An application of Spectral Clustering on Named Entity Recognition

Ali Josue Limon; Michele Cerú
ajl649@nyu.edu ; mc3784@nyu.edu

December 5, 2016

1 Abstract

Named Entity Recognition (NER) is one of the important parts of NLP. It aims to find and classify expressions of special meaning in texts written in natural language. These expressions can belong to different predefined classes such as Person, Organization and Location.

Prior work on NER come in three kinds: supervised learning algorithms, semi-supervised learning algorithms, and unsupervised learning algorithm. Compared with supervised and semi-supervised methods, unsupervised approach for entity recognition can overcome the difficulties on requirement of a large amount of labeled data.

In this project we aim to investigate the possibility of using spectral clustering [2] to a NER task in Spanish. Becuase of the lack of spanish labeled corpora, an unsupervised techique could have important applications. The dataset we are going to use contains some words labeled with a tag that represents whether the word is a place, a person or an organization [1]. The aim of the project is to use the vector that represent the context of each word and apply to them the spectral clustering technique [2], to see weather the words cluster according to their labels or not. A similar approach can be found in [3], where spectral clustering is used for word sentence disambiguation WSD. To represent the context of a target word, we will use the vector formed by the continuous bag of words of the context [4], that is the average of the word2vec [5] representation of all the words in the sentence containing a target word. In this way, the contribution of our project would be the implementation of the word embedding vector representation to compute spectral clustering.

The experimentation of this work will be carried on a corpora that consist of sentences extracted from news articles. The corpora corresponds to the *CoNLL 2002* Shared Task Spanish data, the original source being the *EFE Spanish*

Newswire Agency. The data consists of two columns separated by a single space. The first item on each line is a word and the second the named entity tag. In this dataset, there are four types of entities: person names (PER), organizations (ORG), locations (LOC) and miscellaneous names (MISC) and other (O). This dataset contains 11,752 sentences, 369,171 words and 26,706 Name Entities.

References

[1] http://www.cnts.ua.ac.be/conll2002/ner/

[2] http://www.cims.nyu.edu/∼bandeira/TenLecturesFortyTwoProblems.pdf

[3] Popescu, M., & Hristea, F. (2011). State of the art versus classical clustering for unsupervised word sense disambiguation. Artificial Intelligence Review, 35(3), 241?264. http://dx.doi.org/10.1007/s10462-010-9193-7.

[4] http://crscardellino.me/SBWCE/

[5] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781.