ARABIC NAMED ENTITY RECOGNITION (NER): INVESTIGATION ON FEATURES COMBINATION

HANAN FARAG ABUALI
DR. SABRINATIUN

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRACT

There is no doubt that the name entity recognition playing an important role in the fields of Natural Language Processing (NLP) in multiple domains such as Information Extraction (IE), Sentiment Analysis (SA) and Question Answering (QA). It aims to extract names of people, locations, organizations, currencies, and dates. Arabic is one of the common languages that has been investigated by NER. The researchers have been done several approaches in terms to determine named entities. However, there is still lacking research efforts regarding NER in the Arabic language (Abdallah et al. 2012; Benajiba 2009; Esmaail 2011; Shaalan& Raza 2009). that's mean there is still room for improvement to extract the entity names, despite the availability of many approaches to solving the problem of determining the entity names in the Arabic language. Moreover, the goal of this study is to propose an application that combines a new set of features in order to increasethe performance of the classification based on features impact which is extractingnamed entity. the feature extraction task which contains POS tagging, keyword trigger, Affixes and N-grams. This project will utilize these set of features that will be extracted from the dataset in order to create training and testing setsthat will be utilized in the Support vector machine (SVM) classifier.

Keywords:Named Entity Recognition (NER); Arabic named entity (ANER),S Vector Machine (SVM), Natural Language Processing (NLP).

INTRODUCTION

Recently the named entity has become very important because of the task information extraction, which seeks to classify named entities in the text into predetermined categories such as a person's name, organization name, location name, date or currency. However, there has been a lot of work have been done in this area, especially in English. Borthwick (1999). Several fields have been used Named entity, such as information retrieval (IR), Sentiment Analysis and Question Answering system (Toda &Kataoka 2005). Many researchers have been proposed several approaches to extract and classify named entity, for example machine learning techniques such as Support Vector Machine (SVM), Decision Tree, Naïve Bayes (NB) and etc.

However, choosing good features has a significant impact on the effectiveness (Bender et al. 2003). For example, by using such keywords, Mr, Mrs, Assoc. Professor and Doctor, etc. To determine the people's names appears to be a useful approach. Further, the are some language needs specific features to determine the Name entity such as the names in the English language, there is Capital letters, prefixes and suffixes features that can be used to classify these entities.

In addition, these features maybe not fit when using it with other languages (Huang 2005). From a different perspective sometimes there are areas that require specific features such as a Biology contains entities such as protein, gene and chemical compounds (Narayanaswamy et al. 2003).

In addition, Part of Speech (POS) Tagger can play an essential role in languages to determine the named entity, because, this feature provides a tag for each word that indicated its syntactical type (e.g. Noun, verb, adjective, etc.). Especially in Arabic because it contains a large

ambiguity of some of the names are the names of person or organization's name or location's name.

Hence, using POS to identify the nouns, then classifying those nouns whether they are named entities or typical nouns. This makes a challenge in terms of classification name entities. Thus, this study aims to solve the problem of recognizing entities in the Arabic language.

RELATED WORK

several approaches have been proposed in terms of classifying Arabic named entities for instance, Shaalan (2014) examined the work in Arabic NER. Most previous works include the official MSA style language used in the news area. The list of numerous works widely considered in this review.

In this section, the author focused on the work done for the extraction of Arab named entities of social media contexts. Attempting to recognition the entities from the Arabic tweets is introduced by (Darwish and Gao, 2014). In this paper, the conditional random field (CRF) classifier was used to extract person names, location, and organization, based on names depending on the" language-independent "features.

Moreover, used a set of tweets that have been collected and annotated in previous work (Darwish, 2013), Benajibaet al. (2007). have been the proposed system by a combination of n-gram-based approach and Maximum Entropy classifier. The system provides weights for each feature, for instance, capitalization, suffixes, and prefixes, the maximum entropy is a probability distribution with the selection of the highest probability. In addition, due to the lack of resources of the Arabic oriented tasks of NER Thus, the authors built their own corpora for training and testing. In addition, they are also built to test the effect gazetteers using external sources of information about the system.

On other hand, have been proposed a NER system, which is a combination of several sets of features with machine learning techniques (Benajiba et al. (2008). A set of signs consisting of contextual features of lexical features and directories while the proposed method of machine learning is support Vector Machine (SVM). Context entities feature associated with differences within several contexts. Whereas lexical feature associated with markers of orthographic nature, such as special characters, punctuation, abbreviations and numbers. On the other hand, Gazetteers associated with dictionaries or lists that contain the person, organization and location names. Finally, the SVM is adjusted to the objects and classify the type called it. The proposed method has reached 82,71 F1-sccore.

