Creating Sense Vectors using Cross-lingual Information

Rachita Chhaparia[1], Deepanshu Gupta[2]
Advisor: Dr. Amitabha Mukerjee[1]
{rachitac,dpanshu,amit}@iitk.ac.in
[1] *Dept. Of Computer Science and Engineering*
[2] *Dept. Of Mathematics and Statistics*

October 04, 2015

1 Introduction

Word-sense disambiguation (WSD) is the process of identifying the meaning/sense of a word in a sentence. It is a common phenomenon for a word in English(or any other language) to have more than one meaning. In language processing the task of WSD is to distinguish between the different senses of the words appearing in different contexts. However, this task is daunting as WSD remains to be an open problem.
The significance of the problem can be estimated by its relevance in field of *Semantics*, *Search Query improvement*, et cetera. To give an example of WSD problem consider the three sentences below.

- No one here knew how to plant crops, but the next town over was made up of farm laborers.
- The mining and washing plant is extremely good and largely constructed at Cagliari.
- Tulsi plant is well respected and worshipped in India.

Note that in the first sentence the word *plant* refers to placing something to grow, whereas in second example we have a reference to an industrial site and the third example is the most common used word in the context of a living multi-cellular organism.

Sense Vectors are similar to word embeddings with the underlying difference that a single word has a unique existence in word embedding space, whereas in a sense vector space a word may/maynot have unique vector representation, depending upon the number of meanings/senses the word can take. Also, most of the words in the sense vector space do not have independent existence, they depend on their context for their definition.

2 Related Work

Owing to the age of the problem, WSD has been approached using many different angles. The first few attempts at solving the WSD problem were made using Supervised and Knowledge based methods like Using Dictionary, Expert Human Parsing, and Statistical Modeling[9].

Recently, however usage of Neural Language Modeling(NLM) has shown promising results in Semantic modeling exercise.[2] Unsupervised methods like Word Vector based discrimination has been found to outperform most of the knowledge and human annotated based methods[9].

Several proposed methods that use cross-lingual similarities for WSD include [5], [6] and [7].

The problem of Sense vectors has been approached primarily using dictionaries or monolingual corpora as in [1]. [7] has however tried using parallel corpus to look at the semi-supervised version of the problem.

3 Our Approach

Our approach consists of the following steps:

1. Learn word embeddings for both English and Hindi using the data-set mentioned in the following section. We plan to use word2vec[10] for this purpose.
2. Generate a bilingual dictionary in English and Hindi consisting only of monosemous words using Babelnet[8].
3. Learn the mapping (M_{EH} and M_{HE}: from English to Hindi and from Hindi to English respectively) between the two vector spaces using the bilingual dictionary created above. We use the method proposed by [3] for this purpose. We only use monosemous words as their embeddings are less noisy since, they are not corrupted because of multiple senses.
4. Identify target polysemous words in English (Hindi) using Babelnet for which we want to create the different sense vectors. Next, we translate its each sense into Hindi (English) and project back its embedding to the English (Hindi) word vector space. This is the sense vector in English (Hindi) for the target word and for the target sense.

4 Resources

- Dataset
  - English: English Wikipedia articles[4]
    To train the word embeddings in the English Word Vector Space
  - Hindi: Hindi Wikipedia articles[4]
    To train the word embeddings in the Hindi Word Vector Space
- Babelnet[8]
  We plan to use Babelnet for the following two purposes:
  - To generate a bilingual dictionary in English and Hindi containing only of monosemous words to learn the mapping between the two vector spaces
  - To identify polysemous words in English (Hindi) and to translate its different senses to Hindi (English)

References

[1] Reisinger, Joseph, and Raymond Mooney. "A mixture model with sharing for lexical semantics." Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010.

[2] Mikolov, Tomas, et al. 'Efficient estimation of word representations in vector space.' arXiv preprint arXiv:1301.3781 (2013).

[3] Mikolov, Tomas, Quoc V. Le, and Ilya Sutskever. 'Exploiting similarities among languages for machine translation.' arXiv preprint arXiv:1309.4168 (2013).

[4] Al-Rfou, Rami, Bryan Perozzi, and Steven Skiena. 'Polyglot: Distributed word representations for multilingual nlp.' arXiv preprint arXiv:1307.1662 (2013).

[5] Diab, Resnik 'An unsupervised method for word sense tagging using parallel corpora' ACL '02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics

[6] Bhingardive et al. 'Neighbors Help: Bilingual Unsupervised WSD Using Context' The 51st Annual Meeting of the Association for Computational Linguistics - Short Papers (ACL Short Papers 2013)

[7] Guo, Che et al. 'Learning Sense-specific Word Embeddings By Exploiting Bilingual Resources.' Proceedings of COLING, 2014 - aclweb.org

[8] Ehrmann, Maud, et al. "Representing multilingual data as linked data: the case of BabelNet 2.0." Proc. of LREC. Vol. 14. 2014.

[9] McCarthy 'http://lct-master.org/files/WSD.pdf'

[10] Word2Vec 'https://code.google.com/p/word2vec/'