Development of an Approach for Disambiguating Ambiguous Hindi postposition

Avneet Kaur
Department of Computer Science
Punjabi University, Patiala
Patiala, Punjab, India

ABSTRACT-In this survey paper, we have taken the problem as "***Development of an approach for disambiguating ambiguous Hindi postposition***". Word Sense Disambiguation (WSD) refers to the resolution of lexical semantic ambiguity and its goal is to attribute the correct senses to words in a given context. WSD is a most challenging problem in the area of NLP. We have chosen to develop an efficient algorithm for disambiguating ambiguous postpositions present in the Hindi language. We are taking this problem with the case study of existing Hindi-Punjabi Machine Translation System. Thus the disambiguation will be done from the machine translation point of view. This is mainly used for removing the ambiguity from the corpus.
N-gram algorithm is used for developing the Hindi postpositions. N-gram algorithm is used for extracting the words from the corpus.

Keywords: Word sense disambiguation (WSD; Natural Language Processing (NLP); Postpositions; 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. Many words have more than one possible meaning. For example-

सोना सोना चाहती है।

It can be translated as-

Sona wants gold.
Or
Sona wants to sleep.

So in this way there is ambiguity for 'सोना' because it is being interpreted of as gold means 'सोना' or as sleep means 'नींद' or as Sona (the name) means 'सोना'. When we look up a word in any dictionary, it can be seen that a word may have many meanings some of which are very different from each other.

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 "Mechanical 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 Importance of WSD in Machine Translation:

Machine translation is the original and most obvious application for WSD .WSD is required for lexical choice in MT for words that have different translations for different senses and that are potentially ambiguous within a given domain (since non-domain senses could be removed during lexicon development).For example, in an English- French financial news translator, the English noun *change* could translate to either *changement* ('transformation') or *monnaie* ('pocket money'). In MT, the senses are often represented directly as words in the target language. However, most MT models do not use explicit WSD.The machine translation process requires at least two stages:
1) Understanding the source language and
2) Generating sentences 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 ranslate a text, we need to know which sense is intended in the text.

POSTPOSITIONS

Postpositions- Postpositions are words that come after a noun to indicate a relationship to something else. (English uses prepositions which come before the noun. These are words such as, in, before, about, with). It has been analyzed that in Hindi, there are five most common postpositions like मे, पर, तक, से, and को in the listed below:

- का
- की
- के
- को
- में
- से
- बिना
- तक
- ने
- पर

It has been analyzed that there are only two postpositions that are ambiguous from the machine ranslation point of view. We are taking the case study of Hindi to Punjabi Machine Translation System:

- पर
- से

Following examples will demonstrate these ambiguities:

- से
  पड़ोसियो से ताल-मेल रखना चाहिए |

  बीस साल से कोई मिलने नही आया |

- पर
  किताबें मेज पर पड़ी हैं |

  पक्षी के किसी ने पर काट दिए |

2. APPROACHES-

Word Sense Disambiguation (WSD is the problem of determining in which sense a word having a number of distinct senses is used in a given sentence. A survey of learning methodologies that have been used for WSD is presented in the following section.

- Corpus based approaches
- Knowledge based approaches

2.1 CORPUS BASED APPROACHES-

In corpus based approaches, information is gained from training on some corpus. A corpus provides a set of samples that enables the systems to develop some numerical models. In corpus based approaches, information is gained from training on some corpus. A corpus provides a set of samples that

enables the systems to develop some numerical models. This approach can further be classified into two subclasses based on the training corpus as follow:
i. Supervised techniques
ii. Semi-supervised techniques
iii. Unsupervised techniques

