International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
IJISAE
ISSN:2147-6799 www.ijisae.org Original Research Paper

Unsupervised Word Sense Disambiguation for Marathi language using Word Embeddings

Rasika Ransing*[1,2], Dr. Archana Gulati[3]

Submitted: 29/01/2024 **Revised**: 07/03/2024 **Accepted**: 15/03/2024

Abstract: Word Sense Disambiguation (WSD) is a significant challenge within the field of natural language processing. Marathi is considered to be a language with limited resources. Therefore, there has been less research conducted on the Marathi language. There are many methodologies used in Word Sense Disambiguation (WSD), including Knowledge-based and Machine Learning methods. The Lesk algorithm, which uses WordNet as the sense inventory, is well recognized as a prominent technique within the knowledge-based approach for Word Sense Disambiguation (WSD). Word embeddings are a technique used to encode individual words by using low-dimensional vectors with real-valued components. We introduce a method that leverages word embeddings of the terms in a given phrase, except the ambiguous word. Additionally, it incorporates word embeddings of glosses, synonyms, and instances of the ambiguous words obtained from the Marathi WordNet. The most suitable sense of an ambiguous word is then determined using the context and the word embeddings. From the results of our proposed methodology, we could conclude that including the embeddings of synonyms and examples of ambiguous words increases the accuracy of disambiguating words.

Keywords: *Marathi language, Natural Language Processing, Unsupervised Learning, Word Embeddings, Word Sense Disambiguation*

1. Introduction

Word sense disambiguation (WSD) is the process of computationally interpreting a word in a given context. It's one of the difficult and persisting problems in natural language processing (NLP) [1]. While a significant amount of study is done on the English language, a considerable amount is done on the Hindi language, and very little research is done on the regional Indian languages with fewer resources, such as Marathi [2]. One of the few languages with a rich morphology is Marathi, which permits the formation of several morphological variations from inflections on a single root word. Due to Marathi's free order sequencing and inflectional variations, word sense disambiguation—a challenging problem in natural language processing—becomes considerably more difficult [3].

Based on the research done so far, several researchers have suggested different methods for word sense disambiguation (WSD). These methods are divided into two main categories: Knowledge-based approaches and Machine-learning approaches. The three subcategories of machine learning techniques are supervised, unsupervised, and semi-supervised, and they all need a training corpus. Knowledge-based techniques need lexical resources such as ontologies, WordNets, machine-readable dictionaries, etc. [4]. The task of constructing a sense-annotated corpus presents significant challenges and requires considerable effort for Supervised Machine Learning methods. Moreover, this approach necessitates repetition across several languages, diverse domains, and numerous sense inventories [5]. Therefore, unsupervised methods have garnered significant interest owing to their lack of need for sense-annotated corpora [6]. The unsupervised technique has the benefit of not requiring a substantial annotated corpus, making it more feasible for real-world applications. The utilization of a comprehensive and well-structured public lexicon called WordNet facilitates the achievement of this task.

WordNet offers a collection of cognitive synonyms, also known as synsets, for every word, which are further categorized based on their part-of-speech tags. Each synset is accompanied by a definition, examples of usage, and information regarding its semantic relationships with other words [7]. Marathi WordNet is a lexical database specifically designed for the Marathi language. It was developed by the Center for Indian Language Technology (CFILT) at the Indian Institute of Technology Bombay (IIT Bombay) [8]. The WordNet encompasses several semantic connections, including but not limited to hypernymy, hyponymy, and meronymy [9].

The Lesk technique is designed to resolve the ambiguity of a word by quantifying the degree of overlap between the

1 Datta Meghe College of Engineering, Airoli
Mumbai, India
2 Vidyalankar Institute of Technology, Wadala
Mumbai, India.
3 School of Business Management, SVKM's NMIMS University, Kharghar
Mumbai, India
** Corresponding Author Email: rasikaransing275@gmail.com*

definitions of each possible interpretation of the ambiguous word and the surrounding terms in its context [10]. The meaning of a word that has the greatest overlap with the surrounding terms is regarded as the accurate interpretation of the ambiguous word. The underlying idea of the algorithm is that words within a designated time frame often exhibit a tendency to be associated with a shared context [11]. The subsequent iteration of the Lesk algorithm proposes the retrieval of the several meanings of the ambiguous term from WordNet [12].

