RelANE: Discovering Relations between Arabic Named Entities

Ines Boujelben, Salma Jamoussi, and Abdelmajid Ben Hamadou

Miracl, University of Sfax, Tunisia
Boujelben_ines@yahoo.fr,jamoussi@gmail.com
adelmajid.benhamadou@isimsf.rnu.tn

Abstract. In this paper, we describe the first tool that detects the semantic relation between Arabic named entities, henceforth RelANE. We use various supervised learning techniques to predict the word or the sequence of terms that can highlight one or more semantic relationship between two Arabic named entities.
For each word in the sentence, we use its morphological, contextual and semantic features of entity types. We do not integrate a relation classes predefined in order to cover more relations that can be presented in sentences. Given that free Arabic corpora for this task are not available, we built our own corpus annotated with the required information.
Plenty of experiments are conducted, and the preliminary results proved the effectiveness of our process that allows to extract semantic relation between Arabic NEs. We obtained promising results in terms of F-score when applied to our corpus.

Keywords: relation, named entity, supervised method, Arabic language

1 Introduction

The extraction of Relations involving Named Entities (RNE) task is seen as a step towards a more structured model of the text meaning. Therefore, it presents a fundamental task for the many Natural Language Processing (NLP) and information extraction tasks, such as question answering and automatic summarization. Hence, the NLP community shows a great interest concerning this issue. This interest in RNE is shown for English and European languages. However, a few works are done for the Arabic language. This is due to the complexity of this language morphology and the lack of available resources, notably annotated corpus. Given that the Arabic language is a rich morphological language, the RNE applying to Arabic language task has to be challenging.

In literature, RNE is the task of finding pre-defined semantic relations between two entities or entity mentions from text (e.g. [1,5]). This task requires two main subtasks: relation detection and relation classification. In our work, we aimed at finding all binary relations without any restriction to relation classes. Our main goal is to detect a set of words that predicts relation between NEs. In a subsequent step, we will assign a specific class to each extracted trigger word.

The rest of the paper is organized as follows. We firstly introduce the entity relation extraction task. Then, we survey previous work on relation extraction. The ensuing

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 233–239, 2014.

section is devoted to describe our data in which we depict our annotation guidelines. In the last section, we present the different experiments from which we discuss the reported results.

2 Relation Extraction

In our case, a relation can be expressed directly through one word or a sequence of words which is very common for some family (e.g. ابن العم/cousin, ابن عمه/ his cousin) or functional relations (e.g عام أمين/Secretary-General), , respectively. These words can be depicted in the same context (before, between NEs or after NEs), or each of them can be located in different contexts (e.g. طلبأحمديدسلمى من أبيها للزواج /Ahmed ask Salma's father for her hand in marriage). We assume that support words can offer helpful information to recognize the semantic relations holding between Arabic NEs.

Unlike some recent researches which focused on semantic relation classes, we assume that detecting an infinite number of relations (independently of semantic relations classification) poses a more challenging problem.

3 Related Work

Several methods have been proposed to extract semantic relation between NEs. Some of them are based on linguistic method, which relies on rules that are usually implemented in the form of regular expressions or finite-state transducers. [2,4] have been elaborated local grammars under the linguistic platform NooJ[1] to discover relation between Arabic NEs. In order to automate this task, some researchers have been oriented towards machine learning (ML) methods. We distinct three main methods: (i) the unsupervised methods that make use of massive quantities of unlabeled text. They focalize almost on clustering techniques and similarities between features or context words [8,12]. To overcome the problems encountered by unsupervised approach, some researchers are oriented towards semi-supervised learning methods which rely on a small set of initial seeds. These latter can be depicted as a sample of linguistic patterns or some target relation instances for the purpose to acquire more basics until discovering all target relations such as [14,11]. A last method under the ML methods is the supervised technique which considers the relation extraction as a classification task. This method requires fully labeled corpus. An early attempt to extract relation between Arabic NEs is carried out by [1], who used MaxEnt classifier. Based only on morphological and POS information, his system achieves satisfactory results when applied to ACE 2005 Multilingual training data[2]. In [5], the author combined two supervised techniques which are simply decision trees (DT) and PART decision lists algorithms to extract three semantic relations (role, social and location) between NEs. The author focused on the part-of-speech tag of the context before and between the two entities only, without

[1] NooJ local grammars are typically used in order to describe sequences of words that present meaningful units or entities. They represent a set of rules by means of transducers.

