Unsupervised Word Sense Disambiguation for Low Resource Languages

Sudha Bhingardive and Pushpak Bhattacharyya
IIT Bombay

Introduction

Word Sense Disambiguation (WSD) is one of the most prominent and challenging research problems in the field of Natural Language Processing (NLP). In any natural language, there are many polysemous words, *i.e.*, words having more than one meaning or sense. It is often easy for a human to identify the correct sense of a word in a given context, but the same task when performed by a computer is among the most difficult problems in NLP. Generally, in order to disambiguate a given word, we should have a context in which the word has been used and also the knowledge about the word, otherwise it becomes difficult to get the exact sense of the word. WSD is the process of identifying the correct sense of a word in a particular context. For instance, consider the sentence, 'Ram is playing cricket on a playground', where the word 'cricket' is ambiguous having two senses: 'a game' and 'an insect'. Here, the correct sense of 'cricket' is 'a game' as words 'playing' and 'playground' appears in its context.

Problem Statement

WSD system relies on two very important resources: i) Sense repository and ii) Sense-annotated corpus. Therefore, a WSD system can be easily developed for a language, if it has a sense repository like wordnet and considerable amount of sense-annotated corpus in that language. Resource scarcity acts as a major bottleneck in WSD, as many languages lack these aforementioned resources. Our main objective is to study resource scarcity in WSD for Indian languages and provide resource conscious solution in order to overcome it.

Motivation

Over the years, various WSD approaches have been proposed and shown considerable improvement in their performances. Most of these approaches are devised for resource rich languages like, English, Spanish, German, *etc*. However, very little work has been done for resource scarce languages like, Hindi, Marathi, Bahasa, Malay, *etc*. Development of resources such as sense repository and sense-annotated corpus is costly in terms of time and money, as they are usually created manually. In addition, it is impracticable to create such resources which can cover all language-domain pairs. Consequently, most of the researchers attract towards the unsupervised approaches, as they do not require sense-annotated corpus. Therefore, we focus on development of unsupervised WSD approaches for low resource languages.

Our Approaches

We developed two unsupervised WSD approaches where WSD is performed without relying on sense-annotated corpus. The two approaches devised for unsupervised WSD are as follows:

- **Context-based Bilingual WSD Approach [2]**: This is a bilingual WSD approach, where two resource deprived languages help each others' WSD using a context-based

Expectation Maximization (EM) formulation. This is built over the framework of Khapra et al., [1] where context is introduced in the basic EM-based formulation. Here, WSD is performed by finding the similarity between the target word and the context words in three different ways, i) *using word co-occurrence counts*, ii) *using wordnet similarities*, and iii) *using distributional similarities*. This approach works on the principle that the words with strong semantic relation help each other in disambiguation. We tested the approach on Hindi-Marathi language pair on Health and Tourism domain corpus. An improvement of 17% - 35% in the accuracy of verb WSD is obtained compared to the existing EM-based approach.

- **Most Frequent Sense Detection Approach [3]**: This is a novel approach for finding the Most Frequent Sense (MFS) of a word where the use of *word embeddings* is explored. Word embeddings are increasingly being used in variety of NLP tasks. They represent each word with low-dimensional real valued vector. They work under the assumption that similar words occur in similar context. In this approach, we compared the word embeddings of a word with its sense embeddings and extracted the predominant sense with the highest similarity. Here, word embeddings are trained using word2vec[1] tool. However, sense embeddings are created using various wordnet based features like synset members, content words in the gloss and example sentences, synset members of the hypernymy-hyponymy synsets, *etc*. We observed that this approach easily beats the wordnet first sense baseline when tested on Hindi WSD task. Our approach is language independent and can be easily ported to various domains across languages.

Conclusions

We devised the resource conscious solution(s) for word sense disambiguation problem in Indian languages. In this paper, two unsupervised WSD approaches *viz*., *Context-based bilingual WSD using EM algorithm*, and *Most frequent sense detection using word embeddings* are presented. In both these approaches, WSD is performed by utilizing the knowledge from raw text and from lexico-semantic features of IndoWordNet. In context-based bilingual WSD approach, the problems of 'inhibited progress due to lack of translation diversity' and 'uniform sense assignment, irrespective of the context' are solved which the previous EM-based approach suffered from. In most frequent sense detection approach, it is proved that *word embeddings capture the most frequent sense of a word efficiently*. Approaches proposed in this paper justify the importance of unsupervised techniques and the use of wordnet for WSD as the results are found to be promising.

References

[1] Mitesh M. Khapra, Salil Joshi, Arindam Chatterjee, and Pushpak Bhattacharyya. Together we can: Bilingual bootstrapping for wsd, *In Proceedings of* ACL'11, Stroudsburg, PA, USA, 2011.

[2] Sudha Bhingardive, Samiulla Shaikh and Pushpak Bhattacharyya. Neighbor Help: Bilingual Unsupervised WSD Using Context, *In Proceedings of* ACL'13, Sofia, Bulgaria, 4-9 August, 2013.

[3] Sudha Bhingardive, Dhirendra Singh, Rudramurthy V, Hanumant Redkar and Pushpak Bhattacharyya. Unsupervised Multilinual Most Frequent Sense Detection using Word Embeddings, *In Proceedings of* NAACL'15, Denver, Colorado, USA, May 31 - June 5, 2015.

[1] https://code.google.com/archive/p/word2vec/