In supervised WSD the training data is sense-tagged whereas in unsupervised WSD the training data is raw corpora which have not been semantically disambiguated. These features are of two classes: **collocation and cooccurrence features**.
Collocation features encode information about words of specific positions that are located to left or right of targetword. Typical features include the word, the root form of word, and the word's part-of-speech.
Consider an example:
"An electric guitar and bass player stand off to one side, not really part of scene, just as a sort of nod to gringo expectations."
Here we need to disambiguate word bass, so it is our target word. Collocation feature vector considering 2 words to right and 2 words to left of target words is: *[guitar, NN1, and, CJC3, player, NN1, stand, V V B]*
Co-occurrence features consist of data about neighboring words, ignoring their exact position. In this approach words themselves serve as features. The value of feature is the number of times the word occurs in the region surrounding the target word. The region is often a fixed window with target word as center. For the earlier example, a co-occurrence vector consisting of 12 most frequent words from a collection of bass sentences drawn from WSJ corpus has following features: fishing, big, sound, player, fly, rod, pound, double, runs, playing, guitar, and band.

5.1.1 Supervised Techniques:
Words can be labeled with their senses. Supervised approaches are similar to tagging:
- given a corpus tagged with senses
- define features that indicate one sense over another
- Learn a model that predicts the correct sense given the features.

In supervised approaches, a sense disambiguation system is learned from a representative set of labeled instances drawn from same distribution as test set to be used. Input instances to these approaches are feature encoded along with their appropriate labels. The output of the system is a classifier system capable of assigning labels to new feature encoded inputs. A major problem with supervised approaches is the need for a large sense tagging set. Bayesian classifiers, decision lists, decision trees, neural networks, logic learning system and nearest neighbor methods all fit into this paradigm. But we will discuss only first two because they have been the focus of considerable work in WSD.

5.1.2 Semi-Supervised Techniques:
The semi-supervised or minimally supervised methods are gaining popularity because of their ability to get by with only a small amount of annotated reference data while often outperforming totally unsupervised methods on large data sets. There are a host of diverse methods and approaches, which learn important characteristics from auxiliary data and cluster or annotate data using the acquired information.

5.1.3 Unsupervised Techniques:
Unsupervised approaches to sense disambiguation eschew the use of sense tagged data of any kind during the training. In this technique, feature vector representations of unlabeled instances are taken as input and are then grouped into clusters according to a similarity metric. These clusters are then labeled by hand with known word senses. Main disadvantage is that senses are not well defined.

2.2 Knowledge Based Approaches-
In Knowledge based approach; they provide both the means of constructing a sense tagger and target senses to be used. Attempts to perform large scale disambiguation have lead to the use of Machine Readable Dictionaries (MRD). In this approach, all the senses of a word to be

disambiguated are retrieved from the dictionary. Each of these senses is then compared to the dictionary definitions of all the remaining words in context. The sense with highest overlap with these context words is chosen as the correct sense.
For example: consider the phrase pine cone for selecting the correct sense of word cone and following definitions for pine and cone:

Pine:
1. Kinds of evergreen tree with needle-shaped leaves
2. Waste away through sorrow or illness
Cone:
1. Solid body which narrows to a point
2. Something of this shapes whether solid or hollow
3. Fruit of certain evergreen trees

The most common algorithm is N-gram algorithm:

- N-gram is a sequential list of n words.
- We approximate the probability of a word given all the previous words by the probability of the word given the single previous word.
- The Bigram model approximates the probability of a word all the previous words P (wn|w1n-1) by the conditional probability of preceding P (wn|wn-1).
- The Trigram model is same as a bigram model, except that we condition on two previous words.
- It involves splitting sentence into chunks of consecutive words of length "n".

EXAMPLE

- "I don't know what to say"
- 1-gram (unigram): I, don't, know, what, to, say
- 2-gram (bigram): I don't, don't know, know what, what to, to say
- 3-gram (trigram): I don't know, don't know what, know what to, etc.
- …
- n-gram

- An *n*-gram of size 1 is referred to as a "unigram".
- An *n*-gram of size 2 is a "bigram".
- An *n*-gram of size 3 is a "trigram".
- An n-gram of size 4 or more is simply called an "*n*-gram". Some language model built from ngrams are "(n − 1)-order Markov models".
  - An *n*-gram model is a type of probabilistic model for predicting the next item in such a sequence.

