ISSN:2249-5789
V Ramsundhar et al , International Journal of Computer Science & Communication Networks,Vol 3(3), 147-151

Automatic Word Sense Disambiguation Using Wikipedia Link Structure

V.Ramsundhar
School of computing
Sastra University
Thanjavur- 613401,
Tamilnadu, India
sundar_varathu@yahoo.com

M.Ramkumar
School of computing
Sastra University
Thanjavur- 613401,
Tamilnadu, India
ramkumar.ce1@gmail.com

J.Sivakumar
School of computing
Sastra University
Thanjavur- 613401,
Tamilnadu, India
jpsivas@gmail.com

Abstract

In this project an approach based on Wikipedia link structure for word sense Disambiguation is presented and evaluated. This paper describes a method for creating sense tagged data using Wikipedia as a source of sense semantic annotations. Word sense disambiguation (WSD) is the ability to identify the meaning of words in context in a computational manner. It is essence of communication in a natural language. Through word sense disambiguation experiments and results we show that the Wikipedia-based sense annotations are reliable and can be used to construct accurate sense. Ambiguous words or sentences can be understood multiple ways, though only one meaning is intended. Disambiguation seeks to decipher the intended meaning of words and sentences.. However, the presented approach has several limitations: a small sample and a big number of fine senses in WordNet, many of which are not that distinguishable from each other.

Keywords -Word Sense Disambiguation, Wikipedia, Wordnet, Semantic web

1. INTRODUCTION

All natural languages contain words that can mean different things in different contexts. Ambiguities are essential to human language. In particular, word sense ambiguity is prevailing in all natural languages, with a large number of words in any given languages have more than one meaning. For instance, the English noun play can mean game or sports or drama. The correct sense of an ambiguous word can be selected based on the perspective where it occur, and correspondingly the problem of word sense disambiguation is defined as the task of automatically assigning the most appropriate meaning to a polysemous word within a given context. Word sense disambiguation is an important issue to understand the semantic web/web3.0, because the semantic web can be regarded as addition of a machine-understandable and machine - tractable layer to complement the existing web of natural language hypertext, in order to support automatic communication between web-based applications.

In WSD the Knowledge-Based approaches are many here are methods like LESK algorithm which calculates overlap with respect to dictionary definitions. One of the several approaches proposed in the past is Michael Lesk's 1986 algorithm.

This algorithm is based on two assumptions. First, when two words are used in close proximity in a sentence, they must be talking of a related topic and second, if one sense each of the two words can be used to talk of the same topic, then their dictionary definitions must use some common words.

2. Word Sense Disambiguation

In computational semantics, word sense disambiguation (WSD) is an open problem of natural language processing, which rules the process of identifying which sense (meaning) of a word is used in a sentence or words, when the word has multiple meanings (polysemy). The solution to this problem impacts other computer-related writing, such as discourse, improving relevance of search engines, discours, anaphora resolution, inference, coherence, and others.

Provided a set of sense-annotated examples for a given ambiguous word, the task of a word sense disambiguation system is to automatically learn a disambiguation model that can predict the correct sense for a new, previously unseen occurrence of the word. The disambiguation algorithm starts with a preprocessing step, where the text is tokenized and annotated with part-of-speech tags. Collocations are identified using a sliding window approach, where a collocation is defined as a sequence of words that forms a compound concept defined in WordNet .Next, local and topical features are extracted from the context of the ambiguous word. Specifically, we use the current word and its part-of-speech, a local context of three words to the right and left of the ambiguous word, the parts-of-speech of the surrounding words, the verb and noun before and after the ambiguous words, and a global context implemented through sense-specific keywords determined as a list of at most five words occurring at least three times in the contexts defining a certain word sense.

3. Wikipedia Link Structure

Wikipedia is a free online encyclopedia, representing the outcome of a continuous effort of a large number of volunteer contributors. Virtually any Internet user creates or edits a Wikipedia webpage, and this "freedom of contribution" has a positive impact on both the quality of this online resources. Wikipedia editions are available for more than 200 languages, with a number of entries varying from a few pages to more than one million articles per language. The basic entry in Wikipedia is an *article* (or *page*), which defines and describes an entity or an event, and consists of a hypertext document with hyperlinks to other pages within or outside Wikipedia.

