Journal of King Saud University – Computer and Information Sciences (2014) **26**, 425–440

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

A hybrid method for extracting relations between Arabic named entities

Ines Boujelben *, **Salma Jamoussi, Abdelmajid Ben Hamadou**

Miracl Laboratory, University of Sfax, Tunisia

Available online 28 September 2014

KEYWORDS

Hybrid method;
Relation extraction;
Named entity;
Machine learning;
Genetic algorithm;
Rule-based method

Abstract Relation extraction is a very useful task for several natural language processing applications, such as automatic summarization and question answering. In this paper, we present our hybrid approach to extracting relations between Arabic named entities. Given that Arabic is a rich morphological language, we build a linguistic and learning model to predict the positions of words that express a semantic relation within a clause. The main idea is to employ linguistic modules to ameliorate the results that are obtained from a machine learning-based method.

Our method achieves encouraging performance. The empirical results indicate that the hybrid approach outperformed both the rule-based system (by 12%) and the machine learning-based approaches (by 9%) in terms of the *F*-score, to achieve 75.2% when applied to the same standard testing dataset, ANERCorp.

1. Introduction

Given the enormous amount of Arabic electronic text, we note that there is a high frequency of named entities (NEs) that do not have any linked information. The recognition of these entities represents the first task toward building a semantic analysis and information extraction system. The second task consists of extracting semantic relations between the entities that are useful for a better understanding of human language. Thus, the second task constitutes a crucial move toward natural language processing (NLP) applications. This type of information enables the task of discovering a useful relationship or interaction between two entities from the content of the text. This approach has received a large amount of attention because it is used in many NLP applications, such as automatic summarization, web mining and question–answering (QA).In fact, the NEs' relations extraction can be exploited to extract more precise and correct answers. If we take the example "Where was Taha Hussein born?", the expected answer would be "**Taha Hussein** was born in **Al-Minya Governorate**". The relational triple is born-in (Person, Location), where Person and Location are the NEs.

Therefore, several studies on NE recognition have already been performed in many languages, such as English, French and Chinese. Additionally, many NE recognition systems have been built for the Arabic language. In the literature, three types of approaches have been proposed for Arabic NE recognition systems. Some of the proposed systems rely on handcrafted

* Corresponding author.
E-mail addresses: Boujelben_ines@yahoo.fr (I. Boujelben), jamoussi@gmail.com (S. Jamoussi), adelmajid.benhamadou@isimsf.rnu.tn (A. Ben Hamadou).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2014.06.004

rules, namely, the rule-based approach (Mesfar, 2007) and (Fehri et al., 2011). Other studies use a machine learning (ML)-based approach. They utilize a set of features that were extracted from an annotated corpus. In this context, (Benajiba and Rosso, 2008) and (Abdul-Hamid and Darwidh, 2010) have used Conditional Random Fields sequence labeling. (Benajiba and Rosso, 2008) reported 90%, 66% and 73% *F*-measures for the location, organization and persons, respectively. (Abdul-Hamid and Darwidh, 2010) achieved an improvement in the *F*-measure over (Benajiba and Rosso, 2008) for recognizing persons and organizations, by 9 points and 2 points, respectively. Finally, a few studies in Arabic NE recognition have used a mixed approach. We mention (Shaalan and Oudah, 2014), who concentrated on a hybrid approach. Because of their combination of rule-based and ML-based approaches, these authors achieved a 90% *F*-measure. Their system outperforms the state-of-the-art for Arabic NER in terms of accuracy when applied to the ANERCorp[1] standard dataset.

However, the results reported in the NE relation extraction task were not as good as those achieved in the NE recognition task. For this task, only a few studies have addressed the Arabic language. We notice (Ben Hamadou et al., 2010a), whose approach is based on patterns that were rewritten into local grammar within the linguistic platform NooJ. They aimed to extract functional relations between persons and organizations. Additionally, (Alotayq, 2013) adopted the learning classifier MaxEnt to extract relations between various types of NEs. To the best of our knowledge, there is no study that has adopted a hybrid approach to discover the relations between NEs in the Arabic language. Thus, it will be challenging to adopt this approach for extracting the relations between NEs in the Arabic language.

In this paper, the relations between Arabic NEs are tackled through developing a hybrid system to combine the advantages of ML- and rule-based approaches. Mainly, an ML approach followed by a post-processing rule-based approach is used in an attempt to enhance the overall performance of the ML system. Our aim is to predict the trigger words that express the semantic relations between NEs from Arabic text, relying on a set of rules. First, our system is based on ML algorithms to extract the rules using a decision tree technique and an Apriori algorithm. Then, a genetic algorithm (GA) is used to extract and generate the most significant and interesting rules. After applying ML methods, we added hand-crafted rules to treat both invalid examples and unseen relations.

The remainder of this paper is organized as follows: First, we survey prior studies on relations extraction. Section 2 provides background on relations between NEs. Then, we explain the relation extraction task as well as the different challenges. The fourth section illustrates the architecture of our hybrid process, in which we detail the main steps of our proposed method. Afterward, we present the different experiments from which we discuss the reported results.

2. Related studies

Today, relation extraction that involves NEs is seen as a step toward a more structured model of text meaning. Several methods have been proposed to extract semantic relations between NEs. These methods can essentially be classified into three broad categories: the rule-based approach, ML-based approach and hybrid approach.

2.1. Rule-based approach

In the first approach, the rules are usually implemented in the form of regular expressions or finite-state transducers. From the studies performed in the Arabic language, we mention (Ben Hamadou et al., 2010a) and (Boujelben et al., 2012).These authors extracted a set of linguistic patterns from a training corpus. Subsequently, they rewrote those patterns into finite state transducers within the linguistic platform NooJ,[2] using specifically local grammars.[3] This approach uses a representation of linguistic rules by means of transducers.

(Ben Hamadou et al., 2010a) reported an *F*-score of 70%, while (Boujelben et al., 2012) achieved an *F*-score of 60%. This result is significant because (Ben Hamadou et al., 2010a) is limited to only the functional relations between the NE pairs (PERS–ORG). Thus, they concentrate solely on one NE pair, which enables them to construct more precise and concise rules. In contrast, (Boujelben et al., 2012) are interested in extracting more relations among five pairs of NEs (PERS–LOC, PERS–PERS, PERS–ORG, ORG–LOC and LOC–LOC). To extract the relations between these NE pairs, the authors elaborated five sub-grammars. Each grammar contains the pattern of relations between each pair. The system considers the gender and the number features of the relation triggers when it verifies whether the NEs are related. Because of these NooJ grammars, their process enables the extraction of semantic relations that are predicted through one or multiple word forms that appear before, between, or after the NEs.

The rule-based method offers a significant analysis of the context for each NE and its relations with the other NEs. However, the complexity of Arabic sentences and the high variability in the expressions used make it intricate to detect some of the relations between the NEs. To accomplish that goal, a tangible effort is required to write down all the rules for discovering relations between NEs. To overcome this manual step, some studies, such as (Ezzat, 2010), are oriented to a semi-automatic method for automatically producing recognition grammars for relation detection between NEs. These grammars present a set of patterns that are provided by an algorithm. The algorithm relies on generalizing a large collection of sentences that contain the relevant relation. These sentences are collected by a linguist or a domain expert.

2.2. Machine learning-based approach

To fully automate the relation extraction task, some research studies have been oriented toward ML methods, including un-supervised, semi-supervised and supervised learning techniques.

The un-supervised methods make use of massive quantities of unlabeled text and are based almost entirely on clustering

[1] Available on http://www1.ccls.columbia.edu/~ybenajiba/downloads.html.

[2] Available on http://www.nooj4nlp.net.

[3] NooJ local grammars are typically used to describe sequences of words that present meaningful units or entities. In fact, these grammars can be used to locate syntactic constructions of interest, such as sentences that contain certain grammatical words or syntactic constructs.

techniques and similarities between features or context words. For example, (Hasegawa et al., 2004) focused on clustering NE pairs according to the similarity of the context words that intervene between the NEs. These authors did not account for the context words before and after the NEs. However, these two contexts can introduce helpful information to discover the semantic relations among the NEs. Furthermore, the authors consider relations whose contexts are of the same type. In the same context, (Zhang et al., 2005) computed the similarity between two parse trees, to cluster them using the hierarchical clustering model. Each obtained cluster is labeled, and some bad clusters whose NE pair number is under a predetermined threshold are discarded. In these the authors reported a 90% precision and an 84% recall. Additionally, (Hassan and Emam, 2006) have relied solely on redundancy to select informative patterns for extracting information. Such approaches require a high frequency of NE pairs to be efficiently constructed, which is not the case for the majority of relations that are defined in running text.

To remedy the problems with the unsupervised approach, some studies have been oriented toward semi-supervised learning approaches or bootstrapping methods. This approach relies on a small set of initial seeds. A sample of linguistic patterns or some target relation instances can be used to acquire more basic relations until discovering all the target relations, such as in (Zhou et al., 2009) and (Zhang, 2004).

A last approach under the ML techniques is the supervised method, which relies on a fully labeled corpus. This approach considers relation extraction as a classification task. Among the most often used supervised techniques, we mention Support Vector Machines (SVM), Conditional Random Fields (CRF), decision tree and maximum Entropy (MaxEnt). A recent attempt to extract the relations between Arabic NEs has been made by (Alotayq, 2013), who used a classifier that was based on MaxEnt. Based only on morphologic and part-of-speech (POS) information, this system achieves satisfactory results when applied to the ACE[4] corpus. Other studies are based on a combination of supervised techniques. Indeed, (Celli, 2009) has combined two supervised techniques, namely, the simple decision tree and PART decision list algorithms, to extract three semantic relations (role, social and location) between NEs. These authors relied on the POS of the context before and between the two entities only, without considering the context after the NEs. They reported an *F*-score of 81.2% when applied to the I-CAB[5] data. Finally, some other studies have been based on association techniques to discover patterns from text data. Based on the dependency graph that is generated by syntactic analysis, (Kramdi et al., 2009) adopted the learning pattern algorithm LP^2 that was proposed by (Ciravegna and Wilks, 2003) to generate annotation rules. They obtained an *F*-score of 50%. The resulting patterns that were produced by such a method often suffer from low precision.