Word embeddings are a method used to represent individual words using low-dimensional real-valued vectors. These models function based on the concept that words with similar meanings are likely to appear in comparable linguistic contexts. Textual similarity estimation has been facilitated by the use of word embeddings [13]. The use of contextualized word embeddings has shown to be successful in many Natural Language Processing tasks due to its ability to include valuable semantic information. The use of contextualized word embeddings has shown significant advantages across a wide range of Natural Language Processing applications. The primary factor contributing to this significant achievement is the use of contextualized word embeddings, which capture the semantic meaning sent by the surrounding context in which the words are used [14].

Word embeddings have shown their significance as valuable resources due to their ability to provide a concise collection of attributes that may be used in subsequent natural language processing activities such as machine translation, question answering, information retrieval, word sense disambiguation, and others. In contrast to standard Machine Learning approaches, which often need supervised feature extraction, embeddings may be generated using entirely unsupervised methods [15].

The majority of research in the field of Natural Language Processing (NLP) has been conducted on languages that possess abundant linguistic resources, such as English. Indian languages such as Hindi and Marathi suffer from a significant dearth of extensive linguistic corpora. Several datasets, such as the named entity dataset and word analogy dataset, which are valuable for evaluating word embeddings, are currently unavailable for Indian languages. The lack of resources for Indian languages has resulted in a dearth of NLP tasks specifically tailored for these languages. Consequently, some activities, such as neural machine translation, exhibit lower levels of accuracy. Word embeddings are created by conventional methods such as GloVe, Word2Vec, and FastText. The efficacy of these strategies in Indian languages has not been extensively evaluated [16].

The advent of transfer learning and the subsequent use of pre-trained word embeddings has mitigated the dependence on the magnitude of a dataset. The pre-trained word embeddings based on deep learning effectively capture the semantic links among different words, therefore illustrating word analogies in vector space using an equation representation. When words are represented in a visual manner using word embeddings, it can be seen that words with similar meanings tend to cluster together, suggesting a stronger level of similarity between their respective word vectors. Over the last several years, there has been a progression in the field of embeddings, transitioning from context-free embeddings to contextual embeddings such as CoVe, ELMo, and BERT [3].

This paper proposes a model whereby the context set of the ambiguous word is constructed by using all the words in the phrase, except the ambiguous word. The word embeddings for the words in the context set are computed. A separate sense set is built for every sense of the target ambiguous word. Every sense set comprises the gloss, synonyms, and examples associated with that sense of the target ambiguous term. For every sense set, a sense vector is built which is the computation of word embeddings in that set. For computation of word embeddings, we have used FastText, IndicBERT and MURIL models. The computation of cosine similarity is performed between the context vector and each sense vector. The disambiguation of the ambiguous word is accomplished by identifying the sense that exhibits the highest cosine similarity value.

The subsequent sections of the paper are structured in the following manner. Section 2 presents a concise summary of the literature and the research conducted on the use of word embeddings, the Lesk approach, and IndoWordNet for disambiguation methods for Indian languages. Section 3 provides an overview of the proposed system. The findings of the algorithm testing are presented in Section 4. The conclusion and future work are presented in Section 5.

2. Related Work

Khanuja et al. introduced the MuRIL language model (LM), which stands for Multilingual Representations for Indian Languages. This language model is a multilingual one that has been particularly developed to cater to the diverse range of languages spoken in India. The MuRIL model is exclusively trained on a substantial volume of Indian text corpus. The MuRIL model presently provides support for a total of 17 languages, including English (en), Hindi (hi), Kannada (kn), Malayalam (ml), Marathi (mr), etc. The architecture of MuRIL is specifically optimized to generate contextual word representations for Indian languages. It undergoes evaluation across several natural language processing tasks in order to gauge its efficacy in effectively processing multiple languages. The performance of MuRIL surpasses that of multilingual BERT (mBERT) across all tasks under the cross-lingual XTREME benchmark, demonstrating its superiority [17].

