Procesamiento del Lenguaje Natural, Revista nº 44, marzo de 2010, pp 151-152 recibido 27-01-10 revisado 15-02-10 aceptado 08-03-10

Arabic Named Entity Recognition*

Reconocimiento de Entidades Nombradas en Textos Árabes

Yassine Benajiba
DSIC - Universidad Politécnica de Valencia

Center for Computational Learning Systems (CCLS),
Columbia University, New York City, NY
ybenajiba@ccls.columbia.edu

Resumen: Tesis doctoral en Informática realizada por Yassine Benajiba y dirigida por el doctor Paolo Rosso (Univ. Politécnica de Valencia). El acto de defensa de tesis tuvo lugar en Valencia en Mayo de 2009 ante el tribunal formado por los doctores Felisa Verdejo (UNED), Mona Diab (Columbia Univ.), Imed Zitouni (IBM T.J. Watson Research Center), Horacio Rodriguez (Univ. Politécnica de Cataluña) y Encarna Segarra (Univ. Politécnica de Valencia). La calificación obtenida fue *Sobresaliente Cum Laude*.
Palabras clave: Extracción de información, Reconocimiento de Entidades Nombradas, Procesamiento de idiomas con morfología compleja.

Abstract: PhD thesis in Computer Science written by Yassine Benajiba under the supervision of Dr Paolo Rosso (Univ. Politécnica de Valencia). The author was examined in May 2009 in Valencia by the committee formed by Felisa Verdejo (UNED), Mona Diab (Columbia Univ.), Imed Zitouni (IBM T.J. Watson Research Center), Horacio Rodriguez (Univ. Politécnica de Cataluña) and Encarna Segarra (Univ. Politécnica de Valencia). The grade obtained was *Cum Laude*.
Keywords: Information Extraction, Named Entity Recognition, Complex Morphology Languages Processing.

1. Introduction

In this Ph.D. thesis we have investigated the problem of the recognition and classification of Named Entities (NEs) within Arabic text, i.e. *Arabic Named Entity Recognition (NER)*.

In order to achieve this goal, we have explored a wide range of features including: lexical, morphological and syntactic ones, we have employed three different discriminative Machine Learning (ML) approaches and we have validated our approach on 9 standard data-sets. We have studied the following problems:

1. The difference in performance when using a 2-step approach as an attempt to separate the problem of recognizing the NEs from the one of classifying them;
2. The relevance of using the rich morphology of the Arabic language in order to obtain a high performance NER system;
3. The difference in performance when difference ML approaches are employed; we have studied the possibility of combining them; and
4. Transferring knowledge about NEs from another language, i.e. English, in order to enhance the performance of an Arabic NER system.

Even though our study is focused on Arabic, as a language, and NER, as a task, the obtained results might be easily extrapolated to most of the morphology complex/rich languages and most of the Information Extraction tasks.

2. Thesis overview

In this thesis, we have addressed the challenges raised by the morphologically rich languages. Arabic in our case, to a supervised NLP task: i.e. NER. The document (Benajiba, 2009) is structured as follows:

In Chapter 1, we have introduced basic concepts and we summarize the major contributions of the research work carried out.

Chapter 2 describes the challenges of NLP in general and NER in particular for the Arabic language.

* This PhD thesis was supported by an AECI scholarship

ISSN 1135-5948

In Chapter 3, the NER task is introduced. This includes both presenting the standard definitions of the task and giving an overview on the most influential research works in the NER field in general and in the Arabic NER field in particular. In this chapter, we attempt to make it easier for the reader to see where the contribution of this thesis stands exactly.

Chapter 4 describes the three ML discriminative approaches which are used in our research study, namely: Maximum Entropy (ME), Conditional Random Fields (CRFs) and Supports Vector Machines (SVMs). All of these ML approaches are feature-based and have proved to be efficient for sequence classification problems.

Chapter 5 is dedicated to present our 1-step and 2-steps, ME-based, Arabic NER system. In this section we first report the results obtained when a ME-based classifier is used to build the system. Following, we report our results when the NER task is split into two sub-tasks where the first one determines the spans of the NEs within the text and the second one assigns a class to each one of them. This study is important for it provides empirical proves that enhancing the performance using a 2-step approach is only possible when the first step achieves a performance close to 100 points of F-measure.

In Chapter 6 we have switched our focus to the features and the ML approaches which we might resort to in order to boost our system. We conduct several experiments where ME, CRFs and SVMs are used with different feature-sets. We validate these experiments on 9 different standard data-sets of different genres (newswire, broadcast news and weblogs). The major contribution of this study is that it has shown that different ML approaches might benefit differently from the available features and they also obtain different results for the different NE classes. This has triggered the research work which we present in Chapter 7 where we have combined different classifiers where each classifier is trained for one NE class. Similarly, the feature selection is done for each classifier separately. The combination of the different classifiers is done finally in order to obtain a single outcome.

Chapter 8 introduces a very novel approach where we project knowledge about NEs from another language, i.e. English. This research study has been conducted in collaboration with IBM T.J. Watson Research Center as a six-month internship subject of the Ph.D. student Yassine Benajiba. The results show that a statistically significant improvement is always obtained.

Finally, in Chapter 9 we have drawn our conclusions and discussed some interesting research directions.

3. Thesis contributions

The major contributions of the investigations carried out are:

1. A deep analysis of the difference of behavior of different ML approaches in the context of NER;
2. A multi-classifier approach has been shown to lead to the best results;
3. Statistically significant improvement is obtained across the different data genres when additional knowledge about NEs is projected from a resource-rich language such as English;
4. The study of an effective approach to build an efficient and robust Arabic Named Entity Recognition system;
5. Providing empirical proof that the morphology richness of the Arabic language can be employed to enhance the performance of an NER system;
6. The ANERcorp data-set together with SVMs and CRFs Arabic NER models have been made publicly available for the research community[1].

We have validated our results on 9 data-sets of 4 different genres, namely: newswire, broadcast news, Arabic Treebank and Weblogs. Our experiments are easily replicable for anyone because they are all built on top of publicly available tools. We have also described in details our incremental selection approach which can be easily used in case a new feature is being added to the NER system. To our knowledge, our research study is the most extensive work which has been reported, up to now, on Information Extraction for morphologically rich languages and our results are very competitive on a world wide scale. A summarized description of part of the research work can be found in (Benajiba et al., 2009).

References

Yassine Benajiba, Mona Diab, and Paolo Rosso. 2009. Arabic named entity recognition: A feature-driven study. *The special issue on Processing Morphologically Rich Languages of the IEEE Transaction on Audio, Speech and Language Processing*, July.

Yassine Benajiba. 2009. Named entity recognition. *Ph.D. Thesis dissertation, Universidad Politécnica de Valencia*, May, http://users.dsic.upv.es/∼prosso/resources/BenajibaPhD.pdf.

[1] http://www.dsic.upv.es/grupos/nle/downloads.html