2016 IEEE 28th International Conference on Tools with Artificial Intelligence

Semi-Supervised Pattern-Based Algorithm for Arabic Relation Extraction

Injy Sarhan
College of Engineering and Technology
Arab Academy for Science and Technology
Alexandria 1029, Egypt
injy.sarhan@aast.edu

Yasser El-Sonbaty
College of Computing and Information Technology
Arab Academy for Science and Technology
Alexandria 1029, Egypt
yasser@aast.edu

Mohamed Abou El-Nasr
College of Engineering and Technology
Arab Academy for Science and Technology
Alexandria 1029, Egypt
mnasr@aast.edu

***Abstract*—While several relation extraction algorithms have been developed in the past decade, mainly in the English language, only few researchers target the Arabic language owing to its complexity and rich morphology. This paper proposes a semi-supervised pattern-based bootstrapping technique to extract Arabic semantic relation that lies between entities. In order to enhance the performance to suit the morphologically rich Arabic language, stemming, semantic expansion using synonyms, and an automatic scoring technique to measure the reliability of the generated patterns and extracted relations were used. To further improve performance, a dependency parser was then used to omit negative relations. The proposed system was tested by applying it to two corpora, which differ in both size and genre, scoring a highest F-measure of 75.06%. Furthermore, the effect of adding stemming and synonyms was also experimentally tested. The results show that this bootstrapping methodology achieves higher performance than existing state-of-the-art methods, and can be expanded to include more relations for use in various NLP tasks**

Keywords—Natural Language Processing; Relation Extraction; Arabic; Patterns

I. INTRODUCTION

Different Natural Language Processing (NLP) tasks have been attracting massive attention, being the intersection between linguistics and computational science [1]. One of the most essential tasks is information extraction. Its prerequisites include recognition of named entities and relation extraction, thus motivating NLP researchers to work on extracting semantic relations.

Several other NLP applications like question answering and summarization necessitate relation extraction in order to understand the semantic relation that lies between named entities, attracting more attention to the relation extraction task. Similar to other NLP tasks, preprocessing is a necessary step in relation extraction. This can possibly include tokenization, clause splitter, Part of Speech (POS) tagging, and Named Entity Recognition (NER).

The widespread of the Arabic language in about 24 Arabic-speaking countries, and the presence of a vast amount of Arabic text on the web, has brought working on Arabic texts into focus. While NLP on the Arabic language has become quite necessary [2], other tasks such as extracting Arabic text from images have also been explored as in [3]. The complexity of the relation extraction task on the Arabic language, as mentioned in [4-6], was the primary motivation to extend the current state-of-the-art relation extraction systems to conform to such a morphologically powerful language. Various relations can be extracted from text, including Is-a relation, Synonym-of, Antonym-of and Has-a. In this approach, the focus is on entity relations, for instance, Person-Organization, Person-Location and Location-Organization.

In order to achieve this goal, the following steps were made. First, as a preprocessing step, a stemmer was applied to the corpus in order to extract the word root. Second, semantic expansion using synonyms for verbs was done. Third, an automatic scoring technique was employed in order to measure the reliabilities of the generated patterns and the extracted relations. To the best of our knowledge, this is novel to relation extraction in the Arabic language. Finally, a dependency parser was employed to omit negative relations, thereby further enhancing the performance of the proposed system.

The rest of the paper is structured as follows: Section 2 presents related work in the field of relation extraction, Section 3 briefly describes the Arabic language. The proposed pattern-based algorithm is explained in Section 4, with results and evaluation discussed in Section 5. Finally, Section 6 concludes the paper.

II. RELATED WORK

In this section we focus on previous works done on Arabic relation extraction in the literature. Most research falls under one of the following categories: supervised, semi-supervised or unsupervised. The supervised approaches require large amounts of annotated data that is often expensive to gather. Furthermore, supervised techniques are sensitive to errors since they mainly depend on parse trees [6]. In unsupervised techniques, the learner is provided unlabeled examples, thus the evaluation is challenging at a large scale. However, semi-supervised approach has been one of the most popular approaches in computational linguistics recently. The main reason behind its increasing

2375-0197/16 $31.00 © 2016 IEEE
DOI 10.1109/ICTAI.2016.33

popularity, its ability to overcome the problems associated with both unsupervised and supervised approach.

