Vol. (1) No. (1) March 2018

JK Research Journal in Mathematics and Computer Sciences

Word Sense Disambiguation in Natural Language Processing

Preeti Dubey

Department of Computer Science, Govt. College for Women, Parade Jammu, J&K- India

Abstract: Word Sense Disambiguation has been a research area since the evolution of Natural Language processing. Most of the languages have ambiguous words. Resolution to these ambiguous words is a very importance task in developing any tool for natural language processing which otherwise hampers the efficiency of the developed systems. The accuracy of the output generated depends on the sense of the context a word is being used in a sentence. Word Sense Disambiguation is used for resolving the ambiguous words. This paper presents a brief introduction to Word Sense Disambiguation, an overview of the approaches that can be used solving ambiguity and work done for various Indian languages.

Keywords: *WSD, NLP, Ambiguity, Lexeme*

Introduction

All human languages have ambiguous words, these are the words whose meaning differ with context to their reference. The computational identification of correct sense of a word in a context is known as Word Sense Disambiguation (WSD). It automatically figures out the intended meaning of an ambiguous word when used in a sentence. Few illustrations of ambiguous words are:

English: ***Bark, light, fair, fan, right, fly, watch, bank, bat, lead, live***

Hindi: *सोना, मत, गुलाब, आम*

Illustration of some ambiguous words in English and Hindi is shown below in table I:

Word Sense Disambiguation is applied in various areas such as Machine translation (MT), Information retrieval (IR), speech recognition, computational advertising, text processing, etc. Without WSD, the processing of data is error prone. WSD plays a very important role in increasing the accuracy of the output produced by a system. It preserves the meaning of the word with reference to the context.

Ambiguity A word, phrase, or sentence is said to be ambiguous, if it has more than one meaning. The various types of ambiguities are discussed below:

Preetidubey2000@yahoo.com
*Preeti dubey

ENGLISH	
Bark	The bark of the tree is bown
	I was awakened by the bark of the dog.
Fan	The fan stopped worked.
	I am a die hard fan of Tom Cruise.
HINDI	
मत	मत बोलो । (Don't Talk)
	मत के आधार पर फैसला हुआ। The decision was based on votes.
गुलाब	गुलाब जामुन अच्छा है। Gulab Jamun (Indian Sweet dish) is sweet.
	गुलाब लाल है। The rose is red.

Table I: Examples of Ambiguity

Lexical ambiguity: Words with multiple senses can either be homonymous or polysemous. Two senses of a word are said to be homonyms when they mean entirely different things but have the same spelling. For example, the two senses of the word 'right' may be uses as below:

1) He is a right-handed batsman
2) He is right.

A word is said to be polysemous when its sense has various shades of the same basic meaning i.e. it has different but related meaning. For example, the word accident is polysemous since its two senses – a mishap and anything that happens by chance are somewhat related to each other.

Structural (Syntactic) Ambiguity:

If the ambiguity is in a sentence or clause, it is called structural ambiguity. It occurs when a phrase or sentence has more than one underlying structure e.g

We should be discussing violence on TV.

We should be discussing violence on TV.

Referential ambiguity

Sentences in which pronouns refer to certain words, but it is often difficult to find out, to which word it is referring e.g.

She dropped the plate on the table and broke it. [13]

Here, the ambiguity is that the pronoun 'It' refers to which noun, the table or the plate.

Such references are called anaphoric references or anaphora.

The Multi-Word Constructs:

Ambiguity arising due to multi-word constructs like Idioms and Phrasal verbs. For example, "My *neighbor's house was broken into last night.*" The actual meaning of the idiom can't be identified from its constituent words.

Approaches for WSD

The Word Sense Disambiguation approaches are classified into three main categories:

1) Supervised approach
2) Unsupervised approach
3) Knowledge based approach

Supervised Approach

The Supervised is based on a annotated data/ corpus. The supervised approach s based on trained sense annotated corpus to build classifiers. Initially, annotated corpus is required to build a classifier. The classifier is used to recognized the sense of the word based on their context of use. This approach assumes that the context can provide enough evidence on its own to disambiguate words. A major disadvantage of this approach is he requirement of a large sense-tagged corpora.

The accuracy of the classifier depends on the size and the variety of the data incorporated in the corpus. Larger the corpus, better the results. The creation of corpus itself is a challenge in terms of time and money. Some important WSD algorithms based on supervised approach are: Naïve Bayes Method, Decision List Method, Decision Tree Method, Support Vector Machine (SVM) Method.

Unsupervised Approach

The Unsupervised approach is based on unannotated corpora. It is based on clustering of words. It assumes that sense of the ambiguous word can be induced from the neighboring words. Since this approach is based on unannotated data, it is very essential to have most of the senses of the word in the training corpus. Some algorithms based on this approach are: Context Group Discrimination, Co-occurrence graphs, WSD using parallel corpora.