[2] Available on `http://www.nist.gov/speech/tests/ace/`

considering the context after NEs and he reported an F-score of 81.2% when applied to I-CAB[3] data. Through the above study of different works, we decide to rely on supervised techniques regarding their promising results. Therefore, we examined a set of supervised techniques used in prior work to reach a conclusion. We aimed at discovering trigger words that can explicitly predict a relation between NEs. Consequently, an infinite number of relations will be detected, without being dependent on predefined relations classes.

4 Data

As far as we know, Arabic misses lexical resources, especially free resources available for a RNE purposes. It is mandatory to point out the existence of an Arabic corpus which is annotated with the relation between NEs, namely ACE multilingual training data. Unfortunately, it is not freely accessible. This is the reason why, we have to contract our own corpus to carry out this task. After wards, in sub step, we tend to share our corpus with other researches who are interested to make a comparative study on the RNE task in Arabic language.

Our corpus consists of 870 heterogeneous articles. They were gathered from various sources of Arabic electronic newspapers such as "البيان/AlbyAn", "الجزيرة/Ajazeera", "الشروق/Al$rwq", "الحياة/AlHyAp" and from Wikipedia[1]. The present work focuses on the possible relations between couple of NEs from Person (PERS), Location (LOC) Organization (ORG) and Date (DATE). The choice of these types of NEs is motivated by their high frequency in the majority of electronic texts. Our corpus is composed of a set of 1,245 sentences containing at least a pair of NEs. These sentences were automatically annotated with:

- Morpho-syntactic analyzer which provides the Part of speech (POS) tagging of each word. This information is retrieved using linguistic resources elaborated by (Mesfar, 2008) like Verbs, nouns and adjectives dictionaries as well as some lexical and syntactic grammars elaborated through the linguistic platform NooJ.
- Clauses splitter which is a cascade of finite-state transducers elaborated by [9], proceeds to split long sentences into a set of clauses.
- Arabic NE recognition [10] that is presented as a set of syntactic grammars to recognize different types of NEs.

Otherwise, segmentation, NE tagging and morphological annotation errors were manually rectified in order to obtain an efficient relation recognizer. The NEs distribution along the different types and the characteristics of our corpus is presented in Table 1.

At the relation annotation stage, we just identify the word that can predict a semantic relation between NEs presented within a clause. Therefore, any word on the sentence should be annotated as one of the following tags:

[3] Italian Content Annotation Bank: an Italian corpus annotated with temporal expressions and four named entity types (person, organization, location and geo-political entity).

[4] Available on `http://www.wikipedia.org/`

Table 1. Gold corpus characteristics

Types	Sentences	Words	PERS	LOC	ORG	REL	P-REL
Number	1245	10234	966	764	257	1709	923

- Rel= A word expresses a relation between NEs pair.
- PRel= A word expresses only a part of a relation between two NEs in a given clause.
- N= A word doesn't enclose a relation.

For the sentences that contain more than two NEs, we treated each related pair of entities separately. Unlike proposed ACE annotation guidelines, the negatively defined relations will be taken into account in our annotation (e.g. أحمد ليس في كندا/Ahmed is not in London) and as the output, these relations will be deduced as negatively defined relationship. Then, three Arabic linguistic experts were asked to predict which word or a sequence of words can define semantic relation between NEs within a clause. We provided them with detailed description of our relation extraction task as well as our main goal. The inter-annotator agreements are computed from which we get a promising Cohen kappa of 79%. The focal disagreements came from some examples in which a relation cannot be predicted directly from words. Moreover, some relations are expressed through more than one word, which poses little disagreements between our linguistic experts.

