Hierarchical N-gram Algorithm for extracting Arabic Entities

Eslam Amer
Computer Science Department,
Faculty of Computer & Informatics,
Benha University, Benha,
Egypt
eslam.amer@fci.bu.edu.eg

Heba M . Khalil
Computer Science Department,
Faculty of Computer & Informatics,
Benha University, Benha,
Egypt
heba.khalil@fci.bu.edu.eg

Tarek . El-shistawy
Information System Department,
Faculty of Computer & Informatics,
Benha University, Benha,
Egypt
t.shishtawy@bu.edu.eg

ABSTRACT
Entities Extraction becomes very important for developing many applications of Natural Language Processing (NLP). In this paper, we present a new algorithm to extract entities from Arabic text. The approach uses the semi-structured knowledge source: Arabic Wikipedia to predict the words that constitutes an Arabic entity. Our method is generic and can be applied directly to other languages to extract entities. The proposed method has been designed to analyze Arabic text hierarchically with variable length N-gram. The experimental results have proven that the proposed system is very efficient in detecting entities from large set of Arabic news.

Keywords
Natural Language Processing; Entity; N-gram; Arabic Wikipedia; Information Extraction

1. INTRODUCTION
One of difficult problems in understanding material language is the segmentation problem which is the process of dividing written text into meaningful units such as words, sentences or topics. The task is difficult as it requires information about the language and real world [1]. The recognition of entities and named entities define entities as recognized tokens such as person name, organizations, places, idioms and other entities. It plays an important role in building a syntactic and semantic levels analysis for information extraction systems [2]. It also helps to extract semantic relations between entities that are useful for a better understanding of human language [3].

Recently, many researches work introduced Wikipedia as an external resource to recognize information about the world [4, 5]. Wikipedia is a free online encyclopedia edited collaboratively by large numbers of volunteers [6]. Due to the openness and the editing policy of Wikipedia, Wikipedia has a lot of high-quality documents, which becomes an important external resource for information retrieval [7, 8].

NFOS '16, May 09-11, 2016, Giza, Egypt

DOI: http://dx.doi.org/10.1145/2908446.2908449

Many Entities recognition systems have been built for the Arabic language [9, 10]. Some of the proposed systems rely on extracting categories from Wikipedia [11]. Other studies use a machine learning (ML)-based approach [3, 12]. Another study that adopted the local grammar method was performed by [13] was first used to extract time, date and address expressions from letters. But this method detects special types of the entities.

We propose a new algorithm, which is based on Arabic Wikipedia. It aims to detecting entities from Arabic text through matching titles of Wikipedia articles. Also; the proposed algorithm exploits semantic sets of Wikipedia which matches also the same entity or the same meaning.

The proposed approach is characterized by:

1) Maximize the number of segmented words that describe a single entity. We follow the N-gram concept [14], which is based on segmenting the input sentence into variable number of tokens.
2) The algorithm is language-independent, and can be applied to any language.
3) The proposed algorithm is fast as it exploits search engines to detect the occurrence of the named entities within the article titles of the wiki pages. This does not need downloading all article or wiki databases.
4) The algorithm exploits the semantic sets of Arabic Wikipedia.

This paper is organized as follows: section 2 covers the related work. Section 3 describes the proposed method. Experimental results are shown on section 4. Finally section 5 presents the conclusion and future work.

2. RELATED WORK
Several studies have been carried out in the last few years on Information extraction using Wikipedia article to extract entities or extract the relation between entities [8, 11, 15, 16, 17].The extraction includes searching the title, hyperlinks, categories, info boxes and disambiguation data of Wikipedia articles[6].

Ruiz et al [8] proposed an automatic approach to identify lexical patterns which represent semantic relationships between concepts, from Wikipedia. Then, these patterns can be applied to extract new relationships that did not appear in Word Net originally. They depend on extracting the matched hyperlinks from Wikipedia. The experiments have been performed with the

Simple English Wikipedia and Word Net. This approach extracted 1,224 relations from Wikipedia for three different types of relations (Hyponymy, Holonymy and Meronymy) with a precision from 61 to 69 %.