Knowledge Based Approach

This approach based on using Knowledge bases for Word Sense Disambiguation. The knowledge bases that can be used are: Word Net, Machine readable dictionaries, thesauri etc. This

approach is suitable for domain specific word sense disambiguation. One of the famous algorithm used in this approach is the algorithm proposed by Michael Lesk in1986 and is named Lesk Algorithm.

Literature Review

Lesk Algorithm[5] developed by Micheal Lesk in 1986. It is used to identify senses of polysemy words. He used the overlap of word definition from the Oxford Advanced Learner's Dictionary of Current English (OALD) to disambiguate the word senses.

Banerjee and Pedersen[7] adapted the original Lesk algorithm and used WordNet as the knowledge source. The overall accuracy of their system was evaluated to be 32%. They used Senseval-2 word sense disambiguation exercise to evaluate their system.

Sinha M. et. al.[8] **at** IIT Bombay **developed** automatic WSD for Hindi language using Hindi WordNet.They used statistical method for determining the senses. The system could disambiguate the nouns only. They compared the context of the polysemy word in a sentence with the contexts constructed from the WordNet. They evaluated the system using the Hindi corpora provided by the Central Institute of Indian Languages (CIIL). The accuracy of their algorithm was found in the range from 40% to 70%.

Shrestha N. et. al.[9] developed WSD module based on Lesk algorithm for the disambiguation of Nepali ambiguous words. They collected words to be used as the knowledge resource from synset, gloss, examples and hypernym.[] The number of example for each sense of the target word was only one in the database.

Each word in the context words is compared with each word in the collection of words available in the formed database.

Dhungana and Shakya[12] have also worked for the disambiguation of the polysemous words in Nepali language using the Lesk algorithm. The experiments performed on 348 words (including the different senses of 59 polysemy words and context words) with the test data containing 201 Nepali sentences shows the accuracy of their system to be 88.05%.

Richard Singh and K. Ghosh[14] in 2013 proposed a architecture for Manipuri Language. The system performs WSD in two phases: training phase and testing phase. The Manipuri word sense disambiguation system is composed of the following steps: (i) pre-processing, (ii) feature selection and generation and (iii) training, (iv) testing and (v) performance evaluation.

Haroon, R.P. (2010)[10] has given the first attempt for an automatic WSD in Malayalam. The author used the knowledge based approach.

Rakesh and Ravinder[16] have proposed a WSD algorithm for removing ambiguity from the Punjabi text document. The authors used the Modified Lesk Algorithm for WSD.

Ayan Das and Sudeshna Sarkar[10] have developed WSD system based on unsupervised graph-based clustering approach for Bengali language , which is applied to the Bengali-Hindi machine translation..

Preeti Dubey (2014)[13] has applied the unannotated corpus based WSD algorithm for word sense disambiguation in the copyrighted Hindi-Dogri Machine Translation system developed by her under the guidance of Prof. Devanand and Prof. Shashi Pathania with assistance from Dr.Vishal Goyal. Ambiguity in the Hindi-Dogri Machine Translation System is resolved using the 'n gram approach' for the morphemes 'से' and 'और'. Some use cases of the morpheme 'से' are shown in Table II:

As seen in Table II, the translations from Hindi to Dogri did using the Hindi-Dogri MTS show high accuracy rate for disambiguation of the morpheme से. Out of 9 sentences taken from online Hindi news only two resolutions shown in table III of से are incorrect.

Sentence	Incorrect Translation	Correct Translation
बाहर से आकर	बाह्र कोला आइयै	बाह्रे द्रा आइयै
बीबीसी से	बीबीसी कोला	बीबीसी कन्नै

Table III: shows the incorrect resolutions by WSD module in Hindi-Dogri MTS as well as the correct translations of से.

Conclusion

Word sense disambiguation (WSD) has been a very active area of research in computational linguistics. It plays an important role in the increased accuracy of the output generated by systems. One of the factors that has hampered WSD research is the unavailability of linguistic resources. The state of art is that most researchers are mostly using the knowledge based approach of disambiguation based on Lesk's algorithm. Due to lack of corpus, researchers are using WordNet for disambiguation.

It is also studied that the knowledge based approaches are efficient in resolving the sense of the ambiguous words, but there is a great need to develop lexical resources to be used for the efficiency of the natural language processing tools.

References

[1] Arindam Chatterjee, Salil Rajeev Joshi, Mitesh M. Khapra, Pushpak Bhattacharyya, "Introduction to Tools for IndoWordNet and Word Sense Disambiguation" http://www.cfilt.iitb.ac.in/wordnet/webhwn/IndoWordnetPapers/18_Tools%20for%20IndoWordNet%20&%20WSD.pdf

[2] C. Fellbaum, "WordNet: An Electronic Lexical Database.", MIT Press, 1998.

