Arabic Relation Extraction: A Survey

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—**Being the intersection between lexical and computational science, Natural Language Processing (NLP) has been earning a vast amount of attention in the past years. Relation Extraction is a well-studied subject when it comes to English language. However, due to the complexity of the Arabic language, it is challenging to extract relations from Arabic text. The foremost goal of this paper is to discuss the major techniques used in Arabic relation extraction and investigate their strengths and weaknesses in order to guide future research towards creating an enhanced convenient relation extraction algorithm.**

Keywords-Arabic; NLP; Relation Extraction

I. INTRODUCTION

As the demand for a fast and efficient method to transform unstructured data into structured data increases day by day, researchers are encouraged towards NLP tasks. NLP is a large scale field that includes: information extraction, summarization and question answering. All the aforementioned tasks require relation extraction in order to understand the semantic relation that lies between the named entities.

Relation extraction is the task of extracting relations between named entities. A relation is either a binary relation, for instance: Located-In(Moscow, Russia), or a higher order relation (n-ary), for instance a 3-ary relation between Employee-Position-Company(Adam Smith, Marketing Manager, XYZ Company). Several relation extraction approaches such as Dual Iterative Pattern Relation Expansion (DIPRE) [1], Snowball [2] and TextRunner [3] attain promising results on English language. However, such applications are challenging on morphologically rich languages such as Arabic language, due to its rather complex grammatical functions.

With almost 24 Arabic-speaking countries and the presence of a massive amount of Arabic text on the web, NLP algorithms on Arabic language is necessary [4-5]. For instance, Arabic sentiment analysis studies were carried out in [6-7]. In addition, Ezzeldin et al. proposed an Arabic question answering system [8], Arabic stemming was carried out in [9-11]. Furthermore, Arabic text may also be found in images, Fathalla et al. proposed a method to extracted Arabic words from images [12]. Thus vital NLP tasks such as relation extraction are receiving a large amount of attention.

Recently, a large number of Arabic relation extraction approaches were proposed, each uses a different technique to investigate pairs of named entities and detect the relation between them. Similar to any other NLP task, pre-processing is necessary, that includes:

- Tokenization: Breaking sentences into words.
- Part of Speech (POS) Tagging: Labeling each word in the corpus with the corresponding grammatical tag (Noun-Adjective-Verb etc.).
- Named Entity Recognition (NER): Classifying and labeling words into categories, such as (Person-Location-Organization-Date)

The remainder of this paper is structured as follows; Section 2 briefly describes the Arabic language. Section 3 presents supervised relation extraction, followed by semi-supervised relation extraction in Section 4. Finally, Section 5 concludes the paper.

II. BASIC STRUCTURE OF THE ARABIC LANGUAGE

The Arabic language is a universal language, it's the language of the Holy Quran and the native language of countries of the Arab league and other countries, although its dialects differ from a country to the other. Nonetheless, Modern Standard Arabic (MSA), is the formal written standard Arabic that is used all over the Arab world.

Arabic is composed of 28 alphabets in which all are consonants. Unlike English language, Arabic doesn't have vowels, the vowels are represented above the letter itself by small symbols called diacritics. One of the challenges faced, is the lack of diacritics in most electronic Arabic text. Thus a word could easily be misinterpreted, for instance "ذهب محمد الي العمل" (Mohamed went to work), the word "ذهب" (went) could be misinterpreted as (gold) instead of its original meaning in the sentence (went). However, if it's diacritized "ذَهَبَ", it's original meaning is apprehensible.

Similar to the English language, an Arabic sentence is a combination of one or more sequential words, nevertheless the syntax is more flexible. Thus, an Arabic sentence could have one of the following structures:

- Verb-Subject-Object: "ذاكر احمد الدرس" (Studied Ahmed the lesson).
- Verb-Object-Subject: "ذاكر الدرس احمد" (Studied the lesson Ahmed).
- Subject-Verb-Object: "احمد ذاكر الدرس" (Ahmed Studied the lesson).

