Extracting Named Entities Using Named Entity Recognizer for Arabic News Articles

Tarek Kanan
Department of Software Engineering
AlZaytonah University of Jordan
Amman, P.O.Box 130 , Jordan
tarek.kanan@zuj.edu.jo

Raed Kanaan
Department of Management Information System
Amman Arab University
Amman, P.O. Box: 2234, 11953, Jordan
rk@aau.edu.jo

Omar Al-Dabbas
Department of Computer Engineering
Al-Balqa Applied University
Al-Salt P.O. Box: 19117, Jordan

Ghassan Kanaan
Department of Computer Science
Amman Arab University
Amman, P.O. Box: 2234, 11953, Jordan
Gkanaan@aau.edu.jo

Ali Al-Dahoud,
Edward Fox

AlZaytonah University of Jordan
Amman, P.O.Box 130 , Jordan
aldahoud@zuj.edu.jo

Abstract—This paper describes how to extract, for the Arabic language, named entities and topics from news articles. Indeed, there is a lack of high quality tools for Named Entity Recognition (NER) for Arabic; therefore the authors have built an Arabic NER (RenA). NER involves extracting information and identifying types, such as name, organization, and location. For English language there are effective tools for NER, however these are not directly applicable to Arabic language. As a result, a new method and tool (i.e., RenA) have been developed. For NER evaluation purposes a baseline corpus was built for assessment and comparison with other methods and tools, with help from volunteer graduate students who understand Arabic. RenA produces good results, with accurate Name, Organization, and Location extraction from news articles collected from online resources. A comparison between the RenA results with a popular Arabic NER resulted in a noticeable enhancement.

*Keywords:**Arabic Language; Named Entity Recognizer; Natural Language Processing.***

I. Introduction

1.1. Arabic: Language, Encoding and Morphology

1.1.1. Arabic Language

One of the Arabic language is commonly spoken and a widely used in arc across the world. In fact, it has major differences from any other popular prevalent languages (i.e., English and French). Such differences are include but not limited to the following factors: it has distinguished specifications in terms of grammatical forms, diversities of word synonyms, and word meanings that are changed based on several factors such as word order. Although of such complexities, handful effort and work has been dedicated to natural language processing involving Arabic language, when compared it to the English language, which has been addressed by numerous studies. The majority of software packages and APIs for natural language processing and information retrieval do not address Arabic language requirements. To do so, significant adaptation and more effort need to be done to handle Arabic language data.

Arabic language consists of 28 alphabetical characters as well as diacritics and is written from right to left, with no capitalization. Several forms of the Arabic language are listed below [1]:

- Classical Arabic – This form is used in reading / reciting the holy books.
- Modern Standard Arabic (MSA) –which is usually used in writing, speech, interviewing, broadcasting, etc. implementation in this research is based on MSA.
- Spoken – oral dialects that are vary significantly from region to region.

1.1.2 Arabic Encoding

Recognizing the characters programmatically or via a computer program considered to be significant challenges raised when dealing with Arabic texts. Encoding tends to be problematic; nonetheless this issue can be solved effectively using Unicode (UTF8 for example). Alternatives may include Windows CP-1256 or X-Mac Arabic.

1.2. Named Entity Recognizer – NER

Consider the English quote,

“Go back, Sam. I’m going to Mordor alone.”

— Frodo, The Lord of the Rings: The Fellowship of the Ring

Based on NER “Sam” and “Motor” should be extracted as a name and location respectively. This is very helpful since NER extracts useful keywords in context.

NER for Arabic names of persons, organizations, and locations requires modification on, and creation of, available tools, e.g., the Stanford Named Entity Recognizer (SNER), to extract names of entities from text (i.e. from news articles). Extracting the named entities for any text may help figure out key elements. The authors argued that the main three entities (persons, organizations, and locations) reveal the most important entities in the text and could help as the main features for future work in text summarization for Arabic news article. In order to extract the appropriate Arabic named entities, authors of this paper have made some modifications on Yasine Benajiba’s Arabic NER tool named Arabic Name Entity Recognition (ANER) [2].

