Knowledge-Based Contextual Overlap keen Ideas for Word Sense Disambiguation using Wordnet

Alok Chakrabarty
Department of Computer Science
IIIT Bhubaneswar
alok@iiit-bh.ac.in

Bipul Syam Purkayastha
Department of Computer Science
Assam University Silchar
bipul_sh@hotmail.com

Lavya Gavshinde
Department of Computer Science
IIIT Bhubaneswar
lavya3@gmail.com

Abstract

Word Sense Disambiguation (WSD) is a task of identifying correct sense of a given word especially when it has multiple meanings. WSD acts as a foundation to many AI applications such as Data Mining, Information Retrieval and Machine Translation. It has drawn much interest in the last decade and much improved results are being obtained. For WSD we require a knowledge-base, using which we can resolve the ambiguity and identify the correct sense of a given word in a sentence. The most commonly used computational lexicon for WSD, especially for English, is the English Wordnet, which later inspired the use of Wordnet specifically for WSD for Indo-Aryan languages such as Hindi, Sanskrit, etc. In this paper we explore some ideas that may enhance the efficiency of knowledge-based contextual overlap WSD algorithms when they are used on Wordnets.

1 Introduction

The aim of any Word Sense Disambiguation (WSD) system is to obtain the intended senses of a set of target words, or of all words of a given text against a sense repository using the context in which the word appears. The sense repository can be a machine readable dictionary, a thesaurus, a corpora or a computational lexicon like the Wordnet (Miller, et al., 1993) which, contrary to a dictionary or a thesaurus, contains conceptual-semantic and lexical relations.

There are various approaches to WSD like Knowledge Based (KB) approaches which encompass Overlap Based (OB) approaches, or Selectional Preferences based WSD, Machine Learning based Approaches which include supervised, semi-supervised, and unsupervised approaches, and Hybrid Approaches.

In this paper we focus on the Knowledge-Based Contextual Overlap WSD approach. In the subsequent sections we present some ideas that may enhance the efficiency of knowledge-based contextual overlap WSD algorithms like the Lesk algorithm considerably, for their use on Wordnets.

2 The Knowledge-Based Contextual Overlap Approach

The Knowledge-Based Contextual Overlap Approach for WSD works as follows:

A word is assigned a sense with the use of

(i) the context in which it has been mentioned
(ii) the information in the Wordnet and
(iii) the overlap between these two pieces of information.

The sense with the maximum overlap is declared as the *winner sense*.

For obtaining the context of a polysemous word *w* in a sentence of the text under consideration, needing disambiguation, a *context bag* C is created by collecting a set of context words in its surrounding. Sometimes along with the words in the current sentence in which *w* appears, we also

add the words from previous and the following sentences too in the context bag.
The information in the wordnet related to the word *w* is accumulated for each sense *s* of *w* in another data structure called the *sense bag* B (a bag of words) which includes information obtained from the

(i) Synonyms
(ii) Glosses
(iii) Example Sentences
(iv) Hypernyms
(v) Glosses of Hypernyms
(vi) Example Sentences of Hypernyms
(vii) Hyponyms
(viii) Glosses of Hypernyms
(ix) Example Sentences of Hypernyms
(x) Meronyms
(xi) Glosses of Meronyms

By measuring the overlap between C and B using the intersection similarity measure the approach outputs that sense *s* as the most probable sense, which has the maximum overlap. (Sinha, et al., 2004)

3 Ideas to facilitate Contextual Overlap

In this section we discuss some ideas that when incorporated in Wordnet may facilitate better contextual overlap for efficient WSD using overlap based algorithms such as the Lesk algorithm or its variants.

i) *Use of multiple glosses made up of diverse vocabulary for all concepts*

This idea suggests that for each concept of the wordnet we must keep multiple glosses (explanations) for that sense. The phrases for the glosses must be made up of diverse vocabulary or words that are frequently used with that concept. This is because contextual overlap based WSD algorithms are usually very sensitive to the exact wording used in the definitions. So the absence of some words in the glosses for a sense *s* of a word *w* can radically change the overlap result, as glosses form the primary ingredient to the sense bag.

The idea if implemented will increase the chances of overlap with the most appropriate sense of the concept in question using a Lesk like algorithm for WSD.

However the decision on the choice of most frequently used words and/or phrases may be a tedious task for a lexicographer involved in designing the glosses for the concept.

ii) *Use longer glosses*

The use of longer glosses will result in better overlap rank that eliminates cases where there are equal overlap ranks among different senses. For example consider a word "economy" which has 4 senses:

1. The system of production and distribution and consumption.
2. The efficient use of resources.
3. Frugality in the expenditure of money or resources.
4. An act of economizing; reduction in cost.