The role of the hyperlinks is to guide the reader to pages that provide additional information about the entities or events mentioned in an article

3.1 Interwiki Link The links between Wikipedia pages are called inter wikis.

3.2 Inter language Link This links act as an internationalization mechanism, they are connections between Wikipedia articles on the same topic but in different languages.

3.3 Strong Link*:* we are interested in for WSD are what we called "strong links". We define a strong link as a bidirectional connection between two pages.

4. WORDNET

WordNet is a lexical database of English. Nouns, adjectives, verbs, and adverbs are grouped into sets of reasoning synonyms (synsets), each expressing as a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. For instance, words in WordNet are arranged semantically instead of alphabetically. Synonymous words are grouped together to form synonym sets, or synsets. Each such synset therefore represents a single distinct sense or concept. Thus, the synset {base, alkali} represents the sense of any of various water-soluble compounds capable of turning litmus blue and reacting with an acid to form a salt and water. In WordNet, each word occurs in as many synsets as it has senses. For example the word base occurs in two noun synsets, {base, alkali} and {basis, base, foundation, fundament, groundwork, cornerstone}, and the verb synset {establish, base, ground, found}. WordNet stores information about words that belong to four parts–of–speech: Nouns, verbs, adjectives and adverbs.

Words with multiple senses can either be *homonymous* or *polysemous*. Two senses of a word are said to be homonyms when they mean entirely different things but have the same spelling. For example the two Senses of the word *bark* – tough protective covering of trees and the sound made by a dog are homonyms because they are not related to each other. A word is said to be polysemous when its senses are various shades of the same basic meaning.

For example, the word *accident* is polysemous since its two senses – a mishap and anything that happens by chance are somewhat related to each other. Note that WordNet does not distinguish between homonymous and polysemous words, and therefore neither do we. Thus WordNet does not indicate that the two senses of the word

accident are somewhat closer to each other in meaning than the two senses of the word *bark*.

Parts of Speech	Gloss length in process (in words)		Number of Senses	
	Average	Deviation	Average	Deviation
Noun	11.1	6.3	1.2	0.8
Verb	6.2	3.4	2.2	2.5
Adjective	7.0	3.9	1.4	1.1
Adverb	4.9	2.3	1.2	0.7

.

5. RELATED WORK

The Wikipedia is the largest online collaborative knowledge sharing system, a free encyclopedia. Built upon traditional wiki architectures, its search capabilities are limited to title and full-text search analyzed the link structure in Wikipedia. They tackle the problem of missing links between articles. For doing this they cluster similar pages based on similar link structure and then they examined these cluster to find missing links between them [1].

In [2] used this gold standard for testing three approaches for WSD. The first reported one is called knowledge based approach and takes into account the paragraph where the ambiguous word was found as a representation of the context (LESK algorithm), and a second one called data-driven method imply a classifier and builds a feature vector with words in proximity of an ambiguous word found in the text and in the Wikipedia link of possible senses .A third approach combines these two. Disambiguation process instead focuses only on the strong link analysis.

In [2] and [3] discuss the use of Wikipedia for Word Sense Disambiguation (WSD).This approach selects all paragraphs in Wikipedia which contain a contextualized reference to an ambiguous term in the link label and then maps the different Wikipedia annotations to word senses instead of relying on the Wikipedia disambiguation pages. This is due to the face that sometimes not all meaning is elicited in the disambiguation page. Finally, the labels which describe

6. PROPOSED SYSTEM

The first attempts at automated sense disambiguation were made in the context of machine translation (MT). In his famous Communication, Weaver discusses the need for WSD in machine translation and outlines the basis of

Table 1: Gloss size and number of senses for each part of speech in WordNet

WordNet stores information about words that belong to four parts-of-speech: noun, verb, adjective and adverb. WordNet version 1.7 has 107,930 nouns arranged in 74,448 synsets, 10,860 verbs in 12,754 synsets 21,365 adjectives in 18,523 synsets and 4,583 adverbs in 3,612 synsets .prepositions and conjunction do not belong to any synset. Although our algorithm is general enough to be applied to any lexical database that has a similar hierarchical semantic arrangement, we have done our experiments using WordNet version 1.7, and so we only disambiguate those words that are occur in WordNet.
the possible senses for a word are manually mapped to Word- Net senses. In this way the number of example for each word can increase improving the performance of a classifier [3].