N-GRAM APPROACHES

- Weighted Approach
- Lengths Approach
- Weights with Lengths Approach
- Repetition Approach

Advantages and Disadvantages:

Advantages

- Encode not just keywords, but also word ordering, automatically.
- They are completely dependent on real data.
- Learning features of each affect type is relatively fast and easy.

Disadvantages

- Long range dependencies are not captured.
- Low frequency affects the quality of the n-gram model.

3. 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

3.1 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. 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.

3.2 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.

3.3 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".

3.4 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".

4. PROBLEM DEFINITION-

Word sense disambiguation (WSD) involves the association of a given word in a text or discourse with a definition or meaning which is distinguishable from other meanings potentially attributable to that word.

- The first step is to determine all the different senses for every word relevant to the text or discourse under consideration.
- The second step involves a means to the appropriate sense to each occurrence of a word in context. All disambiguation work involves matching the context of an instance of the word to be disambiguated either with information from external knowledge sources or with contexts of previously disambiguated instances of the word.
- Finally a third step is also involved: the computer needs to learn how to associate a word sense with a word in context using either machine learning or manual creation of rules or metrics. Main focus of our work will be to use the machine learning approaches for WSD. In these approaches, systems are trained to perform the task of word sense disambiguation.

5. 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 rivaling 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%. [1]

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. [2]

Rigau et al. (1997) it correctly states that most WSD algorithms have been developed as stand-alone 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. [3]

Ide et al. (2002) and Tufis et al. (2004) they present a knowledge-based approach which exploits EuroWordNet. Given two aligned words in a parallel corpus, they sense tag them with those synsets of the two words which are mapped through EuroWordNet's interlingual index. The most frequent sense baseline is used as a backoff in case more than one sense of the word in the source language maps to senses of the word in the target language. 75% accuracy is achieved in disambiguating a manually. [4]

"Parallel Texts for Word Sense Disambiguation" Hwee Tou Ng, Bin Wang, and Yee Seng Chan, 2002 has developed the approach to automatically acquire sense-tagged training data *from English-Chinese parallel corpora*, used by English lexical sample (ELS) task. They acquire the sense tagged data. The task of word sense disambiguation (WSD) is to determine the correct meaning, or senses of a word in context.Two approaches were used:

- *Corpus based approach and*
- *Supervised approach*.

In the supervised approach, first collected the corpus in which each word has the correct sense according to the dictionary. The advantage is that it would reduce the performance gap between the two approaches. The accuracy difference between the two approaches is only 14.0%, The main drawback is that, they require the manually sense-tagged data. This problem is particular evere for WSD, since sense-tagged data must be collected separately for each word in a language. [5]

"Word Sense Disamniguation 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 cooccurrence probabilities. [6]

"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. Supervised approaches were used to minimize the potential effects of classification errors on student learning. The Homonyms pane1 has 99.82% accuracy. [7]

"Hindi Word Sense Disambiguation" Manish Sinha Mahesh Kumar Reddy .R Pushpak Bhattacharyya, Prabhakar Pandey Laxmi Kashyap (2004) , They explained that WSD for Hindi words make use of the Wordnet for Hindi developed at IIT Bombay, which is a highly important lexical knowledge base for Hindi. This is the first attempt for an

Indian language at automatic WSD. The main idea is to compare the context of the word in a sentence with the contexts constructed from the Wordnet and chooses the winner. The output of the system is a particular set number designating the sense of the word. The mentioned Wordnet contexts are built from the semantic relations and glosses, using the Application Programming Interface created around the lexical data.The evaluation have been done on the Hindi corpora provided by the Central Institute of Indian Languages and the results are encouraging. Work is on for other parts of speech too. The accuracy is very low and results are not promising.
We use the supervised approach for this disambiguation. That there will be high Overlap between the words in the context and the related words found from the wordnet lexical and semantic relations and glosses.