Another study that adopted the learning rules method was performed by (Boujelben et al., 2013a).These authors seek to use the association rule algorithm Apriori (Agrawal et al., 1993). This mining rule model aims at finding all the rules from a database that satisfy minimum support and minimum confidence values (see Section 5.2). To cover more instances of the training dataset; they further used the decision tree technique C4.5 (Quinlan, 1993). Although they combined these two mining techniques, they added four selection levels including filtering and enrichment of the obtained rules to extract the more interesting rules, and they obtained a low recall rate. As a continuation, (Boujelben et al., 2013a) proposed a genetic process with the aim of extracting the best set of rules. These rules are either provided by learning methods or produced by genetic operators such as crossover and mutation (see Section 5.2). The main advantages of supervised relation NE systems are that they can be applied to other domains and languages. Additionally, their update is conducted with minimal time and effort, in cases in which a sufficient data base is available.

2.3. Hybrid approach

The two categories of approaches described above can be combined to obtain a mixed approach. Recently, research studies have been oriented toward the use of hybrid approaches because such an approach achieves an enhanced performance that is better than either the rule-based approach or the ML-based approach alone. Some studies have been performed on a specific domain, such as in the biomedical field. As an example, (Ben Abacha and Zweigenbaum, 2011) propose a hybrid approach to extract relations between diseases and treatments. These authors combined a supervised learning method with a rule-based technique. For the linguistic method, a set of patterns is constructed manually from the training corpus and from other MEDLINE[6] corpora; in this set, a weight is associated with each pattern. This weight serves to choose the more convenient pattern in the case of multiple extraction candidates in the hybrid method. For the ML method, the authors investigated the SVM classifier, using lexical, morph-syntactic and semantic features. The obtained results of this hybrid approach show an enhancement toward the ML- and pattern-based methods. Recently, (A. Kadir and Bokharaeian, 2013) combined three methods, which are the co-occurrence, rule-based and kernel method, to extract both simple and complex relations in the biomedical domain. The authors used Kernel-based algorithms to map the data into a high-dimensional feature space. Moreover, they relied on the occurrence of two NEs together within the text.

The achieved studies using the hybrid approach were developed in English and some European languages. However, there is no study that was developed in the Arabic language. Drawing inspiration from the main idea of these methods, we propose our novel process, which is based on a hybrid approach and aims at detecting relations between Arabic NEs. Our method is distinct from the mixed proposed approaches in that we did not exploit the entire rule-based method. We added only some handcrafted rules or linguistic constraints to the rules that are produced by ML techniques for the two main objectives. We treat the quiet instances or the noise produced by the proposed ML model by adding some grammatical constraints to exclude ambiguous and invalid relations. Additionally, we plan to enhance the quality and the accuracy of our system output.

[4] http://www.ldc.upenn.edu/projects/ACE/.

[5] Italian Content Annotation Bank: an Italian corpus composed of 525 news documents taken from a local newspaper called "LAdige", annotated with temporal expressions and 4 named entity types (person, organization, location and geo-political entity).

[6] http://mbr.nlm.nih.gov/Download/.

3. Arabic named entity relations challenges

In contrast to the significant achievement concerning English or some European languages, the research progress on Arabic relation discovery is relatively limited. This limitation might be due to the nature of the Arabic language as well as the lack of available linguistic resources. Indeed, the accessible corpora are not annotated with NEs, and the relations do not include a sufficient number of annotated examples to be exploited for learning approaches. Arabic is a Semitic language that presents interesting morphological and orthographic challenges that could complicate the extraction of relations between NEs. In addition to the problems that are related to Arabic NE recognition (Ben Hamadou et al., 2010b) and are cited in (Abdul-Hamid and Darwidh, 2010) and (Mesfar, 2007), the relation extraction task poses some specific challenges. Some of these challenges are listed in (Ben Hamadou et al., 2010a). We present further goals, which are summarized in Table 1.

The challenges listed above should be considered to achieve an efficient system for extracting relations within NEs.

4. Background

First, the NE relation extraction has been considered to be an information extraction task in (MUC-6).[7] Then, a template relation task is proposed in (MUC-7) that is based on only three relations. Subsequently, the Automatic Content Extraction (ACE) introduced the relation extraction as a relation detection and characterization task (RDC). In this task, five general relation types that yield a total of 24 types/subtypes of relations are proposed. In our case, a relation is defined as a semantic interaction that holds between NEs.

In contrast to some prior studies that take for granted that a relation can be recognized through a verb, we assume that a relation can be expressed through context words that can have different POS tagging. Furthermore, a relation among NEs can be either introduced directly through words from a context or expressed implicitly from a context or the mining of a sentence. Additionally, unlike previous studies such as (Zelenko et al., 2003) and (Hasegawa et al., 2004), which account for only the words between NEs, the Arabic relations can be detected before the first NE, between NEs or after the second NE.

For example, consider the following circumstance:

[7] سافر **أحمد حلمي** الى **السعودية**.

sAfr > **Hmd Hlmy** AlY **AlsEwdyp**.
Ahmed Helmi traveled to **Saudi Arabia**.

Here, a relation is given by the verb "سافر/traveled", which is located in the first word between the two NEs. "**Ahmed Helmi**" is the first argument, and "**Saudi Arabia**" presents the second argument. This phrase can be rewritten as (**أحمد حلمي يريد السفر** الى **السعودية**/**Ahmed Helmi** wants to travel to **Saudi Arabia**), and here, a relation can be predicted through the noun "**سفر**/travel", which is situated in the second word after the first NE. The same example can be rewritten as the following clause: (عند **أحمد حلمي** إلى **السعودية**، **التقى بصديقه** ذهاب/End *hAb > Hmd Hlmy <lY AlsEwdyp, AltqY bSdyqh/When he traveled to **Saudi Arabia**, **Ahmed Helmi** met his friend). Hence, the relation can take a variety of categories and positions for the words in the sentence. On the other hand, semantic relations between Arabic NEs are too numerous to be accounted for. For these reasons, we decided to extract first the relation words or triggers by extracting their corresponding positions in a given sentence. We aim at predicting which word defines a semantic relation between automatically recognized NEs. For our case, we focus on extracting binary relations, and a relation is predicted through a word from the sentence. Examples are born-in (Tah Hussein, Pittsburgh) and travel (Ahmed Helmi, Saudi Arabia).

The main goal of our method is to provide rich semantic annotation of Arabic text. Thus, besides the morpho-syntactic annotations of each word and the semantic annotation of Arabic NEs, we plan to identify the trigger words in the sentence that can predict semantic relations that involve a given pair of NEs in a sentence.

5. Proposed method

While most recent research on relation extraction focuses on semantic relation classes that occur within one sentence, our main goal is to extract the position of the word that expresses a semantic relation between NEs. We combined the ML method to automatically extract rules that are based on the GA with some linguistic modules for the goal of improving the overall performance of the ML method.

The proposed hybrid approach is motivated by the peculiarities of the Arabic language. For the sake of simplicity and clarity of our method, we restrict our discussion to binary relations between two NEs (PERS, ORG and LOC). The proposed method is illustrated in the following figure.

Fig. 1 shows our approach's working model divided into three general levels of processing. The first level involves a preprocessing step that produces the training data set. The second level concerns the rule-mining process in which we applied learning algorithms to our training data. These rules present the initial population of the GA, which is used to extract the best set of rules. Afterward, in the third level, we introduced further linguistic models, which are presented as Hand Crafted rules and linguistic constraints to be handled with rules generated from the GA, to improve the consistency and accuracy of the captured relation triggers.

5.1. First level: build training data

There are few publicly available annotated corpora that have the required information, especially for the Arabic language. In fact, many available corpora are neither annotated with NEs nor include a sufficient number of related NEs. This lack of Arabic linguistic resources makes the ML technique, notably the supervised model, difficult to use. As far as we know, the only annotated corpus with Arabic relations is the ACE corpus.[8] However, this corpus is not yet available. Therefore, we drive to construct and annotate our own corpus using the available tools for the Arabic language. We have constructed our corpus from different resources to obtain a representative

[7] Sixth message understanding conference.

[8] http://projects.ldc.upenn.edu/ace/.

Table 1 Challenges of NE relations between Arabic NEs.