Nguyen et al [18] proposed a method to extract relations among entities from English Wikipedia articles. First, they check the appearance of entities in Wikipedia articles. Then, they extracted features from Wikipedia using SVM and subtrees mined from the syntactic structure of text. According to the features they extracted the relationships among entity pairs. They selected 3,300 entities randomly and manually annotate their relations. For training 3,100 entities are used and 200 used for testing. Milne et al [15] developed an approach for calculating semantic relatedness between terms using the links between their corresponding Wikipedia articles. The first step in measuring the relatedness between two terms is to identify the concepts they relate to (the articles which relate to them). Then, extracting the hyperlinks of each article and the relatedness between them was calculated. Al-Rajebah et al [6] indicated that using hyperlinks within article text might lead to ambiguity. Because it cannot be guaranteed that the hyperlinks appeared within Wikipedia's articles may indicate semantic relations between concepts.

Chernov et al [11] presented one of the attempts to automatically extract the semantic information from Wikipedia by analyzing the links between categories. They depend on two measures of strong semantic connections between Wikipedia categories. One measure is the number of links between pages in two categories, and the other is Connectivity Ratio. They can be applied to in links or out links separately. They applied these measures to the English Wikipedia. Yin et al [19] presented a solution to improve the poor retrieval Performance. The solution was based on extract Wikipedia's category knowledge, they believed that if two Wikipedia articles are related, they both belong to some category and also believe that co-occurrences of terms in Wikipedia articles in the same category are also considered.

Wang et al [20] proposed the Positive-Only Relation Extraction (PORE) method to extract relations from Wikipedia using Wikipedia categories hierarchy, info boxes and hyperlinks. It was based on B-POL algorithm which is an extension of Positive-Only Learning algorithm, where negative examples are identified and then classified until they reach convergence. They evaluated their work by 3 human judges. Al-Rajebah et al [6] presented a method to extract the semantic relations from Arabic Wikipedia to build an Arabic ontology. They extracted info box and the list of categories from Wikipedia's article to calculate the semantic relations. To evaluate the quality of their system, they used two measures human judgment and precision. Our proposed system presented: (1) Exploiting Wikipedia to identify all known entities. By known entities, we mean exact expression, named entities, known events …etc. (2) the presented algorithm is general and can be applied to any language.

3. PROPOSED METHOD

The proposed approach is illustrated in Figure.1, which involves the Arabic version of Wikipedia. The input to the algorithm is Arabic sentence to be analysed, and the output is a set of recognized entities.

The system has two main components: Pre-processing, Matching algorithm. Here we describe each one:

3.1 Preprocessing

Sentence preprocessing is the first step in our algorithm process. This step aims to reduce the noise in input text by removing all the unnecessary terms such as non-alphabetical characters, numbers, markers, special characters, etc. In our work, traditional stop words such as prepositions, question tools, pronouns, etc., were not removed as they may play an important role in matching Arabic entities.

3.2 Matching algorithm

The second phase is the Matching algorithm process. The input of the Matching algorithm is the processed sentence. The algorithm makes use Arabic Wikipedia as a resource and N-Gram methodology algorithm.The matching algorithm detailed in Figure 1.

A Table 1 shows the steps of detecting entities using Arabic Wikipedia. Step2 of the algorithm is used to determine the end condition of the algorithm, according to the state of input text: empty or still have unrecognized words. When it has words, the algorithm continues, otherwise it terminates.

Table 1. Steps for matching algorithm

Our Proposed Algorithm for Matching algorithm	
Step1	Detect the input text length and put it equal N

Step2	IF Input text have words, then
	Tokenize the input text to entities by using the N-Gram algorithm
	ELSE
	There is no words in input text and stop

Step3	Matching the output of step2 with Wikipedia titles
Step4	IF step3 matches then remove it from input text ELSE Decrement N IF N=0 then stop ELSE go to step2
Step5	Repeat the same steps from 1:4 until input text finished.

