International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2020 IJSRCSEIT | Volume 6 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT2062135

Ambiguity Resolution : An Analytical Study

Prashant Y. Itankar[1], Dr. Nikhat Raza[2]
[1]PhD Scholar, Department of Computer Science and Engineering, MPU Bhopal, Madhya Pradesh, India
[2]Associate Professor Department of Computer Science and Engineering, MPU Bhopal, Madhya Pradesh, India

ABSTRACT

Natural language processing (NLP) is very much needed in today's world to enhance human-machine interaction. It is an important concern to process textual data and obtain useful and meaningful information from these texts. NLP parses the texts and provides information to machine for further processing. The present status of NLP's computational process of identifying the meaning (sense) of a word in a particular context is ambiguous, where the meaning of word in the context is not clear and may point to multiple senses. Ambiguity in understanding correct meaning of texts is hampering the growth and development in various fields of Natural language processing applications like Machine translation, Human Machine interface etc. The process of finding the correct meaning of the ambiguous texts in the given context is called as word sense disambiguation (WSD).

WSD is perceived as one of the most challenging problem in the Natural language processing community and is still unsolved. It is evident that different ambiguities exist in natural languages and researchers are contributing to resolve the problem in different languages for successful disambiguation. These ambiguities must be resolved in order to understand the meaning of the text and help to boost NLP processing and applications. Objective is to investigate how WSD can be used to alleviate ambiguities, automatically determine the correct meaning of the ambiguous text and help to boost NLP processing and applications.

Resolving ambiguity for translation involves working with various natural language processing techniques to investigate the structure of the languages, availability of lexical resources etc. Word Sense Disambiguation (WSD) in the field of computing linguistics is an area which is still unsolved. This paper focus on the in-depth analysis of such ambiguity, issues in Language Translation, how WSD resolves the ambiguity and contribute towards building a framework.

Keywords : Natural Language, Ambiguity, Word Sense Disambiguation.

I. INTRODUCTION

Language is an efficient medium of communication which represents ideas and expressions of human mind. Various languages exist in the world which depicts linguistics diversity making it difficult for the humans to understand each language. This depicts the need for multilingual translation which is ever growing. [6] This along with modern technology made way for the generation of modern machine translation for lingual framework.

Translation involves converting the text from one language to another language using computer system. In translation, it is evident that different ambiguities exist in natural languages and researchers are contributing to resolve the problem in different languages. The present status of natural language processing computational way of selecting the accurate sense of a word in a particular context is ambiguous, where unclear meaning of the words may point to multiple senses. The situation becomes more ambiguous if considered for any source language to any target language translation.

II. REVIEW OF LITERATURE

A. Machine translation

Survey on Translation systems by Garje and Kharate [6] beautifully presents a brief history of translation systems from the late forties to recent systems. Various approaches for developing machine translation systems include direct MT, statistical MT, transfer based, inter-lingua, example based and hybrid MT achieving satisfactory results. Work by Firal et.al in [7] focuses on multilingual machine translation which enables a single neural translation model to translate between multiple languages, with a number of parameters that grows only linearly with the number of languages. The model was evaluated on a large set of parallel corpora with improved results.

B. WSD

Navigli R. in [8] has beautifully presented a survey of WSD highlighting the motivation for solving the WSD.
1. Knowledge based approach: Lesk's algorithm also called as overlap approach was the first attempt for detecting the accurate meaning of the word using the available dictionary. Context words and dictionary definitions were matched to find the overlap and determine the result. Lack of strong clues resulted in average accuracy of 50-60% on short samples and news stories. Proper nouns if present in the context are not part of dictionary which also leads to poor accuracy [10].

Banerjee and Pederson in [11] modified the original Lesk algorithm to include richer knowledge base called Wordnet instead of relying on standard dictionary. The overall accuracy observed was 31.7%. The advantage was the increased accuracy as compared to Lesk algorithm as wordnet provides rich ontological information such as synset, hypernyms, hyponyms, antonyms etc.

