Word Sense Disambiguation By Using Domain Knowledge

Wei Jan Lee and Edwin Mit
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak,
94300 Kota Samarahan,
Sarawak, Malaysia
weijan87@gmail.com, edwin@fit.unimas.my

Abstract— **Over the decades, lot of studies had been carried out to suggest different approaches for Word Sense Disambiguation (WSD) process. From times to times, different approaches had been suggested to define the sense of a polysemous word. In this paper, a WSD approach with the domain knowledge will be discussed. In this approach, by using Wordnet, domains of each single word will be defined and a process of defining the best domain to be assigned to that particular word will be carried out. A method of calculating the weight of each domain to its corresponding word will be discussed. According to the weight assigned to each domain, the sense of the ambiguous word will be identified.**

Keywords: Word Sense Disambiguation, Domain, Wordnet

I. INTRODUCTION

Word Sense Disambiguation (WSD) process is a process to define the sense/meaning of an ambiguity word. WSD is an essential process as it supports other Natural Language Processing (NLP) applications, such as Machine Translation, Information Extraction and Content Analysis. Throughout the decades, a number of different approaches had been introduced to the research area. These approaches are categorized into two main categories, Supervised WSD and Unsupervised WSD.

Supervised WSD approaches always produce a better accuracy compared to unsupervised WSD approaches. However, supervised WSD approaches required a training corpus which is expensive and required human resources to annotate the corpus. Besides that, supervised WSD had come to a bottleneck situation where the accuracy of the result has no significant improvement.

Unsupervised WSD approaches in the other hands have the potential to overcome the bottleneck of the knowledge acquisition [14]. The idea of unsupervised WSD approaches is based on using the text itself. Instead of using a labeled training corpus to gain a certain score or probabilistic measurement, unsupervised WSD approaches gain the information and measurement from the text itself to assign the senses. The disadvantages of this approach are due to no external knowledge resources involved, it cannot rely on the shared reference inventory of senses [14].

In this paper, a WSD approach with Domain knowledge will be discussed. In this approach, an external knowledge source will be adopted in order to gain the information about the knowledge of domain. WordNet is adopted here, together with the domain. This approach represents a combination between knowledge source WSD approach and Unsupervised WSD approach. The details of the knowledge resource will be described in next section.

For every given corpus or a single text file, certain domains or categories about the corpus or text can be identified. Some properties [1] of the domain identification in a text can be very useful to WSD process. In a text, some portions of the context are a composition of a set of words that belong to the same domain. This property hence reduces the ambiguity of words if the domain of a text can be identified.

Since a text is a composition of few set of words that belong to domains, hence the context of text provides a source to identify the domains of the text. A bag-of-words approach will be suggested by in proposed approach and the domain term-distribution in the given text would be measured and works with the domain information to define the sense of a word.

This paper is structured as follow. Section II will be describing the knowledge resources that adopted in this approach while Section III will be discussing some of the related works. In Section IV, the proposed approach is discussed and Section V describes the evaluation that been carrying out.

II. WORDNET & WORDNET DOMAIN

WordNet [2] is an English lexicon which contains the word and its meanings, structured according to its semantic relation. It is a lexical inheritance system. It encodes the concept in terms of sets of synonyms which is known as synsets (synonym set) in WordNet. For each synset, Wordnet provides the semantic relation information such as the hypernymy (is-a relation), hyponymy, troponymy, meronymy (part-of relation), and similarity.

WordNet Domain [8] however is an extension of WordNet in which each of the synset in WordNet had been annotated with one or more domain labels. The domain set used in WordNet was obtained according to Dewey Decimal Classfication [15] where a set of computational in between these two taxonomies had been done to make sure of the completeness of WordNet Domain. WordNet Domain does

978-1-61284-353-7/11/$26.00 ©2011 IEEE