Moreover, the verb can be omitted, for example the sentence "المدينة جديدة" (City new) constructs a full sentence, it consists of a noun (City "المدينة") and an adjective (new "جديدة"), there's no need for a verb (is). An Arabic sentence follows one of the following structures:

- Nominal: Begins with a noun or pronoun, " هذه مدرسة ممتازة" (This is an excellent school).
- Verbal: Begins with a verb, "درسنا الدرس" (We studied the lesson).

Multiple languages were influenced by Arabic language, including Persian, Kurdish, Hindi, Bengali and more [4]. Being the sixth most-spoken language in the world, Arabic is considered one of the richest languages morphologically.

III. Supervised Relation Extraction

In supervised relation extraction, the task is presented as a classification task. Supervised methods depend on machine learning algorithms and a training set in order to extract semantic relations. It consists of three main phases:

- Select the set of relations to extract.
- Use the appropriate named entities, find and label them in the dataset.
- Divide the corpus into: training, development and test sets.

RelANE is a relation extraction system proposed by Boujelben et al. [13] in 2014 to discover relations between Arabic named entities. Due to the high frequency of specific Named Entities (NE), 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). The authors manually constructed their own dataset. Preprocessing stage includes an Arabic clause splitter, POS tagger and NER. The entities of interest are the ones stated above. In addition, each Arabic word was manually annotated with one of the following flags:

- REL: The word is a relation between two NE.
- PREL: The word is a part of a relation.
- N: The word is neither a relation nor a part of a relation.

The features investigated for each word in a sentence were:

- POS tag of a word.
- POS tag of the three words before and after this word.
- Grammatical structure of each clause.
- NE tags of a word (PERS-LOC-ORG and DATE).
- Numeric features: Position of the word according to NEs and number of characters of each word.

Six different classification techniques were used, PART, Decision Tree, Adaboost, Naïve Bayes, Support Vector Machine (SVM) –which are all available on WEKA[14]- and MaxEnt [15]. SVM achieved the highest performance of 85.23% in terms of F-measure, followed by Adaboost that scored a F-measure of 82.13%. Due to the vagueness in determining the correct POS and the declassification of the NEs, some relations are not extracted, this is the main drawback of RelANE. Furthermore, their system would be evaluated better if any other popular dataset was used instead of the manually constructed dataset. Nonetheless, negative relations were taken into account. In Addition, this approach applied a 10-fold cross validation to overcome the over fitting problem.

In the following year, Mohanaed Falih and Nazila Omar [16] proposed an Arabic grammatical relation extraction based on machine learning classification. The main goal of this approach is labelling each Arabic word with the correct grammatical relation (subject, object or verb). A special training Arabic corpus was created by the authors 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 architecture of the proposed system is shown in Figure 1 below.

In the preprocessing phase, a clause splitter is first used on each Arabic sentence to ease the classification and extraction of the grammatical relation. In addition, POS tagging is also carried out followed by tokenization. The subsequent step is feature extraction, in which every word is transformed into a feature vector using optimized sliding window techniques previously described by [17-19]. Various term weighing could be used, either Term Frequency (TF), Inverse Document Frequency (IDF) or a merger of TF-IDF. Machine learning classification techniques are then used for grammatical relation extraction and classification. This is achieved using either of the three classifiers:

- SVM
- K-nearest neighbor (KNN)
- Combination of SVM and KNN

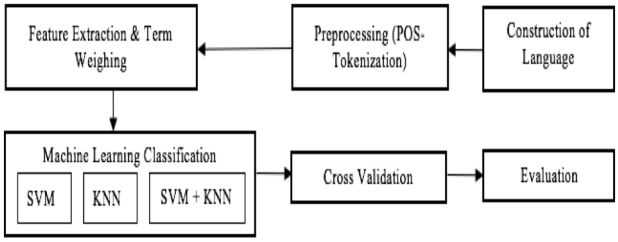

Figure 1. System architecture of Falih and Omar relation extraction model

The best performance was achieved by merging SVM and KNN, resulting in a F-measure of 93.48%, followed by SVM then KNN, each resulting in a F-measure of 82.4% and 62.5% respectively. Finally, cross validation is applied to evaluate the results. The drawback of this approach is the manually assembled test corpus used which might lead to an unfair evaluation of the whole system. However, merging SVM and KNN features enhanced the results of this approach.