An adaptation of Lesk's dictionary based on word sense disambiguation algorithm. Rather than using a standard dictionary as a source of glosses for, the lexical database Wordnet is employed. This provides a rich hierarchy of semantic relationships that our algorithm can exploit [4]

Investigate whether the Wikipedia corpus is amenable to multilingual analysis that aims at generating parallel corpora. We present the results of the application of two simple heuristics for the identification of similar text across multiple languages in Wikipedia. Despite the simplicity of the methods, evaluation carried out on a sample of Wikipedia pages shows encouraging results [5].

Work based on Wikipedia knowledge which searches for synonyms and related terms in the Wikipedia category structure and analyzing hyperlinks between pages. The algorithm could be used to extend queries in a search engine, or as an assistant for forming a dictionary of synonyms [6].

Analyzed the link structure in Wikipedia. They tackle the problem of missing links between articles. For doing this they cluster similar pages based on similar link structure and then they examined these cluster to find missing links between them [7].
an approach to WSD which underlies all subsequent work on the topic.

Considerable amount of effort has been devoted to the development of automatic annotation

Methodologies for the Semantic web during the last few years. Most of the approaches proposed exploit information extraction techniques such as the recognition of named entities, relationships and events. For example, Kogut & Holmes (2001) present a system that generates DAML annotations for most proper nouns and common relationships from web pages using AeroText TM, a commercial information extraction tool. Propose an adaptive information extraction approach where information from structured sources is used to train learning algorithms capable of automating the annotation of domain specific web pages.

Agirre and Rigau uses a measure based on the proximity of the text words in WordNet (conceptual density) to disambiguate the words. The idea that translation presupposes word sense disambiguation is leveraged to disambiguate words using bi-lingual corpora. The idea of constructing a BBN from WordNet has been proposed earlier and forms a motivation for the present work.

However, unlike we particularly emphasize the need for soft sense disambiguation, *i.e.* synsets are considered to probabilistically cause their constituent words to appear in the texts. Also we describe a comprehensive training methodology and integrate soft WSD into an interesting application, *viz.*, QA. Bayesian Balief Network (BBN) is used as the machine for this probabilistic framework.

Wikipedia is an invaluable Web corpus for knowledge extraction. Researches on semantic relatedness measurement are already well conducted. Wiki Relate is one of the pioneers in this research area. The algorithm finds the shortest path between categories which the concepts belong to in a category graph.

7. IMPLEMENTATION AND RESULTS

In this section we propose a computer-based experiment for measuring the quality of our approach in WSD. They have created a gold standard to be used in evaluating WSD algorithms as described in the dataset consist of 112 manually semantic annotated Wikipedia articles. The annotators where also asked to choose words with a corresponded Wikipedia definition to describe the topics of the article. These keywords are used to avoid that inaccuracy in information extraction could influence the WSD task.

Hence, they assumed that the keyword extraction stage produced 100% precision and recall. We decided to apply the same methodology and first evaluate our disambiguation algorithm based on the manual keyword extraction, as proposed in the original experiment settings and consequently by automatically extracting the most important keywords from the articles as we normally do in our approach. Focusing the evaluation on a task at the time of the disambiguation process permits to avoid and then calculate the error Propagation effect.