	Challenges	Examples
Arabic challenges	The agglutination: a stemcan be attached to prefixes (articles,preposition,conjunction,…) and/or suffixes (linked pronouns) in different combinations.	**وبزواجها /?wbzwAjhA/ And with her marriage** Here, the conjunction(and/ و), the preposition (with/ب)and the pronoun (here/) are agglutinated to the noun زواج/ marriage "".
	The absence of capitalization in Arabic texts, which can prevent the NE recognition task.	**تونس /twns/ Tunisia**
	The absence of vowels inArabic text can produce ambiguity in analyzing the word.	The unvowelled word "**حسن**" can refer to a person NE "**Hasan"**, a verb that means "**to improve"**, oranadjective**"good".**
NE relations challenges	**A)**The omission of one element of the relation between NEs (NE1, R, ?), (?, R, NE2) or (NE1,?,NE2).	**[1]** إن المتظاهرين المؤيدين للديمقراطية قداحتفلوا برحيل **صالح بن علي**. *The pro-democracy demonstrators had celebrated the departure of Saleh Ben Ali.* ➔We have a person NE "Salah Ben Ali" and the relation "depart", but we do not know to or from where. Hence, we note the omission of a location NE.
	B) Relation among NEs, but its mere presence does not mean the occurrence of a relation in the given sentence.	**[2]** قال السيد **مات** أن **اليونسكو** كانت تعمل على تعزيز وسائل الاعلام المحلية. [a]qAl Alsyd **mAt**>n **Alywnskw** kAnt tEml ElY tEzyz wsA}l AlAElAm AlmHlyp. **Mr. Maat** said that **UNESCO** aimed at strengthening indigenous media. ➔ «**مات**» and «**اليونسكو**» are not related despite their presence in the same sentence
	C) Implicit Relations: they are not directly recognizedthrough wordsin the text. They are mined from the text using contextual elements.	**[3]** **محمد قاسم,كلية الطب** حصل على الدكتوراة. *Mohammad Qasim, Faculty of Medicinegot doctoral degree*. ➔The relation (Mohammad Qasim, Faculty of Medicine) is not explicitly indicated in the sentence. However, we can deduce that the person NE "Mohammad Qasim" belongsto "the Faculty of Medicine" (Ben Hamadou et al., 2010).
	D) Negatively defined relation.	**[4]** ليست **كريستين**بزوجة**مايكل.** lyst **krystyn**bzwjp **mAykl** **Kristin** is not **Michel's** wife. ➔The triplet wife (Kristin, Michel) is in negative form.
	E) Multiple relations between the same NE pair.	**[5]** **تكلم محمد قاسم** مع أخيه **أحمد.** *Mohammad Qasim speaks with his brother Ahmed*. ➔ Two binary relations are presented between the pair (*Mohammad Qasim, Ahmed),* which are detected through "speaks" *and "brother".*
	F)Ambiguous relations between NE pairs.	**[6]** ذهب **أحمد** و أخي **صالح** إلى المدرسة *hb **>Hmd** w >xy**SAlH**<lY Almdrsp. **Ahmed** and my brother**Salah** are going to the school.

[a] All the examples in this paper are given in Arabic along with their English translation and their transliteration using Buckwalter1.1.

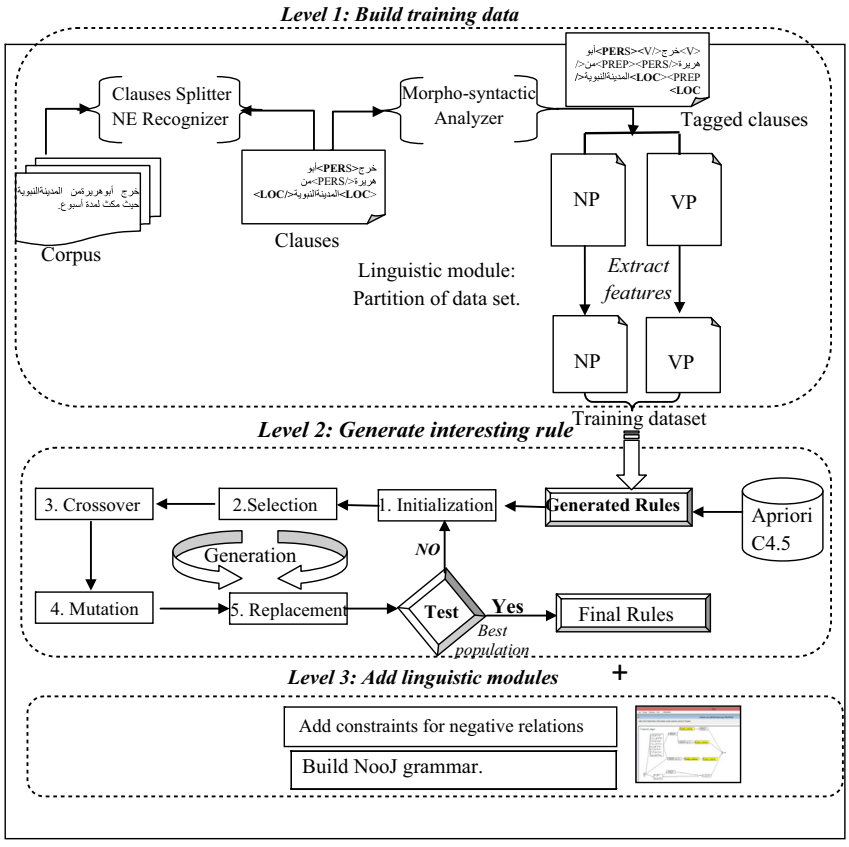

Figure 1 The architecture of our hybrid method.

number of examples. Our training corpus was gathered from various sources of Arabic electronic newspapers, such as "البيان/AlbyAn", "الشروق/Al$rwq" and "الحياة/AlHyAp" and from Wikipedia.[9] Our corpus is composed of 1465 texts, 5930 paragraphs, 17 702 sentences, 90 105 words, 9760 NEs (2200 LOC, 2430 ORG and 1843 PERS, and 3287 for other categories).

As a first step, we extract only the sentences that contain at least two NEs because we aim to discover binary relations between them. Therefore, we utilize the Arabic tool of NE recognition that was elaborated by (Mesfar, 2007). The present work focuses on the possible relations between a couple of NEs from Person (PERS), the Location (LOC) and Organization (ORG). The choice of these three types of NEs is motivated by the importance and the high frequency of these three types in both electronic texts. As a result, we obtain 2450 phrases. When studying these phrases, we observe some NEs that are not related despite their presence in the same sentence (see example [1] Table 1). Such examples can undoubtedly propagate ambiguities to the subsequent processing of our relation extraction task. It is therefore reasonable to avoid this problem by excluding these examples. Indeed, Arabic text is characterized by the lengths of the sentences and by a complex syntax. To alleviate this problem, (Riedel et al., 2010) used a factor graph to verify whether two NEs are related. In this graph, they created a relation variable for each pair of NEs. These entities must be mentioned together in at least one sentence. For each pair, they create one relation mention variable, and they connect it to the corresponding relation variable. For our case, we split sentences into clauses. A clause is composed of a set of words that contains a subject and a predicate. Hence, a clause can be presented also as a sentence. This extraction required Arabic clauses splitter as well as Arabic NE recognition tools. We use the Arabic splitter elaborated by (Keskes et al., 2012), which segments Arabic sentences into clauses based on a cascade of local grammars within the NooJ platform.

Because of this module, we can partially resolve the problem of unrelated NEs and ensure the presence of NE relations in a given clause. For example, if we consider example [1] after splitting the phrase into clauses, we obtain these two clauses: (قال السيد مات/qAl Alsyd mAt/Mr Maat said) and (اليونسكو كانت تعمل على تعزيز وسائل الاعلام المحلية أن/ > n Alywnskw kAnt tEml ElY tEzyz wsA}l AlAElAm AlmHlyp / that UNESCO aimed at strengthening indigenous media.). Because these clauses do not contain two NEs, they will be excluded from our training dataset.

The resulting clauses that were produced by this Arabic splitter contained at least two NEs and were then annotated to extract the relevant features. These features are presented in Table 2.

As mentioned in Table 1, we compiled three types of features to describe the dataset:

- Numeric features that introduce the number of words before, between and after the NEs.
- Morpho-syntactic features that indicate the POS tag of three words of each context. We have added another

[9] http://www.wikipedia.org/.

Table 2 Used features.

Type	Feature		Description	Value
Semantic	NE1		The first named entity tag	PERS, LOC, ORG
	NE2		The second named entity tag	PERS, LOC, ORG
	PAIR		The appearance order of NEs	PERS–LOC, LOC–PERS, PERS–PERS, PERS–ORG, ORG–PERS, PERS–PERS, LOC–LOC, ORG–ORG
Numeric	N-W-C1		The number of terms before NE1	Number
	N-W-C2		The number of terms between NEs	Number
	N-W-C3		The number of terms after NE2	Number
Morpho_syntactic	Clause structure		The clause structure	Nominal clause (NC) and verbal clause (VC)
	C1	POS-W1-C1	The part of speech tag of the first word before NE1	Verb (V), noun (N), adjective (A), determiner (DET), preposition (PREP), punctuation (PONCT), negative particle(NEG), adverb (ADV), pronoun (PRON), pseudo-Verb (PSV), and NE
		POS-W2-C1	The part of speech tag of the second word before NE1	
		POS-W3-C1	The part of speech tag of the third word before NE1	
	C2	POS-W1-C2	The part of speech tag of the first word between NE1	Verb (V), noun (N), adjective (A), determiner (DET), preposition (PREP), punctuation (PONCT), negative particle(NEG), adverb (ADV), pronoun (PRON), pseudo-Verb (PSV), and NE
		POS-W2-C2	The part of speech tag of the second word between NEs	
		POS-W3-C2	The part of speech tag of the third word between NEs	
	C3	POS-W1-C3	The part of speech tag of the first word before the second NE	Verb (V), noun (N), adjective (A), determiner (DET), preposition (PREP), punctuation (PONCT), negative particle(NEG), adverb (ADV), pronoun (PRON), pseudo-Verb (PSV), and NE
		POS-W2-C3	The part of speech tag of the second word before the second NE	
		POS-W3-C3	The part of speech tag of the third word before the second NE	

syntactic feature, which is the clause structure. This feature serves to determine whether the clause is verbal or nominal. The utility of this feature is explained in the evaluation section.

– Semantic features, including the semantic type of the NE and the type of the NE pairs.

After the recognition of all the NEs via the Arabic NEs recognizer (Mesfar, 2007) and its revision, we acquired the POS tags based on the different Arabic resources (Mesfar, 2006) (dictionaries and local grammars). Herein, we have added a NooJ grammar to simplify the POS tags into twelve categories, as mentioned in Table 2.

