IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 11, Issue 3 (May. - Jun. 2013), PP 78-83
www.iosrjournals.org

Role of Machine Translation and Word Sense Disambiguation in Natural Language Processing

Gurleen Kaur Sidhu[1], Navjot Kaur[2]
[1]*(Department of C S E ,Sri Guru Granth Sahib World University, Fatehgarh sahib, Punjab,India),*
[2]*(Department of C S E ,Sri Guru Granth Sahib World University, Fatehgarh sahib, Punjab,India)*

Abstract : *Natural language is most common way to communicate with each other but sometime we can't understand other languages to understand different languages machine translation is needed. Machine translation is the best application which helps to understand any other language in less cost and less time. In this some problems are faced by researchers: words which pronounce same but having different meaning, some words spelled different but having same meaning, in some cases combination of words may change the meaning. To resolve such kind of problems Word Sense Disambiguation is needed. Word Sense Disambiguation is used to understand the correct meaning of the word with respect to context in which that is used; Word Sense Disambiguation is the part of natural language processing. Word Net plays an important role in Word Sense Disambiguation. In this paper, we will discuss about the basic concept of Natural language processing, Machine Translation and Word sense disambiguation.*

Keywords - *Natural Language Processing, Machine Translation, Word Net, Word Sense Disambiguation.*

I. INTRODUCTION

With the growing world and business people move from one state to another and country to country, Mostly data is computerized. There are a lot of Blogs and websites contains the useful information. If we want to access this information then we face a problem of understanding the text. To sort out this problem the concept of Natural language processing is invented. [1] The Most common way to share your views with single or group of people is natural language. When we relate natural language with computer we need some kind of processing that is natural language processing. [2] Various applications comes under the natural language processing : Automatic Summarization , Machine Translation, Named Entity Reorganization, Optical Character Reorganization, Parsing, Sentence Breaking, Sentiment Analysis, Part of Speech Tagging, Word Sense Disambiguation and so on. Machine translation helps us to automatic translation of text using various methods: Direct Method, Transfer Machine Translation, Interlingua Machine translation, Corpus based, example based statistical machine translation and so on. Word Sense Disambiguation is used to remove the ambiguity of the word with respect to sentence. To remove the ambiguity of the word approaches are divide into knowledge based and corpus based. We will discuss in detail in this paper.

II. NATURAL LANGUAGE PROCESSING

The most common way to share our views with single or group of peoples we use natural language. This communication may be in text form or vocal in the form of dialogues. Today time most of the work done is by the computers. If we access any search engine after the query processing there are so many sites in the search some are useful but we are not able to understand due to language problem. There should be need to translate or encrypt the data. Natural Language processing is field of data mining, artificial intelligence and computer science. [2] NLP applications range from querying archives, to accessing collections of texts and extracting information, to report generation, to machine translation.

2.1 Automatic Summarization *(A S): Produce summaries of text of a known type.*

2.2 Machine translation(MT)*:* Automatically translate text from one human language to another.

2.3 Named entity recognition (NER): Determine which items in the text map to proper names, such as people or places, and what the type of each such name is (e.g. person, location, organization).