The concept of selection preference or restrictions is best described by Philip Resnik in [12] where the focus is on the grammatical structure of the context. Results achieved were 44% when tested on Brown corpus. Exhaustive knowledge base required. Conceptual density approach using wordnet proposed by Agirre and Rigau[13] selects a sense based on the close proximity of the word sense with the input. Requirement of labeled corpus is eliminated. When tested on sense tagged version of brown corpus, the observed precision was 47.3%.Fu in [14] proposes building a semantic tree by arranging he words into vector. F-score achieved for this approach is 73.74%. Large amount of raw data is needed to convert the words into vectors for better accuracy. Resnik in [15] captures close proximity using information content of concepts.

Very little work is reported on Indian languages to the best of our knowledge which is summarized in table I. Achieving satisfactory accuracy; this approach has the advantage of not requiring tagged or raw corpus for disambiguation. Knowledge base and external resources leads to successful sense disambiguation.
2. Supervised approach: This approach uses the sense annotated corpus for performing the disambiguation process. Machine is trained using the annotated corpus and classes are formed to assign accurate sense to each designed. Sense tagged or annotated data is a time consuming process. For different application or language which requires different sense distinctions, this solution is infeasible due to high annotation cost. Some technique is desired to make use of sense annotated data in one language to be reused in another language [6].Recent research on supervised approach

in [16] compares two supervised approaches namely naive bayes and decision tree algorithm. Data set of 10 nouns and 5 verbs were used. The overall accuracy observed for Naive Bayes classifier is 62.86% and for decision tree is 45.14%. Naïve bayes gives better performance as it trains on small data whereas decision tree suffers from problem of over fitting of data.

[17] Describes the use of decision tree for performing disambiguation. It uses the DSO corpus and Semcor and results show that Semcor can be acceptable at 0.7 precision for polysemy words. A comparison between Naive Bayes and Exempler based approach for performing disambiguation is discussed in [18]. Experiments were performed on DSO corpus on a dataset of 15 words. The accuracy observed for Naive Bayes classifier is 66.4% and for Exempler based approach, the accuracy observed is around 60%.

Very limited work is reported on Indian languages using the supervised approach as shown in table II for Indian languages. Results are encouraging due to the presence of tagged corpus. But if a lingual framework is designed, creating such a huge corpus for multiple languages is too expensive and time consuming.

3.Unsupervised approach: Due to the unavailability of learning resource or tagged data, unsupervised approach is a challenging approach. Chaplot , Bhattacharya and Paranjape in describes unsupervised approach to WSD using sense dependency and selective dependency. They focused on calculating and maximizing joint probability for all the senses reducing the sense drift problem. Navigli and Lapata in [20] describe graph based methods for unsupervised word sense disambiguation. Various features were considered for analyzing the graph to determine the accurate meaning. Graph based methods are suitable for unsupervised approach as it gives better representation of senses in the ontology.

Table I: Knowledge Based Approach For Indian Language

Title	Dataset	Result
Role of semantic relations in Hindi WSD. [26]	60 Polysemy nouns 7506 instances	Precision- 56% Recall- 51%
Correlation based WSD [27]	60 Polysemy nouns 1824 instances	Precision- 88.92%
Hindi WSD using semantic related measures. [28]	20 polysemy nouns 710 instances	Accuracy 60.65%
Measuring context meaning for open class words in Hindi language. [29]	500 sentences	60.25% using node neighbor Connectivity 41.25% using graph clustering
Evaluating effect of context window size, stemming and stop word removal on Hindi WSD. [30]	10 polysemy nouns	Accuracy 54.81%
Hindi Word Sense Disambiguation [31]	8 text files from hindi corpus	Accuracy 40-70%

Table II: Supervised Approach For Indian Language

Title	Dataset	Result
Role of karaka relations in Hindi WSD. [32]	60 polysemy nouns. 7506 instances	56.56% precision
Disambiguating Hindi words using N grams smoothing models. [33]	10 polysemy words	60-70% with deleted interpolation and 50-60% with back off method
A Supervised Algorithm for Hindi WSD. [34]	60 Polysemy nouns 7506 instances	Precision- 78.98% Recall- 73.41%

The most famous algorithm depicting unsupervised approach was proposed by Yarowsky in [21]. The algorithm is based on one sense per discourse and one sense per collocation - exploited in an iterative bootstrapping procedure. If a polysemy ambiguous word is used more than two times in a discourse, it tends to have same meaning. Similar words grouped together using the information content identified by syntactic dependencies was discussed in [22].

