INTERNATIONAL RESEARCH JOURNAL ON ADVANCED SCIENCE HUB
e-ISSN : 2582 - 4376
Open Access
RSP SCIENCE HUB
(The Hub of Research Ideas)
Available online at www.rspsciencehub.com

Special Issue of First International Conference on Advancements in Management, Engineering and Technology (ICAMET 2020)

Exploring Resources in Word Sense Disambiguation for Marathi Language

Amit Patil[1], Chhaya Patil[2], Dr. Rakesh Ramteke[3], Dr. R. P. Bhavsar[4], Dr. Hemant Darbari[5]
[1,2] Assistant Professor, Department of Computer Application, RCPET's IMRD, Maharashtra, India
[3,4]Professor, School of Computer Sciences, KBC North Maharashtra University, Maharashtra, India
[5]Director General, Centre for Development of Advanced Computing (C-DAC), Maharashtra, India

Abstract

Word Sense Disambiguation (WSD) is one of the most challenging problems in the research area of natural language processing. To find the correct sense of the word in a particular context is called Word Sense Disambiguation. As a human, we can get a correct sense of the word given in the sentence because of word knowledge of that particular natural language, but it is not an easy task for the machine to disambiguate the word. Developing any WSD system, it required sense repository and sense dictionary. It is very costly and time-consuming to build these resources. Many foreign languages have available these resources, that is why most of the foreign languages like English, German, Spanish etc lot of work is done in these Natural languages. When we look for Indian languages like Hindi, Marathi, Bengali etc. very less work is done. The reason behind this is resource-scarcity. In this paper, we majorly focus on Marathi Language Word Sense Disambiguation because of very less work is done in the Marathi Language as compared to Hindi and other Indian Languages. Our main objective is to provide information about various resources available for the Marathi language which will be helpful for researchers who wants to do work for Marathi WSD. This paper also gives a review on work done for Marathi Language WSD and its challenges and problems.

Keywords: WSD, WordNet, Indo WorldNet, Part of Speech(POS), NLTK, iNLTK

1. Introduction

Word sense disambiguation is a process which is automatically recognized which multiple meaning of ambiguous words is being used in a specific sentence. In other Word, WSD is identifying which meaning of a word (i.e. sense) is used in a sentence when the word has multiple senses text. In natural language processing, word sense disambiguation (WSD) is an open problem of computation linguistic. It is a word sense disambiguation is a massive challenge in natural language processing. The human mind is quite talented at word-sense disambiguation. A human language developed in a way that human easily understands the meaning which reflects in the sentence. In the computer, it has been a long-standing challenge to improve the ability of computers to do natural language processing. In the supervised machine learning approach, the classifier is trained for every different word on manually senses annotated. These methods assume that the context can provide sufficient proof on its own to disambiguate the sense.It uses annotated corpus and ambiguity is resolved by finding the nearest or closest word having similarities.

2. WSD Application and Techniques

WSD is required in various areas like Information Retrieval (IR), Sentiment Analysis, Knowledge Graph Construction, Text Mining and Information Extraction (IE), Lexicography and Machine Translation. Solutions to WSD are mostly categorized into knowledge-based, supervised and unsupervised approaches.

Every machine learning approach has many algorithms. In the supervised approach follows many algorithms are Decision List, Decision Tree, Naïve Bayes, Support Vector Machine (SVM), Logistic Regression, Logistic Regression, Random Forest, KNN (k- Nearest Neighbours) Ensemble Methods, Neural Networks, etc. In the Unsupervised approach follows many algorithms are Word Clustering, Context Clustering, Co-occurrence Graph, K-means, Apriori algorithm, etc. The Knowledge Base approach follows many algorithms are Genetic algorithm, Decision Support, Lesk algorithm, Semantic Similarity, Selection Preferences.

The authors [1] explored the work status of WSD in the Marathi Language. Many researchers used different algorithms for disambiguating the Marathi sentence like the graph-based algorithm to resolve ambiguity based on word sense and context domain. The researcher used Genetic Algorithm technique through which they resolve the ambiguity of the words based on their context domain and their senses. Other authors used approach consists of a modified Lesk algorithm with Support Vector Machine. etc. The accuracy of every algorithm is depends on text corpus and different techniques applied to the data set.