5 Features

Each word in a given clause is assigned to a set of learning features, in order to built the learn data base. We investigated:

- The POS information of a word.
- The POS of the three words before and after this word.
- Grammatical structure of clause: To simplify the relation entity extraction task, we focused on a clause rather than a sentence. We added the grammatical structure of of a clause which is provided by the Stanford tagger [6].
- The semantic features concerned the NE tags which can be PERS, ORG LOC and DATE.
- Numeric features included the position of a word in the clause, the position according to the first NE as well as the second NE, the number of words in the sentence, number of words before, between and after the second NEs and the number of characters of each word.

Once these features assignments are done, we build 10,234 instances which is presented as a set of pairs (attribute or feature and its corresponding value) and a class label. We enclosed three classes: ***Rel, PRel*** and ***N***.

6 Experiments and Results

To avoid the over fitting, all the reported experiments are done with 10-fold cross-validation on our entire data-set by means of standard evaluation metrics. In order to

analyze how the learning procedure can be influenced by the instances number, we have computed a learning curve, by dividing our corpus into different learning sets. For each set, we apply the DT algorithm.

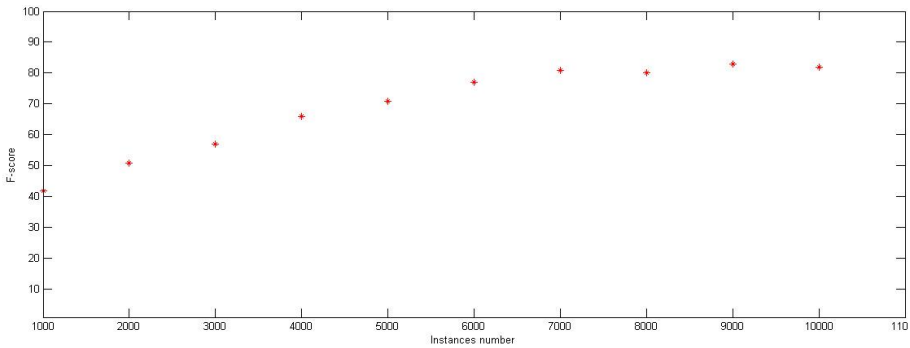

Fig. 1. F-score behavior for each instances number (using Adaboost)

The F-score curve shows that the curve grows regularly between 0 and 7,000 instances while it seems to plateau between 6,000 and 10,000 instances. We can thus conclude that the addition of more than 10,000 instances will only slightly increase the performance of the relation extraction task. Four our learning model, we investigated six ML techniques. They have been examined individually in order to choose the best technique to be applied on our Arabic non standard corpus. Firstly, we adopted the MaxEnt technique since it has been successful used not only in the RNE [1] but in many other NLP tasks, peculiarly NE recognition [3]. Similarly, we applied the DT and PART algorithms which are used also in [5] and the support vector machine (SVM) which is used in [7]. Other algorithms in literature are used to make a comparative study. The DT (C4.5), SVM, Adaboost (with DT), PART and Naïve Bayes are available in WEKA[5]. While, the MaxEnt is available on NLP Stanford tools[6]. Table 2 shows the system's performance in terms of precision, recall and F-measure when applied to our gold corpus.

Table 2. Results of different supervised algorithms

MaxEnt	PART	Decision	Tree	Adaboost	Naïve Bayes	SVM (SMO)
Precision	57.4	82.1	82.36	84.43	53.9	86.5
Recall	64.4	75.3	72.26	80.16	71.2	84
F-score	60.7	78.2	76.56	82.13	58.1	85.23

[5] http://www.cs.waikato.ac.nz/ml/weka/

[6] http://nlp.stanford.edu/software/classifier.shtml

According to the empirical results illustrated in this table, the used algorithms obtained very competitive scores. The highest performance of our system is accomplished by PART and Adaboost classifiers. Experimental study exhibits that SVM significantly outperforms other algorithms in term of precision. To conclude, SVM and Adaboost performed well on the entity relation detection task.