All the numeric, morpho-syntactic and semantic features were automatically extracted from annotated clauses except for the relations between the NEs. Indeed, the relation is manually annotated by three Arabic linguistic experts. We provided them with a detailed description of our relation extraction task as well as our main goal. They were asked for predicting which word can define the semantic relations between the NEs within a clause. The inter-annotator agreements are computed, from which we obtain the promising Cohen kappa of 79%. The main disagreements came from some examples in which a relation cannot be predicted directly from words, namely implicit relations. Furthermore, some ambiguities are raised when a sentence presents multiple relations between the same NE pair. Finally, some relations are expressed through more than one word, which poses little disagreement between our linguistic annotators.

Afterward, the dataset file was built and transformed into XML format. Hence, we have a semi-automatically tagged corpus. Once these features are assigned, we can build our data base, which is presented as a set of pairs (an attribute or feature and its corresponding value) and a class label. We built our training data base, which is composed of a set of instances. Each instance presents a set of pairs (an attribute or feature and its corresponding value) and a class label. We call each pair (an attribute and its value) an *itemset*.

In case a sentence or clause contains more than two NEs, we duplicate such sentences to have multiple clauses that are annotated by only one relation position word and one pair of NEs. This point is illustrated in the following example:

[8] سيصل رئيس **فنزويلاهوغو شافيز**إلى **روسيا البيضاء** هذا اليوم.

sySl r}ys **fnzwylAhwgw $Afyz**<lY **rwsyA AlbyDA**' h*A Alywm.

The president of **Venezuela, Hugo Chavu,** will arrive today to **Belarus**.

As shown in this example, three different NEs are presented: two LOC NEs "فنزويلا/fnzwylA/Venezuela" and "روسيا البيضاء/rwsyA AlbyDA'/Belarus" and the PERS "شافيز هوغو/hwgw $Afyz/Hugo Chavu". In our training corpus, this sentence will be duplicated to obtain two instances: the first instance presents the relation "رئيس/r}ys/president" between "هوغو شافيز/Hugo Chavu" and "فنزويلا/fnzwylA/Venezuela", and the second instance identifies the relation

Table 3 Number of instances of each class.

Relation position	Number of instances
W1C1	**381**
W2C1	**265**
W3C1	42
W4C1	22
W1C2	**414**
W2C2	**236**
W3C2	**60**
W4C2	12
W1C3	**55**
W2C3	26
W3C3	3
Others	24
Total	1540

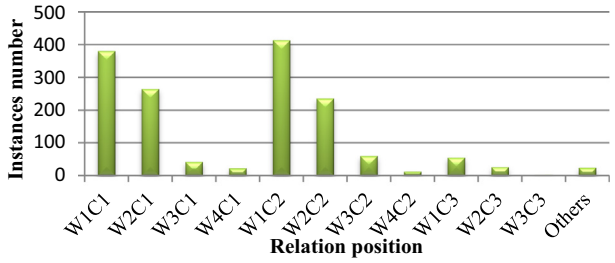

Figure 2 Number of instances of mentions for each class.

expressed through "يصل/arrive" between "هوغو شافيز/Hugo Chavu" and "روسيا البيضاء/rwsyA AlbyDA'/Belarus".

From our initial corpus, we obtained only 8345 sentences (from 17 702 sentences) that contain at least two NEs. After splitting these sentences into clauses, we have only 3302 NEs that are related. Because we are interested in only the NEs of the types (PERS, LOC and ORG), we have at least 1200 NEs (LOC), 880 NEs (ORG) and 1222 NEs (PERS). As a consequence, our data are composed of 1651 instances.

We present in the following table the number of instances that are available for each class. As mentioned previously, our class presents the position of a relation between the NEs in a given clause. The class label is composed of two parts (W*n* C*m*): W*m* designs the word position in the context, and C*m* refers to the context (*m* can be 1 if the context is before the first NE, 2 if the context is between the two NEs, and 3 if the context is after the second NE). For example, the class label of the example [1] is W1C2. That label means that the relation is indicated through the first word (W1) of the second context (C2), which refers to the word "travels".

As shown in the Fig. 2 and Table 3, the W1C1, W1C2, W1C2 and W2C2 classes are the most frequent in our data. There are some classes that have a small number of instances, such as W4C1 and W3C3. Hence, the small number of examples that correspond to these classes does not allow an efficient learning-driven extraction. Because we are basing this step on a clause and not on the entire sentence, we focused on short clauses that were composed of up to 10 words. Therefore, the prediction classes were reduced to six relation positions: W1C1, W2C1, W1C2, W2C2, W3C2 and W1C3.

5.2. Second level: automatic rule extraction

A rule is defined as a conditional statement that can be easily understood by humans and easily used within a database to identify a set of records. Therefore, we rely on automatically extracted rules to discover the trigger word that predicts a relation between NEs. Basically, a rule presents regular expressions that describe a set of target clauses that contain Arabic entities at specific positions in a more or less specific lexical context. Each rule consists of a sequence of itemsets (a feature and its corresponding value) that must be verified to be in accordance with the convenient class. Let us consider the following rule:

Rule1: **If** NE1 = PERS and NE2 = LOC and pos-w1-w1 = V and nb-w-C1 > = 1 and pos-w1-c2 = PREP and nb-w-c2 = 1 **Then,** class = W1C1

This rule can be applied to clauses [1] and [9]:

[9] وقد يذهب أحمد حلمي الى السعودية اليوم

wqd y*hb > Hmd Hlmy AlY AlsEwdyp Alywm.
Perhaps Ahmed Helmi may go to Saudi Arabia today.

When applying this rule (Rule1) to examples ([3] and [4]), we deduce that a relation is located in the first word of the first context (W1C1). This finding means that the words "سافر/sAfr/traveled" and "يذهب/y*hb/go" present the trigger verbs that predict the relation between the NEs.

We associate efficiency measures with each rule, namely, the confidence and support. The confidence shows how frequently the rule head occurs among all the groups that contain the rule body. The support presents the number of instances in which a rule is applicable, regardless of whether it is correct or false.

5.2.1. Generating rules using ML algorithms

For the extraction of such rules, we investigated the Apriori algorithm (Agrawal et al., 1993) to generate the class association rules because of its known performance. The Apriori technique aims at finding all the rules that exist in the database that satisfy some minimum support (minSup) and minimum confidence (minConf) constraints.

In addition to the Apriori algorithm, we explore the decision tree technique to produce more heterogeneous rules. The decision tree C4.5 algorithm (Quinlan, 1993) is preconized because it can match other instances that are not covered by the training data and cannot be provided by Apriori. In addition, the decision tree has been among the most powerful and popular classifiers, as stated in some studies (Jantan et al., 2010; Tso and Yau, 2007). (Celli, 2009) proved its efficiency at extracting semantic relations between Italian NEs because it obtained an *F*-score of 81.2% when applied to the I-CAB data. The decision tree algorithm chooses an attribute to maximize the separation between the classes (using an information gain criterion). This algorithm generates classifiers that are expressed as decision trees.

Similar to association rules, we can derive rules from a decision tree. The results can be converted into a set of rules that have the form of "if Attribute1 = value1 and Attribute2 = value2 and Attribute3 = value3 then class X". We obtained, then, a set of rules by running these learning

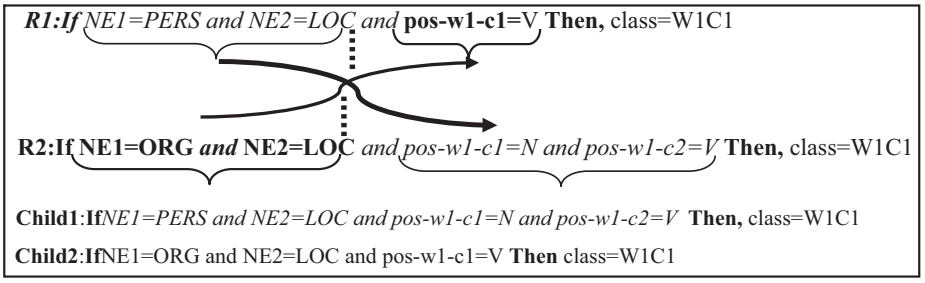

Figure 3 Illustration of the single point crossover operator.

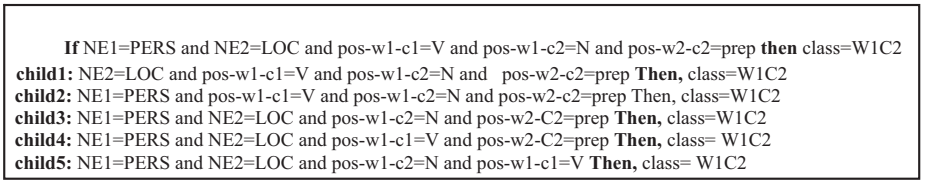

Figure 4 Illustration of the mutation operator.

algorithms in our training data .Nevertheless, the application of these algorithms produces an important number of rules, which can be in some sense interesting or not. Therefore, it is mandatory to filter these multiple rules using a filtering module. To fully solve this problem, we seek to apply genetic operations to these rules, in an attempt to cover further instances and to enhance the precision of our process.

5.2.2. Discovering the best rule sets using a genetic algorithm

Seeing that genetic algorithms (Holland, 1970) have been successfully applied in many research optimization and ML problems, recently, a number of studies have been conducted using evolutionary algorithms for mining rules. Inspired by the ASGARD[10] system (Jourdan et al., 2002), we adopted this process to automatically extract the more interesting rules. In our genetic process, we used the Michigan approach.[11] This approach considers a separate rule as a chromosome or as an individual. The main idea is to progressively improve the quality of the initial rules by constructing new fitter rules until either rules of high quality are found or no further improvements are recorded. The adopted GA process is explained in detail in (Boujelben et al., 2013b); it can be summarized by the following steps:

- Evaluate the fitness of each chromosome in the population in terms of confidence, support and size. The selected rules will participate in producing the next generation.
- Create a new population by repeating the following steps until the new population is completed:

The Crossover operator: We randomly select two parents that have the same class. Given these, we use a classical single point crossover: a position is randomly chosen in each of the parents, and the two corresponding parts of the parents are exchanged to form two children. This genetic operator is illustrated in Fig. 3.