A. Relation Extraction Using Genetic Algorithm

Genetic Algorithm (GA), is a machine learning search model based on concepts from biologic evolutions [20]. Recently, GA popularity increased in NLP field and machine learning as well. For instance, McIntyre and Lapata [21] used GA in story generation, while Echizenya et al., used GA in machine translation in [22-24]. However, to the best of our knowledge very little NLP work has been done on Arabic language using genetic algorithm.

Boujelben et al. [25] proposed a supervised model that automatically extract rules for relation extraction using genetic algorithms. 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 includes almost 2000 named entities. This algorithm is mainly composed of two phases:

1. Generating rules using learning methods.
2. Discovering rules using GA.

The initial step is extracting clauses that only contains two named entities (NE), this was achievable by using two tools, Arabic clause splitter and NER tool. The following step uses the automatically annotated clauses to extract NE and POS tags adjoining those NE as shown in Figure 2 below.

Where C1, C2 and C3 are the words before the first NE, between the two NEs and after the second NE respectively. Next, rules are discovered using genetic algorithm using Michigan approach [26], in which each rule is illustrated by a chromosome. The dominant objective is to enhance the initial rules quality; this can be achieved by building a rules filter. Prior

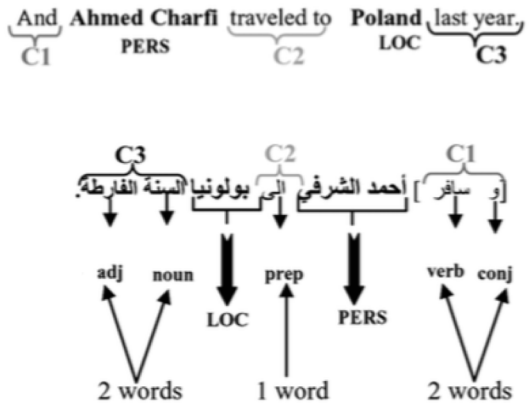

Figure 2. Example of the annotation in the GA approach

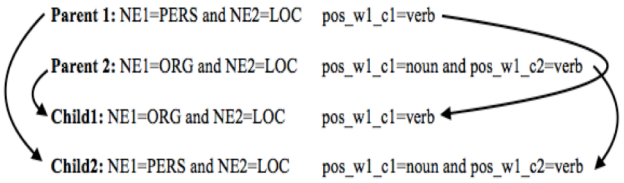

Figure 3. Single point cross-over illustration

to the crossover and mutation process, the filtering module is applied. Using parents' 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. Moreover, a random element in each parent is selected and crossed over to generate two rules, as shown in Figure 3.

ANERCorp [27] corpus was used for evaluation to test the algorithm. This approach scored an F-Measure of 66.1%. The main disadvantage of this approach, its inability to discover a 2-word relation. Furthermore, noise errors that are generated due to the ambiguity of the Arabic language will indeed have a negative impact on the rules generated, for instance a word could be misinterpreted as PERS while it's actually a LOC. On the other hand, interesting and wide range of relations were discovered.

A summary of the discussed supervised approaches is presented in Table 1, along with their advantages and disadvantages.

IV. Semi-Supervised Relation Extraction

Due to the presence of a vast number of unlabeled data on the web, supervised approaches are no longer applicable in this case. In semi-supervised relation extraction one of two methods are used for relation discovery:

- Rule-based (Pattern-based) Method: Search for rules (patterns) that connect entities detect the relation between them.
- Statistical Method: Relies on machine learning process from a reduced annotated corpus.

The main disadvantage of the rule-based method is its inability to handle a large scale of data. Nonetheless, the generation of annotated corpus is the statistical method is rather expensive. However, recently bootstrapping approach became popular, in which small set of seeds are used instead of a training corpus.