The algorithm applies N-gram as large first methodology. First, the algorithm tries to match the maximum number of adjacent words, and therefore N is set to the length of the input sentence.

The segmented words are matched against the Arabic wiki titles. In case of matching, all segmented words are marked as an entity. Otherwise, N is decremented by one, and all permutation of N-tokens are checked for matching as a second iteration. The process is continued until N=0 or all words have been recognized.

Table 2 shows an example, the algorithm starts with N=4 we checked each one of them. If it matches with an article in Wikipedia until terms finished or N=0.

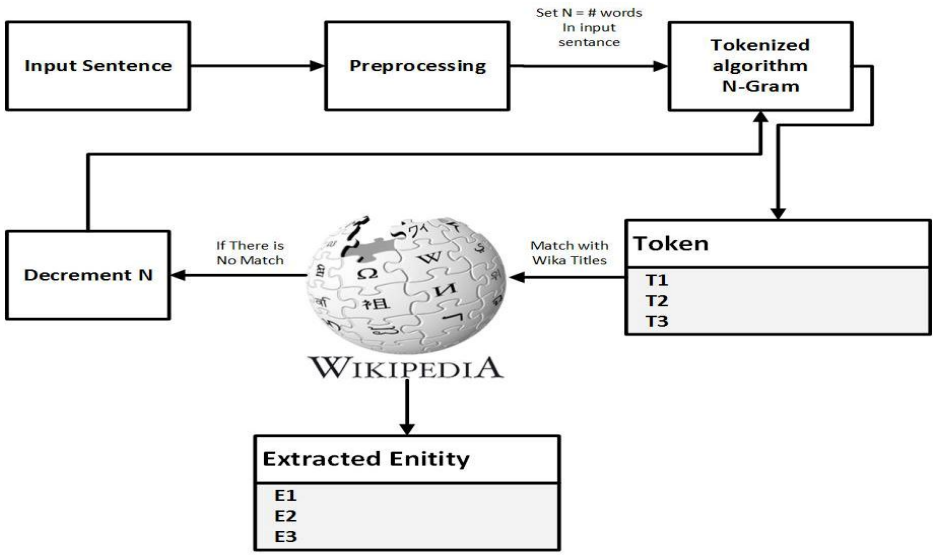

Figure.1 matching algorithm model

Table 2. Shows input sentence

Input Text In Arabic	Input Text In English
"التطور فى الشرق الاوسط"	"Development in the Middle East"

Table 3. shows the output terms matched when N=4:1

Tokens	N-Gram	Matched terms with wiki
التطور فى الشرق الاوسط	4	No Match
التطور فى الشرق	3	No Match
فى الشرق الاوسط		
التطور فى	2	الشرق الاوسط
فى الشرق		
الشرق الاوسط		
التطور	1	No Match

In table 3 when there is no match, the N decremented by one and matches again with Wikipedia. The extracted terms are marked when N=2 and the other words have been taken. The process is continued until N=0. Another example of exploiting wiki semantic set is shown in table 4.

Table 4. shows Wikipedia Semantic Set

Arabic Input Sentence	Extracted Entities	English Equivalent
كوكب الشرق علامه مصريه	**-ام كلثوم** -علامه	Star of the East mark Egyptian
ابوتريكه من نجوم منتخب مصر	-منتخب مصر **-محمد ابوتريكه** -نجوم	Aboutrika of Egypt Star team

4. EXPERIMENTAL RESULTS

4.1 Dataset

For testing our algorithm, we have collected dataset of 300 Arabic random news in different domains. This news belongs to different domains such as Events, Arts and Sports. The dataset is collected from an online Arabic newspaper (http://www.youm7.com/). The reason of building our own dataset is the shortage of common Arabic datasets available.

