A Study of Word Sense Disambiguation in Malayalam

Shilpa Prem[1], Shobhith Chandran[2], Dr. Bijimol T K[3]

[1]Pursuing master's degree program in Santhigiri College of Computer Science, Vazhithala, Kerala, India
Email: *mca2022_silpaprem@santhigiricollege.com*

[2]Pursuing master's degree program in Santhigiri College of Computer Science, Vazhithala, Kerala, India
Email: *mca2022_shobithchandranp@santhigiricollege.com*

[3]Assistant Professor in Santhigiri College of computer Science, Vazhithala, Kerala, India
Email: bijitk@santhigiricollege.com

Abstract: The task of determining the correct sense of a word within the context by the help of computer is known as word sense disambiguation (WSD). Word Sense Disambiguation is basically solution to the ambiguity which arises due to different meaning of words in different context .It includes different approaches towards word sense disambiguation and help to detect the correct meaning of the ambiguous words. The word sense disambiguation provides multilingual feature representation.WSD has been a trending area of research in natural language processing and machine learning. This paper studies the word disambiguation of the Malayalam sentences.

Keywords: Word Sense Disambiguation, Word Sense Induction, Ambiguity

1. Introduction

The task of determining the correct sense of a word within the context by the help of computer is known as word sense disambiguation (WSD). The word sense induction (WSI) is used to detect the sense of an ambiguous word. The identification of word sense induction (WSI) by computational identification is called word sense induction (WSI). This paper is introduced for giving information among people about the confusion spread and the importance for the ambiguous words in people's mind. Most of the words in our language have more than one meaning. So, definitely it will make confusion on people's mind. So, the word sense disambiguation (WSD) is to get rid of this confusion and make it easy for people to read.

For example,

- The bank will not accept cash on Saturday.
- The river overflowed the bank.

It is an example where the people's confusion begins.That is, the occurrence of word 'bank' in both the sentences clearly denotes distinct meaning.

In the first sentence the word 'bank' indicates thatthe commercial bank or financial bank. In the second sentence, the word'bank'indicates the river bank.Most humans don't eventhinkabouttheambiguity of words that occurs. They think thatit is simple, but itissimpleforamachinetoidentifytheexactmeaningofwords. Thecomputationalidentificationofwordiscalled word sense disambiguation.Thispaperincludesalltheimportantaspectsabo utwordsensedisambiguation(WSD) andwordsenseinduction (WSI).

2. Word Sense Disambiguation

In natural language and processing, the word sense disambiguation (WSD) has a very trending area of research. WSD is the main solution for finding the meaning of ambiguous words. The task of determining the correct sense of a word within the context by the help of computer is known as word sense disambiguation (WSD). The word sense induction (WSI) is used to detect the meaning of the ambiguous words.

3. Evaluation of WSD

The evaluation of WSD contains the following two inputs. They are:

i. Dictionary
ii. Test Corpus

The terribly first input for analysis of Word Sense Disambiguation (WSD) is wordbook that is employed to specify the senses to be disambiguated.

Another input required by WSD is the high-annotated test corpus that has the target or correct-senses.

The check corpora may contain of two types:

- Lexical sample – This type of corpora is employed within the system, wherever it's needed to clear up a tiny low sample of words.
- All-words – This type of corpora is employed within the system, wherever it's expected to clear up all the words during a piece of running text.

4. Approaches and Methodsto Word SenseDisambiguation

4.1 Dictionary-based or Knowledge-based Methods

As the name suggests, these methods primarily rely on treasures, dictionaries and lexical knowledge base. There are no use corpora evidences for disambiguation. The Lesk method is the seminal dictionary-based method introduced in 1986 by Michael Lesk. The Lesk algorithm is based on the Lesk definition "measure overlap between sense definitions for all words in context". In 2000, Kilgarriff and Rose Rosensweig simplified the Lesk definition as "measure overlap between sense definitions of word and current context", whichmeans to identify the correct meaning of one word at a time.