When examining the output of our process, we can deduce that some relations are difficult to be extracted from a word or a sequence of words. They need to be understanding from the meaning of previous sentences, or the main subject of the text from which this clause is extracted. Moreover, some words are not recognized even though they express a relation. This can be caused by the non recognition of the right category of NEs or the ambiguity in determining the POS tag of words.

7 Conclusion

Through this paper, we described our supervised process relANE to extract relationship between Arabic NEs. Our main goal was to study various features of each word in the sentence in orderto predict which term can explicit a relationship. Several supervised algorithms were applied.But mainly, the SVM and Adaboost techniques proved to be efficient for the NE relation task.

For a future work, we obviously intend to classify supported words into adequate level of semantic relationship classification. Similarly, we seek to evaluate our process on a standard corpus such as ACEdatawhich is not available yet. We are also considering the possibility of investigating other features to boost the overall performance of our system.

Acknowledgments We thank our developers Sana Trigui and Marwa Ben Hammouda for their efforts and their feedback to design our extractor interfaces.

References

1. Alotayq, A.: Extracting Relations between Arabic Named Entities, In: Proceedings of TSD 2013, pp. 265–271. Pilsen, Springer-Verlag Berlin Heidelberg, (2013)
2. Hamadou, A.B., Piton, O. and Fehri. H.: Multilingual Extraction of functional relations between Arabic Named Entities using NooJ platform. hal-00547940- version 1. (2010)
3. Benajiba. Y. Rosso. P. Benedi, J. M.: ANERsys: An Arabic Named Entity Recognition system based on Maximum Entropy. In:CICLing-2007, Springer-Verlag, Berlin, Heidelberg (2007)
4. Boujelben, I., Jamoussi, S. and Ben Hamadou, A.: Rules based approach for semantic relations extraction between Arabic named entities. In: NooJ 2012, in INALCO-Paris (2012)
5. Celli, F.: for Semantic Relations between Named Entities in I-CAB, (technical report available at `http://clic.cimec.unitn.it/fabio`) (2009)
6. Green, S. and Manning, C.: Better Arabic Parsing: Baseline, Evaluations and Analysis, In: 23rd International Conference on Computational Linguistics COLING2010, Beijing, China (2010)
7. Gumwon, H.: Relation Extraction Using Support Vector Machine. In: IJCNLP 2005, LNAI 3651, pp. 366–377 (2005)

8. Hasegawa, T., Sekine, S. and Grishman, R.: Discovering relations among named entities from large corpora. In: Proceedings of Association for Computational Linguistics. Morris-town, NJ, USA (2004)
9. Keskes, I., Benamara, F. and Belguith. L.: Clause-based Discourse Segmentation of Arabic Texts. In: Language Resources and Evaluation LREC, pp. 2826–2832. Istanbul (2012)
10. Mesfar, S.: Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en Arabe standard, University of Franche-Comté. (2008)
11. Zhang, Z. Weekly supervised relation classification for information extraction. In: CIKM 2004, Washington D.C., USA. (2004)
12. Zhang M., Su, J., Wang, D.Zhou, G., Tan, and C.L.: Discovering Relations between Named Entities from a Large Raw Corpus Using Tree Similarity-Based Clustering. In: Proc. of IJCNLP '05, LNAI, vol. 3651, pp. 378–389 (2005)
13. Zhao, S., Grishman, R.: Extracting Relations with Integrated Information Using Kernel Methods. In: 43rd ACL Meeting, Ann Arbor (2005)
14. Zhou, G., Qian, L. and Zhu, Q.: Label propagation via bootstrapped support vectors for semantic relation extraction between named entities. In: Computer Speech & Language 23(4), pp. 464–478 (2009)