Very little work is reported on Indian languages using the unsupervised approach for performing disambiguation. Hindi WSD was performed using the unsupervised approach based on the concept of network agglomeration [23]. A sentence graph is created for a given sentence. From the sentence graph, interpretation graph is generated for each of the interpretation of the sentence. Network agglomeration is computed to identify the desired interpretation. When tested on health and tourism corpus, this method yields an average accuracy of around 52%.

In [24], most frequent sense detection was performed using word embeddings. The approach was tested on Hindi and English language WSD. F1 score of 62.5 was reported for Hindi data set and 52.34% and 43.28% for English SENSEVAL-2 and SENSEVAL-3 dataset. Bilingual WSD work is reported in [25] using the contextual information. Expected maximization formulization was modified using context and semantic relatedness of neighboring words. An improvement in the accuracy of 17-35% for verbs is reported compared to the existing expected minimization approach.

Unsupervised WSD has the biggest advantage of working with the raw corpus. But designing a multilingual framework requires the corpus in huge amount for multiple languages with if not available needs to be created which is a time consuming and costly process.

With the literature review on machine translation, it is evident that if ambiguity is fully resolved it can lead to successful machine translation. Before processing the source text for translation, machine must acquire knowledge from it. Knowledge acquisition will add new learning capability to the machine and from the acquired knowledge the machines will intelligently make new interpretations for successful translation from one language to another. Various researchers have attempted to resolve the ambiguity for polysemy words for English languages. Very little work is reported on ambiguity resolution for Indian languages.

III. AMBIGUITIES

1. Wordnet ambiguity

Wordnet is a large lexical knowledge base where nouns, verb, adverb and adjectives are grouped into synonym set called synset and it is semantically connected to other synset using semantic relations. Scarcity of connection between connected components can make the disambiguation task difficult. If the components are strongly connected, the accuracy of the disambiguation process can rise.

Another problem is the granularity of senses. Senses of a word are categorized into two parts: Fine grained and coarse grained senses. Fine grained senses make the disambiguation task more difficult as can be seen from the above example that Bank has been finely split into both the above senses referring to the financial institution. For the coarse grained sense of bank, these two senses can be merged together to get more general meaning of the bank referring it as financial institution.

B. Input ambiguity

Input consists of ambiguous word and other words surrounding the ambiguous words that are visible. These words are the clue words or unique features present in the context which help in the disambiguation process. Context size and its associated features extracted can affect the results of WSD. Various types of ambiguities are seen in the context representations which are discussed below.

Lexical Ambiguity: It means a word can be a noun, verb, adjective.

Example:
“She received three silver vessels.”
“Leena gave a silver talk”

Lexical Semantic Ambiguity: single ambiguous word in the context having multiple meanings.
“The crane is loaded.”
“The beak of the crane is very big.”

The ambiguous word “crane” represents two meanings one in the lifting sense and another in the bird sense. To resolve the ambiguities arising from the polysemy words, machine needs the context, world knowledge etc.

Syntactic Ambiguity: Ambiguity arises from the way the sentence is put together.

“He saw a man with binoculars.”

Here it is not clear whether the man was wearing binoculars or he was using binoculars. Hence the meaning of sentence changes.

Discourse level ambiguity: High level of world knowledge is required for interpretation.

Anaphoric Ambiguity: Giving emphasis to certain words. Example:

“Cat went up the hill. It was slippery. It got angry”
The anaphoric reference of “it” in the two situations cause ambiguity.
Pronoun ambiguity: Pronoun usually points to noun which precedes it.
“Alan likes to play cricket. He is a good player.” (“He” points to Alan)
“Lisi is going to office. She has some work.” (“She” points to Lisi)

“Seeta and Geeta are sisters. They both resemble the same.” (“They” points to “Seeta” and “Geeta”).
But presence or two or more nouns makes the disambiguation of pronoun a difficult task especially the pronoun “it”.
“Monkey ate the banana as it was hungry.” In this case “it” points to Monkey due to the presence of the verb “hungry”.
When it comes to processing the above words, humans does not have severe problem because of the world knowledge or the common sense knowledge possessed by them. But same is not the case with the computers. From the machine processing perspective, it is a difficult task as machine needs to process the ambiguity in pronoun.