In supervised relation extraction, the task is presented as a classification task. The approaches discussed below are all binary relations. RelANE is a relation extraction system proposed by [6] to discover relations between Arabic named entities. Due to the high occurrence of certain Named Entities (NEs), the relation of interest is the relation that lies between any pair of the following four NEs, Person (PERS), Location (LOC), organization (ORG) and Date (DATE). Six different classification techniques were used, PART, Decision Tree, Adaboost, Naïve Bayes and Support Vector Machine (SVM) - which are all available on WEKA [7]- and MaxEnt [8]. Adaboost scored the highest performance, followed by SVM.

In the following year, Mohanaed Falih and Nazila Omar [9], proposed an Arabic grammatical relation extraction based on machine learning classification. The main objective of this approach is to label each Arabic word with the correct grammatical relation: subject, object or verb. This is achieved using either of the three classifications: Support Vector Machines (SVM), K-nearest neighbor (KNN) or a combination of both. A special training Arabic corpus was created by Falih and Nazila for their system, it consisted of 80 sentences in which each sentence was manually annotated with its appropriate grammatical relation; subject, object or predicate. The drawback of this approach is that the manually assembled test corpus used consisted of only 80 Arabic sentences which might lead to an unfair evaluation of the whole system.

Boujelben et al. [4], proposed a supervised model that automatically extracts rules for relation extraction using genetic algorithms to discover and generate rules. This technique was used to extract the following relations: PERS-LOC, PERS-ORG, PERS- PERS, ORG-LOC and LOC-LOC. Training data was mainly collected from Arabic journals that include almost 2000 named entities. ANERCorp [10] corpus was used for evaluating and testing the algorithm. In this approach rules are discovered using genetic algorithm using Michigan approach [11], where each rule is illustrated by a chromosome. Before the crossover and mutation process, the filtering module is applied. Using parent's rules crossover children are produced. Mutation probability is calculated for each rule created, rules are discarded if they are below a certain threshold value, the rest are used according to their confidence score.

Pattern based methods fall under either semi-supervised or unsupervised categories, Hearst algorithm [12], was the first to introduce pattern-based bootstrapping approach which inspired succeeding pattern based algorithms. Hearst extracted relations using a set of predefined patterns and used a bootstrapping approach to generate more patterns. Three relations were extracted using Hearst algorithm: Hyponym-Hypernym (cutlery, spoon), IS-A (giraffe, mammal), and Kind-of (Egypt, African Country). The main drawback of Hearst algorithm is that large human intervention was needed in order to create patterns from real examples. Al-zamil and Al-radaideh [5] generated a system for automatic extraction of ontological relations from Arabic text. The main purpose of this system is to generate patterns and extract semantic features of Arabic text in order to extract ontological relations. This algorithm is an enhancement of Hearst algorithm. The improvements that this approach made on Hearst algorithm include pattern filtering, enhancing the quality of the pattern, and pattern assessment.

Ontological enrichment continued in the work by Maha Al-Yahya et al. [13], "Badea" system in 2014, a pattern-based bootstrapping approach to extract semantic relation using a seed ontology. The primary objective of this approach is to extract antonyms pair from text. A small seed of antonyms pairs is used in order to extract patterns from an Arabic text. The initial step is pattern identification, in which the small set of seeds is used on the first corpus (corpus A) to extract patterns. Afterwards, those extracted patterns are transformed into regular expressions. In the following phase, a new corpus (corpus B) is used. A pattern recognition algorithm is applied to corpus B to extract new antonym pairs using the regular expressions that were previously extracted from corpus A. The pairs extracted are manually checked to calculate the precision and pattern score, in order to evaluate each pattern. Corpus A, Arabic corpus ArTenTen [14] was used which contains over 170 million sentences. For corpus B, the King Saudi University Corpus of Classical Arabic (KSUCCA) [15] was used. On one hand, due to the incorrect pattern scoring technique, the precision score of this approach is 0.80%, which is considered extremely low, which is the main drawback of "Badea" system. On the other hand, a large number of patterns were extracted however, some improvements should be made concerning the pattern score in order to improve the precision score.

Moreover, Arabic statistical relation extraction approaches were also carried out. Abdullah Alyotak [16] proposed a machine-learning-based relation extraction algorithm, which resulted in 85% accuracy by testing on 10-folds for ACE 2005 dataset [17]. Weim Lahbib et al. [18] used vocalized texts to minimize ambiguities and proposed a hybrid approach which extracts noun phrases at the first stage, then transforms them into semantic relations. A survey on Arabic relation extraction can be found in [19].