2.4 Natural language generation(NLG): Convert computer databases into readable human language.

2.5 Optical character recognition (OCR): An image representing printed text,determine corresponding text.

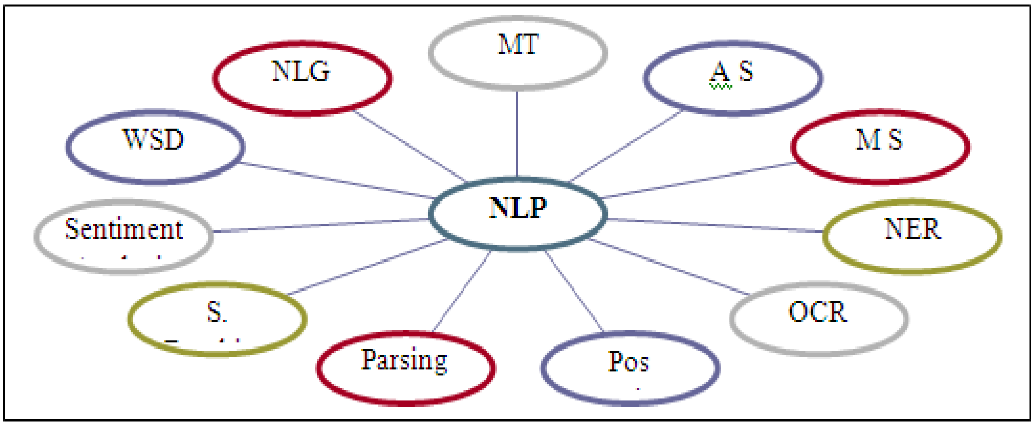


Fig.2.1 Major task in Natural Language processing

2.6 Part-of-speech tagging (Pos tagging): Many words, especially common ones, can serve as multiple parts of speech.

2.7 Parsing: The grammar for natural languages is ambiguous and typical sentences have multiple possible analyses. For a typical sentence there may be thousands of potential parse.

2.8 Sentence breaking (S.breaking) :Find the sentence boundaries. Sentence boundaries are often marked by periods or other punctuation marks, but these same characters can serve other purposes.

2.9 Sentiment analysis: Extract subjective information usually from a set of documents, for the purpose of marketing.

2.10 Word sense disambiguation: Many words have more than one meaning; we have to select the meaning which makes the most sense in context.

III. Machine Translation

MT (Machine Translation) is the use of computers to automate the production of translations from one natural language into another, with or without Human assistance [3]. Machine translation System is used to translate the source text into target text. MT system uses the various approaches to complete the translation. Accurate translated text is that which have two basic properties: adequacy and fluency.

Machine translation is considered as difficult task on the other hand Translators are expensive and time consuming. The source and target languages are natural languages such as English and Hindi, as opposed to man-made languages such as C or SQL.

MT system contains components for analysis, transfer and generation as shown in the fig 3.1. These components incorporate a lot of knowledge about words (lexical knowledge), and about the language (linguistic knowledge). Such knowledge is stored in one or more lexicons, and possibly other sources of linguistic knowledge, such as grammar. The lexicon is an important component of any MT system.

A lexicon contains all the relevant information about words and phrases that is required for the various levels of analysis and generation. A typical lexicon entry for a word would contain the following information about the word: the part of speech, the morphological variants, the expectations of the word some kind of semantic or sense information about the word, and information about the equivalent of the word in the target language [4].

3.1 Approaches of Machine Translation : The MT system uses various approaches to complete the process of translation.

3.1.1 Direct Machine Translation: Rule-based direct mapping is done in this phase. Procedure is as shown below:

• English Sentence	I will go to the market
• Morphological Analysis	I go FUTURE to market
• Constituent Identification	< I > < go FUTURE > < to market >
• Reorder	< I > < to market > < go FUTURE >
• Dictionary Lookup	<मैं> <बाज़ार> <जाना FUTURE>
• Inflect	मैं बाज़ार जाऊँगा

Fig.3.A.1 Direct Machine translation [5]

As shown in fig 3.A.1 we enter English sentence as input, then morphological Analysis is implemented which tell use the given talk belongs to which class, and then pre-processing is done. Then we change the order of words according to language rules. At last maps with database & direct translated sentence is generated.