2. Literature Review

2.1. Named Entity Recognizer – NER

Manning et al. [3] defined Stanford Named Entity Recognizer (SNER) as a Java implementation of an NER. NER labels series of words in a text, that is to say proper nouns, such as gene and protein names, company, and person names .NER has feature extractors for Named Entity Recognition, in addition to several options for defining additional feature extractors. Included with the tool download are good named entity recognizers for English, particularly for the 3 classes (PERSON, ORGANIZATION, and LOCATION) [3]. Benajiba created his ANER system by examining the distinctive aspects of the Arabic language related to NER tasks and the state-of-the-art of NERs [2]. [4] Depict a new method to extract names from Arabic text by Abuleil et al. By building graphs they described the relationships between Arabic words. The findings of the proposed technique reveal that it extracts some names are extracted while others, however, are missing; the authors argue that if more articles are used the missing names would be extracted. Kanaan et al. build lexical entries using an existing tagger in order to identify and recognize proper names and other crucial lexical items [5]. Moreover, using a rule-based approach Shaalan develops a Named Entity Recognition system for Arabic (NERA) [6]. He utilizes a whitelist to represent a dictionary of names, in addition a grammar have been included, in the form of regular expressions. Using special tagged corpora for evaluation purposes, the outcomes that result from NERA was satisfactory in terms of precision, recall, and F1-measure. Kareem Darwish attempts to enhance Arabic named entity extraction for Wikipedia links [7] by using cross-lingual resources (Arabic/English).. Interestingly, a positive impact on recall using his method compared with Benajiba one [2].

3. Methodology

3.1. Arabic Named Entity Recognizer –RenA

3.1.1. Building the NER

To build an NER there are three principal ways, these are as follows:

- Knowledge Base – In this way a collection of words employed to detect entities based on a predetermined dataset; such a collection consists of a set of words mapped to a specific entity.
- Unlike knowledge base, Machine Learning –utilizes statistical models to classify and recognize grammars for the identifying entities, with Conditional Random Fields (CRF) [2] considered a significant choice.
- Training – manually or automatically train a classifier to deterministically identify the entities.

Reviewing these ways, this paper relies on a knowledge base one, but it is later improved by training. This way is described in more details in the following sections.

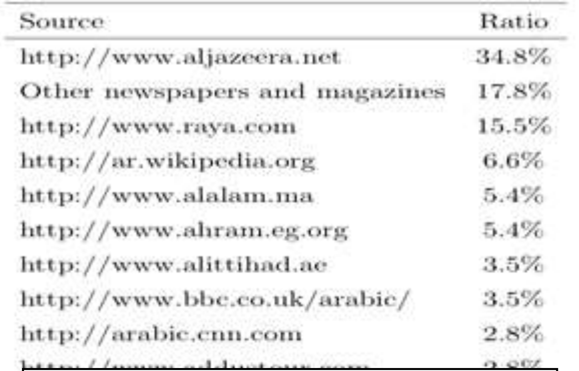
3.1

We ely

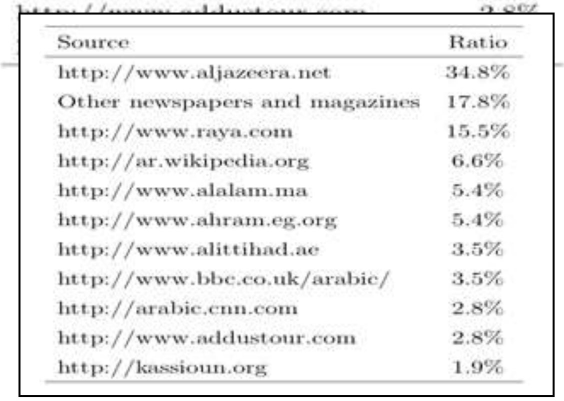
dis hly

150

Ta 8]



