ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 12(46), DOI: 10.17485/ijst/2019/v12i46/147512, December 2019*

Relation Extraction from Arabic Wikipedia

Gehad Zakria[1,*], Mamdouh Farouk[2], Khaled Fathy[2] and Malak N. Makar[1]

[1]Department of Mathematics, Faculty of Science, Assiut University, Egypt; gehad.zakria@aun.edu.eg, malak@aun.edu.eg

[2]Department of Computer Science, Faculty of Computers and Information, Assiut University, Egypt; mamfarouk@fci.au.edu.eg, khaled@fci.au.edu.eg

Abstract

Objectives/Methods: This study aims to extract relations between entities from Arabic text. RelationExtraction is one of the most important tasks in text mining. Relation extraction is considered as a main step for many applications such as extracting triples from the text, Question Answering and Ontology building. However, extracting relations from the Arabic text is a difficult task compared to English due to lack of annotated Arabic corpora. This paper proposes a method for extracting relations from Arabic text based on ArabicWikipedia articles characteristics.The propose system extracts sentences that contain principle entity, secondary entity and relation from Wikipedia article, then we use WordNet and DBpedia to build the training set. Finally Naive Bayes Classifier is used to train and test the datasets. **Finding:** There are few works to extract relations from Arabic text. These works depend on classification, clustering and rule based. **Application/improvement:** The experiments show the effectiveness of the proposed approach which achieves high precision with 89% for classifying 19 type of semantic relations.

Keywords: Relation Extraction, Arabic Wikipedia, Semantic Relation, Arabic language.

1. Introduction

There is a significant increase in information on the Internet. Therefore, the recent works in semantic web aim to access this information semantically to answer queries that exceed the capabilities of the standard search engines.Wikipedia (www.wikipedia.org) is one of the most important sources of information that is a Wikimedia Foundation project. Wikipedia is the largest online encyclopedia in the world. It posts its articles every day by its founders.

The Information Extraction (IE) aims to extract structured information from unstructured text by extracting the entities from the text and identifying relations between them automatically. Relation Extraction (RE) is the main task in Information Extraction (IE) from the text.

The task of relation extraction is considered as the process of finding a relation between two entities in a sentence. Moreover,[1] classified the two entities as the principal entity and secondary entity. In the case of the Wikipedia page, the title of the article is considered as a principle entity, and the secondary entity is other entity that appears on the page. Most of the previous works in this field were in other languages such as English, French and Chinese. However, there are few works in this field related to the Arabic language.

Moreover, the process of extracting relations from text becomes a difficult task in Arabic text because Arabic has many special characteristics. For example, Arabic has the diacritical marks that change the meaning of the sentence. In addition, there are many challenges in Arabic language processing.[2–4] Some of these challenges should be considered in relation extraction task such as:

- Existence of implicit relations between two entities.
- The absence of the actor in the Arabic sentences. This causes some problems such asdetermining the distance between the two entities.
- There is no positive relation in view of negative words in a sentence.

*Author for correspondence

- There are multiple relations between the same entities.

This study aims to deal with the Arabic Wikipedia to extract relations between the principle entity and the secondary entity. In Arabic sentences, the two entities may be mentioned explicitly. Sometimes, the first entity does not appear in the sentence. The proposed system extracts relations from unstructured Arabic text. Both the principle entity and secondary entity are determined as a first step. The features (lexical and syntactic) are extracted. Based on the extracted features, the proposed system detects the correct relations.

This study is organized in the following manner: Section 2 mentions some of the related work on relation extraction especially supervised methods. Section 3, is talking about the characteristics of Wikipedia articles. Section 4,explains the proposed method in detail. Section 5, showsthe results obtained from the experiments. Finally, section 6 concludes this study.