The mutation operator (see Fig. 4): According to the mutation probability ($P_{mut}(R)$ = (1−confidence(R))/10), a new offspring is generated. For each rule, we remove one item (the attribute and its value), and we retain the remainder of the rule to obtain the derived rule. This process is applied for each item. The rules are then used from the most specific to the most general in the relation extraction task.

Both genetic operators appear to be complementary because the first operator enables us to explore new areas while the second covers more instances of our dataset through generic obtained rules.

- The replacement operator: We compare each source rule with its offspring to satisfy two main assumptions: (i) each derived rule that holds with a confidence value of more than a specified threshold and obtains a support that is higher than the support of the top rule will be selected. (ii) In the case in which all the derived rules have confidence values that are below the threshold value, we will conserve only the target rule and eliminate all the derived rules.
- We then re-insert these descendant rules into the initial population to create a new population.

The GA runs to produce solutions over successive generations until either interesting rules are found or a stagnation of the population's evaluation is reached. Otherwise, a fixed maximum number of generations is reached. As a result, the GA generates a population that, in the end, has high quality rules. The rules generated for each generation will be sorted in terms of confidence and support. In case we have similarity between these measures, we select the longer rule in terms of the itemset number, to obtain more accurate rules.

In the next section, we will explain the different modules that are integrated into this ML process based on GA to produce more concise and precise rules.

[10] Adaptive steady state genetic algorithm for association rule discovery.

[11] There are two main approaches for a GA: the Michigan approach, in which a separate rule is handled as an individual, and the Pittsburgh approach, in which each rule set is handled as an individual.

5.3. Third level: enhancement models

To overcome the above problems listed in Table 1, which cannot be resolved by the automatic rules provided by learning algorithms, we propose some auspicious modules that can be integrated within the ML model to boost the overall performance of the ML process.

5.3.1. Partition of the data set into verbal and nominal sentences

Because our goal is to extract the relation between two Arabic NEs, we focus on how to collect useful information that is related to this task. In addition to the POS tagging of context words, the position of the relation word can be dependent on the clause structure in the Arabic language. This approach could provide more concise and precise results. Previous studies such as (Haddad, 2003) proved the effectiveness of using the phrase structure to represent the text's content, which can enhance the efficiency of the information extraction process. Thus, the clause structure provides a more sophisticated representation. In the same way, (Smeaton, 1995) demonstrated that focusing on the sentence allows indexing directly into a vocabulary. He shows that using the sentence as a basis is sufficient for extracting the meaning of the text without the need to refer to its set of words. According to his finding, the set of phrases is richer than the set of words or word senses. From this main idea, we are motivated to add the clause structure to our learning features.

Additionally, unlike the English language, which is characterized by only nominal phrases, the Arabic language is characterized by the presence of other sentence structures. As a result, we can benefit from this peculiarity of the Arabic language. Indeed, the Arabic sentence is generally classified as either a nominal sentence or a verbal sentence. The verbal sentence is defined as a clause that begins with a verb and has the order (Verb-Subject-object), whereas a nominal phrase starts with a noun. The nominal phrase consists of two parts: the subject (called "**مبتدأ**/mbtd>" in Arabic) and the predicate (labeled "**خبر**/xbr" in Arabic). Each part has many cases. The subject can be a noun, a pronoun, a demonstrative, a compound noun or another entity. Each sentence, either nominal or verbal, can be preceded by a conjunction ("و/wa/and", "ثم/thomma/then"), adverb ("عندما/EndmA/when"), negation particle ("لن,لا,لم/ ln, lA, lm/ not") or combination ("وعندما/ wEndmA/ and when"). Similarly, a sentence can be either simple or compound.

For these reasons, we believe that the sentence structure can enhance the results. If we take example [1] again and additionally recall examples [9] and [10], the same NEs pair (PERS-LOC) and the same relation detected through "سافر/sAfr/travel" are presented, and only the structure of the clause is different.

[10] **أحمد حلمي** سافر الى **السعودية.**

>**Hmd Hlmy** sAfr AlY **AlsEwdyp**.
Ahmed Helmi traveled to **Saudi Arabia**.

In the first verbal phrase [1], the relation is situated in the first word of the first context *W1C1*. However, the same relation is located in the first word of the second context *W1C2* for the nominal phrase [9]. Therefore, the clause structure changes the order of the words, which can in turn change the position of the word, expressing the relation (our output). Hence, the position of the trigger word for expressing the relation between the NEs depends on the phrase structure.

This step is accomplished using the Stanford tagger (Green and Manning, 2010), which provides syntactic information about each sentence. Thus, we are interested only in the acquisition type of a given sentence in our data set. Thus, we added another syntactic feature to mention whether we have a nominal or verbal phrase. Then, we partitioned our data according to this feature. As a result, we have two training corpora: one corpus is for the nominal clauses, and the second concerns the verbal clauses. Obviously, the same partition must be applied to both the training data and the test data. We think that the partition of our data set according to the clause structure leads to more efficient and accurate results. This contribution will be evaluated in the next section (see Section 6).

5.3.2. Handcrafted rules

In this module, we added patterns that were proposed by a linguistic expert to resolve some of the problems that were cited previously. Some of the modules are presented as further constraints in an attempt to rectify the output of the ML results. In addition, others were added to lead to more accurate results.

First, we address the negation relation point (example [5], Table 3). To tackle this issue and to obtain a more significant extracted relation, we add constraints to each rule to verify whether there is a negative particle that expresses the negation relation. For example, recall that in example [11] we have two relations between the NE pair; the first relation is detected via the verb "**يتزوج**/ ytzwj/marry", which is in negative form, and the second relation is identified through the noun "**صديقته**/ Sdyqth/friend", which is in positive form. Thus, in the case in which a relation is detected through a verb and the latter is proceeded by a negative particle such as ("ما/mā/not", لن/ ln/not", "لا/lā/not", "لم/lm/not"), then the relation between these NEs is not achieved or is in the negative form.

[11] لم يتزوج **مايكل** صديقته **كريستين**.

lm ytzwj**mAykl** Sdyqth **krystyn**.
Michel did not marry his friend **Kristin.**

Let us consider now the example of a negation relation that is detected through a noun. There, the relation is negated through the incomplete verb that is called in Arabic "أَخَوَاتُ كَانَ/sisters of kaan(a)" like "ليس/lays(a)/ is not", which can be placed just before the noun that expresses the relation or in the first context (before the first NE) [4]. Thus, when we have a relation that is identified through a noun and we have "ليس/*lays(a)*/is not" in the first or the second context [12], our relation is introduced in a negative form.

[12] ان **كريستين** ليست زوجة **مايكل**.

An **krystyn** lyst zwjp**mAykl**.
Kristin is not Michel's wife.

In the same context, we added some generic and intuitive patterns to rectify the output that is generated by our supervised learning method. These patterns concern the possible POS tag of words that can predict the semantic relation between NEs. When studying our training corpus, we note the following statistical results: 39% of the relations are predicted through a verb, 32% are discovered through a noun,

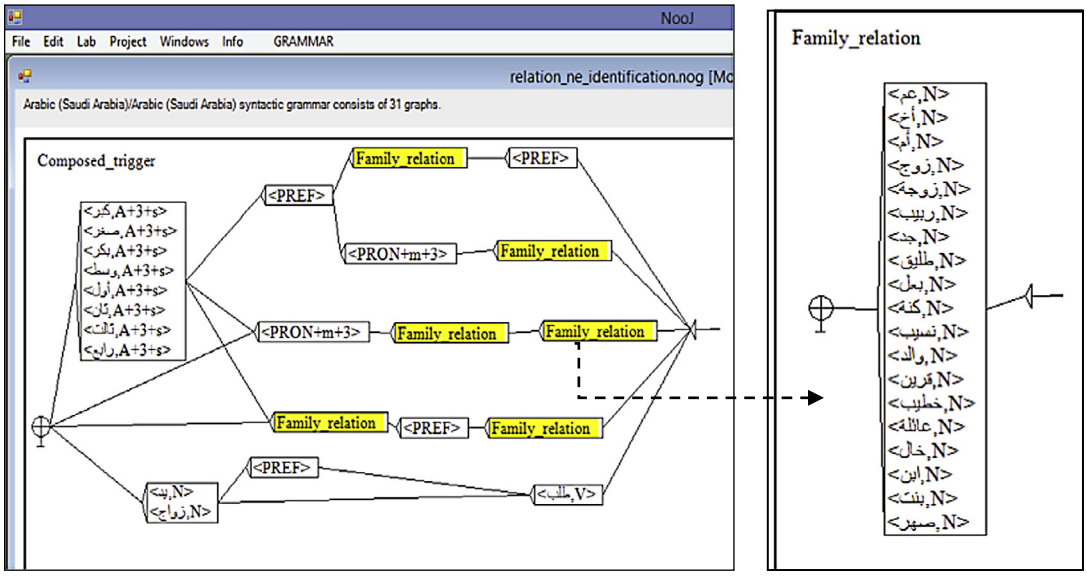

Figure 5 Composed trigger of family relations between the NE pairs PERS_PERS.

14% are expressed through adjectives, 4% though punctuation marks and a rare percentage of relations are introduced through adverbs and conjunctions. Hence, some POS tags could not predict a semantic relation between the NEs, such as a negative particle. We grouped the morphological categories that could not contain a relation in the *NOTR* list. The list of POS tags that could identify a relation is grouped into a POSR list.

From this statistical study, we added generic constraints that serve to verify the POS of the obtained classes. In the case in which we have a relation that is expressed through a word that has a POS tag that belongs to the *NOTR* list and its corresponding instance can be treated by more than one rule, we choose the rule that produces an output that belongs to the *POSR* list.

Finally, we handle the relation that can be predicted through more than one word [13].