Source	Ratio
http://www.aljazeera.net	34.8%
Other newspapers and magazines	17.8%
http://www.raya.com	15.5%
http://ar.wikipedia.org	6.6%
http://www.alalam.ma	5.4%
http://www.ahram.eg.org	5.4%
http://www.alittihad.ae	3.5%
http://www.bbc.co.uk/arabic/	3.5%
http://arabic.cnn.com	2.8%



Source	Ratio
http://www.aljazeera.net	34.8%
Other newspapers and magazines	17.8%
http://www.raya.com	15.5%
http://ar.wikipedia.org	6.6%
http://www.alalam.ma	5.4%
http://www.ahram.eg.org	5.4%
http://www.alittihad.ae	3.5%
http://www.bbc.co.uk/arabic/	3.5%
http://arabic.cnn.com	2.8%
http://www.addustour.com	2.8%
http://kassioun.org	1.9%

Furthermore, a feature set has been built, mainly for the following named entities: PERSON (PERS), ORGANIZATION (ORG), and LOCATION (LOC).

3.1.3. Approach

3.1.3.1. Stage 1 – Building RenA

To guide the implementation of the NER, the knowledge base is employed to classify the entities of the selected texts. To map the words in the knowledge base to their entity types (PERS, ORG, and LOC) it is essential to build a dictionary. As the collection of words has been mapped to the suitable entities, the populated dictionary and a chunker (tokenizer) can be used to classify a collection of text and define the words' entities. The chunker will tokenize based on whitespaces. For each word, the chunker will identify possible results of the tags; for instant, the word, "Washington" can be added to the dictionary, as a location, with a tag (LOC); "Washington" can be added as well to indicate for a person with a tag (PERS).Once applied, every appearance occurrence of the word “Washington” will return both tags , [PERS and LOC].

The produced results are acceptable. Some of the keywords are tagged, while others are not. However, some of the words are tagged incorrectly. It is clear to the authors that the problems in this stage are due to stopwords and Harakat. These issues are addressed in next stage (i.e. stage 2).

3.1.3.2. Stage 2 – Improving RenA

The main concern in stage 2 was to fairly filter out stopwords and normalize words. While eliminating stopwords is considered a simple technique as a total of stopwords list can be used to filter out words of little importance, Normalization, however, helps in reducing the words’ dimensionality and allows identical results as duplicate ones will not appear when chunking.

Applying normalization and filtering out stopwords improve results by reducing the varieties of the words. Nonetheless, issues like words being tagged improperly and/or missing words tags are still of concern. Indeed, As there are complexity in organization names this leads to inaccurate results. Often, it presents confusions with both person and location entities. Some Arabic NERs report low accuracy result for organizations [2, 7]. Interestingly, this issue seems to be a dominant problem for Arabic NER. However, the persons and locations produce acceptable results, with only few minor defects that can be over comes easily.

By training the corpus, of course based on the knowledge base, we can improve the precision of each entity. In this case, the use of inclusion and exclusion lists resulted in a remarkable increase in precision. In our experiment, a random sampling of news articles have been chosen in which it contains a significant number of keywords to help improve our training by enhancing it with names of persons, organizations, and locations. Table 2 shows the list of resources and their countries, which reflect sources used to enhance our knowledge base.

Table 2: News articles used to train the NER collection via inclusion/exclusion

Sites	**Region**
http://www.aljazeera.net	Saudi Arabia
http://alwatan.com	Qatar
http://www.alarabalyawm.net	Jordan
http://www.alanwar.com/	Lebanon

As a result of using inclusion and exclusion lists, our knowledge base begins to improve by either addition or removal of words to their suitable entity type classes. For each training (enhancing) phase, we are relying on the current state of our NER and enriching the knowledge base by adding new tagged or removing wrong tagged entities for each phase. Interestingly, it should be noted that each time the NER trained on a given, it works better on the next set of articles.