2. Literature Survey

Many approaches have been proposed for relation extraction from text. However, few works have been done for extracting relations from Arabic text.[5] Nanda Kambhatla builds a maximum entropy model to apply it on Automatic Content Extraction (ACE) evaluation to extract six explicit relations from a sentence. Her approach extracts 49 relations from text based on a set of extracted features from the sentences are Table 1. She uses the following features: Words between two entities, number of words between entities, type of each entity, Part of Speech (POS), the path of Penn tree and type of sentence. Her study combines between lexical, syntactic and semantic features.[6]

Table 1. Extracted features for the sentence "نفيتس اتنالتأ يف 1973 ويلوي 29 ديلاوم"

POS of words:	[Athlete, NN-1, CD-2, NN-3, CD-4, IN-5, Location]
Syntactic path:	[NNP-0, NN-1, CD-2, NN-3, CD-4, IN-5, NNP-6]
POS for the verb:	No-verb
Principle entity type:	Athlete
Secondary entity type:	Location
Length:	5
Subject:	Person

In[7] proposed an approach for extracting relations from text. His approach depends on Zhu Zhang's work which uses Support Vector Machines to classify relations. However, he improves some points such as:

- Enlarge the data size and increase performance.
- Detect the relation before making the classification.

Moreover, the extracts the explicit relations from a sentence that contains two entities. His approach uses the ACE[8] corpus as an annotated dataset to extract 49 relations from text based on the features extracted from the sentences. He uses thefollowing set of features: Lexical tokens, syntactical structure, semantic entity type, and distance between entities. In addition, DUC segmenter[8] and Charniakparser[9] is used to extract features from sentence.[8]

In[10] introduced a new approachto extracted 13 relations from Wikipedia text. Each sentence should contain at least one entity. The principle entity which appears in the first sentence in the Wikipedia page is considered as the second entity. Their study uses the Naive Bayes as a classifier.[10]

Moreover, Othman Ibrahim et el producedsome rules for extracting relations from nominal and verbal sentences.These rules are appliedto80 sentences. They made evaluation by comparing their system output and human results.[11]

On the other hand, applied the idea of distant supervision in relation extraction from Arabic text.[12] DBpedia is used as the knowledge base along with large unlabeled corpora to make the labelled data. They extracted sentences from Wikipedia and tagged these sentences using DBpedia relations. They combined lexical, syntactic and Arabic specific features to train and test the classifier.

Furthermore, the rule-based approach is also used to extract relations from the Arabic text. AS Rextractor tool[13] extracts and annotates semantic relations between Arabic named entities using the TEI (Text Encoding Initiative) formalism.[14] Transducer cascade is used for extraction and annotation process. However, rule-based needs a large amount of manual works in a specific domain.

3. Characteristics of Wikipedia

Wikipedia articles have many characteristics. Wikipedia pages are very cleaner than typical web pages. This

is because it has a high-quality structure. Moreover, linguistic technologies such as parser can be used in web understanding.[15] In addition, all Wikipedia articles are linked to each other via links.

Moreover, each article talks about a specific entity which is considered as the principle entity. In addition, each page in Wikipedia has a summary part, which contains other entities and relations between these entities and the principle entity. These links produce many interesting relations among Wikipedia articles.[16] So the proposed system exploits these characteristics to get labelled dataset. However, there are some pages that do not contain the summary part.

However, some pages refer to the same entity with different terms. For example, the two terms "ده‌عم ستسوشتاسام ايجولونكت" and "ده‌عم ستسوشتاسام ايجولونكتلل" refer to the same entity.

4. The Proposed System

As shown in Figure 1, the proposed system is divided into two main phases. The first phase is instance sentence extraction in which the sentence in Wikipedia page that expresses a specific relation is extracted and prepared for the feature extraction phase.The second phase extracts (syntactic, lexical) features from the selected sentences. Based on the extracted features the proposed system detects the type of relation.

4.1. Instance Sentence Extraction

The proposed system collects the sentences that contain the relation between a principle entity and a secondary entity. The proposed system pre-processes each sentence to be ready for feature extraction. The principle entity and secondary entity are determined for each sentence.

