CE807 Text Analytics
Assignment 2: Information Extraction
Task 1

Nicolás Reyes Huerta - ID: 1700927

April 2018

1 Introduction

A Name Entity Recognition (NER) system aims to automatically identify and classify a named entity, such as people, locations or organizations, present within a text. Training a NER system requires labeled data, which may not be always available, since the labeling process is highly laborious and expensive to carry out. Therefore, the automation of this process becomes critical to reduce time and expenses on expert annotation required to obtained the data, on one hand and to extend and apply NER systems to other fields, such as medicine, on the other.

A distant supervision strategy seems an effective to automatically create labeling data for training propose. The idea is to employ large databases, which are not domain specific (such as Wikipedia), or free text, which can be found in websites, to automatically extract sentences where entities are present. The assumption behind distant supervision is that any sentence express a relation if two desired entities are present. These assumptions widen the scope of the relation, enhancing feature extraction.

2 Literature Review

The use of Wikipedia, an open and collaborative encyclopedia that contains millions of articles in several languages, has been widely used since the last decade to extract entity descriptors which can be used to train NER systems. One method of doing this is by directly extracting the features from Wikipedia articles. [1] proposed to extract the description of the entity from the first sentence of a Wikipedia article (main text). The authors realized that generally, the noun phrase which follows the verb "to be", is a good category label ideal to be used directly as a feature to train a NER system.

Similarly, [2] base their categorization of entities by using several Wikipedia features which can be found in the English articles, such as category links, article links among others. Although its implementation is to automatically obtain annotated text for any language (not only English), the development is based on English Wikipedia articles. In principle, the idea is to make use of the category structure[1] to map and obtain the category to which it belongs. Therefore, for each article title in the English Wikipedia, a key phrase category is extracted and assigned to an entity with the help of a predefined map. Then, the label of the entity, which appears in a non-English article, is linked by finding its English equivalent or based on associations of category information present in equivalent English articles.

On the lines of [2], [3] creates NER annotated data for particularly, Korean and Bulgarian languages. In addition to [2] methodology, they incorporate parallel-foreign sentences annotation to train a model to project these relations between languages. In this sense, the authors develop three groups of linking annotations which allow them to combine with the tagger proposed by [2].

In addition to what has been discussed above, [4] remark the fact that structured features extracted form Wikipedia's articles contains an important source of information. Therefore, their approach extracts separate features, as a bag of words representation, from three different sources: sentence, paragraph and the title of an article. Additionally, the content of infoboxes, sidebars and taxoboxes.

An alternative way of generating training data is by using an existing relational database, such as freebase. Freebase is a large database with entity pair relations and relation instances of whose

[1] These are those categories in which an entity has been classified according to its content by Wikipedia.

major information source is Wikipedia. The benefit of using this database is to enlarge the context in which an entity relation is used. This is not explored in previous methods, since they just consider a subset of an article. In this sense, by enriching the number of times an entity pair is seen in a relation (sentence), feature extraction can be enhanced. This principle is used by [5], where the authors develop a model whose ability is to combine several sentences in which a given relation occurs, for a particular entity relation.

In terms of state of the art using the WikiGold set as a test set, it was found that [4] archived an 66.6% of accuracy, in 2013. Nevertheless, in 2017 [6] outperformed that results, obtaining an 67.14%. In a similar way, [6] outperformed both by authors achieving an 74.3% of accuracy. The system behind this outstanding results combines features similar to what it *********

Finally, two main approaches to automatically extract training data has been described to be used to train NER systems. On the one hand, those who employ Wikipedia articles directly as data to extract training features([1], [2], [3] and [4]). On the other hand, [5] which use pre-existing relational database (freebase) to generate their training data. Additionally, it was possible to establish a common line of work, where [2] and [4] worked upon the eariler studies / experiments to enhance their work. In this sense, it was observed that the most preferable algorithm to train NER sytems was logistic regression and conditional random fields. Therefore, these two methods should be taken into consideration for training in task 2.

References

[1] Kazama Jun'ichi, Torisawa Kentaro. *Exploiting Wikipeadia as External Knowledge for Named Entity Recognition.*

[2] Richman Alexander, Schone Patrick. *Mining Wiki Resources for Multilingual Named Entinty Recognition.*

[3] Kim Sungchul, Toutanova Kristina and Yu Hwanjo. *Multilingual Named Entety Recognition using Parallel Data and Metadata from Wikipedia.*

[4] Nothman Joel, Ringland Nicky, Radfoard Will, Murphy Tara and Curran James R. *Learning Multilingual Named Entity Recognition from Wikipedia.*

[5] Mintz Mike, Bills Steven, Snow Rion and Jurafsky Dan. *Distant Supervision for Relation Extraction without Labeled Data.*

[6] Agerri Rodrigo and Rigau German. *Robust Multilingual Named Entity Recognition with Shallow Semi-Supervised Features.*

[7] Abbas Ghaddar and Philippe Langlais. *WiNER: A Wikipedia Annotated Corpus for Named Entity Recognition.*