MedCrave
Step into the World of Research

International Robotics & Automation Journal

Review Article

Open Access

CrossMark

Towards relation extraction from Arabic text: a review

Abstract

Semantic relation extraction is an important component of ontologies that can support many applications e.g. text mining, question answering, and information extraction. However, extracting semantic relations between concepts is not trivial and one of the main challenges in Natural Language Processing (NLP) Field. The Arabic language has complex morphological, grammatical, and semantic aspects since it is a highly inflectional and derivational language, which makes task even more challenging. In this paper, we present a review of the state of the art for relation extraction from texts, addressing the progress and difficulties in this field. We discuss several aspects related to this task, considering the taxonomic and non-taxonomic relation extraction methods. Majority of relation extraction approaches implement a combination of statistical and linguistic techniques to extract semantic relations from text. We also give special attention to the state of the work on relation extraction from Arabic texts, which need further progress.

Keywords: relation extraction, arabic nlp, arabic semantic relation extraction, arabic ontology construction

Volume 5 Issue 5 - 2019

Abeer AlArfaj

Department of Computer science, Princess Nourah Bint Abdul Rahman University, Saudi Arabia

Correspondence: Abeer AlArfaj, Department of Computer science, College of Computer and Information Sciences, Princess Nourah Bint Abdul Rahman University, Saudi Arabia, Email aaalarfaj@pnu.edu.sa

Received: December 17, 2019 | **Published:** December 24, 2019

Introduction

Relation extraction is an important aspect of ontology construction. semantic relation extraction between concepts in text used approaches based on the co-occurrence statistics of specific terms and machine learning approaches, as well as more linguistic approaches based on pattern or extraction rules or hybrid approaches which combines these two techniques.

Methods for sematic relations extraction can be classified according to the learning paradigm they employ as supervised and unsupervised. Supervised approaches task is to identify which types of relation hold between concepts using predefined relations. Various machine learning algorithms have been used for relation extraction, including Support Vector Machine, Conditional Random Fields and Maximum Entropy. However, supervised methods require annotated training data and predefined relations. For example, Zhou et al,[1] proposed a semi supervised method that uses labeled and unlabeled relation instances to learn sematic relation between named entities.

In ontology construction we need to extract unknown relations rather than known relations, therefore supervised approaches are ineffective. While unsupervised approaches seeks to find unknown relations which useful for ontology construction.[2]

Several studies have explored unsupervised approaches.[3,4] applied association rules to find relation between concepts. To label the extracted relations, they asked an expert to specify labels for those relations. On other hand,[5,6] used verbs to label extracted sematic relations between concept pairs. Also,[7] utilized the distributions of co occurring concepts and verbs as significant measures to identify verbs as sematic label. Serra et al,[8] proposed PARNT, which is a novel approach that supports ontology engineers in extracting semantic relations from corpora.

The Arabic language compared with the English language has a much more complex syntax. So, the need for new methods to construct ontology from Arabic texts is growing.

The Arabic ontology is a necessary knowledge for applications that process Arabic documents.[9]

For Arabic language, there are fewer references to existing work dealing with relation extraction.

Contribution of this paper are as follows:

a. We present a brief overview of relation extraction from Arabic text.

b. We classify the existing Arabic relation extraction approaches.

c. We discuss the challenges facing researchers in extracting semantic relations from Arabic texts and the way these challenges might be solved.

This paper is organized as follows. After the introduction, Section 2 describes the approaches for taxonomic relation extraction. Section 3 characterizes what is a relation, the techniques used to extract semantic relations between concepts, and Section 4 discusses recent works on relation extraction from Arabic text. Finally, Section 5 presents some concluding remarks.

The following subsections provide a detailed description of the most common approaches for relation extraction.

Taxonomic relation extraction

There are three main approaches for taxonomic relations extraction from text. The first one is the lexico syntactic patterns such as Hearst patterns.[10] Although, this approach have high precision, their recall is very low. This due to that these patterns occur rarely in the corpus.

Thus, we need to process large corpora to find more patterns. For this reason, recently several researchers have attempted to match these patterns on the web. A further drawback of the approach that is based on lexico syntactic patterns is that the patterns are specified in the regular expression form which is difficult to cope with language variety. Also, the learned relations between words forms rather than between senses of concepts.[11] To overcome this we suggest combining Hearst's and statistical methods.