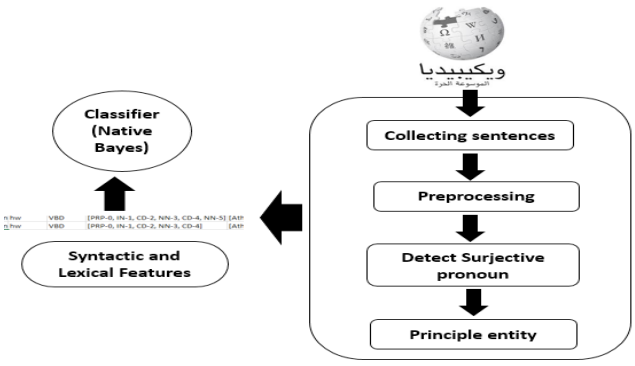

Figure 1. Proposed system architecture.

Moreover, in the Arabic sentence principle entity may be explicit, personal pronoun or absent pronoun.

4.1.1. Collecting Sentences

The sentences are collected according to the summary part of Wikipedia articles.For each entity-relation pair in the summary, a sentence which contains the relation and secondary entity is extracted. The proposed system extracts sentences that reflect on this relation.For example, if the secondary entity is "القاهره" "Cairo" and the associated relation in summary part is "Birthplace" the system will extract the sentence "ولد أحمد في القاهره" "**Ahmed was born in Cairo**" and will not extract "سافر أحمد الي القاهره" "**Ahmed traveled to Cairo**", because the second sentence does not contain any word reflects the relation.

4.1.2. Preprocessing

The boundaries are determined for each sentence, so each sentence contains only one verb. We used punctuations to divide the text into small sentences. Moreover, Latin words, brackets, strange symbols and any word after secondary entity are removed from the sentence. Many names in the Arabic language contain more than one syllable so the proposed system considers only the first syllable of each entity.

4.1.3. Detect Surjective Pronoun

Some sentences in the Arabic language do not contain a subject. In such a case, the proposed system uses Arabic WordNet[17] to determine suitable subjective pronouns. By using Word Net we able to express the subject with the right conscience, whether feminine or masculine. There are three types of sentences in the Arabic language: Nominal, verbal and sentence-like. In the case of the verbal sentence, the proposed system determines the subjective pronoun according to the verb and puts the pronoun after the verb. In other cases, the proposed system uses the principle entity instead of the absent subject at the beginning of the sentence.

4.1.4. Principle Entity Detection

In this step principle entity of Wikipedia page is detected. The following steps explain the used steps to detect the terms used as the principle entity. The output of these steps is a list of named entities that can be used as a principle entity.

We have empty Principle_entity set{}.

i. The first two names on Wikipedia page (the title of the article and the first name in the first sentence) can be considered as the principle entities. If one of them founded add it to the Principle entity list. If there is no entity on the page, stop.
ii. Each name derived from names defined in the initial step (i) is added to Principle_ entity list. For example, if the principle name is "محمد ابوتريكة" which translated to "Mohamed Abotrika", then the names "محمد" (Mohamed) and "ابوتريكة"(Aboutrika) are also added to the list.
iii. If a sentence contains personal pronoun we can consider it an alternative to principle entity and add it to the list.[8]

4.2. FeatureExtraction

In this step, a set of features is extracted from the sentence to be used for the training phase. DBpedia is used to determine the type of each entity such as (Date, Location, and Time period), Moreover, Stanford parser[18] is used to generate parse tree and determine part of speech (POS) for each word in the sentence. The following subsections explain the different features used in the proposed approach.

4.2.1. Lexical Feature

Lexical features are used as a comprehensive description of the two entities and words surrounding them in the sentence.