In 1992, Hearst [28], debuted pattern-based bootstrapping approach which inspired succeeding pattern based algorithms. Hearst used a set of predefined patterns to extract relations, and used a bootstrapping approach to generate more patterns. Three relations were extracted using Hearst algorithm, Hyponym-

TABLE I. SUMMARY OF SUPERVISED APPROACHES, THEIR ADVANTAGES, AND DISADVANTAGES

Method	Extracted Relation	Acquired preprocessing	Dataset	Results (%)	Advantages	Disadvantages
Boujelben et al. [13]	Relation that lies between PERS-LOC-ORG-DATE	• POS-NER-Clause Splitter	Arabic Electronic News (1245 Sentences)	P=86.5 R=84 F=85.23	• Negative relations detection. • Investigated six different classifiers.	• A standard corpus should have been used instead of random news. • No comparison with previous research.
Falih and Omar [16]	Grammatical Relations (Subject-Object-Verb)	• POS-Tokenization	Manually constructed (80 Sentences)	P=94.44 R=93.33 F=93.48	• Combining SVM and KNN to enhance the results.	• Small corpus was used with only 80 sentences, which might lead to an unfair evaluation.
Boujelben et al. [25]	PERS-LOC, ORG-LOC, PERS-ORG, PERS-PERS, LOC-LOC	• POS-NER-Clause Splitter	ANERCorp (25000 sentences)	P=74.1 R=59.6 F=66.1	• Several relations are extracted.	• Negative relations not taken into account.

Hypernym (Purple, Color), IS-A (lion, mammal) and Kind-of (Germany, European Country). Large human intervention is required in order to create patterns from real examples this is considered the main disadvantage of Hearst algorithm.

In 2014, Al-zamil and Al-radaideh [29], enhanced Hearst algorithm by generating a system for automatic extraction of ontological relations from Arabic text. The objective of this approach is to generate patterns and extract semantic features of Arabic text in order to extract ontological relations. The enhancements that this approach made on Hearst algorithm include:

- Pattern filtering.
- Enhancement of the patterns' quality and assessment.
- Increased the number of relation extracted: Cause-Effect, Has-a and Part-whole.

The architecture of the enhanced system is shown in Figure 4.

In this approach the semantic relations are composed of positive and negative rules, the classifier role is to look for occurrences that are made up of the positive rules and don't hold any of the negative ones. Furthermore, in order to ease the text analysis task some preprocessing is done that includes POS and stemming. Initially, patterns are extracted from an Arabic corpus, afterwards, the extracted patterns are converted into queries and new terms are extracted. To overcome redundant patterns, this method takes synonyms of the extracted patterns into consideration, those synonyms are taken from Arabic WordNet tool. For evaluation, three different Arabic corpora were collected, classic Arabic from the Holy Qur'an, modern standard Arabic form newspapers and unstructured Arabic text from social blogs. The highest F-measure of 74.70% was achieved using the newspaper corpora, while Hearst scored an F-measure of 48.48% using the same corpora. Few classification errors were detected that had a negative impact on the overall performance of the system. In contrast, one of the main advantages of this approach, is that a filtering task is applied to overcome the ambiguity created by stemming and POS. In addition, negative patterns are used to enhance the accuracy of the system, along with a coverage metric to avoid covering the same data by more than one pattern.