Now for a sentence like "Economy of a state depends on the demand and supply of the goods in the market" if we enhance the gloss for sense 1 of "economy" as "the system of production and distribution and consumption which is governed by the demand and supply principles" then the use of this longer gloss will increase the overlap count and ease out the choice of the winner sense.

iii) *Introduction of proper nouns in wordnet and keeping an indicator for the category*

The introduction of proper nouns and keeping an indicator to its associated category can lead to faster disambiguation of polysemous words. If the declared proper noun is related or associated with a concept A (an indicator to the noun's category) then all keywords adjacent to it must be compared with the senses corresponding to A.

For example we know that "Isaac Newton" is associated with Physics and/or Mathematics, hence all the polysemous words adjacent to Isaac Newton will usually correspond to Physics or Mathematics. Similarly, we can have more proper nouns such as Sachin Tendulkar who is associated with Cricket (indicator/category). Thus the word "ball" nearby Sachin Tendulkar will refer to the sense of a "cricket ball" rather than to the sense of "ball dance".

iv) *Introduction of a new field in the synset structure of Wordnet to store information related to the "frequently used" or "highly expected" words for that concept*

The introduction of a "frequently used" or "highly expected" field in the synset structure of

wordnets can scale-up the efficiency in determining winner sense of a polysemous word, as these highly related words will enrich the sense bag with more information, thereby enhancing the chances of appropriate overlap. For example the following list of terms may be considered as "highly expected" or "frequently used" with the concept of "computer":

1: Central Processing Unit.
2: Keyboard.
3: Mouse.
4: Monitor.
5: Universal Serial Bus.
6: USB Stick.

Thus putting the above list in the synset structure of the most appropriate sense of "computer" will result in attaining high degrees of overlap with sentences comprising of the word "computer" and several of the words from the above list.

v) *Preparing gloss for a sense using most frequently used terms for that sense*

The extent of overlapping will significantly increase if the gloss of a sense *s* is prepared by using the most frequently used terms with that sense *s*. For example consider a polysemous word "break" (noun) which has 16 senses as per the English wordnet (Fellbaum, 1998; Miller, 1995). Out of these 16 senses one of the senses is "the act of breaking something". As we know that the act of breaking something usually results into "smaller pieces" of an object. So, we may include the word "piece(s)" in the gloss "the act of breaking something" to obtain the gloss for "break" in the same sense as "the act of breaking something into pieces". We do this for increasing the overlap count, as the word "piece" is usually a frequently used term with the word "break" used in the sense of "breaking something". Use of such rich glosses will result in a better overlap using Lesk like algorithms.

vi) S*torage of information related to the "distributional constraints" and constantly enriching it*.

This idea suggests that we should try to capture the distributional constraints and other such relationships between different concepts of wordnet into account.

For example the concepts of "cigarette" and "ash" have a relationship between them. If we keep information about this relationship in the wordnet then it will certainly be helpful in sense disambiguation of concepts of "ash" and "cigarette".

If required, we may keep on enriching information related to such relationships between concepts in the wordnet. Incorporation of such relationships in the wordnet will increase the efficiency of knowledge-based contextual overlap WSD algorithms.

vii) *Exhaustive pre-processing for morphology and designing of the gloss accordingly*

Exhaustive pre-processing for morphology can result in better matching of senses, which is the fundamental requirement of contextual overlap based methods. For example "biscuit" (noun) is defined as "small round bread leavened with baking-powder or soda". Use of this gloss will not give a good overlap count, if words adjacent to "biscuit" in a sentence (in context to which the word "biscuit" is to be disambiguated) describe taste or its purpose or something about its history. Hence an analysis of the morphology for the word "biscuit" and accordingly enriching the gloss for it will give better results in this case. A suggested gloss in this case for the concept of "biscuit" can be: "small piece of bread leavened with baking-powder or soda that comes in different shapes (like circular, rectangular, square-shaped), flavours (sweet or salty) and colours (brown, red, yellow) which is generally used as a morning or evening snack with tea or coffee".

4 Conclusion

In the present paper several ideas to facilitate WSD using knowledge-based contextual overlap algorithms have been discussed. The ideas may seem attractive but at the same time implementation of such ideas may be quite a tedious task for the lexicographer engaged with the designing of the wordnet for his/her respective language.

However the implementation of the discussed ideas will certainly enrich the sense bag with more information leading to high degrees of overlap for the most appropriate sense of a word in question thereby achieving better quality word sense disambiguation of senses.

References

Fellbaum, C. (ed.) 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1993. Five Papers on WordNet. MIT press. http://www.mit.edu/~6.863/spring2009/readings/5papers.pdf

George A. Miller. 1995. English WordNet - A Lexical Database for English. *Communications of the association for Computing Machinery*, 38(11):39-41.

Manish Sinha, Mahesh Kumar, Prabhakar Pande, Lakshmi Kashyap and Pushpak Bhattacharyya. 2004. Hindi Word Sense Disambiguation. International Symposium on Machine Translation, Natural Language Processing and Translation Support Systems, Delhi, India, November, 2004. http://www.cse.iitb.ac.in/~pb/papers/HindiWSD.pdf