International Journal of Engineering Science Invention
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org Volume 2 Issue 10ǁ October. 2013 ǁ PP.01-04

Word Sense Disambiguation

Mukti Desai, Mrs. Kiran Bhowmick
Department of Computer Engineering Dwarkadas J. Sanghvi College of Engineering
Mumbai University, India.

ABSTRACT: Ambiguity and human language have been tangled since the rise of philological communication. One of the long established problems of Natural Language Processing (NLP) is Word Sense Disambiguation (WSD). Researchers have been diligently trying to deal with this problem since the birth of Machine Translation. Word sense disambiguation (WSD) can be defined as the aptitude to recognize the meaning of words in the given context in a computational manner. WSD is an AI-complete problem, that is, a problem having its solution at least as hard as the most difficult problems in the field of artificial intelligence.

KEYWORDS: *word sense disambiguation, WSD, natural language processing, NLP, polysemy*.

I. INTRODUCTION

Human language is fairly ambiguous; hence, numerous words can be portrayed in several different ways based on the context in which they occur. Most of the words in natural languages are polysemous, having multiple possible meanings or senses.

The fact that computers do not possess the privilege of vast experience of the world of Natural language, which is possesses by only human beings, the problem of arbitrating the correct sense of a polysemous word by reflex becomes an arduous problem.
Word sense disambiguation is defined as the task of finding the sense of a word in a context. [4]

II. PROBLEM STATEMENT

The identification of the specific meaning that a word assumes in the context is only apparently simple. The reason behind this is that human brain is capable of determining the correct meaning of the word in the given context in a spontaneous way. While most of the time humans do not bother regarding the equivocation of language, machines have to process disorderly textual information and transmute them into data structures which must be analyzed in order to perceive the underlying meaning. The computational identification of meaning for words in context is called word sense disambiguation (WSD). [2]

In English language, we have a variety in types of ambiguities. Let us consider the ones that we have to deal with in order to solve the problem of word sense disambiguation.

Words with same spelling, same pronunciation but either same or different meaning are homographs:
Eg. minute (extremely small, measure of time)

Words with same spelling, same pronunciation but different meaning are homonyms:
Eg. rose (flower, past tense form of verb 'rise')

Words with same spelling but different pronunciation and different meaning are heteronyms:
Eg. dove (bird, past tense of verb 'dive')

Now consider the following sentences in order to understand the word sense disambiguates found in daily communication:

a) Harry took Sarah to a café on a date.
b) Mary's favorite fruit to eat is a date.
c) John's date of birth is November 18, 1992.

As naturally intelligent human beings, we tend to resolve this conflict by integrating a wide variety of gigantic semantic and syntactic hints and using a rich reasoning aptitude which is established on the basis of causality and relatedness.

But the artificially intelligent machines invented by us suffer from word sense disambiguation due to the multiple contexts found in the above mentioned statements regarding the word 'date', and hence, they interpret the sentences in the following way.

a) Harry took Sarah to a café on a date {romantic meeting / fruit / day of month}.
b) Mary's favorite fruit to eat is a date {romantic meeting / fruit / day of month}.
c) John's date {romantic meeting / fruit / day of month} of birth is November 18, 1992.

III. RELATED WORK

Removal of WSD in the field of Natural Language Processing is not only chronic but also a deep rooted issue. It was comprehended as an elementary problem in late 1940s (Weaver, 1949) in the domain of Machine Translation.

Consider the elementary working of WSD. It is completely dependent on knowledge. The fundamental procedure of any WSD system can be enunciated as follows: [2]

a) A set of words are given.
b) A technique is applied which makes use of one or more sources of knowledge to associate the most appropriate senses with words in context.

Knowledge sources can vary considerably from the agglomeration of texts, either unlabeled or annotated with word senses, to more structured resources, such as machine readable dictionaries, semantic networks, etc.

There are various approaches that are used in order the solved WSD. We shall understand the broad classification of these strategies according to the following classification:

1. Knowledge Based Approaches [6]

This approach relies on knowledge resources like WordNet (Banerjee and Pedersen 2003; Navigli and Velardi, 2005) Thesaurus etc.

