Approaches for Disambiguation in Hindi Language

Pankaj Kumar[1], Atul Vishwakarma[2] and Ashwani Kr. Verma[3]
[1]Assitant Professor, Computer Science dept., Shri Ramswaroop Memorial Group of Professional Colleges, Lucknow, UP, India
[2,3]Graduate Scholar, Computer Science dept., Shri Ramswaroop Memorial Group of Professional Colleges, Lucknow, UP, India

Abstract

Hindi is National Language of India spoken by about 500 million people and ranking 4th among the majority spoken language in the world. But the ambiguities present in this language create hindrance in usage of Information technology for native users. So, there is the need of effective measures to perform natural language processing thereby making the native users utilize these technologies to the fullest. Language translator is important tool to resolve this problem. Word sense disambiguation is an important concept that is to be evaluated for performing machine translation and a tool to perform disambiguation. In this research paper the different approaches for disambiguation in Hindi Language are discussed and their comparative study is made to reach the conclusion.

Keywords

Word Sense Disambiguation (WSD), Natural Language Processing (NLP), Machine Translation (MT).

1. Introduction

1.1 Word Sense Disambiguation

The task of selecting the correct sense for a word is called word sense disambiguation, or WSD. A word can have number of senses, which is termed as ambiguity. Something is ambiguous when it can be understood in two or more possible ways. This word sense disambiguation is an intermediate task, but rather is necessary at one level to accomplish most natural language processing tasks. In this way, word sense disambiguation is the problem of selecting a sense for a word from a set of predefined possibilities. Here the sense inventory comes from a dictionary or thesaurus. Many words have more than one possible meaning e.g.

हर वर्ग के लोग कीमतों में वृद्धि से पीड़ित है|

Here 'वर्ग' is interpreted as 'class'.

सात का वर्ग उनचास है|

Here 'वर्ग' is interpreted as 'square of the number'.

यह 5cm का एक वर्ग है|

Here 'वर्ग' is interpreted as 'square-shaped figure'.

So in this way there is ambiguity for 'वर्ग'.

1.2 Machine translation

Machine translation (MT) is an application of computers to the task of translating texts from one natural language to another. Machine translation (MT) is also known as "Automatic Translation" or "Machine Translation". MT is multidisciplinary field of research. It uses the ideas from linguistics, computer science, artificial intelligence, statistics, mathematics, philosophy and many other fields. There are at least two stages:

1) Understanding the source language and
2) Generating sentences in the target language.

WSD is required in both stages since a word in the source language may have more than one possible translation in the target language. For example, the English word *"drug"* can be translated into Turkish as *"ilaç"* for its sense of *"medicine"* or as *"uyuşturucu"* for its sense of *"dope"* depending on the context. In order to be able to correctly translate a text, we need to know which sense is intended in the text.

1.3 Role of Word Sense Disambiguation in Machine Translation

The sense disambiguation is essential for the proper translation of words such as the Hindi 'सोना', which, depending on the context, can be translated as 'Gold', 'Sleep', 'Sona (the name)' etc.[1]
Example-

सोना सोना चाहता है|

It can be translated as-
Sona wants gold.
or
Sona wants to sleep.
or
Gold wants to sleep.

or

Gold wants sona. etc.

So, in this way there is ambiguity for the word 'सोना' because it has different senses.

2. Related work

Some of the methods and their approaches for word sense disambiguation will be discussed. We will discuss works done by various researchers in this particular area and problem.

"Unsupervised word sense disambiguation rivalling supervised methods", Yarowsky, D. (1995), this paper presents an unsupervised learning algorithm for sense disambiguation. The algorithm is based on two powerful constraints - one sense per discourse and one sense per collocation- exploited in an iterative bootstrapping procedure. Tested accuracy exceeds 96%.

Dekang Lin. (1997) in this paper "Two different words are likely to have similar meanings if they occur in identical local contexts" is adopted in this paper. D*isambiguation* is done based on syntactic dependency and *sense* similarity.

Rigau et al. (1997) it correctly states that most WSD algorithms have been developed as standalone and investigate the possibility of combining them. The methods in the study include those used by Pedersen et al. and some baseline methods such as using the most frequent sense. Test results indicate approximately 8 % increase in precision for the combination of disambiguation methods.

"Word Sense Disambiguation by Web Mining" Peter D. TURNEY has developed the NRC (National Research Council) Word sense Disambiguation (WSD) system, which is applied to English Lexical Sample (ELS). In which, we used the Supervised approach for machine learning problem. Familiar tools are used such as the Weka machine learning software and Brill's rule-based part-of-speech tagger. They represented as features like semantic features and syntactic features. The main motive in the system is the method for generating the semantic features, based on word co-occurrence probabilities.