IV. ANALYSIS AND OBSERVATIONS

Research in MT has progressed to a point where system translates the information from source to target language. With worldwide information available in lingual contexts, MT can be an efficient tool for facilitation and direct transfer of such comprehensive information between people with language diversities throughout the world.

Many factors contribute to the difficulty in performing successful MT. Several problems unnoticed by human translators are encountered by machine thus making the MT task more challenging. It is evident that for successful translation, MT systems require both linguistics and world knowledge about the source and target language. As well as target language should be known to the machine. Linguistics knowledge includes the morphological, syntactic and semantic category information. Morphological knowledge tells how certain words are derived from their root words. Knowledge of the source text is important to understand the meaning of the words in the context intended for translation. Without proper knowledge of target language, the system may produce some meaningless or unacceptable output.

Lexical semantic ambiguity if occurs while translation can change the meaning of the ambiguous words in the context of target language text. Unambiguous words in the source language may have several possible meanings in the target language depending upon the context. Also ambiguous word in the source language may be unambiguous in the target language. In both the cases, the most important thing is to keep the original meaning. A perfect understanding of the source text is essential which MT lacks as they don't have the global vision of acquiring the world knowledge about it. To let the machine decide the correct interpretation, we have to provide the accurate meaning of the senses.

When a sentence contains two or more words in the context with multiple meanings, ambiguity tends to multiply. Analysis of the word in the context depends on how ambiguous the word is. If context fails to provide sufficient clues for disambiguation, WSD algorithm fails. In this case, plain raw corpus is used for providing sufficient clues for disambiguation process.

Sentences with multiple grammatical structures also pose a problem in MT. Consider the word 'silver' assigned to more than one syntactic category. It is both an adjective as well as verb in the examples mentioned in the lexical ambiguity. Rules are fed to the machine to determine the correct POS category. Based on the rules applied, it may lead to different syntactic analysis of the sentence. Feeding the machine with rules without telling the actual meaning of the category may lead to incorrect translation. These different analyses may lead to different translation. Selecting from these syntactic analyses requires the knowledge about the correct meaning.

Syntactic ambiguity also occurs with phrases where one sentence has more than one possible structure. It is attached to more than one position in a sentence. In the example described in syntactic ambiguity it is not clear whether the man was wearing binoculars or he was using binoculars. If the correct meaning of the phrase 'with the binoculars' is associated with its intended subject or object, the translation task becomes more accurate.

Pronoun translation is a difficult task as it is uncertain about what a pronoun refers to. The rule of thumb says that pronoun generally refers to its closest antecedent which is a noun the pronoun replaces. But if there is more than one antecedent, then difficulty arises in knowing the actual meaning of the pronoun. Real world knowledge is essential in pronoun disambiguation. Take for example the pronoun 'it'. It normally points to three entities namely living, non-living and facts/events. Living entities are again categorized into animate and humans. Each of these categories, when represented in any language has gender associated with it to which the pronoun is associated with. For example “It is a chair” where the pronoun “it” refers to noun “chair”. “Chair” has a neutral gender in English language but when translated in target language say ‘Marathi’, it occupies feminine gender. Here world knowledge is essential to determine which gender of each particular noun is suited for the translation. Consider the following two translations

“It is a banana” (English) “हे केळ आहे” (Marathi)

“These are the bananas” (English) “ही केळी आहेत” (Marathi)

When the noun “banana” (singular) gets translated into Marathi language in the first sentence, it has neutral gender. When the same noun is translated in its plural form, it becomes feminine in the target language. If we replace the noun banana with mango/mangoes, following translation occurs.

“हा आंबा आहे”.

“हे आंबे आहेत”.

The pronoun “it” in the source language is changed to “हे” ’for “banana” and “हा”’ for mango. Similarly pronoun “these” in the source language is changed to “ही” for “banana” and “हे” for “mangoes” in the target language. If the machine acquires this knowledge and constructs its knowledge base, disambiguation becomes an easy task and will result in successful MT.