[13] **هوانغ** يصمم شعارات غريبة الأشكال لشركة **"غوغل"** .

hwAng ySmm $EArAt grybp Al>$kAl $rkp " gwgl ".

Huang designed logos with strange shapes for **the "Google" company**.

To treat this issue, we have elaborated a linguistic grammar to extract the relations between the NE pairs. To accomplish that goal, it is mandatory to collect first a list of compound words that can express a semantic relation between NE pairs. We have reused a part of the local grammar of extracting relations between NEs (Boujelben et al., 2012). In fact, through the composed word triggers, we can identify a complex relation. Thus, we intend to apply some of these grammars to our training corpus to extract this type of relation. Thereafter, these examined instances will be excluded from our dataset. Hence, we will address only the relations that are expressed through one word. Fig. 5 shows a sample of a local grammar that is elaborated through the NooJ linguistic development environment that extracts the family relation that can hold between two person NEs.

The sub-grammar presented in Fig. 5 concerns only a sample of a relation "Family" that can hold between the pair PERS_PERS; it is composed of five main paths. The first path is intended to extract the case in which we have a family trigger followed by an adjective such as "أحمد هو الأخ الأكبر لفاطمة/>Hmd hw Al>x Al>kbr lfATmp/Ahmed is the oldest brother of Fatma". The second path treats the case in which a trigger is attached to a pronoun followed by an adjective (قالت فاطمةلأخوها البكر أحمد/qAlt fATmp l>xwhA Albkr>Hmd/Fatma said to her older brother Ahmed.). Herein, we must verify the pronoun that is attached to the trigger through the addition of the constraint ("PRON+3+s+m") or ("PRON+3+s+f"), to treat some of the ambiguities that appear in some cases. These gender and number features are captured from the Arabic dictionaries that are constructed by (Mesfar, 2006). This finding means that our system can express a family relation between two arguments that are not truly related semantically, as shown in the following sentence:

[12] ذهب **أحمد** و أخي الصغير **صالح** إلى المدرسة.

*hb>Hmd w>xy AlSgyr SAlH<lY Almdrsp.

Ahmed and my little brother **Salah** went to school.

In the example above, we have two arguments (PERS), and although we have the trigger "brother" of the family relation, they are not related because "صالح/Salah" is "أخي/>xy/my brother" and not the brother of Ahmed. Therefore, we must verify the gender (masculine or feminine) and the number (singular or plural) features that are used in the extraction of such a relation.

In fact, this constraint verifies the gender (masculine or feminine) and the number (singular) of the trigger. The third and fourth path are used in the example of two triggers of family relations, such as (ابن العم/Abn AlEm/cousin, ابن عمه/Abn Emh/ his cousin), respectively. The fifth path is exploited to extract some composed verbs of this relation type, as mentioned in the graph ("طلب يدها/Tlb ydhA/ask for her hand").

These rules are added to the initial rules to provide deep analysis of the context of each entity and relation. For each pair of NEs, we built a sample grammar that contains the composed trigger to treat the case of relations that are expressed through more than one word. Having these composed triggers, (Boujelben et al., 2012) elaborated syntactic grammar to extract the different relations between the NEs. As mentioned previously, these trigger relations can appear in the first context (before the first NE), in the second context (between the NE pair) or in the third context (after the second NE). At least 24 grammars have been elaborated. A NooJ grammar is built for each class type to recognize the semantic relations between the NEs. These local grammars will be applied first to our corpus, to treat some cases. Then, the treated sentences will be excluded from our training corpus to automatically study the remaining clauses. Because of the proposed NooJ grammar, we can overcome the problem of relation predicted through more than one word. Similarly, the examination of these compound triggers in terms of their gender and number enables recognizing some unrelated NEs within a clause.

All the modules that have been proposed can undoubtedly contribute to the automatic extraction of both the undetected and wrong extracted relations that are extracted by the ML model.

6. Evaluation

6.1. Experimental setting

The input data used in our experiments consists of Arabic texts collected from the ANERCorp corpus (Benajiba et al., 2007). We decided to utilize this corpus because it is annotated by NEs. This corpus is composed of more than 316 articles, which contain more than 150,000 words and 3206 labeled NEs. We have discovered only 840 NEs that are related in a clause, given that we focused on three NE types (PERS, ORG and LOC). First, we extracted the clauses that were composed of two NEs using the clause splitter of (Keskes et al., 2012). The morphological analysis is accomplished using the Arabic resources of (Mesfar, 2006). The word that predicts a relation within a clause is identified manually by a linguistic expert. We believe that by integrating the corpus and using rich linguistic processing strategies with an expert revision of different tags, the approach can achieve effective results, in terms of both accuracy and coverage. After applying these resources and obtaining annotated clauses, we extract our learning features to build our test dataset. As a result, we obtained 420 instances in our test database.

In Table 4, we summarize the characteristics of both the training and test corpus, and we mention the final number of clauses, NEs and tokens that compose our corpus.

A part of our training and test corpus is available on the net.[12] In this website, we present a part of our corpus: non-annotated examples, examples after annotation and the final instance generated after extracting the learning features.

Many experiments have been conducted to assess the quality of our system. In this section, we describe the different experimental settings that we used, and we present the obtained results.

[12] https://sites.google.com/site/inesboujelben85/corpus.

Table 4 Training and test corpus.

	Training corpus	Test corpus (ANERCorp)
Clauses number	1411	420
Tokens	12,192	1856
NEs	2828	840
LOC	913	320
PERS	1154	217
ORG	761	303

For the ML technique, we exploited the Association rules algorithm Apriori and the C4.5 classifier, which are available in the WEKA[13] tool. We used the classical measures of precision, recall and *F*-measure. The precision is the number of relevant instances of the system among all the treated instances. The recall is the number of relevant instances that are retrieved divided by the number of reference instances. The *F*-score is a combination of the precision and recall, which is used to penalize the very large inequalities between these two measures.

6.2. Experimental results and analysis

All the reported experiments in this paper were performed on the ANERCorp corpus by means of standard evaluation metrics.

6.3. Experiment 1: training corpus setting

In this experiment, we are interested in evaluating our ML method based on a learning algorithm combined with a GA, to automatically extract the more interesting rules. First, we have computed a learning curve by dividing our training corpus into different learning sets to analyze how the learning procedure can be influenced by the number of annotated sentences. For each set, we apply the Apriori and decision tree algorithms to obtain the initial rules that are considered to be the input population of the GA.

The *F*-score curve (see Fig. 6) shows that the curve grows regularly between 0 and 600 instances, while it seems to plateau between 900 and 1400 instances. We can thus conclude that the addition of more than 1400 instances will only slightly increase the performance of the relation extraction task.

6.4. Experiment 2: evaluation of the ML method

The purpose of this second experiment is to assess the performance of the proposed features. Hence, we intend to choose a small subset of features that is sufficient to correctly predict the class. Here, we attempt to learn which features are better. We apply some recognized selection algorithms. After obtaining the results of these algorithms, we evaluate the performance of each combination of features when applying Apriori, which presents our first baseline **B1**. We envisage evaluating the influence of the considered number of words before, between and after the NEs. The obtained results are presented in the following table.

As is indicated in Table 5, the reported results show that using the context features improves the overage performance

[13] Available on www.cs.waikato.ac.nz/ml/weka/.

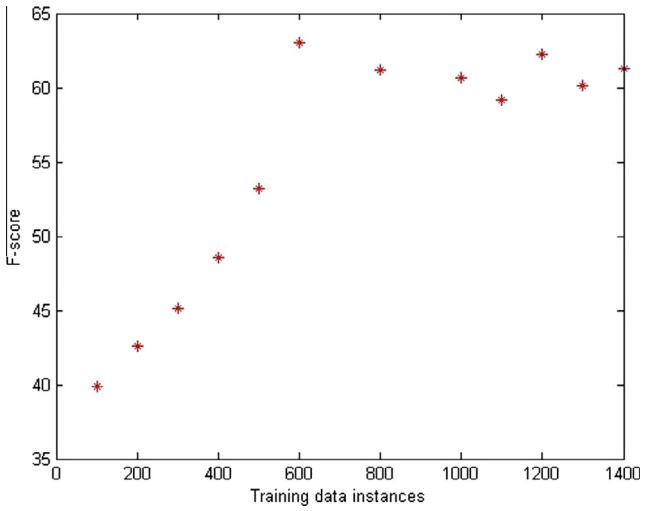

Figure 6 Learning curve depending on the number of instances.

of our process. In addition, the consideration of the two words before, between or after the NEs achieves better results. It is observed that accounting for one or three words for each context gives slightly low results in terms of both the precision and recall. Hence, using one context word alone is not sufficient to extract efficient rules. This finding can be explained by the fact that we focused on a clause and not the whole sentence. Therefore, it is sufficient to work with only two context words. For the next experiment, we will utilize these selected features. After obtaining the best combination of our proposed features, we then compare different ML methods while using the techniques separately or combined.

- B2: The second baseline is based on the combination of B1 with the decision tree technique.
- E1: (Boujelben et al., 2013a): B2 + four selection levels to extract the more interesting rules.
- E2: (Boujelben et al., 2013b): B2 + GA for the rule selection and generation: For the GA parameter setting, the mutation probability is taken as 0.04, the crossover probability is taken as 1, the initial population size is 190 (rules generated by Apriori and C4.5), and the maximum number of generations is fixed at100.

Therefore, we provide a performance comparison with previous studies based on supervised learning methods (Boujelben et al., 2013a) and (Boujelben et al., 2013b) against two baselines, B1 and B2.

As is cited in Table 6, our first baseline (B1) obtains a high precision value with a very low recall. It is significant because the rules that are produced by this algorithm satisfy some minimum confidence (0.6) and minimum support (2) constraints. For this purpose, we are expected to add the decision tree technique, which presents the second baseline method. The combination of these two algorithms achieves an improvement in terms of the coverage of our data set byapproximately11%, while the precision is decreased by 8%.This arrangement can be justified by the fact that the decision tree technique allows matching other instances that are not covered by the training data. Although this combination leads to cover more instances, it implies noisy rules, which reduces the precision of our system.

