sunna Quran	131.282	5.533	7.516	Co-Hyponym
17	ذهب-فضة	Silver Gold	88.609	11.498	8.919	Co-Hyponym
18	صيام-صدقة	Charity Fasting	34.565	10.061	7.825	Co-Hyponym
19	الرجال-النساء	Women Men	288.626	9.339	10.497	Antonym
20	الرجل-المرأة	Woman Man	350.652	8.009	10.144	Antonym
21	الصفا-المروة	Marwa Safa	61.602	15.407	10.937	Antonym
22	السنة-الشيعة	Shiites Sunni	179.667	8.105	9.047	Co-Hyponym
23	الشك اليقين	Trust Doubt	76.292	10.425	9.141	Antonym
24	مؤيد-معارض	Supporter Opponents	48.461	11.625	9.031	Antonym
25	الماضي-الحاضر	Past Present	159.751	8.373	8.517	Antonym

Journal of Theoretical and Applied Information Technology
31st December 2014. Vol.70 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 5. Co-occurrences of antonym patterns for the 25 extracted pairs.

#	Pairs X-Y (Arabic)	Pairs X-Y (English)	Occurrence of Pairs using Antonyms Patterns				
			من س الى ص From X to Y	من س أو ص From X or Y	بين س و ص Between X and Y	في س أو ص in x or In	لا س ولا ص Niether X nor Y
1	المحيط-الخليج	Gulf Ocean	Yes	Yes	Yes	Yes	No
2	الظلمات - النور	Lightness Blackness	Yes	No	Yes	No	Yes
3	البحر - النهر	River Sea	Yes	Yes	Yes	Yes	No
4	العدم- الوجود	Presence Absence	Yes	No	Yes	No	No
5	السماء -الأرض	Earth Sky	Yes	Yes	Yes	Yes	No
6	الألف - الياء	A Z	Yes	No	Yes	No	No
7	اليمين - اليسار	Left Right	Yes	Yes	Yes	Yes	Yes
8	المهد - اللحد	Death Birth	Yes	No	Yes	No	No
9	الأعلى- الأسفل	Lower Upper	Yes	Yes	Yes	Yes	No
10	البداية- النهاية	End Begin	Yes	Yes	Yes	Yes	No
11	قريب-بعيد	Far Close	No	Yes	Yes	Yes	Yes
12	قول-فعل	Doing Saying	Yes	Yes	Yes	Yes	Yes
13	ذكر-أنثى	Female Male	No	Yes	Yes	No	Yes
14	ليل-نهار	Day Night	Yes	Yes	Yes	Yes	Yes
15	الداخل-الخارج	Outside Inside	Yes	Yes	Yes	Yes	Yes
16	الكتاب-السنة	Sunna Quran	Yes	Yes	Yes	Yes	Yes
17	ذهب-فضة	Silver Gold	No	Yes	No	Yes	Yes
18	صيام-صدقة	Charity Fasting	No	Yes	No	No	Yes
19	الرجال-النساء	Women Men	Yes	Yes	Yes	Yes	Yes
20	الرجل-المرأة	Woman Man	Yes	Yes	Yes	Yes	Yes
21	الصفا-المروة	Marwa Safa	Yes	Yes	Yes	Yes	No
22	السنة-الشيعة	Shiites Sunni	Yes	Yes	Yes	No	Yes
23	الشك اليقين	Trust Doubt	Yes	Yes	Yes	Yes	Yes
24	مؤيد-معارض	Supporter Opponents	Yes	Yes	Yes	No	Yes
25	الماضي-الحاضر	Past Present	Yes	Yes	Yes	Yes	Yes

5 EVALUATION

Sketch Engine offers a diversity of Arabic corpus tools and association scores. Using Sketch Engine in this study limits the time needed to process the data and reduces human effort in developing a stand-alone application to conduct the experiment. However, Sketch Engine is unable to analyze semantic relationships of specific pairs. The tool provides a variety of similarity distribution measures, with many sub-tools such as collocations, word sketch, Sketch-difference and a thesaurus. These tools are not useful for measuring semantic contrast relationships, but are ideal for finding near-synonym sets.

LogDice association scores appears to offer promising results in classifying extracted pairs by assigning a higher score to antonym pairs. The LogDice score measures the likelihood of pair X and Y sharing a similar context. It confirms that a candidate pair appears simultaneously in a specific context and thus might share a semantic relation. However, the LogDice score cannot confirm the relation type. The relation might be antonymous, hyponymous or synonymous. Therefore, LogDice cannot be used to confirm the antonym co-occurrence hypothesis.

To extract greater numbers of antonym pairs, we used a pattern-based approach, using patterns to

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

extract pairs (x, y) with a semantic relationship. This approach yielded promising results in the first iteration. However, it did not improve the precision of the results and was not consistent. Some patterns were too general, resulting in the extraction of too many odd pairs, such as the pattern مســألة س أو ص . Others were too idiomatic, resulting in the retrieval of only one pair, such as the pattern ظهر س و ز هق ص Synonyms, antonyms, co-hyponyms and hyponym relationships occur as pairs in similar contexts. Thus, it is difficult to distinguish the type of relationship automatically. Results of this experiment show that the proposed process classifies antonyms and co-hyponyms pairs as antonyms.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we presented the first set of experiments for an Arabic antonymous pair-finding algorithm. The main goal of this research was to use the Sketch Engine query tool to detect and analyze Arabic antonym pairs and the patterns associated within Arabic text. Sketch Engine offers different Arabic corpora of different sizes and domains. The Arabic corpus ArTenTen was used in this research. In addition, Sketch Engine features an advanced CQL, which allows the user to run a complicated query using the Regular Expression language.