Kakwani et al. provide a comprehensive introduction to Natural Language Processing (NLP) resources for 11 prominent Indian languages. These resources include pre-trained word embeddings and pre-trained language models. The word embeddings used in this study are derived from FastText, making them well-suited for effectively managing the morphological intricacies present in Indian languages. The pre-trained language models are derived from the ALBERT model, which is known for its compact architecture [18].

Kharate and Patil provide a proposed strategy for word sense disambiguation (WSD) in Marathi language. Their solution utilizes the Lesk algorithm and incorporates Marathi WordNet [8] for disambiguating ambiguous terms in Marathi. The input to their suggested system consists of a statement including a term that has several interpretations, leading to ambiguity. The term "context set" pertains to the collection of words that surround an ambiguous word, while the term "gloss set" refers to the collection of words derived from the glosses of all semantic connections associated with ambiguous words found in Marathi WordNet. The system compares the context in which an ambiguous word is used (known as the Context Set) with the different contexts associated with the ambiguous word (known as the Gloss Set) derived from the Marathi WordNet to determine the most suitable meaning for the ambiguous word [19].

Kumari and Lobiyal have utilized word embedding techniques in order to address the challenge of Word Sense Disambiguation (WSD) in Hindi literature. This challenge comprises of two stages: the generation of word embedding using Word2Vec architectures like Skip-Gram and Continuous Bag-Of-Words models, and the use of cosine similarity to choose a suitable interpretation of the word. The researchers have gathered the corpus and implemented the necessary pre-processing techniques. Word embeddings are then generated for pre-processed data. The sense definition is accessed using various knowledge bases or machine-readable dictionaries. Subsequently, the researchers have used the generated word embeddings to produce Context-Vector and Sense-Vectors for the ambiguous word mentioned in the test phrase. Cosine similarity is used in order to determine the semantic relatedness of these vectors, hence enabling the selection of the sense with the greatest semantic score for the purpose of disambiguating a word. They have conducted extensive tests were conducted on substantial corpora, resulting in an attained accuracy rate of around 52% [20].

Bhingardive et al. introduced an unsupervised approach to determine the Most Frequent Sense (MFS) from untagged corpora, using word embeddings. The researchers conducted a comparison between the word embedding of a given word and all of its corresponding sense embeddings. Through this analysis, they determined the predominant sense by identifying the sense with the greatest degree of similarity. A significant improvement in performance was seen for Hindi Word Sense Disambiguation (WSD) compared to the baseline method of WordNet First Sense (WFS). Their approach has been limited only to nouns. Sense embeddings are generated by using WordNet-based characteristics inside the framework of the extended Lesk algorithm. The cosine similarity was used as a similarity metric in order to assess the accuracy of these word embeddings in relation to sense embeddings [21].

Orkphol and Yang used the Word2vec algorithm to generate a contextual sentence vector as well as sense definition vectors. The researchers proceeded to assign a score to each word sense by using cosine similarity to calculate the similarity between the sentence vectors of those word senses and the context sentence vector. The notion of sense was further broadened by including sense relations obtained from WordNet. In cases where the score did not above a certain threshold, it was aggregated with the probability derived from the sense distribution acquired from SEMCOR, a comprehensive corpus containing sense annotations. The findings of their study demonstrate that the obtained result (50.9% or 48.7% without considering the probability of sense distribution) surpasses the performance of the baselines, namely the original, simplified, adapted, and LSA Lesk methods [7].

In their research, Laatar et al. introduced a novel approach using word embeddings to address the challenge of Arabic Word Sense Disambiguation. The primary objective was to use vectors as word representations within a multidimensional space to effectively capture the semantic and syntactic characteristics of words. The technique described by the authors utilizes the Arabic dictionary for the purpose of word sense selection. The sense ascribed to a word with several meanings is determined by its semantic closeness to the immediate context. The proposed approach involves first training word vectors from the corpus using the Skip-Gram model. Subsequently, the context of a word to be disambiguated, together with all its senses, is represented as a vector in a multidimensional space. The process of Word Sense Disambiguation (WSD) involves using cosine similarity as a measure to assess the similarity between the context vector and the target word sense vectors. The sense with the greatest similarity is then assigned as the disambiguated sense. The experimental results showed that their approach attained a level of accuracy up to 78% [22].