The first feature is the ***Name Entity Type.*** Determining the entity type is a very important task in relation extraction. There are some tools to extract entity type from the Arabic text. However, these tools do not recognize all types. In some systems, English DBpedia is used to identify the type of each entity.However, there is no Arabic DBpedia to find types of Arabic entities. The proposed system uses English DBpedia. After all, entities are extracted from Wikipedia articles,the proposed system extracts the hyperlink for each entity from the HTMLpage that refersto other Wikipedia articles, then accesses this link to get the English article. Finallythe type of entity id determined by using DBpedia.

The second feature is the ***Length of words between the two entities.*** Determining the number of words between the principle entity and the secondary entity is an important feature in relation classification. The maximum number allowed in the proposed system is (length=7).

The third lexical feature is ***Part of speech (POS).*** The proposed system uses Stanford parser to determine part of speech for all words between the two entities.

4.2.2. Syntactic Feature

Syntactic analysis is a process of text grammatical analysis. It is used to understand the sentence and divide it into segmentations. The proposed approach uses Stanford parser[18] to extract the syntactic path from the principle entity to the secondary entity.

All features are grouped into a vector to classify them by using the Naive Bayes Classifier.

For example:

« ستيفن نفرود فلاوم يوليو 29 **1973** في أتلانتا»

"Stephen Dorf was born July 29, 1973 in Atlanta"

After preprocessing, determining the principle and secondary entity send each of them to DBpedia to recognize the type.

« ستيفن نفرود فلاوم يوليو 29 **1973** في أتلانتا »

"Stephen was born July 29, 1973 in Atlanta"

5. Experiment

Naive Bayes Classifier has been used to solve the problem of relations extraction from the Arabic Wikipedia articles. In order to measure the effectiveness of the proposed approach, a new dataset is constructed based on Arabic Wikipedia.

5.1. Datasets

The constructed dataset is collected from 6,491 articles from Arabic Wikipedia. In[19] is used to train the dataset and make a model by using Naïve Bayes. This model is trained to extract 1**9** types of relations. The collected dataset contains 19823 instancessentences which divide them into 80% for training and 20% for testing. The proposed system can classify 3646 correctly and 319 incorrectly class from the testing model.

5.2. Results

The following results were obtained from testing proposed system:

We evaluated the model by collecting a number of sentences from different pages of Wikipedia randomly

regardless of the summary of each page. The number of test instances is 610 from 183 articles collected. The proposed approach achieves a good result in the test dataset. The achieved F-measure is 0.89 while precision and recall are 0.89 and 0.9 respectively.

The proposed system is compared with other systems such as.[12] Our system achieved high F-measure 89% in extracting 19 types of relations from the Arabic text which contains at least one entity, but ArabRelat extracts 97 types of relations with F-measure 70%. In ArabRelat, sentences must contain two entities. In addition, the proposed system uses fewer features compared to ArabRelat. The comparable results are shown in Table 2.

As shown in Table 2, the proposed system achieves better results. This because the adopted set of features in the proposed system is sufficient to capture the important features of instance sentence. However, ArabRelat uses many features which in some cases are noisy features. For example, when including the feature *type of sentence* (nominal, verbal and sentence-like), the number of sentences classified as error increased because of inaccuracy in Stanford parser.

Sometimes, Stanford parser deals with nouns as verbs and vice versa. The proposed system solves this problem by merely using the first name of each entity to avoid errors. Moreover, the proposed system is compared with the AS Rextractor tool.[13] AS Rextractor is depended on finite-state transducers to extract 18 Semantic relations. But their system depends on rule-based. Occasionally ruled based apply on a specific domain. This means for each domain, rules must be extracted. The compared results are shown in Table 2.

6. Conclusion

In this study,a new method was proposed to extract semantic relations from the Arabic text. The proposed approach exploits the properties of the Wikipedia pages to determine relations and secondary entities. The proposed system extracts all sentences that contain at least one entity (secondary entity) and a relation. This method extracts a set of syntactic and lexical features. The used features are carefully selected to avoid noisy features. Moreover, the Naive Bayes Classifier is used to train and test datasets. The results of the experiment achieve high precision 0.89 for classifying 19 types of semantic relations.