The second approach is based on Harris distributional hypothesis in which concept hierarchies have been extracted from text using hierarchal clustering algorithms.[12] In clustering approaches we can accomplish two tasks: concept formation and concept hierarchy induction. Because clusters of similar words have been created to represent concepts and further order these clusters hierarchy. Several problems are raised when applying similarity based clustering techniques. One of them is that due to the sparse data, some similarities don't correspond to sematic similarities. Nevertheless, the distributional similarity hypothesis provides a useful model for ontology learning tasks.[11]

The third approach relies on the hypothesis that the occurrence of some words implies the occurrence of some other words in the same sentence, paragraph or documents indicated relations between both words.[13] The statistical based approach needs user intervention at validation phase to label relations and concept's cluster. However, this approach needs less preparation data than lexico-syntactic methods that need an expert for pattern preparation and construction.

Non-taxonomic relation extraction

Relation learning defined by Cimiano[11] as "a task of learning relation identifiers or labels r as well as their appropriate domain and range".

Relation extraction is an important aspect of ontology construction. Most of the existing approaches focus on the taxonomic relations extraction. There have only a few approaches addressing the issue of learning non taxonomic relations from text. Non-taxonomic relations is the relation between concept pairs except IS-A relation. For example, the meronomic relation that holds between two concepts where one concept is a part of the other concept (part-whole or part-of relation).[14]

In the current research concerning non taxonomic relation extraction, the existing approaches can be classified into the following:

a. Statistical approach relies on the distributional hypothesis through co occurrence distribution of words. In order to find associations between words, we look for the strong associations between words within a certain window of words, a sentence, paragraph or document.[11]

b. lexico syntactic approach relies on patterns matching to extract non taxonomic relations between concepts. Hearst's pattern can be used for learning the domain relations such as part-of, cause, purpose, etc. Other researchers learning labeled relations by exploiting syntactic dependencies between verbs and its argument. The interaction between the participants specified by verbs that usually express relations between them.[15]

c. Hybrid Approach

In order to overcome the deficiencies of using linguistic approach alone or statistical approach, the current approaches use both pattern matching and statistical analysis based on co-occurrence.

Review of semantic relation extraction from arabic text

Most of the existing relation extraction studies have been proposed for English language, For Arabic language, there are fewer references to existing work dealing with relation extraction.

Pattern based

For Arabic, the most existing studies based on the Hearst patterns,[7] in which a set of basic domain independent patterns for relation extraction and a methodology for obtaining new patterns are proposed.

Mazari et al[16] used repeated segment technique to determine the concepts that are relevant to the specific domain. The authors assumed that the more repeated concepts or phrases the more related to the domain. Also, they used filtering mechanism to remove incorrect segment.

Imam et al[17] used the method described in[16] for the relation extraction to build ontology based summarization system.

AL-Zamil & Al-Radaideh[18] used an enhanced version of Hearst's pattern to an Arabic corpus. Their enhanced algorithm include: pattern enrichment, pattern filtering, the application of negative patterns and pattern evaluation. Their evaluation results reached 78.57% average precision and 80.71% average recall.

Al-Yahya et al[19] presented a pattern-based and seed ontology method for extraction of antonyms from Arabic corpus. The extracted patterns then used used to discover new antonym pairs to enrich ontology. Their evaluation results showed that the system enriched ontology with 400% increase in size. For extracting new antonyms, their result showed only 2.7% of the patterns were useful. One disadvantage of this method is the cost for obtain a high recall is very expensive. The method can be integrated in a hybrid framework to increase their recall using statistical method.

Boudabous et al[20] proposed a hybrid method for Arabic ontology construction based on Wikipedia. They used a linguistic method based on morpho lexical patterns to improve AWN (Arabic WordNet) by adding sematic relations between synonymy sets. They first define morpho lexical patterns then use it for sematic relation enrichment.

Sarhan et al[21] proposed a semi supervised pattern based bootstrapping technique to extract semantic relations between entities. They experimented their method with two corpora which differ in size and genre, reaching a highest F measure of 75.06%.

Statistical

Another studies for Arabic used the statistical approach that based on co-occurrence technique and machine learning algorithms to discover relations. Harrag et al[22] used association rule mining technique to extract relations among concepts in hadith text collection.

Alotayq[23] proposed a relation extraction algorithm based on MaxEnt classifier, which resulted in 85% accuracy.

El-salam et al[24] presented a semi supervised method for relation extraction from web. Their method is an iterative process consisting of pattern extraction and instance extraction.

A supervised method for relation extraction is proposed in,[25] which is a cross language method that considered the lexical and syntactic features. The proposed method relied on the Universal Dependency (UD) parsing and the similarity of UD trees in different languages. Their result showed that 63.5% F1 for Arabic data set.

Hybrid Approach

Hybrid approaches combine statistical measures with linguistic features and takes the advantages of both. Lahbib et al[26] proposed a distributional approach for calculating similarities, which is based on syntactic dependencies to extract semantic relations. They first extracted noun phrases then transformed them into semantic relations. They used a morphological analyzer, syntactic analyzer and statistical measures to compute similarity between terms and syntactic relations. Their experimental results showed that, their method outperformed the co-occurrence method. They achieved 60% as the most decreased rate compared to 67% as the best result for the co-occurrence method. They observed that their approach and co-occurrence method can extract the same relations in some cases. And complement each other in other cases. Thus the syntactic dependencies are complementarily with co-occurrence method.