"Word Sense Disambiguation for Vocabulary Learning" Anagha Kulkarni, Michael Heilman, Maxine Eskenazi and Jamie Callan (2006) have developed the word sense disambiguation for vocabulary learning. It is designed to assist English as a Second Language (ESL) student to improve their English vocabulary, to operate at the level of the word meaning pairs being learned and not just the words being learned, for several reasons. The *supervised and unsupervised approaches* were used. Supervised approaches were consistently more accurate than using unsupervised approaches.

3. Approaches

3.1 Approaches for Disambiguation

Following are the approaches of disambiguation: [2]

3.1.1 Knowledge Based Disambiguation

3.1.1.1 WSD using selection preferences (or restrictions)

They have frequently been cited as useful information for WSD but it has been noted that there use is limited and that additional sources of knowledge are required for full and accurate WSD.

For example, 'खाना' can be treated as 'food' or 'to eat', only first sense is available in the context of 'वह आम खाना चाहता है.', only second sense is applicable here as 'आम.' species the selection restriction to eat in the context.

3.1.1.2 Overlap Based Approaches

These require a machine readable dictionary (MRD). They find 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). There are many algorithms used for overlap based approaches. The major algorithms used are:

- Lesk's algorithm
- Walker's algorithm
- WSD using Conceptual Density

3.1.2 Machine Learning Based Approaches

These approaches can be divided into three approaches:

3.1.2.1 Supervised Approaches

Supervised methods are based on the assumption that the context can provide enough evidence on its own to disambiguate words (hence, world knowledge and reasoning are deemed unnecessary)**.** These supervised methods are subject to a new knowledge acquisition bottleneck since they rely on substantial amounts of manually sense-tagged corpora for training, which are laborious and expensive to create.

3.1.2.2 Semi-supervised Algorithm [3]

Its example is the bootstrapping approach. The bootstrapping approach starts from a small amount of seed data for each word: either manually-tagged training examples or a small number of sure-fire decision rules (example 'आम' in the context of 'फल' almost always indicates the fruit). The seeds are used to train an initial classifier, using any supervised method.

3.1.2.3 Un-ssupervised Algorithm

They are the greatest challenge for WSD researchers. The underlying assumption is that similar senses occur in similar contexts, and thus senses can be induced from text by clustering word occurrences using some measure of similarity of context. It is hoped that unsupervised leaning will overcome the knowledge acquisition bottleneck because they are not dependent on manual efforts.

3.1.3 Hybrid approach

These approaches are the hybrid between different methods like statistical based and rule based methods of machine learning approaches. By this approach, we can also combine the advantages of corpus based and knowledge based methods. For example Sin-Jae Kang [4] has applied the previously secured dictionary information to select the correct senses of some ambiguous words with high precision, and then use the ontology to disambiguate the remaining ambiguous words.

4. Comparative study of different approaches

4.1 Comparison of Knowledge Based Approaches[2]

Table 1

S .No	Algorithm	Accuracy
1.	WSD using restrictions	44% on brown corpus
2.	Lesk's algorithm	50-60% on short samples of "Pride and Prejudice" and some "news stories".
3.	WSD using conceptual density	54% on brown corpus.
4.	Walker's algorithm	50% when tested on 10 highly polysemous English words.

4.2 Comparison of Machine Based Approaches

4.2.1 Comparison of supervised approaches

Table 2

S .No	Approach	Average precision	Average recall	Corpus	Average baseline accuracy
1.	Naïve Bayes	64.13 %	Not reported	Senseval-3 all Words Task	60.9%
2.	Exemplar based	68.6%	Not reported	WSJ6 containing 191 contents words	63.7%
3.	Decision lists	96%	Not applicable	Tested on a set of 12 highly polysemous English words.	63.9%
4.	SVM	72.4%	72.4%	Senseval 3-lexical sample task used for disambiguation of 57 words.	55.2%
5.	Perceptron trained HMM	67.6%	73.74%	Senseval 3-all words task	60.9%

4.2.2 Comparison of Semi-supervised approaches

Table 3

S .No	Approach	Average precision	Corpus	Average baseline accuracy
1.	Supervised decision lists	96.1%	Tested on a set of 12 highly polysemous English words.	63.9%
2.	Semi-supervised decision list	96.1%	Tested on a set of 12 highly polysemous	63.9%

			English words.	

4.2.3 Comparison of Semi-supervised approaches

Table 4

S.No	Approach	Average precision	Average recall	Corpus	Average baseline accuracy
1.	Linn's algorithm	68.5%	Not reported	Trained using WSJ corpus containing 25 million words.	64.2%
2.	Hyperlex	97%	82%	Tagged on a set of 10 highly polysemous French words.	73%
3.	WSD using Roget's thesaurus	92%(average degree of polysemy was 3)	Not reported	Tagged on a set of 12 highly polysemous English words.	Not reporter.