[3] Lokesh Nandanwar, Kalyani Mamulkar, "Supervised, Semi-Supervised and Unsupervised WSD Approaches: An Overview", International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064

[4] Kuhoo Gupta, Manish Shrivastava, Smriti Singh and Pushpak Bhattacharyya, Morphological Richness Offsets Resource Poverty- an Experience in Building a POS Tagger for Hindi, COLING/ACL-2006, Sydney, Australia, July, 2006.
https://pdfs.semanticscholar.org/1daa/370563deb1e9f5772d6ada1edfb4ec4cd7b4.pdf

[5] M. Lesk. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In Proceedings of the 5th annual international conference on Systems documentation, SIGDOC '86, pages 24–26, New York, NY, USA. ACM

[6] Udaya Raj et.al, "Word Sense Disambiguation using WSD specific Wordnet of polysemy Words" https://arxiv.org/ftp/arxiv/papers/1409/1409.3512.pdf

[7] Banerjee and Pederson "An adapted Lesk algorithm for Word Sense Disambiguity using Word Net" http://www.cs.cmu.edu/~banerjee/Publications/cicling2002.pdf

[8] Sinha M. et. Al, " Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity"
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.62.3892&rep=rep1&type=pdf

[9] Shrestha N et.al "Resources for Napali WSD" in proceedings of IEEE conference on Natural Language Processing and Knowledge Engineering, 2008. NLP-KE '08. 2008

[10] Alok Ranjan Pal and Diganta Saha, "WORD SENSE DISAMBIGUATION: A SURVEY", International Journal of Control Theory and Computer Modeling (IJCTCM) Vol.5, No.3, July 2015

[11] Z.Z. Merhbene Laroussi, " Unsupervised system for Lexical Disambiguation of Arabic Language using a vote procedure," IEEE-ICoAC, 2011

[12] Raj Dhungana, Subarna Shakya ,et.al" Word Sense Disambiguation Using Wsd Specific Wordnet Of Polysemy Words" Udaya International Journal on Natural Language Computing (IJNLC) Vol. 3, No.4, August 2014

[13] **"**Study and Development of Machine Translation System: an important tool to bridge the digital divide". Ph.D. Thesis submitted in the Department of Computer Science & IT, University of Jammu, 2014.

[14] Shallu et.al "A Survey of Word-sense Disambiguation Effective Techniques and Methods for Indian Languages", JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 5, NO. 4, NOVEMBER 2013

S.No	Hindi	Dogri
1	वो कौन से देश हैं, जहां ज़िंदगी शानदार है.	ओह् केह्ड़े देश न , जित्थै जिंदगी शानदार ऐ
2	बेहतरीन सुविधाओं से लैस है?	बेहतरीन सुविधाओं कन्नै लैस ऐ ?
3	उत्तरी ध्रुव के क़रीब स्थित डेनमार्क तीन नॉर्डिक देशों में से एक है	उत्तरी ध्रुव दे करीब स्थित डेनमार्क त्रै नार्डिक देशें बिच्चा इक ऐ
4	जर्मनी से डेनमार्क आकर बसीं एनी स्टीनबाख कहती हैं कि सिर्फ़ अपने नागरिकों के लिए ही नहीं, डेनमार्क बाहर से आकर बसने वालों के लिए भी जन्नत है.	जर्मनी कोला डेनमार्क आइयै बसीं एनी स्टीनबाख आखदियां न जे सिर्फ अपने नागरिकें दे आस्तै गै नेईं , डेनमार्क **बाह्र कोला आइयै** बसने आह्लें दे आस्तै बी जन्नत ऐ .
5	न्यूज़ीलैंड और इसका क़रीबी देश ऑस्ट्रेलिया रहने के लिहाज़ से शानदार देश हैं	न्यूज़ीलैंड ते एह्दा क़रीबी देश ऑस्ट्रेलिया रौह्ने दे लिहाज़ कन्नै शानदार देश न
6	पूरे देश से लोग एक पर्चे पर ये मांग करते हुए एक-दूसरे से जुड़ते गए. जैसे-जैसे दस्तख़त बढ़े, अर्ज़ी की लंबाई भी बढ़ती गई. आज ये अर्ज़ी न्यूज़ीलैंड के एक म्यूज़ियम में रखी है.	पूरे देश शा लोक इक पर्चे पर एह् माँग करदे होए इक - दुए कन्नै जुड़दी गे . जि'यां - जि'यां दसखत बधे , अर्ज़ी दी लंबाई बी बधदी गेई . अज्ज एह् अर्ज़ी न्यूज़ीलैंड दे इक म्यूज़ियम च रक्खी ऐ .
7	यहां का जीवन-स्तर बहुत से पश्चिमी देशों से बेहतर है.	इत्थें दा जीवन - स्तर बड़े हारे पश्चिमी देशें कोला शैल ऐ .
8	बीबीसी से रूबरू हुए सैफ़ ने कहा कि, "बोर्डिंग स्कूल में पलना शहरों में पलने से बेहतर होता है.	बीबीसी कोला रूबरू होए सैफ नै आखेआ जे , बोर्डिंग स्कूल च पलना शैह्रें च पलने कोला शैल होंदा ऐ .
9	तब से लोग इस ठंडी जगह में बसे हैं.	अदूं दा लोक इस ठंडी थाह्र च बस्से न .

Table II: Use cases of word sense disambiguation in Hindi-Dogri MTS