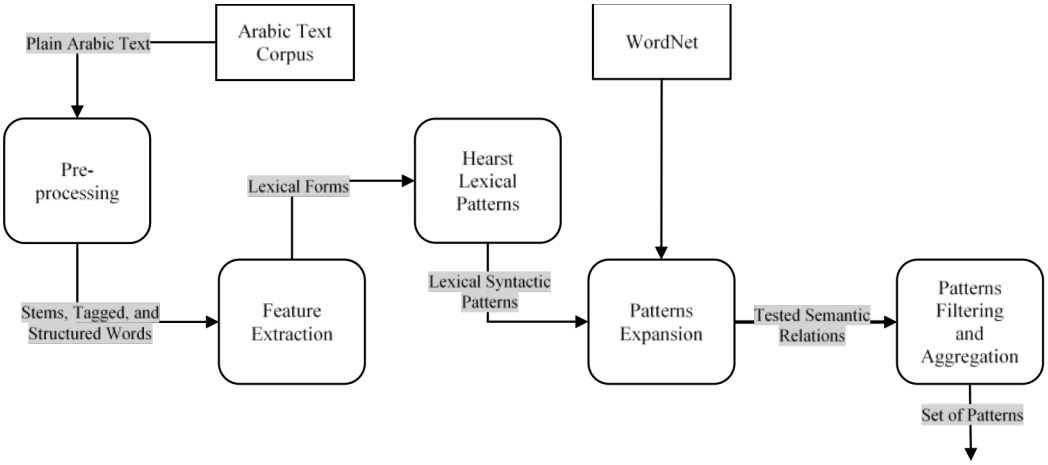

Figure 4. Architecture of the enhanced Hearst algorithm

In 2010, Ben Hamadou et al. [30] proposed a rule-based multilingual extraction of functional relations between Arabic named entities using NooJ platform [31]. The relation of interest is PERS-ORG relation. This approach is composed of three main steps:

1. NER to detect PERS and ORG entities
2. Recognition of relation between the selected entities.
3. Generation of the predicate form illustrating the relation

NooJ resources were used to generate patterns and convert them into rules. Figure 5 shows an example of one of the patterns identified in the training corpus. First Order Logic is used to extract explicit relation in form of a predicate along with PER argument and ORG argument, for example: مديرعام (احمد اليمني، الشركة العالمية للصناعة), General Manager (Ahmed Alyamani, International Company for Industry). On the other hand, if the relation is implicit it's automatically classified as ينتمي - الي (Belongs_to) relation, for example (احمد السيد، كلية الهندسة), (Ahmed Elsayed, College of Engineering). For evaluation a journalistic corpus is used, the system achieved the following scores 63%, 78%, 70% for precision, recall and F-measure respectively. Due to the persence of long and complex organization names, the system did not perform well. However, this system avoided many problems including discontinuity of multiple relations regarding the same NE and discovery of implicit relations.

In 2014, Maha Al-Yahya et al. [32] proposed "Badea" system, a pattern-based approach to extract semantic relation using a seed ontology. The primary objective of this technique is the extraction of antonym pairs from a given corpus using a small set of antonym pairs (seeds). The architecture of "Badea" approach is shown in Figure 6. The initial step is pattern identification, in which the small set of seeds is used on corpus A to extract patterns. In the following phase, regular expressions are generated from the extracted patterns. Afterwards, a new corpus, B, is used. A pattern recognition algorithm is applied to B using the regular expression to extract new antonym pairs. The pairs extracted are manually checked to calculate the precision and pattern score, in order to evaluate each pattern. An existing Arabic language OWL ontology, "SemTree" [33] was used, it is composed of more than 100 Arabic synonym pairs and 70 antonym pairs. Corpus A, Arabic corpus arTenTen [34] was used which contains over 170 million sentences. For corpus B, the King Saudi University Corpus of Classical Arabic (KSUCCA) [35] was used. Almost 913 patterns were generated from corpus A, and, 733 correct antonym pairs were extracted

{<Title>} <PersName> {<P>} <REL> < ORG >
المهندس علي التويجري مدير المجمع الكمياني
Engineer Ali Al-Touijri Director of Chemical group
<Title> ***Engineer*** المهندس
<PersName> ***Ali Al-Touijri*** علي التويجري
<REL> ***Director-of*** مدير
<ORG> ***Chemical group*** المجمع الكمياني