Ambiguity resolution in a bilingual framework is a difficult task due to the difficulties associated with the words, phrases and parts of speech category. Ambiguity with the source as well as target language needs to be understood by the machine to perform successful machine translation. These ambiguities in a bilingual framework tend to multiply when we encounter more than one ambiguous word in the context or more than one type of ambiguity in the context or both. These ambiguities affect the translation scenario for target language resulting in unacceptable output.

These ambiguities will tend to increase by a huge margin in multilingual translation. The kind of analysis which needs to be done for multilingual translation will be number of languages multiplied by the ambiguities associated with each language. This problem needs to be resolved in such a way that ambiguities tend to decrease and boost the framework for multilingual translation.WSD play an effective role in the translation task for successful machine translation. Resolving ambiguity requires extraction of correct senses of the word. Single sentence input lack sufficient clues to perform disambiguation. Hence discourse level context is needed for the machine to learn. Resolving the ambiguity will lead to building a framework for multilingual translation where information can be directly translated from any source language to any target language.

V. CONCLUSION

The observations presented here argue that it will be difficult task to use the existing WSD algorithms to obtain substantial improvements in the lingual translation process. Also working in a multilingual scenario requires huge parallel corpus for different languages obtaining which is a challenging task. Another interesting observation is that supervised approach is a challenging task to work in a lingual setting due to the limitations of the tagged corpus. WSD in lingual translation being a difficult task and various factors contribute to the difficulty in translation including ambiguity, very small amount of work is reported on WSD in lingual translation to the best of our knowledge. Lingual MT can be an efficient.

VI. REFERENCES

[1]. Sharma P., Joshi N.; " Knowledge based Method for word sense disambiguation by using Hindi wordnet", Engineering, Technology and Applied science research, Volume 9, No. 02,2019.

[2]. Salodkar A., Nagwanshi M., Gopchandani B.; "Supervised Approach to word sense disambiguation", published semantic scholar, 2019.

[3]. Saif A., Omar N., Zainodin U., Aziz J.; "Building sense tagged corpususingwikipedia for supervised word sense disambiguation", Elsevier, 2018

[4]. Marvin R., Koehn P.; "Exploring word sense disambiguation abilities of Neural Machine translation systems", Proceedings of AMTA 2018, Vol 1, pp125-131.

[5]. Chaplot Devendra Singh , Ruslan Salakhutdinov, "Knowledge- based Word Sense Disambiguation using Topic Models", Association for the Advancement of Artificial Intelligence (www.aaai.org), 2018,

[6]. Garje G., Kharate G., " Survey of machine translation systems in India. "International Journal on Natural Language computing. Vol. 02, No.4, October 2013.

[7]. Firal O, Cho K., Bengio Y; " Multiway multilingual neural machine translation with a shared attention mechanism",Proceedings of the NAACL-HLT 2016, pp. 866-875.

[8]. Navigli Roberto; "Word sense disambigunation: A survey", Computing surveys, Vol. 34, No. 2, Article 10, 2009.

[9]. Pal A., Saha D.; " Word sense disambiguation: A survey", International Journal of Control theory and computer modelling, Vol. 5, No. 3,

[10]. Lesk Michael; "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an icecream cone", Procesdings of the 5th Annual International conference on Systems Documentation, NewYork, pp. 24-26, 1986.

[11]. Banerjee Satanjeev and Ted Pedersen; "An adapted Lesk algorithm for word sense disambiguation using WordNet." Computational Linguistics and intelligent text processing. Springer Berlin Heidelberg, 2002. 136-145.

[12]. Resnik Philip Stuart; "Selection and information: a class-based approach to lexical relationships." IRCS Technical Reports Series, 1993.

[13]. Agirre Eneko and German Rigau; " A proposal for word sense disambiguation using conceptual distance" arXiv cmplg/9510003.1995.

[14]. Fu Ruiji, "Learning semantic hierarchies: A continuous vector space approach." IEEE Transactions on Audio, Speech, and Language Processing 23.3, 2015 pp. 461-471.

[15]. Resnik Philip; "Using information content to evaluate semantic similarity in a taxonomy." arXiv preprint cmp-lg/9511007, 1995.

[16]. Al-Bayaty, Boshra F. Zopon, and Shashank Joshi; "Comparative Analysis between Naïve Bayes Algorithm and Decision Tree to Solve WSD Using Empirical Approach." Lecture Notes on Software Engineering 4.1, 2016.