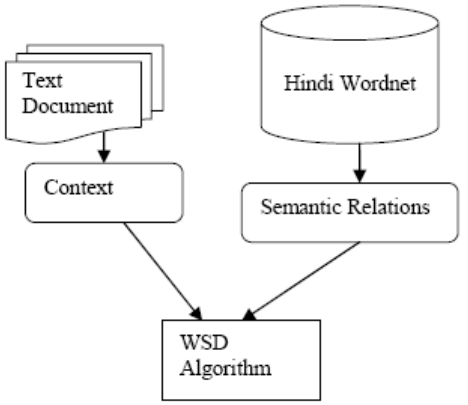

Figure: Extracting semantic relations from Wordnet and building context from the text for WSD. The accuracy value ranges from about 40% to about 70%. The obstacle there is the shallowness of the lexical Network for non-noun words. [8]

"MRD-based Word Sense Disambiguation" Timothy Baldwin, Su Nam Kim, Francis Bond, Sanae Fujita, David Martinez and Takaaki Tanaka, 1986 have developed the Machine Readable Dictionary (MRD), which is applied on the *Japanese Senseval-2*. Japanese is a non-segmenting language. The *unsupervised approach* and *supervised approaches* were used for the knowledge acquisition bottleneck and the Senseval evaluation.
Quiet dog ACC want to keep
"(I) want to keep a quiet dog"
In Japanese, each word has one or more senses. They improve results to a small degree; the best overall results are produced for the weighted combination of all ontological relations. Unsupervised and supervised approaches are used. WSD methods achieve the 0.624 over all the target words (with one target word per sentence). It is compared with an error rate reduction of 21.9% for the best of the WSD systems in the original Senseval-2 task. [9]

6. Conclusion

Word Sense Disambiguation (WSD) refers to the resolution of lexical semantic ambiguity and its goal is to attribute the correct senses to words in a given context. WSD is a most challenging problem in the area of NLP. We have chosen to develop an efficient algorithm for disambiguating ambiguous postpositions present in the Hindi language. In the concepts of Word sense disambiguation, its approaches, its importance and its history has been done. Various Word sense disambiguation approaches have been studied. We are taking this problem with the case study of existing Hindi-Punjabi Machine Translation System. Thus the disambiguation will be done from the machine translation point of view.

7. REFERENCES

[1] Yarowsky, D.1995. "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods," Proceedings of the 33rd annual meeting on Association for Computational Linguistics.
[2] Dekang Lin, 1997. "Using Syntactic Dependency as Local Context to resolve word sense Ambiguity.
[3] Rigau, G., 1997. Combining unsupervised lexical knowledge methods for word sense

disambiguation. Proceedings of the 35th annual meeting on Association for Computational Linguistics.

[4]Ide and Veronis (1998) "Performance Metrics for Word Sense Disambiguation".

[5] Hwee Tou Ng, Bin Wang, and Yee Seng Chan, 2002. "Parallel Texts for Word Sense Disambiguation". In Proceeding of the 2002 Conference on Empirical Methods in Natural Language Processing.

[6] Peter D.Turney. 2003. Coherent key phrase extraction via Web Mining. In Proceeding of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03).

[7] Anagha Kulkarni, Michael Heilman, Maxine Eskenaz and Jamie Callan, 2006. "Word Sense Disambiguation for Vocabulary Learning". Proceedings of the Ninth Proceeding of the 2002 Conference on Empirical Methods.

[8] Manish Sinha, Mahesh Kumar Reddy , Prabhakar Pande, Laxmi Kashyap & Pushpak Bhattacharyya, 2004. "Hindi Word Sense Disambiguatio", International Symposium on machine translation, NLP and Translation support system.

[9] Timothy Baldwin, Su Nam Kim, Francis Bond, Sanae Fujita, David Martinez and Takaaki Tanaka, 1986. "MRD-based Word Sense Disambiguation".