3.1.2 Transfer Machine Translation: In Transfer Machine Translation phase it involves three stages to achieve the target language. Procedure is as shown below:

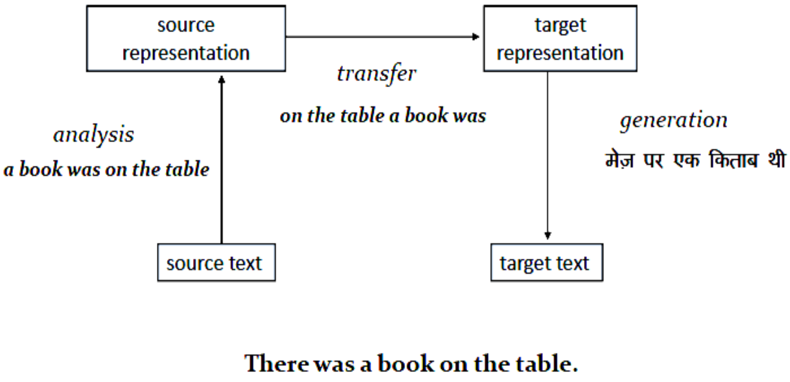


Fig.3.A.2 Transfer Machine Translation [5]

As shown in fig3.A.2 we enter the source text, analysis the text, shows the text representation according to source text language rules. Tokenization of text into S= subject, O = object and v= verb, transfer the representation of text according to the target language rules. Than map with database and will show the result.

3.1.3 Interlingua machine : intermediate is used to get the Target text from source text

3.1.4 Statistical Machine Translation: Translation is done by using the statistical models, in which target language is e & source language is f, then whose p(e/f) is maximum[9]

3.1.5 Example-Based Machine Translation *:* First developed in contrast with rule-based machine translation, translate adapting by examples rather than by linguistic rules and a sentence is translated by using the closest match in parallel data

3.2 Difficulties in Machine Translation[5]:

3.2.1 Lexical Ambiguity: Input words are spelled same but having different meanings.

1. ਯੋਗ ਕੁੜੀ ਦੀ ਭਾਲ ਨਾਸਾ.   2. ਯੋਗ ਆਸਣਾਂ ਨਾਲ ਸਿਹਤ ਠੀਕ ਰਹਿੰਦੀ ਹੈ

3.2.2 Differing word orders: SUBJECT VERB OBJECT or SUBJECT OBJECT VERB– Ram saw Sham or ਰਾਮ ਨੇ ਸ਼ਾਮ ਵੇਖਿਆ |

3.2.3 Syntactic Structure is not Preserved Across Translations

3.2.4 Syntactic Ambiguity Causes Problems

3.2.5 Pronoun Resolution –

ਰਾਜੇ ਨੇ ਬੀਰਬਲ ਨੂੰ ਚੁਣਿਆ.ਉਹ ਹੁਸ਼ਿਆਰ ਹੈ|

3.2.6 Boundary Friction

IV. WORD SENSE DISAMBIGUATION

WSD (Word Sense Disambiguation) is the use of computers to remove the problem of ambiguity of words in natural language. Word Sense Ambiguity is present the characteristic of Natural Language (NL). For Example we have a word "Tank" with multiple meanings; it may refer to a container, a vehicle. The actual sense of the particular word is defined by the textual context in which that particular word is used. In "I saw a military tank" the vehicle sense is intended, while in "The tank was full of water" the container sense is meant. Few word to understand the concept of ambiguity.

Ambiguous words		
Sr. No.	Word	Meaning
1	Cold	Disease, Temperature
2	Palm	Tree , Hand

3	Bank	Financial , edge of river
4	Crane	Bird , Machine
5	Lead	Be in front, a type of metal
6	Consult	Give advice, To take advice
7	Plant	Living , Factory

Table 1: Some Common Examples of Ambiguous words

There are many words in dictionary which have many meaning, it is important for computer to correctly determine the correct meaning of the word in context. Word Sense Disambiguation (WSD) is the resolution of lexical semantic ambiguity and its correct senses to words in a given context.