In addition to Hearst algorithm, some notable work in English relation extraction using pattern based approach was carried out. In 1998, DIPRE (Dual Iterative Pattern Relation Expansion) [20] technique was used for relation extraction to extract author-book pairs. Initially, a small set of seeds are used for pattern generation process, which are later used to extract new patterns thus new relations. Snowball's [21] architecture is inspired from DIPRE, the relation of interest is Organization-Location. The system associates a score with each relation extracted and each pattern generated, the pair extracted is later discarded if it is less than a certain threshold. The main improvement that Snowball made over DIPRE system, is that it avoided long patterns that resulted in decreasing the probability of finding a pattern match thus decreasing the overall performance of the system. Furthermore, Bunescu and Mooney [22] worked on extracting protein interactions from biomedical dataset using subsequence kernel.

III. Basic Arabic Language Structure

Arabic is a complex language and is native to countries of the Arab league and other neighboring countries, albeit in different dialects. However, Modern Standard Arabic (MSA), is the formal written standard Arabic that is used all over the Arab

world. In this section, only a brief explanation of the Arabic language, with emphasis on what is required for the scope of this paper, is represented. A more detailed explanation of the language can be found in [2].

An Arabic sentence is a combination of one or more sequential words as in English language, however the syntax is more flexible. Thus, an Arabic sentence could have one of the following structures: verb-object- subject, verb-subject-object or subject-verb-object for example, ذهب احمد', احمد ذهب الي المدرسة', ذهب احمد الي المدرسة', الي المدرسة ذهب احمد' respectively, they all translate to "Ahmed went to school". Also, the verb can be omitted and the sentence would solely consist of a subject and an adjective. An Arabic phrase is either:

- A **noun** phrase: starts with a noun or pronoun.
- A **verbal** phrase: includes a verb in the present, past or imperative tense.
- A **prepositional** phrase.

One of the challenges in electronic Arabic text is that text is not diacritized, for example, كتب احمد الدرس', كتب' could easily be misinterpreted for "books" instead of "wrote". Being the sixth most-spoken language in the world, Arabic is considered one of the richest languages morphologically.

IV. Proposed Model

In this section the system's workflow is first presented, followed by a detailed description of the proposed relation extraction algorithm.

A. System Overview

The proposed system's architecture, which was inspired by both Snowball [21] and "Badea" [13], consists of two main phases: pattern extraction and generation, and relation extraction. The relations of interest are: PERS-ORG, ORG-LOC, PERS-LOC. Fig. 1 shows the general workflow of the system.

Initially, a small set of random pairs (seeds) of a particular relation are used for pattern extraction. A sample of the seeds is shown in Table I. The algorithm iteratively searches the dataset for patterns matching the occurrences of those seeds. The challenging part of this algorithm is pattern generation as the patterns generated should not be too general nor too specific. Thus, given a reliable set of patterns we can extract reliable pairs. This is explained later in details.

B. Preprocessing Phase and Pattern Extraction

Prior to pattern generation phase, the dataset used undergoes some preprocessing that includes stemming, POS tagging, and NER using Microsoft's Arabic toolkit [23] to identify PERS, ORG, LOC and DATE entities. The algorithm then searches for instances of those seeds and extracts the whole sentence.

Stemming is the process of minimizing large space words into smaller ones (roots) [24]. It was carried out using "The Information Science Research Institute" (ISRI) stemmer [25]. Stemming relies on detaching the longest prefix and suffix present in the input word. Using a root library, the remaining part of the word is matched with a known verb and noun patterns [26]. Furthermore, ISRI stemmer returns a normalized version

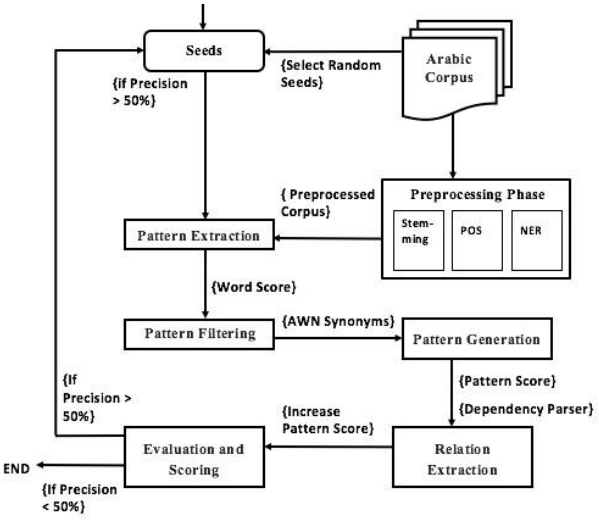

Fig. 1. System's Workflow

of the word if it fails to root it. It detaches specific determinants and end patterns instead of leaving the word unchanged.