PROPOSED METHOD

The Research Methods Framework section formulated to explain, extend existing knowledge within the limits of this study. Furthermore, the Java language has been chosen to build the system, thus most of the tools used were from Stanford and Java library and the data that used in this study is ANERcorp.

The theoretical framework is the structure that can hold or support the idea of this research study Figure 1 shows the diagram of this study, is divided into four phases as described

in the following subsections: the first phase, is pre-processing which mean clean the data from the nose to prepare it for the next step, then the second phase, is applying a set of features, then will combine them to find the best combination. The third phase is the classification approach and has been chosen SVM. Finally, the evaluation, which is the fourth phase.

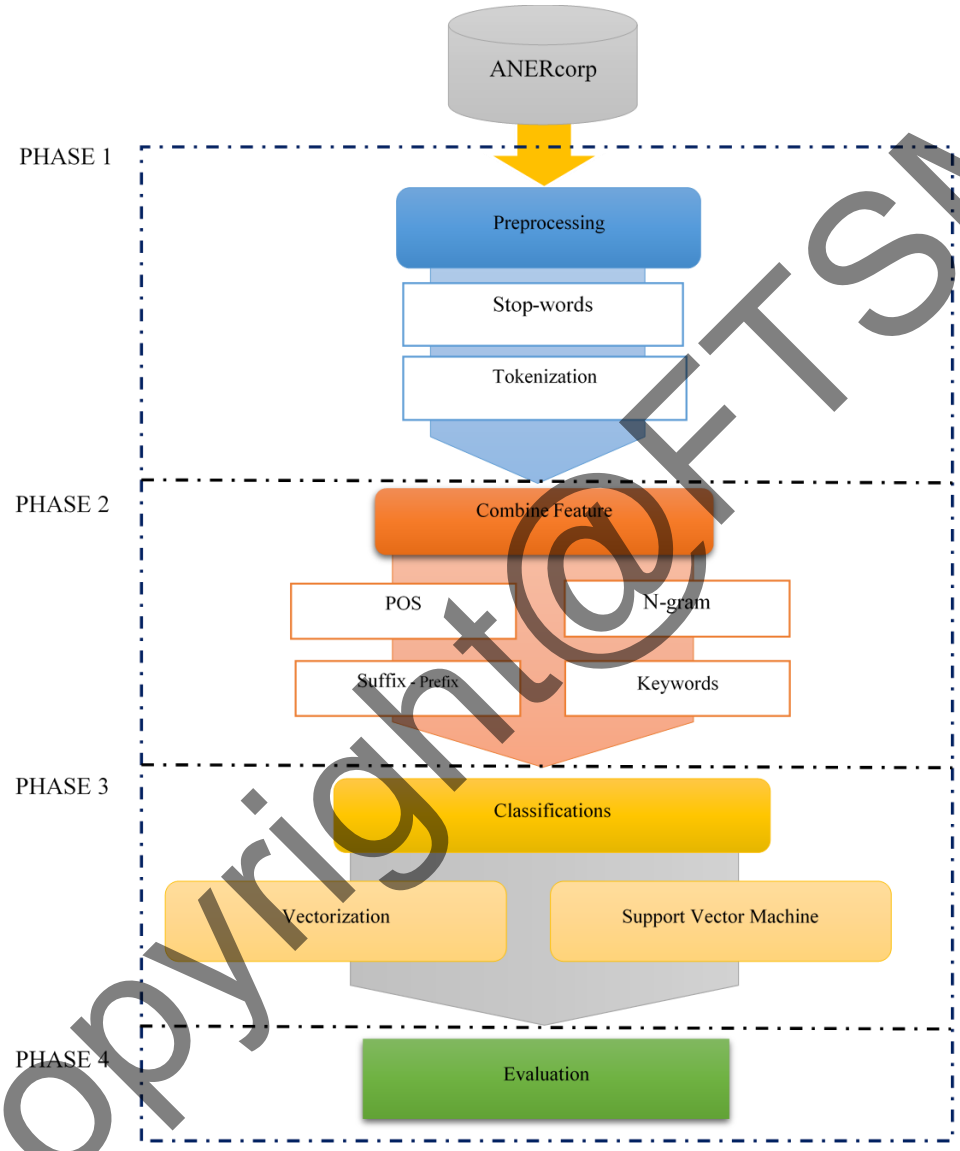

FIGURE 1Research Framework

EVALUATION

DISSCUION ON RESULTS

This thesis aims to contribute towards to find the best features impact to extract NEs from Arabic documents. However, the Arabic language contains linguistic intricacies that make it difficult to recognize the data. The four features (POS, N-gram, affixes, and keywords) have been used separately, and then the sequentially combined features into four sets such as (Quadruple features, Triple features,

Double features and Single) in this research. The study shows that the results obtained were satisfactory and the technique used to be appropriate.