4.3 Comparison of Hybrid Approaches

Table 5

S.No	Approach	Average precision	Average recall	Corpus	Average baseline accuracy
1.	Iterative approach	92.2%	55%	Trained using 179 texts from SemCor.	Not reported.
2.	Sense Learner	64.6%	64.6%	SenseEval-3 all words task.	60.9%
3.	SSI	68.5%	68.4%	SenseEval-3 disambiguation task	Not reported.

5. Applications

Word sense disambiguation a task of removing the ambiguity of word in context, is important for many WSD applications using NLP such as:

- Information retrieval
- Machine translation
- Speech processing and part of speech Tagging
- Text Processing

5.1 Information retrieval

As proposed by WSD helps in improving term indexing in information retrieval has proved that word senses improve retrieval performance if the senses are included as index terms [5]. Thus, documents should not be ranked based on words alone, the documents should be ranked based on word senses, or based on a combination of word senses and words.

For example: Using different indexes for keyword "Java" as "programming language", as "type of coffee", and as "location" will improve accuracy of an IR system. Apart from indexing, WSD also helps in query expansion. Short queries are expanded using words that belong to same sets. Retrieval using expanded queries gives better results than original queries. Thus, WSD is crucial for improving accuracy of IR as it eliminates irrelevant hits.

5.2 Speech Processing and Part of Speech Tagging

Speech recognition i.e. when processing homophones words which are spelled differently but pronounced the same way. For example: "base" and "bass" or "sealing" and "ceiling".

5.3 Machine Translation

WSD is important for Machine translations. It helps in better understanding of source language and generation of sentences in target language. It also affects lexical choice depending upon the usage context.

5.4 Text Processing

Text to Speech translation i.e. when words are pronounced in more than one way depending on their meaning. For example: "lead" can be "in front of" or "type of metal".

6. Conclusion

Here, we have done comparison of different approaches that are being used for WSD; these approaches are as knowledge based, machine based and hybrid approach.

We found that the best approach is knowledge-based approaches. We have also found that our approach is successfully able to resolve the synonymy, antonymy, hypernymy, hyponymy, maronymy and holonymy relations for the different categories of parts-of-speech [6].

In this way, these approaches (e.g. Lesk's algorithm) can be used for disambiguation and its application will encourage and enable knowledge sharing and translation. If knowledge sharing between Hindi and other languages will be possible, it will help to cross the language barrier among the regional people.

References

[1] Mark Stevenson, "Word Sense Disambiguation: Natural Language Processing Group", University of Sheffield, UK. http://research.microsoft.com/india/nlpsummerschool/data/files/MarkStevenson%20-%20WSD%20tutorial.pdf.

[2] Rada Mihalcea and Ted Pedersen,"Slides from AAAI tutorial- advances in Word Sense Disambiguation." http://www.d.umn.edu/~tpederse/WSDTutorial.html.

[3] Sin-Jae Kang, "Corpus based Ontology for Word Sense Disambiguation". http://dspace/wul/waseda.ac.jp/dspace/bitstream/2065/12296/1/PACLIC17-399-407.pdf.

[4] Eneko Agirre and Oier Lopez de Lacalle and Aitor Soroa, "Knowledge-Based WSD on Specific Domains: Performing Better than Generic Supervised WSD", Informatika Fakultatea, University of the Basque Country 20018, Donostia, Basque Country.

[5] Avneet Kaur,"Development of an approach for disambiguating ambiguous Hindi postposition", Department of Computer Science, Punjab University, India.

[6] Parul Rastogi and Dr. S.K. Dwivedi, "Performance comparison of Word Sense Disambiguation (WSD) Algorithm on Hindi Language Supporting Search Engines", BabaSaheb BhimRao Ambedkar University Lucknow, UP, India.

Dr. Pankaj Kumar has received his Doctor of Philosophy (PhD) degree in Computer Application from Integral University, Lucknow. His Area of Expertise is Parallel Computing and Memory Architecture of Parallel Computer. Many of the valuable research papers of Dr. Pankaj Kumar have been published in various national/international journals and IEEE proceeding publication in the area of Parallel Computing. He is reviewer for six different International Journal and member of editorial board for two different International Journals.

Atul Vishwakarma is pursuing his Bachelor of Technology in Computer Science and Engineering. Currently he is working on final year project "Word Sense Disambiguation in Hindi Language using Web Mining." His areas of interest are Data Base Management Systems (DBMS), Operating Systems and Data Structures.

Ashwani Kr. Verma is pursuing Bachelor of Technology in Computer Science and Engineering. For his final year project he is working on "Word Sense Disambiguation in Hindi Language using Web Mining" with his project partner Atul Vishwakarma under the Guidance of Dr. Pankaj Kumar. His areas of interest are Operating Systems and Data Base Management Systems (DBMS).