Stemming is necessary in Arabic language, particularly with Arabic adjectives, due to their tendency to have a complex position in basic sentences [27]. Adjectives have both masculine and feminine forms, they always agree with the noun they modify in terms of number (singular, dual or plural), they have a grammatical case (prepositional, subject, or direct object) and gender.

The ISRI stemmer has been proven to be effective in many NLP tasks on morphologically rich languages such as Arabic in comparison to a non-stemmed approach [28-29]. Hence, using a root stemmer to reduce words enhances the relation extraction task [28].

C. Pattern Generation

Subsequent to pattern extraction process, all named entities are eliminated from the patterns, PERS, ORG, LOC, and DATE, each word is then assigned a weight based on its number of occurrences in all the extracted patterns. A grid-search was performed to select a reliable pattern threshold, and 30% was selected as the minimum percentage of patterns in which a word should appear. A word is discarded, if it appears in less than 30% of the total extracted patterns. This is done to avoid specificity

TABLE I. Intial Sample of PERS-ORG seeds with their English translation

PERS	ORG
كوفي عنان *"Kofi Annan"*	الامم المتحدة *"United Nations"*
لاري بيدج *"Larry Page"*	شركة جوجل *"Google Co."*
جيف بيزوس *"Jeff Bezos"*	شركة امازون *"Amazon Co."*
جيري يانج *"Jerry Yang"*	ياهوو *"Yahoo"*
ماسارو ابوكا *"Masaru Ibuka"*	سوني *"Sony"*

of a pattern. In order to measure the reliability of a pattern, a pattern score is computed, which is the cumulative sum of the words' scores.

Lexical verbs are the most vital POS, if the verb is absent, there cannot be a sentence or even an independent clause. Verbs are the building blocks of the language. They denote strong actions which are considered extremely effective in extracting relations of interest. In order to avoid the generality of a pattern, verbs were assigned a higher score than any other word, that ensures that the verb is always a part of the pattern and will never be discarded. In addition, upon using the generated patterns in the dataset, Arabic WordNet (AWN) [30] is used to expand the words semantically by adding the synonyms for each verb in the pattern. AWN is a lexical database with a similar configuration as Euro WordNet [31]. AWN only accepts diacritized Arabic words, Microsoft's Arabic Diacritizer [23], was used to enable us to extract synonyms. Each synonym extracted was assigned the same score as the verb that generated it. New patterns are now added to the set of generated patterns.

D. Relation Extraction

In this phase the pattern recognition algorithm is used to extract the relations using the generated patterns. Each pattern is then evaluated according to the number of relations it has extracted. This is done by doubling the pattern's score every time it extracts another relation. Consequently, an extracted pair of relation that has been generated by a highly-scored pattern can be considered reliable. In order to avoid error propagation, we assign the same score of the pattern to the relation pair extracted. If a relation was extracted by more than one pattern its score is cumulated. As a result, we only use highly-scored pairs in the following iteration to avoid error propagation.

E. Negative Relations

Negative relations are instances containing entity elements that can never occur together in a valid relation. We cannot simply check if a phrase contains a negative word and deduce the invalid relations, since a single sentence can be recursive by including more than one phrase, where each phrase might contain a relation. For example, consider the sentence: 'السودان في احمد ليس في مصر، لكنه في', (Ahmed is not in Egypt, however he is in Sudan). This sentence has Ahmed-in-Egypt as a negative relation and Ahmed-in-Sudan as a positive relation. Although 'not' appears in the sentence, it does not mean the all relations are negative. To further avoid extraction of negative relations, a Stanford's dependency parser [32] was used. Fig. 2 [33] gives an example of the dependency parse tree for the sentence 'Ahmed is not in Egypt, however he is in Sudan'. The shown dependencies contain negative tag "Neg(not, Egypt)" that's connected to the NE (Ahmed), which is indicative of negative relation between two entities (Ahmed and Egypt).

V. Results and Evaluation

The performance of the proposed algorithm was tested on two different datasets. Different numbers of seeds were tested (5, 10, and 20 seeds). "Badea" approach [13] was also implemented and tested on both datasets using the same number of seeds.

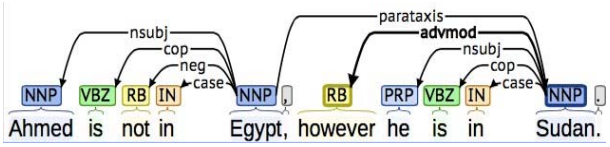

Fig. 2. Dependency Parser Example
(**NNP**: Proper pronoun; **VBZ**: Verb, 3rd person singular in present tense; **RB**: Adverb; **IN**: Preposition; **PRP**: Personal Pronoun; **nsubj**: Nominal subject;**cop:** Copula; **neg**: negation; **advmod**: Adverbial modifier)

A. Datasets

Two different datasets were used for testing: Arabic NER corpus (ANERCorp) [10], and Open Source Arabic Corpus (OSAC) [34], both in MSA form.