4.1 Application of Word Sense Disambiguation

4.1.1 Machine Translation: The most common application of WSD is Machine Translation (MT). In machine translation, process completed at least in two steps: understanding the source language and translate it into target language. In both cases WSD plays an important role because source as well as target language words have the ambiguous property. For example an English word consult translate into Punjabi than we have multiple meanings as shown below Rule-based direct mapping is done in this phase. Procedure is as shown below:

Word	**First meaning**	**Second Meaning**
Consult (English)	Give Advice	To take advice
(Punjabi)	ਮਸ਼ਵਰਾ ਕਰਨਾ	ਸਲਾਹ ਲੈਣੀ

Example: machine translation English to Punjabi with ambiguous word

4.1.2 Information Retrieval: In information retrieval queries are used to extract the information from any big databases. Sometimes ambiguous words in queries are problematic. Hence, WSD filters are necessary for retrieval engines.

4.1.3 Speech Processing: To generate the speech same as the natural sound we determine the correct pronunciations of words. In this process some words sounds same but the meanings are different for example the word "Lead" is "a metal" or "be in front". Some words are spelled differently but sound same.

4.1.4 Text Processing: Word sense disambiguation is necessary for spelling correction in lexical access of Semitic languages. For example we have a sentence "HE VISIT BATHINDA" ➔ "He visit Bathinda" [6]

V. APPROACHES OF WSD

To use all the previously defined applications, the various approaches are followed by the researchers, teachers, and students. These approaches are classified on the bases of information acquire from different resources.

5.1 Corpus Based Approach

The MT The acquired information is arranged in the form of samples. Set of these samples helps systems to develop the numerical model. These sample sets known as Corpus. This approach is classified into two subclasses as shown in fig 5.1:

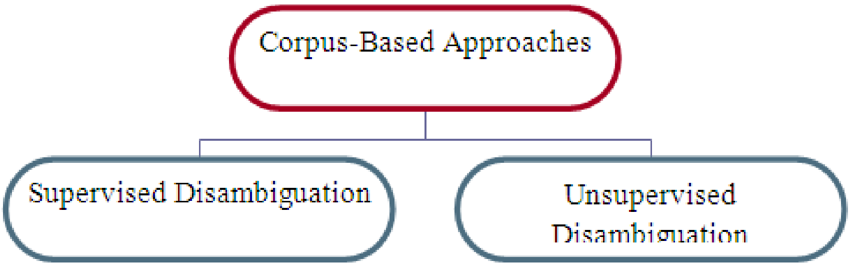
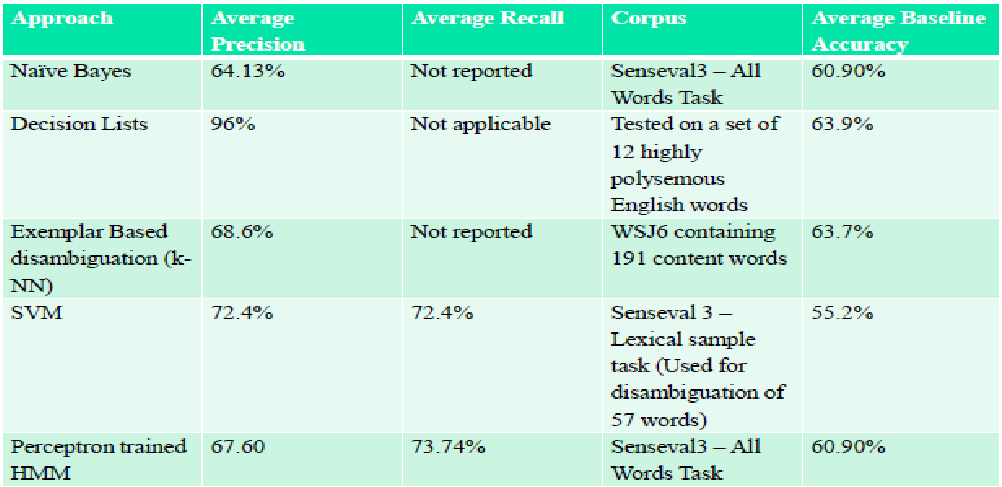


Fig.5.A.1 Classification of Corpus-Based Approach.