[17]. Agirre Eneko, and David Martinez; "Exploring automatic word sense disambiguation with decision lists and the Web." Proceedings of the COLING-2000 Workshop on Semantic Annotation and Intelligent Content. Association for Computational Linguistics, 2000.

[18]. Escudero, Gerard, Lluís Màrquez and German Rigau; "Naive Bayes and exemplar-based approaches to word sense disambiguation revisited." arXiv preprint cs/0007011, 2000.

[19]. Chaplot Devendra Singh, Pushpak Bhattacharyya and Ashwin Paranjape;

"Unsupervised Word Sense Disambiguation Using Markov Random Field and Dependency Parser." In AAI, pp. 2217-2223, 2015.

[20]. Navigli Roberto and Mirella Lapata; "Graph Connectivity Measures for Unsupervised Word Sense Disambiguation." IJCAI, 2007.

[21]. Yarowsky David; "Unsupervised WSD rivaling supervised methods."Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, Massachusetts Institute of Technology, Cambridge, MA., 1995.

[22]. Lin Dekang; "Automatic retrieval and clustering of similar words."Proceedings of the 17th international conference on Computational linguistics-Volume 2. Association for Computational Linguistics, 1998.

[23]. Jain Amita and D. K. Lobiyal; "Unsupervised Hindi word sense disambiguation based on network agglomeration." Computing for Sustainable Global Development (INDIACom), 2nd International Conference on. IEEE, 2015.

[24]. RudraMurthy V, Sudha Bhingardive Dhirendra Singh, Hanumant Redkar, an Pushpak Bhattacharyya;"Unsupervised Most Frequent Sense Detection using Word Embeddings.", ACL, 2015.

[25]. Bhingardive Sudha, Samiulla Shaikh and Pushpak Bhattacharyya; "Neighbors Help: Bilingual Unsupervised WSD Using Context." ACL,2013.

[26]. Singh Satyendr and Tanveer J. Siddiqui; "Role of Semantic Relations in Hindi Word Sense Disambiguation." Procedia Computer Science 46 , 2015 pp. 240-248.

[27]. Agarwal Mohini and Jyoti Bajpai; “Correction based Word Sense Disambiguation” Contemporary Computing (IC3), 2014 7th International Conference on IEEE,2014.

[28]. Singh S., Singh V, Siddhiqui T., “Hindi WSD using semantic related measures”, Springer, 2013, pp. 247-256.

[29]. Jain Abhishek , Suneel Yadav and Devendra Tayak; “ Measuring context-meaning for open class words in Hindi language"Contemporary Computing (IC3), 2013 6th International Conference on IEEE, 2013.

[30]. Singh Satyendr and Tanveer J. Siddiqui; "Evaluating effect of context window size, stemming and stop word removal on Hindi word sense disambiguation. “ Information Retrieval and Knowledge Mangement (CAMP), 2012 International Conference on. IEEE, 2012.

[31]. Sinha M., Kumar M. Pande P., Kashyap L., Bhattacharya P., “ Hindi Word Sense disambiguation. “ International Symposium on Machine Translation, Natural Language Processing and Translation Support Systems, Delhi, India 2014.

[32]. Singh Satyendr and Tanveer J. Siddiqui; "Role of Karaka Relations in Hindi Word Sense Disambiguation." Journal of Information Technology Research (JTTR) 8.3(20150, 21-42.

[33]. Umrinder Pal, Vishal Goyal and Anisha Rani; "Disambiguating Hindi Words Using N-Gram Smoothing Models." International journal of Engineering Science, Issue, 2014,pp. 26-29.

[34]. Singh Satyendr and Tanveer J. Siddiqui; "A Supervised Algorithm for Hindi word Sense Disambiguation." International Journal of Systems, Algorithm Applications 3, 2013, pp. 29-32.

Cite this article as :

Prashant Y. Itankar, Dr. Nikhat Raza, "Ambiguity Resolution : An Analytical Study", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 2, pp. 471-479, March-April 2020. Available at doi : https://doi.org/10.32628/CSEIT2062135
Journal URL : http://ijsrcseit.com/CSEIT2062135