The first dataset, ANERCorp, a news corpus, consists of more than 150,000 words and more than 3000 NEs (PERS, ORG, LOC and Date), of which only 800 are related. This corpus was also used by Boujelben, I. et al. [4] in relation extraction using genetic algorithms, which enabled us to compare the achieved results with the results obtained by Boujelben, I. et al. [4]. A small sample of this dataset is shown in Table II along with its English translation.

The second dataset, OSAC, is clustered into many domains including sports, science and technology, middle east new, business and commerce. Only news and sports were used due to their high popularity. This corpus is composed of more than 12,000 words and 2000 NEs (PERS, ORG, LOC and DATE), of which only 500 are related.

B. ExperimentalResults

Three experiments were carried out to measure and compare the performance of the proposed pattern-based approach. Four evaluation metrics were used to measure performance:

- **Recall (R):** Is the ratio of the number of relevant instances retrieved to the total number of existing relevant instances, defined as:

TABLE II. Sample From ANERCorp Dataset

ان وزير الخارجية فرانك فالتر شتاينماير سيتوجه السبت إلى الشرق الأوسط، يأتي ذلك في وقت حذر فيه وزير الخارجية الفرنسي فيليب دوست بلازي بما يجري الان في البلاد. في السياق طلب الرئيس الفرنسي جاك شيراك من الاتحادالأوروبي ' تفويض ' الممثل الأعلى ل لسياسة الخارجية للاتحاد خافيير سولانا في العمل. وكان هناك بعض الحلول التي تقدم بها الأمين العام ل الأمم المتحدة كوفي أنان لحل الأزمة. وقال المتحدث باسم وزارة خارجية روسياأندريه بوبوف ان هناك الكثير من الحلول لعرضها.
"Foreign Minister Frank-Walter Steinmeier will travel Saturday to the Middle East, this comes at a time when French Foreign Minister Philippe Douste-Blazy warned of what is happening now in the country. Due to the ongoing events, French President Jacques Chirac asked the European Union to 'delegate' High Representative for Foreign Policy Chief Javier Solana. And some of the solutions were put forward by the Secretary-General of the United Nations Kofi Annan to resolve the crisis. A spokesman for the Russian Ministry of Foreign Affairs said that there are plenty of other solutions to offer."

$$R = \frac{TP}{TP+FN} \quad (1)$$

- **Precision (P):** Is the ratio of the number of relevant instances retrieved to the total number of irrelevant and relevant records retrieved, defined as:

$$P = \frac{TP}{TP+FP} \quad (2)$$

- **F-Measure (F):** Is the combination of the precision and recall for penalizing the very large in equalities between these two measure, defined as:

$$F = 2 \times \frac{P \times R}{P+R} \quad (3)$$

- **Accuracy (A):** Is the fraction of true results against the total number of cases evaluated, defined as:

$$A = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}}, \quad (4)$$

Where TP: the number of true positives, FN: the number of false negatives, TN: the number of true negative, FP: the number of false positives

All the aforementioned measures were expressed as percentages throughout the experiments.

1) Experiment 1: The proposed pattern-based approach, and the "Badea" approach were tested on the ANERCorp dataset. Table III illustrates the results of both approaches on the ANERCorp dataset.

For the same number of seeds, the proposed approach achieves significantly higher recall and higher precision than "Badea". Using 20 seeds, the highest recall of 69.31% was achieved versus only 35.80% using "Badea". The highest accuracy of 78.46% was scored when 20 seeds were used, versus 68.58% using "Badea". However, the highest precision was achieved using 5 seeds 87.47%, while "Badea" achieved 85.38%. The reason behind this increase, is that the 5 seeds used extracted fewer relations thus decreasing the number of false positives extracted as well.

The proposed pattern-based approach was compared with the genetic algorithm proposed by Boujelben, I. et al. [4] on the ANERCorp dataset. Since their approach also included LOC-LOC and PERS-PERS relations, for a fair comparison, these relations were also added to the set of extracted relations to both, the proposed approach and "Badea" approach. While Boujelben, I. et al. [4] yielded a significantly higher recall than "Badea", 59.6% and 42.94% respectively, it was still significantly lower than the recall achieved by the proposed model (69.31%).