Below Figure 1 depicts the sequence of steps employed to build the proposed NER (RenA). The inside part of RenA deals with building the knowledge base (dictionary) by importing, after normalizing process, the ANER Corp knowledge base, then adding our inclusion/exclusion list, all together will build our NER dictionary. The researchers used the normalizer and stopwords removal to preprocess our article then the chunker will act as a tokenizer and mapper. The chunker tokenize the preprocessed article based on the articles whitespaces and for each token it will attempt to check if an entity exists in the dictionary based on the token. Finally, the entities extraction function will produce the results from the mapped tokens.

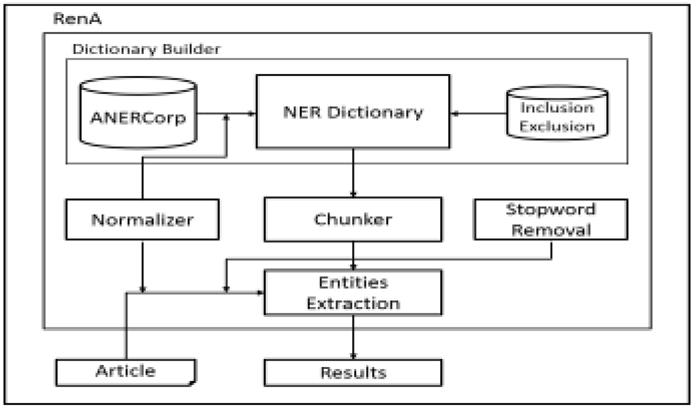
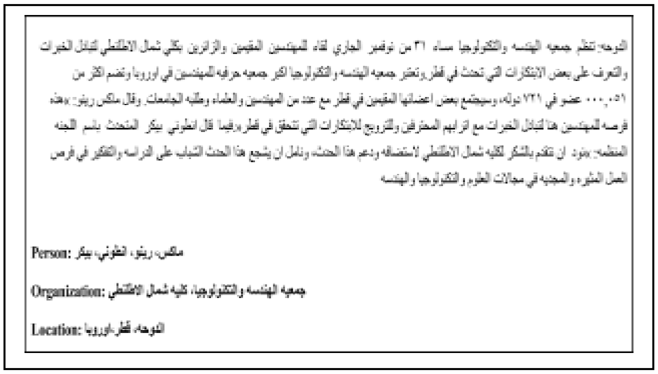


Figure 1: RenA Architecture

3.1.4. Result

Results are appearing in bold after extracting the named entities in the selected article.



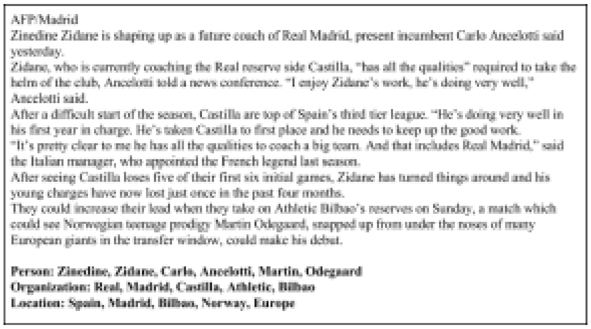
الدوحة: تنظم جمعيه الهندسه والتكنولوجيا مساء ٢١ من نوفمبر الجاري لقاء للمهندسين المقيمين والزائرين بكليه شمال الاطلنطي لتبادل الخبرات
والتعرف على بعض الابتكارات التي تحدث في قطر وتعتبر جمعيه الهندسه والتكنولوجيا اكبر جمعيه حرفيه للمهندسين في اوروبا وتضم اكثر من
١٥٠,٠٠٠ عضو في ١٢٧ دوله، وسيجتمع بعض اعضائها المقيمين في قطر مع عدد من المهندسين والعلماء وطلبه الجامعات. وقال ماكس رينو: يعتبر هذا
فرصه للمهندسين هنا لتبادل الخبرات مع نظرائهم المحترفين والترويج للابتكارات التي تتحقق في قطر وفيما قال انطوني بيكر المتحدث باسم اللجنه
المنظمه: نود ان نتقدم بالشكر لكليه شمال الاطلنطي لاستضافه ودعم هذا الحدث، ونامل ان يشجع هذا الحدث الشباب على الدراسه والتفكير في فرص
العمل المثيره والمجزيه في مجالات العلوم والتكنولوجيا والهندسه