Table 2. Results of comparison with other systems

	Proposed system	**Arab Relate**	**ASRextractor**
Relations	19	97	18
No. dataset	19822	12415	–
No. feature	7	16	–
F-measure	.89	.70	.86

References

1. Culotta A, McCallum A, Betz J. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In: Proceedings of the human language technology conference of the North American chapter of the ACL; 2006.P. 296–303.
2. Boujelben I, Jamoussi S, Ben Hamadou A. A hybrid method for extracting relations between Arabic named entities. J King Saud Univ Comput Inform Sci. 2014;26:425–40.
3. Abd El-salam SM, El Houby EMF, Al Sammak AK, El-shishtawy TA. Extracting arabic relations from the web. Int J Comput Sci Inform Technol (IJCSIT). 2016;8:85–102.
4. Multilingual extraction of functional relations between Arabic named entities using NooJ platform. [cited 2010 May]. https://www.researchgate.net/publication/269167188_Multilingual_Extraction_of_functional_relations_between_Arabic_Named_Entities_using_NooJ_platform.
5. Ontology-based semantic representation for Arabic text: a survey. [cited 2017]. https://www.rgnpublications.com/journals/index.php/jims/article/view/1029.
6. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. [cited 2004]. https://dl.acm.org/citation.cfm?id=1219066.
7. Zhang Z. Weakly-supervised relation classification for information extraction. In: Proceedings of the 2004 ACM CIKM international conference on information and knowledge management;2004.P. 581–8.
8. Hong G. Relation extraction using support vector machine. In: Proceedings of the second international joint conference on natural language processing;2005.P. 366–77.
9. Charniak E. A maximum-entropy-inspired parser. In: Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference; 2000.P. 132–9.
10. Nguyen DPT, Matsuo Y, Ishizuka M. Subtree mining for relation extraction from Wikipedia. In: Proceedings human language technology conference of the North American Chapter of the Association of Computational Linguistics; 2007.P. 125–8.

11. Hammadi OI, Ab Aziz Aziz MJ. Grammatical relation extraction in Arabic language. J Comput Sci. 2012;8:891–8.
12. Mohamed R, El-Makky N, Nagi KM. ArabRelat: Arabic relation extraction using distant supervision. In: The 7th international conference on knowledge engineering and ontology development (KEOD 2015);2015. vol.2.P. 410–7.
13. Mesmiaa FB, Zidb F, Haddarb K, Maurelc D. ASRextractor: A tool extracting semantic relations between Arabic named entities. In: 3rd international conference on Arabic computational linguistics; 2017. vol.117.P. 55–62.
14. TEI P5: guidelines for electronic text encoding and interchange. [cited 2016]. https://tei-c.org/Vault/P5/2.4.0/doc/tei-p5-doc/en/html/.
15. Giles J. Internet encyclopaedias go head to head. Nature. 2005;438:900–901.
16. Gabrilovich E, Markovitch S. Overcoming the brittleness bottleneck using wikipedia: enhancing text categorization with encyclopedic knowledge. In: The twenty-first national conference on artificial intelligence and the eighteenth innovative applications of artificial intelligence conference; 2006. vol.2.P. 1301–06.
17. Elkateb S, Black B, Vossen P, Farwell D, Pease A, Fellbaum C. Arabic WordNet and the challenges of Arabic. In: Proceedings of Arabic NLP/MT conference; 2006. P. 15–24.
18. Better Arabic parsing: baselines, evaluations, and analysis. [cited 2009]. https://nlp.stanford.edu/pubs/coling2010-arabic.pdf.
19. Weka: a machine learning workbench. [cited 1994 Dec 29]. https://ieeexplore.ieee.org/document/396988.