Figure 5. Example of one of the main patterns identified in the learning corpus

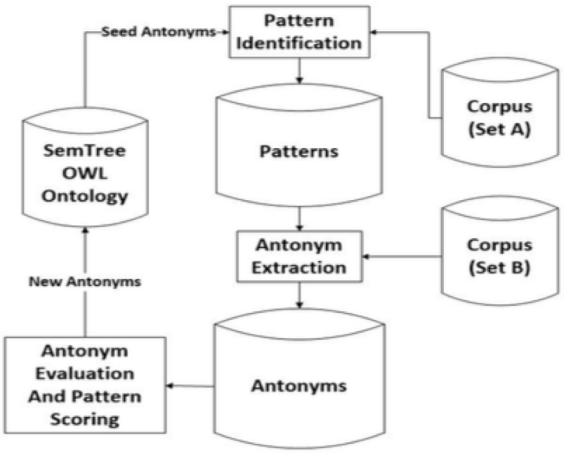

Figure 6. "Badea" system architecture

from corpus B. The precision score of this approach is 0.80%. The reason behind this low precision is the incorrect pattern scoring technique, this is considered the main draw back of "Badea" system. On the other hand, a large number of patterns were extracted but nevertheless, some improvements should be made concerning the pattern score in order to improve the precision score.

In the same year, Maha Al-Yahya et al. extended their previous work in [32] and proposed a pattern-based bootstrapping approach to automatically extract Arabic antonyms [36]. This is achieved using corpus analysis tool Sketch Engine [37]. Sketch Engine tool includes many Arabic corpora, in addition to Corpus Query Language (CQL) algorithm that is used to extract new antonym pairs. A new measure for pattern reliability was calculated from the number of antonym pairs that each pattern generates, and the co-occurrences of each pattern in the corpus. Human intervention was required to evaluate the antonym seeds collected before applying the bootstrapping approach to extract more pairs. The corpus used in this approach is arTenTen [34], which is the largest Arabic corpus available on Sketch Engine. In addition to the features and services provided by Sketch Engine, it also provides some statistics and scores, including, association, dice and LogDice [38] scores. LogDice score is considered a very reliable measure for relation detection between two words X and Y.

$$LogDice = 14 + Log_2 \frac{2frequency\ \mathrm{XY}}{frequency\ \mathrm{X} + frequency\ \mathrm{Y}} \quad (1)$$

Initially, an antonym seed set composed of 57 pairs is used on arTenTen corpus using CQL. The top ten occurring patterns are used in the subsequent step. Afterwards, the LogDice score is calculated on the initial seed set in order to set a threshold, the threshold calculated was 7.0. Next, patterns are extracted, only good patterns that co-occurred with many distinct antonym pairs are used. Then, pattern recognition is run on the arTenTen corpus, thus extraction more antonym pairs. Finally, bootstrapping approach is applied to extract more pairs. Any pair with a LogDice score below 7.0 were removed. Another score

was added in order to improve the precision, the co-occurrence of antonym patterns, if an extracted pair is generated by two or more patterns, thus it is considered a good pair. The results obtained were promising, 359 patterns were generated, their occurrence frequency ranged from5 to 4763, where patterns with frequency less than 100 were ignored. Using LogDice score is one of the advantages since it enhanced the performance of this algorithm. However, some patterns extracted were either too general or idiomatic patterns, which had a negative impact on the overall quality of the system.

A summary of the discussed supervised approaches is presented in Table 2, along with their advantages and disadvantages.

V. Unsupervised Relation Extraction

In unsupervised techniques, the learner is provided unlabeled examples, thus the evaluation is challenging at a large scale. A popular approach is building clusters of patterns expressing the same relation as in [39-42]. However, it's difficult to obtain a reliable set of patterns, that's due to the semantic representation of relational patterns and scalability to large data [43]. To the best of our knowledge, no work has been done in Arabic relation extraction using unsupervised techniques yet.

Nevertheless, multiple unsupervised techniques were carried out in English language. Takase et al. [43], applied approximate frequency counting and efficient dimension reduction to speed up unsupervised relation extraction. In addition, Eichler et al. [44], developed an information extraction system to produce a new information extraction system automatically using unsupervised relation extraction from web documents. Furthermore, Tseng et al. [45], proposed a Chinese open relation extraction for knowledge acquisition, to extract entity-relation triples from Chinese corpus.