Person: ماكس، رينو، انطوني بيكر

Organization: جمعيه الهندسه والتكنولوجيا، كليه شمال الاطلنطي

Location: الدوحه، قطر، اوروبا

Entities



AFP/Madrid
Zinedine Zidane is shaping up as a future coach of Real Madrid, present incumbent Carlo Ancelotti said yesterday.
Zidane, who is currently coaching the Real reserve side Castilla, "has all the qualities" required to take the helm of the club, Ancelotti told a news conference. "I enjoy Zidane`s work, he`s doing very well," Ancelotti said.
After a difficult start of the season, Castilla are top of Spain`s third tier league. "He`s doing very well in his first year in charge. He`s taken Castilla to first place and he needs to keep up the good work.
"It`s pretty clear to me he has all the qualities to coach a big team. And that includes Real Madrid," said the Italian manager, who appointed the French legend last season.
After seeing Castilla loses five of their first six initial games, Zidane has turned things around and his young charges have now lost just once in the past four months.
They could increase their lead when they take on Athletic Bilbao`s reserves on Sunday, a match which could see Norwegian teenage prodigy Martin Odegaard, snapped up from under the noses of many European giants in the transfer window, could make his debut.

Person: Zinedine, Zidane, Carlo, Ancelotti, Martin, Odegaard
Organization: Real, Madrid, Castilla, Athletic, Bilbao
Location: Spain, Madrid, Bilbao, Norway, Europe

3: English News Article and Extracted Named Entities

The extracted entities might be of interest as it reflects the context of the article.. For instant, a football fan who see the key elements in Figure 3 will be able to realize that this article is talking about Real Madrid team (Organization named entity) whereas the focus is on a player called Zidane (Person named entity).

3.1.5. Evaluation

Table 3 shows us the results between RenA and the basic NER from LingPipe [9]. In terms of recall, precision,.

Table 3: Recall, Precision, and F1 Values for RenA and LingPipe NERs

	RenA NER		**LingPipe Toolkit NER**	
	Recall	**Precision**	**Recall**	**Precision**
PERSON	0.826	0.497	0.582	0.371
ORGANIZATION	0.813	0.421	0.39	0.377
LOCATION	0.77	0.558	0.55	0.338
Average	0.803	0.492	0.507	0.362

In Figure 4 below, the precision values of both NERs are displayed for each entity. It shows that RenAproduces higher precision results for each entity.

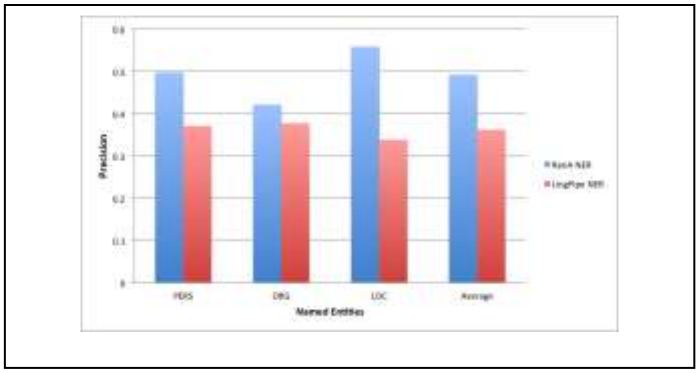


Figure 4: Precision Values for RenA and LingPipe NER

In Figure 5 below, the recall of both NERs is displayed for each entity. RenA showed better results than its counterpart NER in terms of recall.