WordNet is defined as a machine-readable lexical database organized according to meanings by itself. In WordNet, English nouns, verbs and adverbs are organized into synonym sets representing lexical concepts. These sets are linked by relations such as synonym, antonym and so forth itself. [17]

This approach usually picks the sense whose definition is most similar to the context of the ambiguous word, by means of textual overlap or using graph-based measures (Agirre, De Lacalle, and Soroa 2009). Consequently, most dictionary-based methods are sensitive to the exact wording of the definitions, as they have not realized the potential of combining the limited information in such definitions
with the abundant information extractable from text corpora (Cuadros and Rigau 2006).

It allows the use of grammar rules as well as hand coded rules for disambiguation. The overlap based approach strategized by knowledge based technique requires a Machine Readable Dictionary (MRD). It follows the Lesk's algorithm which maintains two data banks:
Sense Bag: contains the words in the definition of a candidate sense of the ambiguous word.
Context Bag: contains the words in the definition of each sense of each context word.

The overlap between the features of different senses of an ambiguous word (sense bag) and the features of the words in its context (context bag) is found. The sense which has the maximum overlap is selected as the contextually appropriate sense.

It can also make use of Walker's algorithm which is thesaurus based methodology. For each meaning of the target word, thesaurus category is found in order to which that sense belongs to in the context. Calculate the score for each sense by using the context words and the highest scorer wins.

2. Machine Learning Based Approaches

The machine learning approach generally includes building a classifier with co occurrence features and using it to assign senses to unseen examples (Chklovski and Mihalcea 2002; Ng, Wang, and Chan 2003). To perform well, it needs large training annotated sets that are extremely expensive to create (Edmonds 2000). [9] These approaches make use of supervised, semi supervised as well as unsupervised algorithms. They are dependent on corpus evidence which is used to train a model using tagged or untagged corpus. Some of the probabilistic / statistical models are also used. Wikipedia is also used by machine learning algorithms (Bunescu and Pasca 2006; Cucerzan 2007). Additionally, recent enrichment methods utilize Wikipedia and map senses to corresponding articles [9] (Mihalcea 2007; Ponzetto and Navigli 2010).

The unsupervised corpus-based methods of WSD are knowledge-lean, and do not rely on external knowledge sources such as machine readable dictionaries, concept hierarchies, or sense-tagged text.

In supervised disambiguation method, the system is trained with manually created examples of correctly disambiguated words in context. [15]

There are various models used in supervised learning approaches: The Naïve Bayesian Classifier, Decision Lists and Trees, Neural Networks etc.

Let us consider the exemplar based classifier using KNN strategy based on supervised learning approach. [6] From each sentence that is marked with its sense, find an ambiguous word to construct a training example. This makes the use of part of speech tagging of the word as well as its neighboring words, along with local collocations, co-occurrence vector, morphological features and subject – verb syntactic dependencies. When a test sentence containing the ambiguous word is given, a test example is similarly constructed. The test example is then compared to all training examples and the k-closest training examples are selected. The sense which is most dominant amongst these k examples is then selected as the correct choice.

3. Hybrid Approaches

They make use of corpus evidence as well as the semantic rules. Few examples of hybrid approaches are bootstrapping approach, Yarowsky algorithm (1995), [16] Semi-automatic Dictionary Drafting: SADD [Kilgarriff and Rychl´y, 2010] etc.

IV. APPLICATIONS

We enunciate a number of real-world applications which might get advantage from WSD and on which experiments have been and are being conducted. [2]

In Information Retrieval (IR), an accurate disambiguation of the document and the query words will eliminate documents containing the same words used with different meanings and to retrieve documents expressing the same meaning with different wordings.

Information Extraction (IE) domain requires WSD in order to solve the problem of named – entity recognition.

Automatic identification of the correct translation of a word in context, called as machine translation (MT) requires WSD translation, based on the intuitive idea that the disambiguation of texts should help translation systems choose better candidates, since depending on the context; words can have completely different translations.

