Artificial Intelligence 238 (2016) 63–82

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Robust multilingual Named Entity Recognition with shallow semi-supervised features

Rodrigo Agerri *, German Rigau

IXA NLP Group, University of the Basque Country (UPV/EHU), Donostia–San Sebastián, Spain

ARTICLE INFO

Article history:
Received 1 September 2015
Received in revised form 11 May 2016
Accepted 15 May 2016
Available online 19 May 2016

Keywords:
Named Entity Recognition
Information Extraction
Clustering
Semi-supervised learning
Natural Language Processing

ABSTRACT

We present a multilingual Named Entity Recognition approach based on a robust and general set of features across languages and datasets. Our system combines shallow local information with clustering semi-supervised features induced on large amounts of unlabeled text. Understanding via empirical experimentation how to effectively *combine* various types of *clustering features* allows us to seamlessly export our system to other datasets and languages. The result is a simple but highly competitive system which obtains state of the art results across five languages and twelve datasets. The results are reported on standard shared task evaluation data such as CoNLL for English, Spanish and Dutch. Furthermore, and despite the lack of linguistically motivated features, we also report best results for languages such as Basque and German. In addition, we demonstrate that our method also obtains very competitive results even when the amount of supervised data is cut by half, alleviating the dependency on manually annotated data. Finally, the results show that our emphasis on clustering features is crucial to develop robust out-of-domain models. The system and models are freely available to facilitate its use and guarantee the reproducibility of results.

1. Introduction

A named entity can be mentioned using a great variety of surface forms (Barack Obama, President Obama, Mr. Obama, B. Obama, etc.) and the same surface form can refer to a variety of named entities. For example, according to the English Wikipedia, the form 'Europe' can ambiguously be used to refer to 18 different entities, including the continent, the European Union, various Greek mythological entities, a rock band, some music albums, a magazine, a short story, etc.[1] Furthermore, it is possible to refer to a named entity by means of anaphoric pronouns and co-referent expressions such as 'he', 'her', 'their', 'I', 'the 35 year old', etc. Therefore, in order to provide an adequate and comprehensive account of named entities in text it is necessary to *recognize* the mention of a named entity and to *classify* it by a pre-defined type (e.g, person, location, organization). Named Entity Recognition and Classification (NERC) is usually a required step to perform Named Entity Disambiguation (NED), namely to link 'Europe' to the right Wikipedia article, and to resolve every form of mentioning or co-referring to the same entity.

Nowadays NERC systems are widely being used in research for tasks such as Coreference Resolution [51], Named Entity Disambiguation [19,27,29,38,26] for which a lot of interest has been created by the TAC KBP shared tasks [32], Machine

* Corresponding author.
E-mail addresses: rodrigo.agerri@ehu.eus (R. Agerri), german.rigau@ehu.eus (G. Rigau).

[1] http://en.wikipedia.org/wiki/Europe_(disambiguation).

http://dx.doi.org/10.1016/j.artint.2016.05.003
0004-3702/

Translation [3,33,5,35], Aspect Based Sentiment Analysis [37,11,50,49], Event Extraction [22,2,31,20,30] and Event Ordering [41].

Moreover, NERC systems are integrated in the processing chain of many industrial software applications, mostly by companies offering specific solutions for a particular industrial sector which require recognizing named entities specific of their domain. There is therefore a clear interest in both academic research and industry to develop robust and efficient NERC systems: For industrial vendors it is particularly important to diversify their services by including NLP technology for a variety of languages whereas in academic research NERC is one of the foundations of many other NLP end-tasks.

Most NERC taggers are supervised statistical systems that extract patterns and term features which are considered to be indications of Named Entity (NE) types using the manually annotated training data (extracting orthographic, linguistic and other types of evidence) and often external knowledge resources. As in other NLP tasks, supervised statistical NERC systems are more robust and obtain better performance on available evaluation sets, although sometimes the statistical models can also be combined with specific rules for some NE types. For best performance, supervised statistical approaches require manually annotated training data, which is both expensive and time-consuming. This has seriously hindered the development of robust high performing NERC systems for many languages but also for other domains and text genres [45,54], in what we will henceforth call 'out-of-domain' evaluations.

Moreover, supervised NERC systems often require fine-tuning for each language and, as some of the features require language-specific knowledge, this poses yet an extra complication for the development of robust multilingual NERC systems. For example, it is well-known that in German every noun is capitalized and that compounds including named entities are pervasive. This also applies to agglutinative languages such as Basque, Korean, Finnish, Japanese, Hungarian or Turkish. For this type of languages, it had usually been assumed that linguistic features (typically Part of Speech (POS) and lemmas, but also semantic features based on WordNet, for example) and perhaps specific hand-crafted rules, were a necessary condition for good NERC performance as they would allow to capture better the most recurrent declensions (cases) of named entities for Basque [4] or to address problems such as sparsity and capitalization of every noun for German [23,7,8]. This language dependency was easy to see in the CoNLL 2002 and 2003 tasks, in which systems participating in the two available languages for each edition obtained in general different results for each language. This suggests that without fine-tuning for each corpus and language, the systems did not generalize well across languages [46].

This paper presents a multilingual and robust NERC system based on simple, general and shallow features that heavily relies on word representation features for high performance. Even though we do not use linguistic motivated features, our approach also works well for inflected languages such as Basque and German. We demonstrate the robustness of our approach by reporting best results for five languages (Basque, Dutch, German, English and Spanish) on 12 different datasets, including seven in-domain and eight out-of-domain evaluations.

1.1. Contributions

The main contributions of this paper are the following: First, we show how to easily develop robust NERC systems across datasets and languages with minimal human intervention, even for languages with declension and/or complex morphology. Second, we empirically show how to effectively use various types of simple word representation features thereby providing a clear methodology for choosing and combining them. Third, we demonstrate that our system still obtains very competitive results even when the supervised data is reduced by half (even less in some cases), alleviating the dependency on costly hand annotated data. These three main contributions are based on:

1. A simple and shallow robust set of features across languages and datasets, even in out-of-domain evaluations.
2. The lack of linguistic motivated features, even for languages with agglutinative (e.g., Basque) and/or complex morphology (e.g., German).
3. A clear methodology for using and combining various types of word representation features by leveraging public unlabeled data.

Our approach consists of shallow local features complemented by three types of word representation (clustering) features: Brown clusters [10], Clark clusters [15] and K-means clusters on top of the word vectors obtained by using the Skip-gram algorithm [39]. We demonstrate that *combining* and *stacking* different clustering features induced from various data sources (Reuters, Wikipedia, Gigaword, etc.) allows to cover different and more varied types of named entities without manual feature tuning. Even though our approach is much simpler than most, we obtain the best results for Dutch, Spanish and English and comparable results in German (on CoNLL 2002 and 2003). We also report best results for German using the GermEval 2014 shared task data and for Basque using the Egunkaria testset [4].

We report out-of-domain evaluations in three languages (Dutch, English and Spanish) using four different datasets to compare our system with the best publicly available systems for those languages: Illinois NER [52] for English, Stanford NER [24] for English and Spanish, SONAR-1 NERD for Dutch [21] and Freeling for Spanish [47]. We outperform every other system in the eight out-of-domain evaluations reported in Section 4.3. Furthermore, the out-of-domain results show that our clustering features provide a simple and easy method to improve the robustness of NERC systems.

Finally, and inspired by previous work [34,9] we measure how much supervision is required to obtain state of the art results. In Section 4.2 we show that we can still obtain very competitive results reducing the supervised data by half (and