The present study used a pattern-based approach to extract Arabic antonymous pairs and patterns in the Arabic corpus. However, pattern-based techniques tend to extract more noisy pairs, potentially increasing the coarseness of the classification. The present study showed that Sketch Engine alone is not sufficiently reliable to find semantic relationships, especially antonymous relations. Merging Sketch Engine's capabilities with a semantic annotation tool that analyzes antonymous relations might facilitate tagging and evaluation of semantic annotations. In future work, we plan on using deep-learning methods to create sets of antonymous noun pairs that can then be analyzed in more detail with a pattern-based technique. We are additionally looking at optimizing the classification results using prior probabilities for pairs and patterns.

REFERENCES

[1] W. Black, S. Elkateb, and P. Vossen, "Introducing the Arabic WordNet Project," *IN PROCEEDINGS OF THE THIRD INTERNATIONAL WORDNET CONFERENCE (GWC-06*, 2006.

[2] H. Rodríguez, D. Farwell, J. Farreres, M. Bertran, M. Alkhalifa, M. A. Martí, W. Black, S. Elkateb, J. Kirk, A. Pease, and others, "Arabic wordnet: Current state and future extensions," in *Proceedings of The Fourth Global WordNet Conference, Szeged, Hungary*, 2008.

[3] E. Riloff and J. Shepherd, "A Corpus-Based Approach for Building Semantic Lexicons," in *In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, 1997, pp. 117–124.

[4] Ellen M. Riloff and Jessica Shepherd, "Corpus-based bootstrapping algorithm for semi-automated semantic lexicon construction," 1999.

[5] S. Mohammad, B. Dorr, and G. Hirst, "Computing word-pair antonymy," presented at the Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA, 2008.

[6] A. Lobanova, T. Van der Kleij, and J. Spenader, "Defining antonymy: A corpus-based study of opposites by lexico-syntactic patterns," *International Journal of Lexicography*, vol. 23, no. 1, pp. 19–53, 2010.

[7] G. Schropp, E. Lefever, and V. Hoste, "A Combined Pattern-based and Distributional Approach for Automatic Hypernym Detection in Dutch.," in *RANLP*, 2013, pp. 593–600.

[8] S. Elkateb, W. Black, P. Vossen, D. Farwell, A. Pease, and C. Fellbaum, "The Challenge of Arabic for NLP/MT Arabic WordNet and the Challenges of Arabic," in *Proceedings of Arabic NLP/MT Conference*, London, UK, 2009.

[9] H. M. Harmain, H. El Khatib, and A. Lakas, "ARABIC TEXT MINING," 2004.

[10] B. Hammo, H. Abu-Salem, and S. Lytinen, "QARAB: A question answering system to support the Arabic language," in *Proceedings of the ACL-02 workshop on Computational approaches to semitic languages*, 2002, pp. 1–11.

[11] Al-saif Amal and K. Markert, *The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic*. .

[12] N. Kambhatla, "Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations," in *Proceedings of the ACL 2004 on Interactive*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

poster and demonstration sessions, 2004, p. 22.

[13] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, "Introduction to WordNet: an On-line Lexical Database*," *Int J Lexicography*, vol. 3, no. 4, pp. 235–244, Jan. 1990.

[14] D. B. Lenat, M. Prakash, and M. Shepherd, "CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks," *AI magazine*, vol. 6, no. 4, p. 65, 1985.

[15] M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora," in *Proceedings of the 14th conference on Computational linguistics - Volume 2*, Stroudsburg, PA, USA, 1992, pp. 539–545.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

[28] A. Kilgarriff, P. Rychly, P. Smrz, and D. Tugwell, "Itri-04-08 the sketch engine," *Information Technology*, vol. 105, p. 116, 2004.

[29] A. Al-Zahrani, M. Al-Dalbahie, M. Al-Shaman, N. Al-Otaiby, and W. Al-Sultan, "SemTree: Analyzing Arabic Language Text for Semantic Relations." Unpublished Thesis, IT Department, KSU, 2012.

[30] A. Kilgarriff, S. Reddy, J. Pomikálek, and P. V. S. Avinesh, "A Corpus Factory for Many Languages.," in *LREC*, 2010.

[31] A. Kilgarriff and I. Kosem, "Corpus tools for lexicographers," *Granger, S. and M. Paquot (Eds.)*, vol. 2012, pp. 31–55, 2012.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[32] C.-R. Huang, A. Kilgarriff, Y. Wu, C.-M. Chiu, S. Smith, P. Rychly, M.-H. Bai, and K.-J. Chen, "Chinese Sketch Engine and the extraction of grammatical collocations," in *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, 2005, pp. 48–55.

[33] M. Pearce, "Investigating the collocational behaviour of man and woman in the BNC using Sketch Engine 1," *Corpora*, vol. 3, no. 1, pp. 1–29, 2008.

[34] 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.

[35] 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.

[36] P. Rychlỳ, "A lexicographer-friendly association score," *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN*, pp. 6–9, 2008.