The performance of the proposed model surpassed that of "Badea's" and Boujelben, I. et al. [4]. The best results of all three methods are illustrated in Table IV.

Different seed set sizes, up to a maximum of 20 seeds were tried, as observed in Table III, the performance of the proposed algorithm increases as the number of initial seeds increases. The algorithm would benefit from a larger seed size if it was feasible. However, since ORG-LOC has only 36 true relations, increasing the number of seeds was infeasible, since we would have to rule-out ORG-LOC relation, and thus the comparison with Boujelben, I. et al. [4] would be invalid.

2) Experiment 2: The second dataset, OSAC, was used in the experiment, the proposed algorithm achieved a maximum recall of 47.59% when 20 seeds were used, while "Badea" scored a maximum recall of 34.3% with the same number of seeds. The results of this experiment are shown in Table V. The F-measure followed the same trend, the highest was scored using 20 seeds (65.1%), and "Badea's" using 20 seeds (47.8%).

Nonetheless, the precision of the proposed algorithm using 5 seeds (86.25%) was higher than the precision achieved using both 10 seeds (78.46%) and 20 seeds (76.03%) for the same reason mentioned in experiment 1. While "Badea" scored the highest precision when 10 seeds were used (82.40%).

As in experiment 1, the accuracy increased as the number of seeds increased, in both, the proposed approach and "Badea", scoring 72.58% and 69.32% respectively, using 20 seeds.

In comparison to the ANERCorp dataset, the overall performance on OSAC was lower. A possible reason behind that is the use of 2 different genres in the OSAC dataset. Since the patterns generated from sports seeds did not perform well in extracting relations from the news genre and vice versa.

The proposed algorithm achieved high precision, by extracting only a few "False Positives", owing to:

TABLE III. SUMMARY OF EXPERIMENT 1 RESULTS ON ANERCORP DATASET

	5 Seeds		10 Seeds		20 Seeds	
	Proposed Approach	**"Badea" Approach**	**Proposed Approach**	**"Badea" Approach**	**Proposed Approach**	**"Badea" Approach**
Recall	56.10%	35.80%	63.99%	38.20%	**69.31%**	42.94%
Precision	**87.47%**	85.38%	85.28%	80.32%	82.72%	75.77%
F-Measure	67.66%	50.44%	72.76%	50.54%	**75.06%**	54.83%
Accuracy	67.85%	55.09%	73.39%	60.14%	**78.46%**	68.58%

TABLE IV. SUMMARY OF EXPERIMENT 1 RESULTS

	Proposed Approach	Boujelben, I.	"Badea"
Recall	69.31 %	59.60%	42.94%
Precision	82.72%	74.10%	75.77%
F-Measure	**75.06%**	66.10%	54.83%

- Detection of negative relations using dependency parser.
- The generated patterns gravitate more towards specificity rather than generality due to the way the patterns' and words' scores are calculated.
- Prior to adding new pairs to the set of extracted relations, we check the pairs' NE tags to ensure that the appropriate relation is extracted.

3) Experiment 3: In this experiment, the significance of each Stemming, and the use of synonyms were examined by testing the proposed model, once with stemming eliminated from the pre-processing phase, and another time with excluding the synonyms. Both tests were conducted on the ANERCorp dataset using 20 seeds, where the highest results were achieved.

The results of this experiment are shown in Table VI, with the four performance measures compared against those of the original proposed model.

As demonstrated in Table VI, the recall decreased substantially from 69.31% to 53.95% when stemming was excluded. This is due to the specificity of the patterns generated. As a consequence, it is challenging to extract new relations using the generated patterns.

On one hand, after eliminating synonyms, the recall declined from 69.31% to 53.95%. The elimination of synonyms resulted in decreasing the amount of patterns generated. As a result, the probability of finding a pattern match was affected resulting in decreasing the total number of instances retrieved.

On the other hand, the precision increased from 82.72% to 83.54% and 84.81% after excluding stemming and synonyms respectively. The rationale behind this increase, in case of excluding stemming the instances retrieved were lessened, thereby the false positives decreased as well. In the case of excluding synonyms, as it was previously mentioned, a limited amount of patterns was generated, yielding a slight number of false positives, leading to the increase in the precision.

TABLE VI. SUMMARY OF RESULTS OF EXPERIMENT 3

	Proposed Approach	Excluding Stemming	Excluding Synonyms
Recall	69.31%	35.47%	53.95%
Precision	82.72%	83.54%	84.81%
F-Measure	**75.06%**	48.85%	65.95%
Accuracy	78.46%	64.84%	72.40%

The original proposed model scored the highest accuracy of 78.46%, while excluding stemming and synonyms scored an accuracy of only 64.84% and 72.40% respectively.