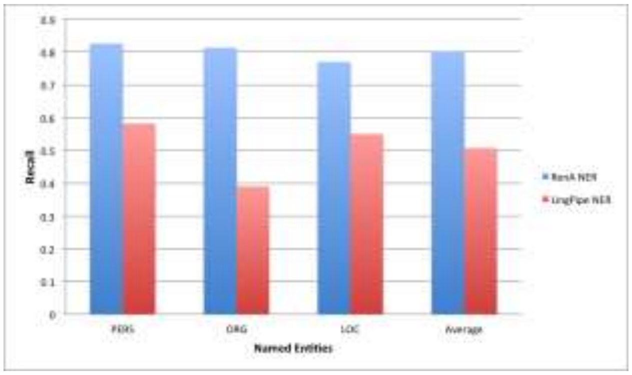


Figure 5: Recall Values for RenA and LingPipe NER

4. Conclusion

Compared to other languages, conducting research on Natural Language processing for Arabic language Natural is relatively hard, particularly there are very few free NLP tools and resources. The value of this research lies in the fact that it is unique in addressing the extraction of named entities from Arabic news articles.

In this research, the aim was to develop a Named Entity Recognizer to be able to extract, with noticeable accuracy, named entities from Arabic news articles, called RenA. The researchers build a corpus for the purpose of RenA evaluation and later to be utilized by other researchers. This process was not being possible without the assistant of graduate students fluent with Arabic language. Building of such corpus was because there is a lack of free resources for a judged news articles corpus.

to evaluate and compare proposed RenA NER with another NER that is available through the LingPipe toolkit, Information Retrieval evaluation measures have been utilized. Researchers of this paper take into consideration three types of named entities, these are: Person, Organization, and Location. The results demonstrated that applying the proposed RenA enhances the named entity extraction results for Person, Organization, and Location entities.

5. Future Work

Our future plan is to expand this research by using the RenA results to fill in templates. We also are planning to extract more attributes, like generating topics using the popular Latent Dirichlet Allocation (LDA)algorithm to fill in templates, towards generating improved Arabic news article summaries. In addition, we aim to use another Arabic stemmer and to compare the ALDA results with the two stemmers.

Acknowledgements

Thanks go to the US National Science Foundation for their support through grants DUE-1141209 and IIS-1319578. We would also like to acknowledge QNRF for their support by NPRP grant # 4-029-1-007.Special thanks go to Al-Zaytonah University of Jordan for their help and support.

References

[1] Habash, Nizar Y. "Introduction to Arabic natural language processing." Synthesis Lectures on Human Language Technologies 3, no. 1 (2010): 1-187.

[2] YasineBenajiba.“Arabic named entity recognition”, PhD dissertation. Universidad Politécnica de Valencia. Valencia, Spain. 2009.

[3] Christopher D. Manning, PrabhakarRaghavan, and HinrichSchütze.Introduction to information retrieval. Vol. 1: Cambridge University Press. Cambridge. 2008.

[4] SaleemAbuleil and Martha Evens.“Extracting Names from Arabic Text for Question-Answering Systems”.In proceedings of RIAO'2004, pp. 638–647, France. 2004.

[5] GhassanKanaan, Reyad Al-Shalabi, and MajdiSawalha.“Fully automatic Arabic text tagging system”.In the proceedings of the International Conference on Information Technology and Natural Sciences, Amman, Jordan. 2003.

[6] KhaledShaalan and HafsaRaza.NERA: Named entity recognition for Arabic. Journal of the American Society for Information Science and Technology. 60(8): pp. 1652-1663. 2009.

[7] Kareem Darwish.“Named Entity Recognition using Cross-lingual Resources: Arabic as an Example”. In proceedings of the 51st Annual Meeting of the Association of Computational Linguistics (ACL), pp. 1558-1567.Sofia, Bulgaria, August 4-9 2013.

[8] The Center for Computational Learning Systems, Columbia University. YasineBenajiba, “ANERCorp”. 2010. [Cited 02/15/2015].http://www1.ccls.columbia.edu/~ybenajiba/downloads.html

[9] LingPipe, a toolkit for processing text using computational linguistics. Alias-i. LingPipe 4.1.0. October 1, 2008. [Cited 02/15/2015]. http://alias-i.com/lingpipe.