The word embeddings for the Hindi language were produced by Gaikwad and Haribhakta, who then evaluated their performance via the use of word analogies and similarity datasets. Their study focuses on the development of word embeddings for the Hindi language via the use of the Adaptive GloVe model and the Adaptive FastText model. The AGM model exhibited superior performance compared to the original GloVe model and FastTextWeb in terms of word analogies and word similarity datasets. The use of FastText embeddings trained on a Hindi monolingual corpus (FastTextHin) has shown a substantial improvement in performance when compared to FastTextWeb. The implementation of the Adaptive FastText model resulted in a significant improvement in the performance of the FastText model when evaluated on word analogies and similarity datasets [16].

In their study, Laatar et al. introduced a word sense disambiguation technique that leverages a hybrid approach, integrating a lexical knowledge-base with contextualized embeddings. The researchers conducted an investigation of the efficacy of contextual word embeddings, specifically stacked embeddings, for the purpose of disambiguating Arabic words. The objective of their approach is to use the Flair architecture to train a neural embedding model that can effectively discern the intended sense of an ambiguous word within the specific contextual framework of a given text. Various embeddings, including layered embeddings, were evaluated in order to determine the optimal combination that yields the most favorable outcomes. Their experimental results demonstrate that the integration of word2vec and Flair word embeddings yields accuracy rates of 70.86%, 52.2%, and 55.34% for Modern Arabic, Old Arabic, and Middle-age Arabic, respectively [23].

3. Proposed Approach

This work builds upon previous research on word embeddings in English and Indian regional languages and presents a novel model for Marathi Word Sense Disambiguation using word embedding techniques.

We propose the model illustrated in Figure 1, that utilizes word embeddings and extended Lesk approach to disambiguate words with multiple senses.

The system receives a statement written in the Marathi language as input, together with the target ambiguous word within the text. The input phrase is pre-processed by the removal of special symbols, stopwords, and lemmatization of words, which involves reducing the words to their root forms. The ambiguous word is also subjected to lemmatization and transformed into its root form. One benefit of doing lemmatization on all words is the facilitation of straightforward comparison and matching of root words. The ambiguous word is now removed from this pre-processed sentence. Hence, the context set consists of the words in the pre-processed sentence except the ambiguous word. Let the input sentence be,

Input sentence = $\{x_{-3}, x_{-2}, x_{-1}, x, x_1, x_2, x_3\}$ where,

x = ambiguous word. Then,

Context set = $\{x_{-3},x_{-2},x_{-1},x_1,x_2,x_3\}$

The word embeddings (WE) for the words in the context set are computed using FastText, IndicBERT and MURIL. Hence, Context Vector is,

CV = WE ($\{x_{-3},x_{-2},x_{-1},x_1,x_2,x_3,x_4\}$)

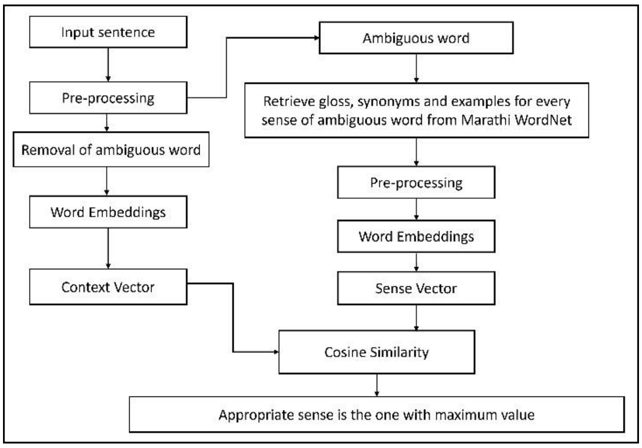

Fig. 1. Proposed system for Word Sense Disambiguation of Marathi words using Word Embeddings and IndoWordNet.

Now, the ambiguous word is considered. All the senses of this ambiguous word are retrieved from pywin, a Python-based API used to access WordNets of Indian Languages [24]. A sense set (SS) is built for every sense of the ambiguous word. Sense set consists of the synonyms, gloss and example sentences of that particular sense. The gloss and example sentences are also pre-processed to have only root forms of content words. If the ambiguous word has 3 senses S1, S2, S3, then,

SS1 = Synonyms, Gloss and examples of S1

SS2 = Synonyms, Gloss and examples of S2

SS3 = Synonyms, Gloss and examples of S3

A sense vector (SV) is built by computing Word embeddings for every sense set using FastText, IndicBERT and MURIL. Word emebeddings for each sense set are,

SV1 = WE (SS1)

SV2 = WE (SS2)

SV3 = WE (SS3)

The cosine similarity metric is used to compute the semantic similarity between Context Vector and every Sense Vector [21].