The analysis of the general content of text in terms of its ideas, themes, etc., for instances, the classification of blogs within the Internet community, social network analysis etc. will definitely avail the solutions to WSD.
Few other application domains for word sense disambiguation are Word Processing, Lexicography, and Semantic Web etc.

V. CONCLUSION & FUTURE SCOPE

In this paper, we have inspected the field of word sense disambiguation. WSD is a hard task of AI since it deals with the complexities of natural language. It aims to identify a semantic structure from unstructured sources of text. Primary reason for its inherent difficulty is that it is highly reliant on knowledge obtained from several sources, which is very difficult for a computational model to process.

A further problem which concerns the representation of senses is their ever-changing nature. [2] A sense might be missing due to a new usage, a new word, a usage in a specialized context that the lexicographers did not want to cover, or simply an omission.

Research in the field of WSD has been conducted since the early 1940s. Significant research is undergoing in the fields of knowledge representation and acquisition today, which leaves abundant amount of scope of improvement in the utilization of resources in a very knowledge-reliant task like WSD.

REFERENCES

[1]. Mihalcea, Rada, and Dan I. Moldovan. "A method for word sense disambiguation of unrestricted text." *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Association for Computational Linguistics, 1999.

[2]. Navigli, Roberto. "Word sense disambiguation: A survey." *ACM Computing Surveys (CSUR)* 41.2 (2009): 10.

[3]. Resnik, Philip, and David Yarowsky. "A perspective on word sense disambiguation methods and their evaluation." *Proceedings of the ACL SIGLEX workshop on tagging text with lexical semantics: Why, what, and how*. 1997.

[4]. Ramakrishnan, Ganesh, et al. "Soft word sense disambiguation." *Proceedings of GWC*. Vol. 4. 2004.

[5]. Satapathy, Shruti Ranjan. *Word Sense Disambiguation*. Diss. Indian Institute of Technology, 2013.

[6]. Thottempudi, Sree Ganesh. "WORD SENSE DISAMBIGUATION."

[7]. Dandala, Bharath, Rada Mihalcea, and Razvan Bunescu. "Word Sense Disambiguation Using Wikipedia." *The People's Web Meets NLP*. Springer Berlin Heidelberg, 2013. 241-262.

[8]. Fernandez-Ordonez, Erwin, Rada Mihalcea, and Samer Hassan. "Unsupervised Word Sense Disambiguation with Multilingual Representations." *LREC*. 2012.

[9]. Raviv, Ariel, Shaul Markovitch, and Sotirios-Efstathios Maneas. "Concept-Based Approach to Word Sense Disambiguation." *AAAI*. 2012.

[10]. Mihalcea, Rada -"Word Sense Disambiguation". (2010): 1027-1030.

[11]. Lee, Wei Jan, and Edwin Mit. "Word Sense Disambiguation by using domain knowledge." *Semantic Technology and Information Retrieval (STAIR), 2011 International Conference of* IEEE, 2011.

[12]. Charhate, Sayali, et al. "Adding intelligence to non-corpus based word sense disambiguation." *Hybrid Intelligent Systems (HIS), 2012 12th International Conference on*. IEEE, 2012.

[13]. Chatterjee, Niladri, and Rohit Misra. "Word-Sense Disambiguation using maximum entropy model." *Methods and Models in Computer Science, 2009. ICM2CS 2009. Proceeding of International Conference on*. IEEE, 2009.

[14]. Bruce, Rebecca, and Janyce Wiebe. "Word-sense disambiguation using decomposable models." *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1994.

[15]. Kulkarni, Manasi, and Suneeta Sane. "An ontology clarification tool for word sense disambiguation." *Electronics Computer Technology (ICECT), 2011 3rd International Conference on*. Vol. 1. IEEE, 2011.

[16]. Diana McCarthy - "Word Sense Disambiguation" - Lexical Computing Ltd., University of Melbourne, July 2011

[17]. Legrand, Steve, and J. R. G. Pulido. "A hybrid approach to word sense disambiguation: Neural clustering with class labeling." *Knowledge Discovery and Ontologies workshop at 15th European Conference on Machine Learning (ECML)*. 2004.