In [5] Mihalcea and Csomai report an Assessment on a set of 85 pages while the same dataset we used consists of 111 articles. The aim of this evaluation was not only focused on comparing our disambiguation process with other WSD approaches, rather to evaluate all the assumptions at the base of our algorithm such as the importance of strong links for determining semantic relevance of articles

7.1 Our disambiguation algorithm follows the following Steps:

1. First, looks up in the database version of Wikipedia for all the article starting with an ambiguous term, in case of Aida for example it retrieves the following pages: Aida (1953 film), Aida (caf), Aida (camp), Aida (film), Aida (musical),Aida (name) and Aida (opera).

2. The second step of the disambiguation process analyzes for each of the candidate definition its strong links and it compares them with the keyword extracted from the original article, for the first evaluation task we skip the keyword extraction but we used the used the given keywords instead. On a second evaluation task the manual annotation will be substituted with keywords automatically extracted from the text in exam. The strong links lookup is done using the online version of Wikipedia.

3. The third and optional step is executed if the disambiguation process using strong link method could not determine the correct sense among the candidate definitions; in this case the disambiguation process is re-iterated taking into account all links.

The result of the first task of the evaluation was encouraging. We could guess the correct page for the entire 111 Wikipedia article but for 9 of them we had to consider the entire link. Only using the strong link analysis we could reach a precision of 90% and considering all links the precision reach the 100%. This means that in the 90% of the cases with the strong link analyses with could guess the correct sense of an article.

Method	Precision	Recall
Knowledge-based	80.63	71.86
Feature-based learning	92.91	83.10
Combine(Knowledge + Feature)	94.33	70.51
Strong Link based	90.01	91.81
Link based	100	100

Table 2. WSD Performance Comparison

8. CONCLUSION

In this paper we have described and evaluated an approach for WSD based on the knowledge extracted from Wikipedia. Wikipedia as a knowledge resource for WSD. In this project, we described an approach for using Wikipedia as a source of sense annotations for word sense disambiguation. Starting with the hyperlinks available in Wikipedia, we showed how we can generate a sense annotated corpus that can be used to train accurate sense classifiers. We have adapted Lesk's algorithm for word

An important outcome of the experiment is the proof that strong link structure among articles is important for drawing the semantic domain in which a topic resides. Moreover, strong link permits with high precision to discard pages that are not relevant. Our algorithm takes into account all pages starting with an ambiguous name since in many NLP tasks the ambiguous name could be stemmed or have other declination with a connected meaning. In this way we can take consider more candidates, anyway the strong links structures helps us distinguish only the relevant pages of a predefined domain.

sense disambiguation using dictionary definitions to the electronic lexical database WordNet. We applied this WSD approach in a digital library environment for automatically annotating and enabling searches and navigation through an unstructured multimedia. The good results of the evaluation suggest that our approach might be applied in different scenarios such as text categorization and document classification, where it is crucial to automatically extract semantic information from content.

9. REFERENCES

[1] S. Chernov, T. Iofciu, W. Nejdl, and X. Zhou. Extracting semantic relationships between Wikipedia categories. In 1st Workshop on Semantic Wikis: June December 2006.

[2] R. Mihalcea and A. Csomai. Wikify: linking documents to encyclopedic knowledge. In CIKM '2007: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pages 233–242, New York, NY USA, 2007. ACM.

[3] R. Mihalcea. Using Wikipedia for automatic word sense disambiguation. In Proceedings of NAACL HLT 2007, pages196–203, 2007

[4] S. Bannered and T. Pedersen. An adapted Lesk algorithm for word sense disambiguation using Word-Net. In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, February 2002.

[5] S. F. Adafre and M. de Rijke. 2006. Finding similar sentences across multiple languages in Wikipedia. In Proceedings of the EACL Workshop on New Text, Trento, Italy.

[6] A. Krizhanovsky. Synonym search in Wikipedia: Synarcher. Arxiv.org. Search for synonyms in Wikipedia using hyperlinks and categories

[7] S. F. Adafre and M. de Rijke. Discovering missing links in Wikipedia. In Link KDD '05: Proceedings of the 3rd international workshop on Link discovery, pages 90–97, New York, NY, USA, 2005. ACM.

[8] S. Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In EMNLP 2007: Empirical Methods in Natural Language Processing, Prague, Czech Republic, pages 708–716, June 28-30, 2007.