With the F-measure being the breakthrough performance measure, scoring 75.06% versus only 48.85% without stemming, and 65.95% without synonyms. This verifies the importance of including both in the proposed pattern-based relation-extraction algorithm.

It's important to note that, not only did stemming improve the recall and increased the number of true positives, but also the number of extracted synonyms per word increased after stemming.

VI. CONCLUSION

Using only 20 input seeds, the semi-supervised pattern-based bootstrapping algorithm presented here delivers higher performance than existing state-of-the-art algorithms. These goals were achieved by relying on the use of stemming and synonymic relation.

Stemming the verb to its base form allows the pattern to be more general rather than specific, thereby increasing the number of extracted relations. The semantic expansion achieved from getting the verb's synonyms allowed more patterns to be generated, which further increased the number of extracted relations. The importance of these two additions was experimentally verified.

TABLE V. SUMMARY OF EXPERIMENT 2 RESULTS ON OSAC DATASET

	5 Seeds		10 Seeds		20 Seeds	
	Proposed Approach	**"Badea" Approach**	**Proposed Approach**	**"Badea" Approach**	**Proposed Approach**	**"Badea" Approach**
Recall	39.80%	20.8%	43.34%	30.42%	**47.59%**	34.30%
Precision	**86.25 %**	80.8%	78.46%	82.42%	76.03%	78.80%
F-Measure	53.64 %	33.10%	51.97%	44.4%	**58.46%**	47.80%
Accuracy	65.49 %	57.24%	67.54%	64.28%	**72.58%**	69.32%

In addition, the proposed algorithm is also capable of inheriting any relation by simply changing the seeds entered. Furthermore, the dependency parser used to detect negative relations, the words' and patterns' scores increased the overall precision of the system.

On the other hand, there is room for improvement in the proposed approach. For example, the criteria upon which certain words are eliminated from the patterns based on their score could be enhanced, this would in turn improve pattern filtering. In addition, the adaptability of the proposed method can be evaluated by testing it on a larger dataset or possibly on another language.

The presented work can be further extended to be able to extract tri-nary relations, for instance PERS-LOC-ORG relation with a larger seed set as it was shown earlier that increasing the number of seeds seemed promising and worth exploring. Finally, this approach can be used in other NLP tasks such as question answering and summarization.

REFERENCES

[1] M. El-Defrawy, Y. El-Sonbaty, and N. Belal, "CBAS: Context based Arabic stemmer," International Journal on Natural Language Computing, vol. 4, No. 3, 2015.

[2] K. Katzner, The Languages of the World. Routledge, London, 3rd edition, 2002.

[3] R. Fathalla, Y. El-Sonbaty, and M. Ismail, "Extraction of Arabic words from complex color image," 9th IEEE International Conference on Document Analysis and Recognition, vol. 2, pp. 1223-1227, Brazil, 23-26 September 2007.

[4] I. Boujelben, S. Jamoussi, and A. Ben Hamadou, "Genetic algorithm for extracting relations between named entities," 6th Language and Technology Conference, Poznan, Poland, pp. 484-488, 2014.

[5] G. Al Zamil, and Q. Al-Radaideh, "Automatic extraction of ontological relations from Arabic text," Journal of King Saud University-Computer and Information Sciences, pp. 462-472, 2014.

[6] I. Boujelben, S. Jamoussi, "Relane: Discovering relation between Arabic named entities," International Conference, TSD, Brno, Czech Republic, 2014.

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. Witten, "The WEKA data mining software: An Update; SIGKDD Explorations," vol. 11, issue 1, 2009.

[8] C. Manning, and D. Klein, "Optimization, Maxent Models, and conditional estimation without magic," Tutorial at HLT-NAACL 2003 and ACL, 2003.

[9] M. Falih, and N. Omar, "A comparative study on Arabic grammatical relation extraction based on machine learning classification Middle-East," Journal of Scientific Research, 2015.

[10] Y. Benajiba, P. Rosso, and J. Benedi, "ANERsys: An Arabic named entity recognition system based on maximum entropy," CICLing, Springer-NNerlag, Berlin, Heidelberg, pp. 143-153, 2007.

[11] J. Holland, and J. Reitman "Cognitive systems based on adaptive algorithms," in D.A. Waterman and F. Hayes-Roth (eds.), Pattern-Directed Inference Systems, Academic Press, NY, 1978.