4.2 Supervised Methods

For disambiguation, machine learning methods use sense-annotated corpora to train. These methods assume that the context can give required evidence to disambiguate the sense on its own. In these methods, the words knowledge and reasoning are considered unnecessary. The context is represented as a set of "features" of the words. It also includes the information about the surrounding words. The most successful supervised learning approaches to Word Sense Disambiguation are Support vector machine and memory-based learning. These methods depend on remarkable amount of manually sense-tagged corpora, which is extremely expensive to create.

4.3 Semi-supervised Method

Most of the word sense disambiguation algorithms use semi-supervised learning methods as a result of lack in training and the reason is semi-supervised methods use both labeled and unlabeled data. These methods require limited amount of annotated text and big amount of plain unannotated text. The technique used by semi-supervised methods for this is bootstrapping from seed data.

4.4 Semi-supervised Method

These are methods which assume that the similar senses occur in similar context. And this is why the senses can be induced from text by clustering word occurrences by using some measure of similarity in the context and this task is called Word Sense Induction or Discrimination. Unsupervised methods have extreme potential by which it can overcome the knowledge acquisition bottleneck due to the non-dependency on manual efforts.

5. Applications Of Word Sense Disambiguation

Word sense disambiguation (WSD) is used in almost every application which utilizes the language technology. Now let's take a look at the scopes of WSD −Machine Translation: Machine translation or MT is the main application of WSD. Lexical choice for the words with distinct translations for different senses is done by WSD in MT. In MT, the senses are represented as words in their target language. Most of the Machine translation systems doesn't use WSD module.

5.1 Information Retrieval (IR)

Information retrieval (IR) is defined as a program that deals with the storage, organization, retrieval and evaluation of information from document repositories of textual information. The system assists its users in finding the information they requires but it does not return the answers of the questions asked. WSD is used to solve the ambiguities of the queries which is provided to IR system. Like MT, current IR systems do not use WSD module and they depend on the concept that the user would type enough context in the query to only get relevant or required documents.

5.2 Text Mining and Information Extraction (IE)

The main and most used application WSD is that necessary to do perfect and analysis of text. For example, WSD helps in intelligent gathering system to do flagging of the correct meaning of words.

6. LEXICOGRAPHY

WSD and writing will work along in loop as a result of fashionable writing is corpus based. With writing, WSD provides rough empirical sense groupings additionally as statistically vital discourse indicators of sense.

Followings are some difficulties faced by word sense disambiguation (WSD):

- Differences between dictionaries:

The major drawback of WSD is to determine the sense of the word as a result of completely different sense is terribly closely connected. Even {different/totally completely different/completely different} dictionaries and thesauruses will offer different divisions of word into senses.

- Different algorithms for various applications:

Another drawback of WSD is that utterly completely different algorithmic rule may well be required for various applications as an example, in MT, it takes the shape of target word selection; and in data retrieval, a way inventory isn't needed.

- Inter-judge variance:

Another drawback of WSD is that WSD systems ar usually tested by having their results on a task compared against the task of kinsmen. This is often referred to as the matter of inter-judge variance.

- Word-sense severalty:

Another issue in WSD is that words can't be simply divided into distinct sub meaning

7. Literature Review

According to SruthiSankar, et. al. [5] the aim of this work is to develop a WSD system for South Dravidian, a language spoken in India, preponderantly employed in the state of Kerala. The projected system uses a corpus that is collected from varied South Dravidian internet documents. For every doable sense of the ambiguous word, a comparatively little set of coaching examples (seed sets) are known

thatrepresents the sense Collocations and most co-occurring words are thought of as coaching examples. Seed set growth module extends the seed set by adding most similar words to the seed set components. These extended sets act as sense clusters. the foremost similar sense cluster to the input text context is taken into account because the sense of the target word.

According to Junaida M K et.al.[6]describes 2 algorithms Conditional Random Field (CRF) and Margin Infused Relaxed (MIRA) during a CRF framework for Malayalam WSD. This framework makes use of the discourse feature info together with the elements of speech tag feature so as to predict the assorted WSD categories. For coaching set, range of ambiguous words has been annotated with twenty five WSD categories. The experimental results of the tenfold cross validation shows the appropriateness of the projected CRF primarily based Malayalam signified tagger. Alternative machine learning techniques like Naive Thomas Bayes classifier, ME, Neural Networks etc will be applied during this study and also the results thus obtained will be compared with the present works.