The table 1 shows the Comparison the Results of all Features and clearly can see that the Quadruple features (F1, F2, F3, and F4) were a very good combination. The proposed features achieved the very high accuracy of 95% average. Also, have been obtaining a very good combination in the triple feature set (F1, F3, and F4) by recorded the second highest average 90%.

TABLE 1Comparison the Results of all Features

Features Set	Shortcut Features	Average
Single set	F1	61%
Single set	F2	52%
Single set	F3	67%
Single set	F4	67%
Double set	F1,F2	68%
Double set	F1,F3	61%
Double set	F1,F4	63%
Double set	F2,F3	51%
Double set	F2,F4	51%
Double set	F3,F4	67%
Triple set	F1,F2,F3	84%
Triple set	F1,F3,F4	84%
Triple set	F1,F2,F4	95%
Triple set	F2,F3,F4	67%
Quadruple set	F1,F2,F3,F4	97%

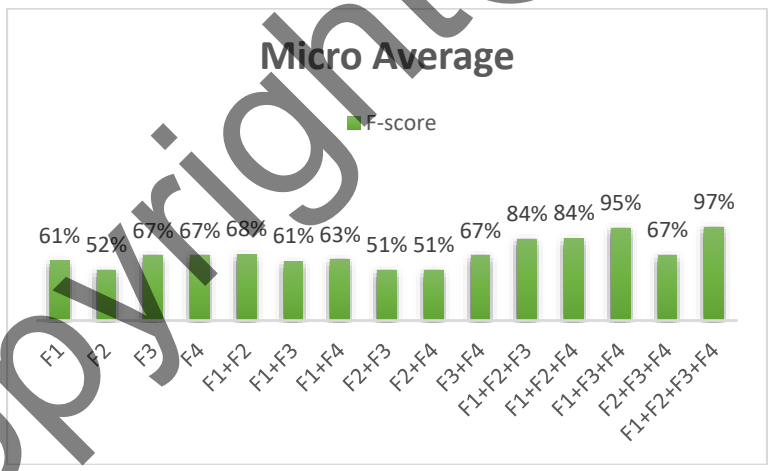

FIGURE 2The features Average Comparison

Figure 1.2 shows the comparison between the features based on the average. Furthermore, the chart shows the accuracy of each feature when determined the highest average at the Quadruple feature.

CONCLUSION

In summary, the Java language was chosen to build the system, so most of the tools used were from the Stanford and Java libraries, and the data used in this study is ANERcorp.

However, in order to achieve the goal of this proposed method, the approach was divided into four phases in this thesis, as described in the following subsections: the first phase is pre-processing, which means clearing the data from the nose to prepare them for the next step, then the second phase, applying a set of features, and then combine them to find the best combination impact. The third stage is the classification approach and SVM was chosen. Finally, the evaluation, which is the fourth stage utilized the three standard measure (precision, recall and f-measure) to evaluate the proposed method. Thus, the evaluation of the proposed approach showed that the combined features have achieved the objectives of this study.

REFERENCES

Andrew Borthwick. *1999. A Maximum Entropy Approach to Named Entity Recognition. Ph.D. thesis, Computer Science Department, New York University.*

Benajiba, Y., Rosso, P. &Benedíruiz, J. M. 2007. *Anersys: An arabic named entity recognition system based on maximum entropy. Dlm. (pnyt.). Computational Linguistics and Intelligent Text Processing, hlm. 143-153. Springer.*

Bender, O., Och, F. J. & Ney, H. 2003. *Maximum entropy models for named entity recognition. Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4, hlm. 148-151.*

Huang, F. 2005. *Multilingual named entity extraction and translation from text and speech. TesisCiteseer.*

Kareem Darwish and Wei Gao. 2014. *Simple Effective Microblog Named Entity Recognition: Arabic as an Example. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), pages 2513–2517, Reykjavik, Iceland, May. European Language Resources Association (ELRA).*

Kareem Darwish. 2013. *Named Entity Recognition using Cross-lingual Resources: Arabic as an Example. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, volume 1, pages 1558–1567, Sofia, Bulgaria. Association for Computational Linguistics.*

Khaled Shaalan. 2014. *A survey of arabic named entity recognition and classification. Computational Linguistics, 40(2):469–510, June.*

Narayanaswamy, M., Ravikumar, K., Vijay-Shanker, K. & Ay-shanker, K. V. 2003. *A biological named entity recognizer. Pac SympBiocomput, hlm. 427.*

Toda, H. &Kataoka, R. 2005. *A search result clustering method using informatively named entities. Proceedings of the 7th annual ACM international workshop on Web information and data management, hlm. 81-86.*