[12] M. Hearst, "Automatic acquisition of hyponyms from large text corpora," Proceedings of the 14th conference on Computational linguistics, vol. 2, pp. 539–545, 1992.

[13] M. Al-Yahya, L. Aldhubayi, and S. Al-Malak, "A pattern-based approach to semantic relation extraction using seed ontology," IEEE Conference on Semantic Computing, 2014.

[14] Y. Belinkov, N. Habash, A. Kilgarriff, N. Ordan, R. Roth, and V. Suchomel, "arTenTen: a new, vast corpus for Arabic," WACL'2 Second Workshop on Arabic Corpus Linguistics, 2013.

[15] 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," Workshop on Arabic Corpus Linguistics, Lancaster University, UK, 2013.

[16] A. Alotayq, "Extracting relations between Arabic named entities", Springer-Verlag, Berlin Heidelberg, Pilsen, pp.265-271, 2013.

[17] Mitchell, Alexis, et al. ACE 2005 Multilingual Training Corpus LDC2005T09. Web Download. Philadelphia: Linguistic Data Consortium, 2005.

[18] W. Lahbib, I. Bounhas, B. Elayeb, F. Evrard, & Y. Slimani, (2013, "A Hybrid Approach for Arabic Semantic Relation Extraction". In The Twenty-Sixth International FLAIRS Conference, May 2013.

[19] I. Sarhan, Y. El-Sonbaty, M. Abou El-Nasr, "Arabic Relation Extraction: A Survey", International Journal of Computer and Information Technology, vol. 5, issue 5, 2016.

[20] S. Brin, "Extracting patterns and relations from the world wide web," WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT, 1998.

[21] E. Agichtein, L. Gravano, "Snowball: Extracting relations from large plain-text collections," Proceedings of the Fifth ACM International Conference on Digital Libraries, 2000.

[22] Raymond J. Mooney and Razvan C. Bunescu. "Subsequence kernels for relation extraction." Advances in neural information processing systems. 2005.

[23] Microsoft Research, Arabic Toolkit Service (ATKS), http://atks.microsoft.com (accessed March 25,2015.)

[24] M. El-Defrawy, Y. El-Sonbaty, and N. Belal, "A rule-based subject-correlated Arabic stemmer," Arabian Journal for Science and Engineering - Springer, pp. 1-9, Feb. 2016.

[25] K. Taghva, R. Elkoury, and J. Coombs, "Arabic stemming without a root dictionary," Information Science Research Institute. University of Nevada, Las Vegas, USA, ,2005.

[26] S. Oraby, Y. El-Sonbaty, M. Abou El-Nasr, "Exploring the effects of word roots for Arabic sentiment analysis," 6th International Joint Conference on Natural Language Processing, Nagoya, Japan, pp 14-18,2013.

[27] S. Oraby, Y. El-Sonbaty, M. Abou El-Nasr, "Finding opinion strength using rule-based parsing for Arabic sentiment analysis," Advances in Soft Computing and its Applications, Springer Lecture Notes in Computer Science, Vol. 8266, pp. 509-520, 2013.

[28] A. Magdi, Y. El-Sonbaty, M. Kholief, "Exploring the effects of root expansion," Sentence Splitting and Ontology on Arabic Answer Selection, 11th International Workshop on Natural Language Processing and Cognitive Science, Venice, Italy, October 27-29, 2014.

[29] A.Magdi, M. Kholief, Y. El-Sonbaty, "ALQASIM: Arabic language question answer selection in machines," CLEF: Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science, vol. 8138, Valencia, Spain, 23-26 September, 2013.

[30] W. Black, S. Elkateb, H. Rodriguez, M. Alkhalifa, "Arabic WordNet project," in Proceedings of the Third International WordNet Conference, 2006.

[31] G. Miller, R. Beckwidth, C. Fellbaum, D. Gross, and K. Miller, "Introduction to WordNet: an online lexical database," International Journal of Lexicography, Vol 3, No.4 Winter , pp. 235-244, 1990.

[32] G. Miller, R. Beckwidth, C. Fellbaum, C. Gross, S. Green and D. Christopher, "Better Arabic Parsing: Baselines, Evaluations, and Analysis," COLING (2010)

[33] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," Association for Computational Linguistics (ACL) System Demonstrations, pp. 55–60, 2014.

[34] M. Saad, and W. Ashour, "OSAC: Open source Arabic corpus," 6th ArchEng International Symposiums, EEECS10 the 6th International Symposium on Electrical and Electronics Engineering and Computer Science, European University of Lefke.