VI. Conclusion

In this paper we have surveyed Arabic relation extraction researches, and provided a detailed analysis of the techniques applied and the results acquired. In addition, we explored the limitations and advantages of each approach. We started by reviewing supervised approaches, Mohanaed Falih and Nazila Omar [16] approach stood out with a F-measure of 93.48%, compared to RelANE system proposed by Boujelben et al. [13] that scored 85.23% F-measure. In Addition, Boujelben et al. [25] used genetic algorithm to extract relations, however, the previous approaches out preformed the genetic algorithm method. The main challenge in supervised approaches is generating the appropriate training set.

Semi-supervised approaches rely on a small set of seed instead of a training set, to extract the appropriate relation. Al-zamil and Al-radaideh [29] extract grammatical relations while Ben Hamadou et al. [30] relation of interest was PERS-ORG. Even though bootstrapping approach results were quite promising, error propagation due to wrong or too general patterns is a huge disadvantage as it affects the precision. "Badea" approach [32] suffered deeply from this problem. Maha Al-Yahya et al. avoided this problem in [36] by using LogDice score and calculating a score for each pattern based on the co-occurrence of the pattern.

In conclusion, with the ongoing advancements in the field of NLP, relation extraction gained a massive amount of attention in the past years. Nonetheless, there is still a room for improvement in Arabic relation extraction task. All the aforementioned studies discussed above are binary relations, future work can focus on extracting higher order relations. It is not an easy task to detect and extract a relation between Arabic named entities, there are still some challenges like relation discontinuity and implicit relations.

TABLE II. Summary of Semi-Supervised Approaches, their advantages, and disadvantages

Method	Extracted Relation	Acquired Preprocessing	Dataset	Results (%)	Advantages	Disadvantages
Al-zamil and Al-radaideh [29]	• Hyponym-Hypernym • Kind-of • Cause-effect • Has-A • Part-whole	POS-Stemming	Arabic Electronic News (1000 documents)	P=89.77 R=84.49 F=87	• Filtering is applied to overcome ambiguity caused by stemming and POS.	• The performance is negatively affected due to classification errors.
Ben Hamadou et al. [30]	• PERS-ORG	NER	Manually Constructed	P=63 R=78 F=70	• Ability to discover implicit relations.	• The performance negatively affected due to long and complex organization names.
Maha Al-Yahya et al. [32]	• Antonym Pairs	None	KSUCCA (50 Million Tokens)	P=0.8	• Ontological enrichment with over 400%.	• Unreliable calculation of pattern score. • No comparison with previous research.
Maha Al-Yahya et al. [36]	• Antonym Pairs	None	ArTenTen (6 Billion Tokens)	P=76	• Merged Sketch Engine with a semantic annotation tool to enhance the performance.	• Idiomatic patterns resulted in extracting the wrong relations. • No comparison with previous research.

REFERENCES

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

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

[3] M. Banko, M. Cafarella, S. Soderland, M. Broadhead, & O. Etzioni, "Open information extraction from the web," IJCAI '07: Proceedings of the 20th International Joint Conference on Artificial Intelligence. Hyderabad, India, 2007.

[4] K. Katzner, "The languages of the world," Routledge, London, 3rd edition. , 2002.

[5] A.E. 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)

[6] 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, October 14-18, 2013.

[7] 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.

[8] A.E. 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.

[9] 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.

[10] 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.

[11] M. El-Defrawy, Y. El-Sonbaty, and N. Belal, "Enhancing root extractors using light stemmers,", 29th Pacific Asia Conference on Language, Information and Compuatation (PACLIC 29), Shanghai, China, October 30- November 1, 2015.

[12] 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.

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

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. Witten, "The WEKA Data Mining Software: An Update; SIGKDD Explorations," Volume 11, Issue 1, 2009.

[15] C. Manning and D. Klein, "Optimization, Maxent Models, and Conditional Estimation without Magic." Tutorial at HLT-NAACL 2003 and ACL, 2003.

[16] 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.

[17] G. Jesus, and M. Lluis "SVM Tool: A general POS tagger generator based on Support Vector Machines," Proceeding of the 4the International Conference on Language Resources and Evaluation, 2004.

[18] M. Albared, N. Omar and M. Abd Aziz, "Classifiers combination to Arabic Morphosyntatic disambiguation," Proceeding of International Conference on Electrical Engineering and Informatics, Selangor, Malaysia, 2009.