Table 6 Comparative performance of different ML techniques.

ML methods	Number of rules	Precision (%)	Recall (%)	*F*-score (%)
B1	120	86.23	23.55	33.46
B2	334	60.4	31.4	42.8
E1	160	62.03	54.52	58.03
E2	190	**74.1**	**59.6**	**66.1**

Overall, the results shown above prove the effectiveness of combining various rules that were produced by two different mining algorithms: precise association rules produced by Apriori and classification rules generated by C4.5. Furthermore, the comparison between the selection levels added by (Boujelben et al., 2013a) and the selection module based on GA (Boujelben et al., 2013b) demonstrates the performance of our genetic process because it boosts the overall results. This approach yields the best results in terms of both the precision and recall, which shows an 8% increase.

After evaluating our proposed ML method, we next evaluate the different modules that were added to this learning technique.

6.5. Experiment 3: effect of the clause structure in the relation detection

In this section, we evaluate the effect of using the clause structure in the relation extraction process. To study the effect of the clause structure, we present the evaluation of our learning method based on the GA when it is applied to the two sub-corpora: C1, which is composed of only nominal sentences, and C2, which is composed of only verbal sentences. As a result, we obtained some rules that were specific to the verbal phrases and other rules that were specific to the nominal phrases.

The results shown in Table 7 demonstrate that the consideration of the clause structure in our learning process increases the overall performance of our extraction process when applied to the test corpus. The first experiment demonstrated that the partition of our corpus into two corpora according to the clause structure increased the precision by approximately 6.6 %, up to 80.7%, and increased the recall by 3%, up to 62.35%, and thus contributed to relation detection.

The obtained results proved the effectiveness of data partitioning according to the clause structure. This finding explains

Table 5 Evaluation of learning features.

Method	Features	Precision (%)	Recall (%)	*F*-score (%)
B1	POS tag (three context words) + Numeric + NEs tags	79	19.79	31.65
	POS tag (two context words) + Numeric + NEs tags	86.23	23.55	33.46
	POS tag (one context word Numeric + NEs Tags	95.2	20.3	**37**
	POS (no context word) Numeric + NEs Tags	64.5	17.54	27.57

Table 7 Evaluation of the structure clause.

Corpus	Number of Rules	Precision (%)	Recall (%)	*F*-score (%)
Initial corpus	190	74.1	59.6	**66.1**
C1	170	81.01	57.55	67.2
C2	153	80.45	67.16	73.2
C1 + C2	323	80.73	62.35	70.2

the fact that morpho-syntactic features are a basis in our learning method. Additionally, given that the relation is predicted through the position of the trigger word, the latter is dependent on the word order in the clause, which in turn can be predisposed by the structure clause. Hence, the clause structure can be introduced to facilitate the identification of the trigger position.

6.6. Experiment 4: comparison with previous studies

After identifying the learning feature that was selected in our GA method, we further compared in Table 8 the performance of our mixed approach against both the ML-based method and the rule-based method (Boujelben et al., 2012) when applied to the same test corpus ANERCorp because these results are produced in a similar setting, they can be compared fairly.

In this experiment, we seek to elaborate a comparative study in which we follow three main approaches: the rule-based method, ML-based method and, finally, our mixed process, in which we add gradually handcrafted rules and other modules. In our hybrid method, we elaborate an evaluation of each proposed linguistic module to verify the effectiveness of each contribution.

As shown above, the rule-based method achieves a low recall, although its precision is promising. We mention some mistreated relations that were caused by the absence of grammars that extract the relations between the NE pairs "ORG-PERS" and "ORG-LOC". In fact, NooJ grammars are based on syntactic information that is combined with some relation triggers, and therefore, they cannot capture certain potentially relevant relations between Arabic NEs. Thus, some examples are not detected because the rules that are rewritten into NooJ grammars do not cover all the possible cases that are caused by the absence of some trigger words in our grammars.

To overcome this limitation, the ML algorithm appears to be a good solution given that it extracts automatically the trigger words regardless of the word meaning. The reported results showed that compared with the rule-based method (Boujelben et al., 2012), the ML-based method that utilizes GAs provides better performance in terms its recall, which had an 8% improvement, while the precision was slightly lower for this method. Additionally, this table reports that our system achieves the best improvement to obtain 84.8%, 67.6% and 75.2% in terms of the precision, recall and *F*-score, respectively. This finding is because some previous relations are not discovered because they are composed of more than one trigger word.

Now, we move on the analysis of error sources. First, we can mention the recall loss because of the untagged NEs, which in turn excludes their associated examples from the data set. Second, the wrongly annotated NEs make up 11% of the overall precision low. Indeed, the influence of NE recognition ambiguity can lead to the application of the inappropriate rule. For example, ("آسيا/Asia/Asia") could be either identified as a name of a person or a name of a location. As a consequence, some of the errors will be produced when applying an unsuitable rule to the associated instance. This issue can be illustrated by the following rule when applied to two different phrases:

Rule A: **If** EN1 = PERS and EN2 = LOC and W1C1 = V and W2C2 = prep **Then,** class = w1C1

[13] سافرت **فاطمة** إلى آسيا.

sAfrt fATmp < lY |syA.

Fatma traveled to **Asia**.

[14] حاولت **فاطمة** التحدث الى آسيا.

HAwlt fATmp AltHdv AlY |syA.
Fatima tried to speak to Asia.

As illustrated in the following example, when applying rule (A) to example [14], the relation is predicted through "W1C1" via the word "سافرت/sAfrt/travel". Nevertheless, because "Asia" was identified as a location NE, the deduced relation is predicted through the verb "حاولت/HAwlt/tried", which is not the correct relation trigger. Here, the ambiguity of the NE type recognition causes the application of an inappropriate rule, which in turn produces erroneous outputs.

Moreover, some ambiguous relations are caused by the morphological ambiguity of the Arabic language. Indeed, in some cases, an NE can be analyzed as a part of speech of a given word. For example, the Arabic proper name ("أكرم/Akram") can be treated as either a verb that means "to immortalize" or the superlative adjective "the most immortalized" [15]. Similarly, in some cases, because a primordial argument of relation extraction is omitted, which is caused by the non-recognition of this word as an NE, the concerned clause will not be treated.

أكرم شاب يعيش **في** السعودية**[15]**.

>krm Al$Ab yEy$ **AlsEwdyp**.

The rule (A) can be applied to this sentence [15]. As a result, we obtain the relation that is predicted through "> krm/أكرم/ immortalized". However, this result is not the correct relation. Moreover, this sentence can have a different semantic analysis, depending on the voyellation. For example, it can be analyzed as "Akram is a young man who lives in Saudi Arabia", in which "> krm/أكرم" is a person NE and" السعودية/AlsEwdyp"

Table 8 Comparative performance of different methods for relation extraction.

Methods		Precision (%)	Recall (%)	*F*-score (%)
Rule-based method Boujelben et al. (2012)		82	51.5	63.26
(B)ML-based method (GA) Boujelben et al. (2013b)		74.1	59.6	66.1
Hybrid method	**B** + base partition	80.73	**62.35**	**70.2**
	+ NooJ grammars: compound trigger relation	82.7	65.1	72.85
	+ Correction rules	84.8	67.6	75.22

is a LOC NE. In this case, a relation is predicted through the verb "يعيش/lives".

Additionally, "akram" can be a superlative adjective, which would mean "The most immortalized man lives in **Saudi Arabia**". In this situation, there is no NE pair. Thus, we cannot extract a relation from this sentence. Finally, it can be treated as a verb, to express that "A young man who lives in Saudi Arabia has been immortalized". Therefore, such ambiguities affect the precision of our rules.

For these reasons, the short fall of the ML method (Boujelben et al., 2013b) was compensated by the use of manually constructed patterns that were proposed by a domain expert, to recognize the relations in their negative forms and to extract relations that were expressed through a compound noun. The partition of our dataset into two parts according to the structure of Arabic phrases achieves an improvement of 9% in the overall ML performance.

In conclusion, we note that a rule-based method can discover a relation between NEs if the relation trigger belongs to our trigger words list in spite of its position in the sentence. However, the problem is disengaged when a trigger word did not exist in the trigger words list. On the one hand, the pattern-based method offers good precision values but can be weak when faced with heterogeneous vocabulary and sentence complexity. On the other hand, the ML technique is more efficient in terms of its coverage of our dataset.

7. Discussion

In this paper, the problem of relation extraction between Arabic NEs is tackled through integrating the supervised learning method with linguistic modules to improve the overall performance. Based on this hybrid approach, we can combine the advantages of ML and rule-based methods.

Several semantic relation extraction approaches detect only whether a relation type occurs in a given sentence. Thus, if we have an NE pair that is not linked by a predefined relation type and subtype, this pair will be discarded. In contrast to other research that is based on a fixed number of relations classes such as in (Zhang et al., 2009) and (Alotayq, 2013), we consider the word position that predicts the relation as an output class. This word position can occur in different contexts according to the position of the NEs: before, between or after the NEs. Thus, we can extract an infinite number of relation instances without being limited to a given type of relation class. For our case, we envisaged extracting first the trigger word that expresses the semantic relation between the given NEs. In a subsequent step, we match each trigger word that was extracted automatically with its appropriate class.

The extraction of relations among Arabic NEs has encountered many problems. Some of these problems can be resolved by an ML method, whereas others need the intervention of certain handmade patterns in an attempt to accomplish more accurate results.