Bounhas et.al[27] used syntactic relations derived from in the structure of multi word terms to link terms. Then the graph of syntactic dependencies is transformed by distributional analysis. A clustering algorithm using the number of circuits in the graph is employed to cluster terms using Hierarchical Small-Worlds Networks to connect and group terms. They compared their approach to co-occurrence and derivational based approach and they concluded that the syntactic based approach is more cost efficient. However, their approach needs a syntactic parser, which is costly and make the method less robust especially in ambiguity language like Arabic language.[28,29]

Table 1 shows a summary of some works on relation extraction from Arabic texts. Based on the conducted studies, existing works of Arabic relation extraction can be classified into the following approaches: Pattern based approach, statistical approach and hybrid approach.

Table 1 Classification of Arabic Relation extraction approaches

Approach	Research	Extraction method
Pattern based	Mazari et al[16]	Repeated segment technique and filtering mechanism to remove incorrect segment
	Imam et al[17]	The method described in[16]
	Boudabous et al[20]	Morpho-lexical patterns definition and semantic relations enrichment
	AL-Zamil & Al-Radaideh[18]	An enhanced version of Hearst's Algorithm
	Al-Yahya et al[19]	Pattern-based and seed ontology
	Sarhan et al[21]	Semi-supervised pattern-based Bootstrapping technique
Statistical approach	Harrag et al[22]	association rule
	Alotyak[23]	Machine-learning-based algorithm based on MaxEnt classifier, which uses morphological and POS information
	El-salam et al[24]	A semi-supervised pattern extraction and instance extraction
	Taghizadeh et al[15]	Supervised learning used the training data of other languages and trains a model for relation extraction from Arabic text.
Hybrid approach	Bounhas et al[27]	Syntactic parser clustering algorithm to cluster terms using Hierarchical Small-Worlds Networks
	Lahbib et al[26]	Distributional approach for similarity calculus syntactic dependencies to extract semantic relations

Conclusion

In this paper, we have reviewed a number of methods that address the relation extraction problem in terms of their strengths and weaknesses in extracting semantic relations between concepts. Majority of relation extraction approaches implement a combination of statistical and linguistic techniques.

Extracting relations between concepts is an important layer for ontology construction from texts. Several methods have been proposed to extract semantic relations. Techniques for relation extraction can be classified as Lexico-syntactic, Statistical approach or a hybrid of both.

Linguistic methods provide a high precision but their recall is very low. The use of linguistic patterns allow named relation to be discovered, however the patterns are specified in the regular expression form which is difficult to cope with language variety.

To identify indirect and implicit relation, a statistic based approach such as co-occurrence and clustering analysis are used. Co-occurrence techniques based on the analysis of large domain corpora which are not always available for specific domains.

We have briefly discussed the importance of relation extraction from Arabic texts. Further, we have provided a brief overview of

some works on Arabic relation extraction followed by a summarizing comparison of them in Table 1. Based on the conducted studies, existing works on Arabic relation extraction can be classified into the following approaches: pattern based approach, statistical approach and hybrid approach.

A growing trend in relation extraction from Arabic texts that exploit Arabic WordNet to add label for relation between concepts. However, this method is unable to handle new terms which do not exist in this resource. We need a method to extract Arabic semantic relations between concepts and to enrich the existing one.

Funding

None.

Acknowledgments

We would like to thank Prof. AbdulMalik AlSalman for his valuable comments.

Conflicts of interest

The author declares that there was no conflicts of interest.