According to Sherly Elizabeth et. al[7] a hybrid approach using a multi-class SVM and corpus based in order to conduct Malayalam word sense tagging. This framework makes use of the contextual feature information along with the parts of speech tag feature in order to predict the various WSD classes. For training set, limited number of ambiguous words has been annotated with 16 WSD classes. The experimental results of the 10 fold cross validation shows the appropriateness of the proposed multi-class SVM of Malayalam word sense tagger with one against one approach for both word only and word +POS.

Ac According to Jisha P Jayanet, et. al. [8] to implement a semi supervised machine learning techniques combined with statistical approach mainly Maximum Entropy is experimented for Malayalam, which shows promising result for a set of corpus trained

8. CONCLUSION

This paper describes about the study of word sense disambiguation in Malayalam that words have totally different meanings supported the context of its usage within the sentence. If we tend to point out human languages, then they're ambiguous too as a result of several words is understood in multiple ways that relying upon the context of their incidence. Word sense disambiguation (WSD) is the main solution for finding the meaning of ambiguous words. The task of determining the correct sense of a word within the context by the help of computer is known as word sense disambiguation (WSD).

References

[1] Brody, S., Lapata, M.: Good neighbors make goodsenses:ExploitingdistributionalsimilarityforunsupervisedWSD.In:Proceedingsofthe22ndInternational Conference on Computational Linguistics(COLING). pp. 65–72. Manchester, UK(2008)

[2] Brody,S.,Lapata,M.:BayesianWordSenseInduction. In:Proceedingsofthe 12th Conference oftheEuropeanChapteroftheAssociationforComputationalLinguistics(EACL).pp.103–111.Athens, Greece (2009)

[3] Chan,Y.S.,Ng,H.T.,Zhong,Z.:NUS-PT:ExploitingparalleltextsforWordSenseDisambiguationintheEnglishall-wordstasks.In:Proceedings of the Fourth International Workshop onSemantic Evaluations (SemEval-2007), Prague, CzechRepublic. pp. 253–256 (2007)

[4] deCruys,T.V.,Apidianaki,M.:Latentsemanticwordsenseinductionanddisambiguation.In:Proceedingsofthe49thAnnualMeetingoftheAssociationforComputationalLinguistics:HumanLanguage Technologies (ACL-HLT). pp. 1476– 1485.Portland, Oregon, USA(2011)

[5] SruthiSankar K Pa, P C ReghuRajb, JayanVc,Unsupervised Approach to Word Sense Disambiguation in MalayalaminInternational Conference on Emerging Trends inEngineering, Science and Technology (ICETEST - 2015).

[6] Junaida M K,Jisha P Jayan,Elizabeth Sherly,word sense disambiguation for Malayalam in a conditional random field frameworkinSBandyopadhyay, D S Sharma and R Sangal. Proc. of the 14th Intl. Conference on Natural Language Processing, pages 495–502, Kolkata, India. December 2017. c 2016 NLP Association of India (NLPAI)

[7] JISHA P JAYAN,JUNAIDA M K,SHERLY, Malayalam Word Sense Disambiguation using Yamcha in 2015 International Conference on Computing and Network Communications (CoCoNet).

[8] Jisha P Jayan, Junaida M K, Sherly Elizabeth, Malayalam Word sense disambiguation using maximum entrophy model in INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT)

Author Profile

SilpaPrem received Degree in Bachelor of Computer Applications from Mahatma Gandhi UniversityKottayam,Keralain 2020.

ShobithChandran P received Degree in Bachelor of Computer Applications from Kannur University Kannur,Keralain 2019.

.

Dr. Bijimol TK has been working as an Assistant Professor in the department of Computer Science sine June 2002. She has completed her PhD from Bharathiar University Coimbatore, Tamil Nadu. She has published a book and has 10 more journal and proceedings publications.