2.1 Marathi Language and its Word Categories

Marathi is the Indo-Aryan language. This language of Sanskrit origin. The Marathi language is the official language of Maharashtra state, a state in India. The language is most speakers in word wide. Approximately 90 million people in India speak this language. Maharashtra is a Southern state in India, the dialects of Marathi include Varhadii, Gawdi of Goa, Nagpuri Marathi, Dangii, Malwani, Kudali, Kasargod, Kosti, Ahirani of Khandeshi, etc. The Marathi language follows the Subject, Object, Verb, Nouns inflect for gender, number, etc. The Marathi language is eight main POS (Part of Speech). These are Noun, Verb, Adjective, Adverb, Pronoun, Postposition, Conjunction and Interjection [3].

3. Marathi Language Resources

One of the challenges in researching Marathi Language WSD is a lack of resources. Still, some peoples started research work for Marathi Language and developed some Marathi Language WSD. In this section, we try to give information about resources available for Marathi Language which will useful for researchers, who want to work on this problem.

3.1 Marathi WorldNet

This is a machine readable dictionary based English WordNet. It is not just a traditional dictionary, but more than this. This dictionary gives different relations between synsets or synonym sets represented as unique concepts. It is developed by Dr. Pushpak Bhattacharya with his team at IIT, Bombay. Marathi WorldNet is organized as a semantic network of large electronic databases.

Paradigmatic relations such as synonymy, hyponymy, antonymy and entailment etc. are used to construct it. It is widely used lexical database today for research in NLP for Marathi language, the different senses called synonym sets or synsets for each open-class word like nouns, verbs, adjectives, and adverbs are listed by Marathi WordNet. It has the index_txt file to Provides information about all words present in Wordnet, the data_txt file for Providing the details of every word in the index file and the onto_txt file which Provides ontology details of the words in data file [2].

Example:

data_txt File-
Structure of Data_txt: Example 00054554 03 02 फसवणे:चकवणे 0001 0400 00000183 | फसेल असे करणे:"नकली मालाची विक्री करून दुकानदार लोकांना फसवतात."

Synset id=00054554
POS=03 number of words present in synset=02
Synsets= फसवणे:चकवणे
Number of relations lexical as well as semantic=0001
Four-digit code relation id=0400 synset_id for which that relation exists=00000183 gloss= फसेल असे करणे
Example sentence=:"नकली मालाची विक्री करून

दुकानदार लोकांना फसवतात."

Here part of speech (pos): 1(noun), 2(adjective), 3(verb), 4(adverb) first four-digit relation type represented and the second four-digit represents the order of words from two synsets for which relation holds.

index_txt File-

Consider the example from index_txt file: गीत 01 01 0400 01 00004897

In the example: word= गीत pos=01 {*pos: 1(noun), 2(adjective), 3(verb), 4(adverb)} number of relations exists for word in all its senses=01 number of senses=01 and sense id is=00004897 in the data_txt.txt file.

onto_txt File-

Structure of Onto_txt : 00000051 0001 00000044 |वेळ (Time) {TIME उदाहरणे :- सकाळ, दिवस इत्यादी}

onto id= 00000051 0001 indicates that parent exists parent onto id= 00000044 onto description=(Time).

3.2 Indo WordNet

Based on EuroWordNet dictionary, IndoWordNet is developed. Eighteen scheduled languages of India, namely Marathi, Hindi, Malayalam, Telugu, Kashmiri, Bodo, Bangla, Gujarati, Kannada, Odia, Konkani, Manipuri, Assamese, Punjabi, Nepali, Tamil, Sanskrit and Urdu represent the lexical linked knowledge base of IndoWordNet. It is an online interface, which users can get outcomes according to needs in various organizations. The Look and feel of IndoWordNet are same as a customary word reference keeping the user versatility. IndoWordNet database structure is imported from English WordNet which is present on Princeton University site [4].