4.2 Measuring Performance

To measure the matching algorithm performance, we separately prepared evaluation sets for human annotators to evaluate correctness of extracted. We asked two annotators with good Arabic language skills to judge the algorithm. We consider the accuracy of the proposed approach within a domain is given by the total number of correct matched sentences to the total number of sentences within each domain. The sentence is correctly matched when the system succeeds to extract all entities of the given sentence. Our approach performance is measured using two measurements a) Recall b) Precision.

a) Recall: which is calculated by the equation (1):

$$\text{Recall} = \frac{\text{CP}}{\text{CP+FN}} \qquad (1)$$

Where CP is the number of correct positive and FN is the number of false negative (false detection returns as correct detection).The results of recall for all Sports domain are 95%, for all Events domain are 85% and 96% for all Arts domain. The average results of dataset for all domains are 92%.

b) Precision: is calculated by the equation (2):

$$\text{Precision} = \frac{CP}{CP+FP} (2)$$

Where CP is the number of correct detection and FP is the number of false detection. The result of dataset for all Sports domain for precision are 96%, for all Events domain are 89% and 97% for all Arts domain. The average results of precision for all domains are 94% where the # correct statements =282.We note that, our algorithm executed a high performance in the two domains Sports and Arts. But it executed a low performance in the domain of Events as the Arabic Wikipedia articles in Events are little.

4.3 The Results

The selected entities have the evaluated intent as shown in table 5.

Table 5. Matching algorithm evaluation

Extraction terms correctness	Accuracy
First annotator	93%
Second annotator	92%
Matching algorithm	95%

Examples of some kinds of the dataset that we used are shown in Table 6. The first column describe the input sentence to our algorithm, the second column shows the equivalent meaning of input sentence in English. The third column presented the entities extracted by Matching Algorithm.

Table 6.

Some results for trying our algorithm with different dataset

Example	English Equivalent	Idioms Extracted
عاصمه جمهوريه مصر العربيه	Capital of the Republic Egypt	- جمهوريه مصرالعربيه - عاصمه
ما هى اهم اعمال نجيب محفوظ	What are the main works of Naguib Mahfouz	- نجيب محفوظ -اعمال
ابوتريكه من نجوم منتخب مصر	Aboutrika of Egypt Star team	-منتخب مصر -محمد ابوتريكه -نجوم
كوكب الشرق علامه مصريه	Star of the East mark Egyptian	-كوكب الشرق -علامه

5. CONCLUSION

In this paper, we presented an algorithm to extract Arabic entities from Arabic input sentence. By known entities, we mean exact expression, Named entities, known events …etc. To detect the entities our algorithm depends mainly on N-Gram and Arabic Wikipedia. The proposed algorithm begins with pre-processing input sentence to remove any noises. Then the Matching Algorithm applies to determine the entities that match with Arabic Wikipedia articles. The final results show that the system accuracy is approximately 94% for precision.

In our future research, we intend to use our algorithm to extract the relation between entities. We will return the concerned concept (features) for each entity and detect the relation between the entities that belong to same input sentence. The feature extraction can be exploited as a query expansion and it plays an important role to enhancement the Arabic information retrieval.

6. REFRENCES

[1] R. Kosala and H. Blockeel.2000. Web mining research: a survey. SIGKDD Explor. Newsl. vol. 2. pp. 1-15.

[2] S. A. Babar and P. D. Patil.2015. Improving Performance of Text Summarization. Procedia Computer Science, (vol. 46). pp. 354-363.

[3] I. Boujelben, S. Jamoussi, and A. Ben Hamadou.2014. A hybrid method for extracting relations between Arabic named entities. Journal of King Saud University - Computer and Information Sciences, vol. 26, pp. 425-440, 12.

[4] S. Auer, and J. Lehmann .2007. What Have Innsbruck and Leipzig in Common? Extracting Semantics from Wiki Content. presented at the Proceedings of the 4th European conference on The Semantic Web: Research and Applications, Innsbruck, Austria.