5.1.1 Supervised Disambiguation: Supervised disambiguation is an application of supervised learning for creating a classifier which correctly classifies new cases based on their context of use. Decision list and Bayesian classifier are the most popular algorithms in supervised disambiguation.A major problem with this approach is large sense tagged training set. Due to large corpora, manually sense-tagging is very difficult and very few sense-tagged data are available. The largest corpora that are available are SemCor corpus and SENSEVAL corpus. The SemCor corpus is a subset of the English Brown corpus containing more than 700,000 running words. all the words are tagged by part of speech and more than 200,000 content words are also lemmatized and sens-tagged. SENSEVAL corpus is derived from the HECTOR corpus and dictionary project:



Approach	Average Precision	Average Recall	Corpus	Average Baseline Accuracy
Naïve Bayes	64.13%	Not reported	Senseval3 – All Words Task	60.90%
Decision Lists	96%	Not applicable	Tested on a set of 12 highly polysemous English words	63.9%
Exemplar Based disambiguation (k-NN)	68.6%	Not reported	WSJ6 containing 191 content words	63.7%
SVM	72.4%	72.4%	Senseval 3 – Lexical sample task (Used for disambiguation of 57 words)	55.2%
Perceptron trained HMM	67.60	73.74%	Senseval3 – All Words Task	60.90%

Fig.5.1.A.1 Comparison between the various algorithms of Supervised disambiguation [7]

5.1.2 Unsupervised Disambiguation: In this information is gathered from raw corpora which have not semantically disambiguated. Unsupervised disambiguation is not possible for word sense since sense tagging requires characterization of the senses. WSD can be divided into two sub-problems: sense discrimination and sense labelling. Sense labelling: it is a part of task in this necessary to define the sense of outside source of knowledge. Sense disambiguation is completely done in unsupervised way. It divides the word into number of classes whether they belong to same sense or not.

5.2 Knowledge-based approach: Knowledge-based methods use the lexical and semantic knowledge bases such as machine-readable dictionaries (MRDs), thesauri, computational lexicons. The efforts to automatically create knowledge bases, Word Net, the most widely used one, are created by hand.

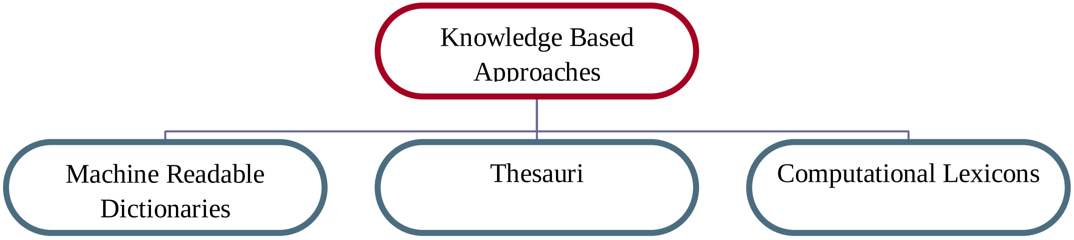
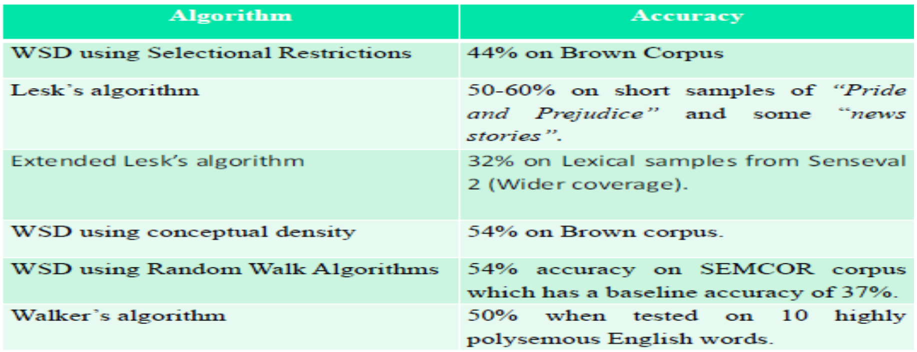