References

1. Zhou G, Li J, Qian L, et al. Semi–supervised learning for relation extraction. *Proceedings of international joint conference on natural language processing (IJCNLP08)*. 2008:32–38.
2. Shen M, Liu D, Huang Y. Extracting Semantic Relations to Enrich Domain Ontologies. *Journal Intelligent Information System*. 2012;39(3):749–761.
3. Maedche A, Staab S. Discovering Conceptual Relations from Text. *Proceeding of the 14th European Conference on Artificial Intelligence(ECAI)*. 2000:321–325.
4. Maedche A, Volz R. The Ontology Extraction Maintenance Framework Text–To–Onto. *Proceedings of the IEEE International Conference on Data Mining*. 2001:1–12.
5. Kavalec M, Maedche A, Svatek V. Discovery of lexical entries for non–taxonomic relations in ontology learning. *International Conference on Current Trends in Theory and Practice of Computer Science*. 2004;293:249–256.
6. Weichselbrauna A. Wohlgenannta G, Scharl A. Refining non–taxonomic relation labels with external structured data to support ontology learning. *Data and Knowledge Engineering*. 2010;69(8):763–778.
7. Punuru J, Chen J. Learning non–taxonomical semantic relations from domain texts. *Journal of Intelligent Information Systems*. 2011;38(1):191–207.
8. Serra I, Girardi R, Novais P. PARNT: A Statistic based Approach to Extract Non–Taxonomic Relationships of Ontologies from Text. Proceeding of 10th International Conference on Information Technology: New Generation (ITNG); 2013 Apr 15–17; Las Vegas: IEEE; 2013. p. 561–566.
9. Al–Arfaj A, Al–Salman A. Towards Ontology Construction from Arabic Texts– A Proposed Framework. Proceeding of The 14th IEEE International Conference on Computer and Information Technology (CIT 2014). 2014:737–742.
10. Hearst MA. Automatic acquisition of hyponyms from large text corpora. The 15th Conference on Computational Linguistics. 1992. p. 539–545.
11. Cimiano P. Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. *Theoretical Computer Science*. 2006.
12. Cimiano P, Hotho A, Staab S. Learning Concept Hierarchies from text corpora using Formal Concept Analysis. *Journal of Artificial Intelligence Research (JAIR)*. 2005;24:305–339.
13. Cimiano P, Madche A, Staab S, et al. Ontology Learning. In: Staab S, Studer R, editors. *Handbook on Ontologies*. Germany: Springer–Verlag Berlin Heidelberg; 2009. p. 245–267.
14. Wong M, Raza Abidi S, Jonsen I. A multi–phase correlation search framework for mining non–taxonomic relations from unstructured text. *Knowledge and Information Systems*. 2014;38(3):641–667.
15. Schutz A, Buitelaar P. Relext: A tool for relation extraction from text in ontology extension. *International Semantic Web Conference*. 2005. p. 593–606.
16. Mazari A, Aliane H, Alimazighi Z. Automatic Construction of Ontology from Arabic texts. *Proceedings of International Conference on Web and Information Technologies (ICWIT)*. 2012:193–202.
17. Imam I, Nounou N, Hamouda A, et al. An Ontology–based Summarization System for Arabic Documents (OSSAD). *International Journal of Computer Applications*. 2013;74(17):38–43.
18. AL–Zamil M, Al–Radaideh G. Automatic Extraction of Ontological Relations from Arabic Text. *Journal of King Saud University – Computer and Information Sciences*. 2014;26(4):462–472.
19. Al–Yahya M, Aldhubayi L, Al–Malak S. A Pattern–Based Approach to Semantic Relation Extraction Using a Seed Ontology. *2014 IEEE International Conference on Semantic Computing*. 2014;1:96–99.
20. Boudabous M, Kammoun N, Khedher N, et al. Arabic WordNet Semantic relations enrichment through morpho–lexical patterns. 1st International Conference on Communications, signal processing, and their Applications(ICCSPA); 2013Feb 12–14; Sharjah: IEEE; 2013. p.1–6.
21. Sarhan I, El–Sonbaty Y, El–Nasr M. Semi–Supervised Pattern–Based Algorithm for Arabic Relation Extraction. *proceeding of IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*. 2016:177–183.
22. Harrag F, Alothaim A, Abanmy A, et al. Ontology Extraction Approach for Prophetic Narration (Hadith) using Association Rules. *International Journal on Islamic Applications in Computer Science And Technology*. 2013;1(2):48–57.
23. Alotayq A. Extracting relations between Arabic named entities. *International Conference on Text, Speech and Dialogue*. 2013;8082:265–271.
24. El–salam S, El Houby E, Al Sammak A, et al. Extracting Arabic relations from the web. *International Journal of Computer Science & Information Technology (IJCSIT)*. 2016;8(1):85–102.
25. Taghizadeha N, Failia H, Malekib J. Cross–Language Learning for Arabic Relation Extraction. *Procedia Computer Science*. 2018;142:190–197.
26. Lahbib W, Bounhas I, Elayeb B, et al. A Hybrid Approach for Arabic Semantic Relation Extraction. *Proceeding of The 26th International Florida Artificial Intelligence Research Society (FLAIRS)*. 2013. p. 315–320.
27. Bounhas I, Elayeb B, Evrard F, et al. Arab Onto: Experimenting a new distributional approach for building Arabic ontological resources. *International Journal of Metadata, Semantics and Ontologies (IJMSO)*. 2011;6(2):81–95.
28. Snchez D, Moreno A, Vasto–Terrientes L. Learning relation axioms from text: An automatic Web–based approach. Expert Systems with Applications. 2012;39(5):5792–5805.
29. Harris ZS. In Papers in structural and transformational linguistics. 1970;775–794.