Because the output of our process is the trigger word that identifies a given relation, the latter can take various POS tags of words; it can be detected through a verb, a noun or a preposition, for example. This approach allows us to capture some implicit relations, such as relations that are detected via a punctuation mark. Indeed, a comma, when presented between two NEs, can indicate the presence of a relation between the NEs [5]. Herein, a relation is expressed through a comma (which belongs to the relation (Ben Hamadou et al., 2010a)) between the NE mentions "Mohammad Qasim" and "the Faculty of Medicine".

Additionally, some ambiguities arise when more than one possible relation exists within the same pair of entities. This issue is neglected in most of the current studies. In our case, this problem can be treated by our proposed supervised learning method by considering the following hypothesis: when an instance of our data set is treated by two rules that have the same value of confidence, then we have two possible relations between the same pair of NEs.

A second aspect of our research is the contribution of hybrid approaches in which we intend to add linguistic modules to our ML method. As shown in the evaluation section, the hybrid method effectively outperforms both the pattern-based and ML approaches in terms of the *F*-score. The experimental results show that our hybrid model gives the best results, which is ~75% for the *F*-score.

Indeed, these further linguistic modules achieve improvements to the overall performance of our process. The obtained results show that these linguistic modules (especially when we partitioned our corpus into noun clauses and verbal clauses) contribute substantially when they are combined. In addition, when using some proposed constraints, we have corrected some relations that are extracted and joined with negative particles. Some other intuitive patterns are added to rectify some of the outputs that are generated by the ML method. These constraints are added after some statistics are generated from our training dataset in which we exclude some of the cases. This finding implies that those linguistic constraints, when applied for each of the rules, are very useful and contribute much better when they match each of the generated rules. Overall, the experimental results exhibit that our mixed method significantly outperforms the rule-based and ML methods when both are applied to the ANERCorp corpus.

Based on rules that are either automatically extracted through the supervised learning technique or manually added, our hybrid method achieves encouraging results. Although it has promising performance in terms of precision and recall, our process cannot extract some of the relations that are present among words that are a long distance from the NEs' positions, notably, in the case of long and complex sentences.

8. Conclusions

In this paper, we combine the advantages of ML techniques and rule-based methods to extract relations between Arabic named entities. We rely on manual patterns when the given relation examples are complicated or expressed through more than one word. Our approach obtains an overall 75.22% for the *F*-score.

The obtained results are promising and motivate the strategy of combining both types of methods to boost the overall performance of our process. We showed the impact of each used linguistic module to produce significant gains against previous results. Similarly, we also studied the effect of nominal and verbal clauses on the performance of our system. Finally, further constraints that were added to the automatic rules that were generated from our proposed genetic algorithm yield more concise and accurate results.

For future work, we intend to classify the trigger word into an adequate level of semantic relation classification. Additionally, we plan to evaluate our approach with other NE types and different corpora languages and domains. It would also be mandatory to test on other languages other learning models (such as SVM and MaxEnt), which have been used in prior relation extraction tasks, to provide a performance comparison with our process. Similarly, we intend to apply our process to the standard ACE data set, to provide a comparative study in our upcoming work.

References

Abdul-Hamid, A., Darwish, K., 2010. Simplified feature set for Arabic named entity recognition. In: Proceedings of the Named Entities Workshop, pp. 110–115.

Agrawal, R., Srikant, R., Imielinski, T., Swami, A., 1993. Mining association rules between sets of items in large databases. In: ACM SIGMOD Conference on Management of Data, Washington.

A. Kadir, R., Bokharaeian, B., 2013. Overview of Biomedical Relations Extraction using Hybrid Rule-based Approach. J. Ind. Intell. Inf. 1 (3), 169–173.

Alotayq, A., 2013. Extracting relations between Arabic named entities. In: TSD2013. Springer-Verlag, Berlin Heidelberg, Pilsen, pp. 265–271.

Ben Abacha, A., Zweigenbaum, P., 2011. A hybrid approach for the extraction of semantic relations from MEDLINE abstracts. In: 12th International Conference on Intelligent Text Processing and Computational Linguistics CICLING2011, Tokyo, Japan, pp. 139–150.

Ben Hamadou, A., Piton, O., Fehri, H., 2010a. Multilingual extraction of functional relations between arabic named entities using Nooj platform. In: hal-00547940, version 1.

Ben Hamadou, A., Piton, O., Fehri, H., 2010b. Recognition and translation Arabic-French of Named Entities: Case of the Sport places. In: Arxiv preprint arXiv:1002.0481.

Benajiba, Y., Rosso, P., Benedi, J.M., 2007. ANERsys: an Arabic named entity recognition system based on maximum entropy. In: CICLing-2007. Springer-Verlag, Berlin Heidelberg.

Benajiba, Y., Rosso, P., 2008. Arabic Named Entity Recognition using Conditional Random Fields, In: Proceeding of Workshop on HLT and NLP within the Arabic World, LREC'08.

Boujelben, I., Jamoussi, S., Ben Hamadou, A., 2012. Rules based approach for semantic relations extraction between Arabic named entities. In: NooJ2012. INALCO, Paris, pp. 123–133.

Boujelben, I., Jamoussi, S., Ben Hamadou, A., 2013a. Enhancing machine learning results for se-mantic relation extraction. In: NLDB, Manchester, UK, pp. 337–342.

Boujelben, I., Jamoussi, S., Ben Hamadou, A., 2013b. Genetic algorithm for extracting relation between named entities. In: 6th Language and Technology Conference, LTC, Poznań, Poland, pp. 484–488.

Celli, F., 2009. Searching for Semantic Relations between Named Entities in I-CAB. Available from: <http://clic.cimec.unitn.it/fabio> (technical report).

Ciravegna, F., Wilks, Y., 2003. Designing adaptive information extraction for the semantic web in Amilcare. In: Handschuh, S., Staab, S. (Eds.), Annotation for the Semantic Web. IOS Press.

Ezzat, M., 2010. Acquisition de grammaires locales pour l'extraction de relations entre entités nommées. In: TALN2010, Montréal.

Fehri H., Haddar, K., Ben Hamadou, A., 2011. Recognition and Translation of Arabic Named Entities with NooJ Using a New Representation Model. In: FSMNLP 2011, France, pp. 134–142.

Green, S., Manning, C., 2010. Better Arabic parsing: baseline, evaluations and analysis. In: 23rd International Conference on Computational Linguistics COLING2010, Beijing, China.

Haddad, H., 2003. French Noun Phrase Indexing and mining for an Information Retrieval System. In 10th international Symposium, SPIRE 2003 Manaus, Brazil, pp. 277–286.

Hasegawa, T., Sekine, S., Grishman, R., 2004. Discovering relations among named entities from large corpora. In: Association for Computational Linguistics. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA.

Hassan, H., Emam, O., 2006. Unsupervised information extraction approach using graph mutual reinforcement. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Sydney, Australia, pp. 501–508.

Holland, J.H., 1970. Robust algorithms for adaptation set in a general formal framework. In: Proceedings of the IEEE Symposium on Adaptive Processes-Decision and Control.

Jantan, H., Hamdan, A.R., Othman, Z.A., 2010. Human Talent Prediction in HRM using C4.5 Classification Algorithm. Int. J Comput. Sci. Eng. 2 (8), 2526–2534.

Jourdan, L., Dhaenens, C., Talbi, E.G., 2002. ASGARD: un algorithme génétique pour les règles d'association. In: ECA.

Keskes, I., Benamara, F., Belguith, L., 2012. Clause-based discourse segmentation of Arabic texts. In: Language Resources and Evaluation LREC, pp. 2826–2832.

Kramdi, S.E., Haemmerl, O., Hernandez, N., 2009. Approche générique pour l'extraction de relations partir de texts. In: Ingnierie des Connaissances IC, Tunisia.

Mesfar, S., 2007. Named Entity Recognition for Arabic Using Syntactic Grammars. In: Natural Language Processing and Information Systems, 12th International Conference on Applications of Natural Language to Information Systems, NLDB 2007, Paris, France, pp. 305–316.

Mesfar, S. 2006. Standard Arabic formalization and linguistic platform for its analysis, In: The challenge of Arabic NLP/MT conference, Londres – Angleterre, Eds BCS – British Computeur Society, pp. 84–94.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

Riedel, S., Yao, L., McCallum, A., 2010. Modeling relations and their mentions without labeled text. In: ECML PKDD'10 Proceedings of the 2010 European conference on Machine Learning and Knowledge Discovery in Databases: Part III, pp. 148–163.

Shaalan, K., Oudah, M., 2014. A hybrid approach to Arabic named entity recognition. J Inform. Sci. 40 (1), 67–87.

Smeaton, Alan F., 1995. NLP & IR a tutorial presented at EACL, Dublin City University.

Tso, G.K.F., Yau, K.K.W., 2007. Predicting electricity energy consumption: a comparison of regression analysis, decision tree and neural networks. Energy 32, 1761–1768.

Zelenko, D., Aone, C., Richardella, A., 2003. Kernel methods for relation extraction. J. Mach. Learn. Res. 3, 1083–1106.

Zhang, Z., 2004. Weekly supervised relation classification for information extraction. In: Proceedings of ACM 13th Conference on Information and Knowledge Management CIKM2004, Washington D.C., USA.

Zhang M., Su, J., Wang, D., Zhou, G., Tan, C.L., 2005. Discovering Relations Between Named Entities from a Large Raw Corpus Using Tree Similarity-Based Clustering. In: the second International Joint Conference on Natural Language Processing IJCNLP05, LNC (LNAI), vol. 3651, pp. 378–389.

Zhang, J., Ouyang, Y., LI, Y., Hou, Y.X., 2009. A novel composite approach to Chinese relation extraction. In: Proceedings of the 22nd International Conference on Computer Processing of Oriental Languages ICCPOL'09, Hong Kong, pp. 236–247.

Zhou, G., Qian, L., Zhu, Q., 2009. Label propagation via bootstrapped support vectors for semantic relation extraction between named entities. Comput. Speech Lang. 23 (4), 464–478.