[19] J. Dehdari, L. Tounsi and J. Van Genabith, "Morphological features for parsing morphologically-rich languages: a case of Arabic," In Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages, pp: 12-21, 2011.

[20] J.H Holland and J.S Reitman, "Cognitive systems based on adaptive algorithms," ACM SIGART Bulletin(63):49. 773, 1997.

[21] N. McIntyre and M. Lapata, "Plot induction and evolutionary search for story generation," Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Stroudsburg, pp. 1562-1572, 2010.

[22] H. Echizen-ya, K. Araki, Y., Momouchi, K. Tochina, "Machine translation method using inductive learning with genetic algorithms," Proceeding of the 16th COLING, Vol. 2, pp. 1020-1023, 1996.

[23] H. Echizen-ya, K. Araki, Y., Momouchi, K. Tochina, "Machine translation using recursive chain- link-type," Systems and Computers in Japan, Vol. 35, No. 2, pp. 1-15, 2004.

[24] H. Echizen-ya, K. Araki, Y., Momouchi, K. Tochina, "Study of practical effectiveness for machine translation using recursive chain-link-type learning," Proceeding of the 19th COLING, Vo. 1, pp. 1-7, 2002.

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

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

[27] Y. Benajiba, P. Rosso, J. Benedi, "ANERsys: An Arabic Named Entity Recognition system based on maximum entropy," CICLing, Springer-NNerlag, Berlin, Heidelberg, pp. 143-153, 2007.

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

[29] M. Al Zamil, Q. Al-Radaideh, "Automatic extraction of ontological relations from Arabic text," Journal of King Saud University - Computer and Information Sciences, 26(4):462- 472, 2014.

[30] A. Ben Hamadou, O. Piton, H. Fehri, "Multilingual extraction of functional relations between Arabic Named Entities using NooJ platform," International Conference and Workshop, Komotini, Greece, 2010.

[31] M. Silberstein "NooJ's dictionaries". Actes de la conférence internationale LTC, Poznan, Pologne, 2005.

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

[33] A. Al-Zahrani, M. Al-Dalbahie, M. Al-Shaman, N. Al-Otaiby, W. Al-Sultan, "SemTree: analyzing Arabic language text for semantic relations." , 2012.

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

[35] M. Alrabiah, A. Al-Salman, 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.

[36] M. Al-Yahya, L. Aldhubayi, "Automated Arabic antonym extraction using corpus analysis tool," Journal of Theoretical and Applied Information Technology, Vol.70 No.3., 2014.

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

[38] P. Rychly, "A lexicographer-friendly association score," Proceedings of Recent Advances in Stavonic Natural Language Processing, RASLAN, pp. 6-9, 2008.

[39] T. Hasegawa, S. Sekine, R. Grishman, "Discovering relations among named entities from large corpora," In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL), pages 415–422, 2004.

[40] Y. Shinyama, S. Sekine, "Preemptive information extraction using unrestricted relation discovery," In Proceedings of the Main Conference on Human Language Technology Conference of the North American

[41] Chapter of the Association of Computational Linguistics (HLT-NAACL), pages 304– 31, 2006.

[42] L. Yao, S. Riedel, A. McCallum, "Unsupervised relation discovery with sense disambiguation," In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL), pages 712–720, 2012.

[43] B. Min, S. Shi, R. Grishman, C. Lin, "Ensemble semantics for large- scale unsupervised relation extraction," In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 1027–1037, 2012.

[44] S. Takase, N.Okazaki, K. Inui, "Fast and large scale unsupervised relation extraction," PACLIC, 2015.

[45] K. Eichler, G. Neumann, "Unsupervised relation extraction from web documens," In 6th Conference on Language Resoruces and Evaluation (LREC'08)m Marrakech, Morocco, 2008.

[46] Y. Tseng, L. Lee, S. Lin, B. Liao, M. Lui, H. Chen, O. Etzioni, A.Fader, "Chinese open relation extraction for knowledge acquisition," In EACL, pages 12-16, 2014.