Fig.5.B.1 Approaches of Knowledgebased approach

5.2.1 Thesauri: it shows the relationship between among the words. It helps in semantic categorization by a thesaurus or a dictionary with subject categories. Thesauri is also not much affective same as MRD, because thesaurus is also a resource for human use.

5.2.2 Machine Readable Dictionaries: Lesk (1986) use Machine Readable dictionaries first time because the simple idea that a words dictionary definitions are likely to be good indicators of the senses they define. By using the Oxford Advanced Learners dictionary (OALD), he counts the number of overlaps in sense definition of ambiguous words & in definitions of context words occurring nearby. After that various researchers apply different methods to improve Lesk algorithm: By optimizing the overlap of all words in a single sentence, simulated annealing, by normalizing the contribution of a word of the overlap count, glosses contained in Word Net. There are some inconsistencies remains because these dictionaries are created for human use not for computers.

5.2.3 Computational Lexicons: the usefulness of lexical relations in linguistic, psycholinguistic and computational research gives the idea of electronic databases of such relations. Since the Word Net is the most popular lexicon.



Algorithm	Accuracy
WSD using Selectional Restrictions	44% on Brown Corpus
Lesk's algorithm	50-60% on short samples of *"Pride and Prejudice"* and some *"news stories"*.
Extended Lesk's algorithm	32% on Lexical samples from Senseval 2 (Wider coverage).
WSD using conceptual density	54% on Brown corpus.
WSD using Random Walk Algorithms	54% accuracy on SEMCOR corpus which has a baseline accuracy of 37%.
Walker's algorithm	50% when tested on 10 highly polysemous English words.

Fig.V.2.1 comparisons between the various algorithms of Knowledge-based Approach [8]

VI. CONCLUSION

This paper concludes that Word Sense disambiguation is helpful in machine translation to understand the actual meaning of the word with respect to text. And accuracy in machine translation met the natural language processing target. In this we discuss important approaches of word sense disambiguation and machine translation. In future work we try to make a new approach which inherits the best functionalities of mostly used and accurate algorithms.

REFERENCES

Journal Papers:

[1] M Ozaki, Y. Adachi, Y. Iwahori, Fabio Ciravegna, Sanda Harabagiu , IEEE Computer Society 2003. Recent Advances in Natural Language Processing.

[2] Available: http://en.wikipedia.org/wiki/Natural_language_processing

[3] J. Hutchins and H. Somers. An introduction to Machine Translation. Academic Press, 1992.

[4] Durgesh D Rao, "Machine Translation",pp.61-70, July1998Available: www.ias.ac.in/resonance/**July1998**/pdf/**July1998**p61-70.pdf

[5] Available:http://www.cse.iitb.ac.in/~cs626-460-2012/lecture_slides/cs626-460-lect9-MT-2012-1-19.pdf

[6] Nancy Ide, Jean Veronis "Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art", 1998J.

[7] Pushpak Bhattacharyya, "CS460/626 : Natural LanguageProcessing/Speech, NLP and the Web (Lecture 25– Knowledge Based andSupervised WSD)", *IIT Bombay*, 6th March, 2012, p.24.

[8] Pushpak Bhattacharyya, "CS460/626 : Natural LanguageProcessing/Speech, NLP and the Web (Lecture 25– Knowledge Based andSupervised WSD)", *IIT Bombay*, 6th March, 2012, p.36

Theses:

[9] R.Harshawardhan, Rule Based Machine Translation System For English To Malayalam Language, Centre for Excellence in Computational Engineering and Networking, December 2011.