[5] P.-M. Ryu, M.-G.Jang, and H.-K. Kim. 2014. Open domain question answering using Wikipedia-based knowledge mode. Information Processing & Management, vol. 50, pp. 683-692, 9.

[6] N. I. Al-Rajebah, H. S. Al-Khalifa, and A. M. S. Al-Salman .2011. Exploiting Arabic Wikipedia for automatic ontology generation: A proposed approach. in Semantic Technology and Information Retrieval (STAIR), 2011 International Conference on, pp. 70-76.

[7] Mahgoub, A. Y., M. A. Rashwan, H. Raafat, M. A. Zahran, and M. B. Fayek.2014. Semantic Query Expansion for Arabic Information Retrieval. Arabic Natural Language Processing Workshop, Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, ACL.

[8] M. Ruiz-Casado, E. Alfonseca, and P. Castells.2005. Automatic Extraction of Semantic Relationships for WordNet by Means of Pattern Learning from Wikipedia. in Natural Language Processing and Information Systems. vol. 3513, A. Montoyo, R. Muńoz, and E. Métais, Eds., ed: Springer Berlin Heidelberg, pp. 67-79.

[9] F. Ataa Allah, S. Boulaknadel, A. El qadi, and D. Aboutajdine .2006. Arabic Information Retrieval System Based on Noun Phrases. in Information and Communication Technologies ICTTA '06. 2nd, pp. 1720-1725.

[10] A. M. AlAwajy and J. Berri.2013. Combining semantic techniques to enhance Arabic Web content retrieval. in Innovations in Information Technology (IIT), 2013 9th International Conference on, pp. 141-147.

[11] S. Chernov,T. Iofciu,W. Nejdl and X. Zhou.2006. Extracting semantic relationships between wikipediacategories.,n 1st International Workshop: "SemWiki2006 - From Wiki to Semantics" (SemWiki 2006), co-located with the ESWC2006 in Budva.

[12] Benajiba, Y., Rosso, P., 2008. Arabic Named Entity Recognitionusing Conditional Random Fields, In: Proceeding of Workshop onHLT and NLP within the Arabic World, LREC'08.

[13] H .Traboulsi, .2009. Arabic Named Entity Extraction: A Local Grammar-Based Approach . Proceedings of the International Multiconference on Computer Science and Information Technology, vol.4, pp. 139 – 143.

[14] M. N. Kumar, M. V. V. B. Vemula, D. K. Srinathan, and D. V. Varma.2010. EXPLOITING N-GRAM IMPORTANCE AND ADDITIONAL KNOWEDGE BASED ON WIKIPEDIA FOR IMPROVEMENTS IN GAAC BASED DOCUMENT CLUSTERING," ed.

[15] D. Milne, I. H. Witten, and A. Workshop. 2008. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. AAAI Workshop Tech. Rep. AAAI Workshop - Technical Report, vol. WS-08-15, pp. 25-30.

[16] J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. R. Curran.2013. Learning multilingual named entity recognition from Wikipedia. Artificial Intelligence, vol. 194, pp. 151-175.

[17] F. Wu, R. Hoffmann, and D. S. Weld.2008. Information extraction from Wikipedia: moving down the long tail. presented at the Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, Las Vegas, Nevada, USA.

[18] D. P. T. Nguyen, Y. Matsuo, and M. Ishizuka.2007. Relation extraction from wikipedia using subtree mining. presented at the Proceedings of the 22nd national conference on Artificial intelligence - Volume 2, Vancouver, British Columbia, Canada.

[19] Z. Yin, L. Wu, L. Kai, and H. Lianen.2012. Improving Retrieval Performance with Wikipedia's Category Knowledge. in Computational and Information Sciences (ICCIS), 2012 Fourth International Conference on, pp. 449-452.

[20] G. Wang, Y. Yu, and H. Zhu.2007.PORE: Positive-Only Relation Extraction from Wikipedia Text. in The Semantic Web. vol. 4825, K. Aberer, K.-S.Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon, et al., Eds., ed: Springer Berlin Heidelberg, pp. 580-594.