In IndoWordNet there are 32829 synsets of Marathi Language are available [5]. The following table gives details of it.

Table.1. Synset count of IndoWordnet for Marathi

Marathi Language	Noun	Verb	Adj.	Adv.	Total
	23599	3345	5325	559	32829

3.3 NLP Libraries

Most of the NLP applications required pre-processing of the text. The libraries that are more useful while using python for text processing are the Indic NLP Library and Natural Language Toolkit for Indic Languages (iNLTK). These two libraries support many Indian languages including Marathi.

3.3.1 Indic NLP Library

This Library is intended to build Python-based libraries for common text processing and Natural Language Processing in Indian languages. Most of the Indian languages have similarities in terms of script, phonology, language syntax, etc. and this library gives a general solution commonly required toolsets for Indian language text.

Using this library you can do Text Normalization, Script Information, Word Tokenization and De-tokenization, Sentence Splitting, Word Segmentation, Syllabification, Script Conversion, Romanization, Translation, etc. for the Marathi Language.

3.3.2 Natural Language Toolkit for Indic Languages (iNLTK)

The iNLTK library is equivalent to the NLTK Python package. This library provides features that an NLP application developer required. It provides Tokenization, Generates similar sentences from given text input, Identifies the language of a text, Text completion, Word Embedding, and Text Generation in 13 Indic Languages including the Marathi Language. iNLTK is an open-source NLP library that support Marathi Language also [6-10].

3.4 Marathi Nominal Morpheme Segmenter

Marathi Nominal Morpheme Segmenter is a noun segmenter for Marathi Language developed as M.Phil project and available on Computational Linguistics R & D, JNU [7].

3.5 IndicCorp

One of the largest publicly-available corpora for Indian languages is IndicCorp. This corpora is created for thirteen Indian Languages, Marathi is one of them. IndicCorp corpora consist of thousands of web sources - primarily news, magazines, and books. The format of the corpus is a single large text file containing one sentence per line. The following table gives information about Marathi Language IndicCorp Corpora [11].

Table.2. Statistics of IndicCorp Corpora

News Articles	Sentences	Tokens
2.31 M	34.0 M	551 M

4. Conclusion

Though Word Sense Disambiguation (WSD) is one of the most challenging problems in the research area of natural language processing. This paper explores the methods, algorithms and technique of Word Sense Disambiguation. In this paper, we try to elaborate on resources that will helpful for working in Marathi Word Sense Disambiguation. We provide information about the Indian Language Libraries and Tools.

References

[1] Amit Patil, Chhaya Patil et. al. "Different Approaches of Marathi Language Word Sense Disambiguation", International Journal of Emerging Trends in Engineering and Basic Sciences (IJEEBS), ISSN (Online) 2349-6967, Volume 7, Special Issue 3 (May-June 2020), PP. 001-003

[2] Gauri Dhopavkar, et. al., "Exploiting Rules For Resolving Ambiguity In Marathi Language Text", International Journal of Research in Engineering and Technology, Volume: 04 Issue: 12 , Dec-2015.

[3] Gauri Dhopavkar, et. al., "Resolving Ambiguity In Marathi Language Text: A Rule Based Solution", International Journal of Current Engineering And Scientific Research (ijcesr), volume-2, issue-10, 2015

[4] Ritika Sharma, et. al., Indo-Word-Net Dictionary: A Review, International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 5, Issue 6, June 2018

[5] http://www.cfilt.iitb.ac.in/wordnet/webhwn/iwn_stats.php

[6] https://github.com/anoopkunchukuttan/indic_nlp_library

[7] http://sanskrit.jnu.ac.in/index.jsp

[8] https://github.com/anoopkunchukuttan/indic_nlp_library#indic-nlp-library

[9] https://github.com/goru001/inltk

[10] Gaurav Arora, iNLTK: Natural Language Toolkit for Indic Languages, arXiv:2009.12534v1 [cs.CL] 26 Sep 2020

[11] https://indicnlp.ai4bharat.org/corpora/