Cosine Similarity1 = cos(CV,SV1)

Cosine Similarity2 = cos(CV,SV2)

Cosine Similarity3 = cos(CV,SV3)

The sense that exhibits the greatest cosine similarity with the collection of context words is chosen as the suitable meaning of the ambiguous word in relation to the provided context.

If the value of Cosine Similarity3 is highest among all the cosine scores, sense S3 is the most appropriate sense for the ambiguous word in the given input sentence.

Marathi is a language that has a very limited amount of resources. The development of semantic tools for the Marathi language is yet incomplete. In this study, we used example phrases from two corpora: the Marathi Monolingual Text Corpus ILCI-II and the Hindi-Marathi Tourism Text Corpus ILCI. Nevertheless, we were able to identify very few phrases, including terms that have ambiguous meanings. Therefore, ambiguous words have been found from Marathi WordNet, and also sentences have been constructed using these ambiguous words. In order to conduct an assessment, we have generated dataset consisting of 200 Marathi sentences. These sentences are derived from 12 ambiguous terms that were collected from various Marathi websites and publications, including those related to news, philosophy, sports, and fiction. The dataset has around 42 different senses. The set of ambiguous terms under consideration comprises of 2 verbs, 2 adverbs, 2 adjectives, and 6 nouns.

4. Results and Discussions

In this section, we present and discuss the results of experiments performed on our system for Marathi Word Sense Disambiguation. We test the proposed algorithm on our dataset described in the previous section.

For comparison purposes, we have also implemented the algorithm proposed by Kumari and Lobiyal [20] and tested it on our dataset. Kumari and Lobiyal have proposed to use the word embeddings of all words in the input sentence as context vector and the word embeddings of only the gloss of every sense as sense vectors for every sense. The results obtained are illustrated in the following Table1.

Case 1 is the methodology as proposed by Kumari and Lobiyal where Context Vector = Words in the input sentence and Sense Vector = Gloss of senses of ambiguous word [20]. Case 2 is our proposed methodology where Context Vector = Words in input sentence and Sense Vector = Gloss, Synonyms and Examples of senses of ambiguous word.

Table 1. Accuracy obtained for various word embeddings in both cases.

Case	FastText	IndicBERT	MuRIL
Case 1:	51%	38%	53%
Case 2:	69%	51%	53%

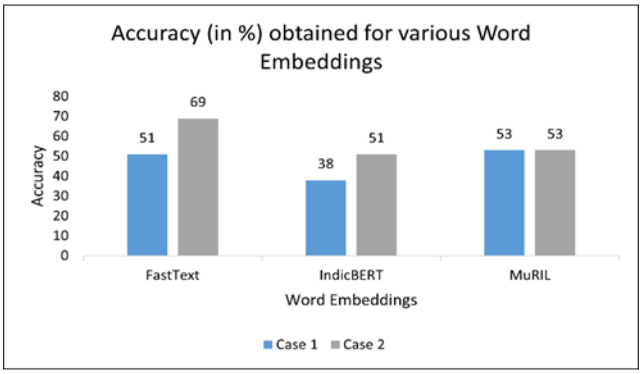

Fig. 2. Accuracy obtained for various Word Embeddings in Case 1 and Case 2.

As seen in Table1 and Figure 2, using FastText embeddings for case 2 gives the highest accuracy of 69%. The accuracy of IndicBERT also increases in case 2 as compared to case 1, whereas the accuracy of MuRIL remains the same in both cases. Also, the MuRIL model takes a comparatively long time to generate the required disambiguated sense.

5. Conclusion

This paper presents a methodology that employs word embeddings for Marathi Word Sense Disambiguation in an unsupervised manner. The Marathi WordNet is used for the purpose of retrieving the meanings, synonyms, and illustrative instances of the desired terms. The performance of the suggested system is enhanced by taking into account the gloss, synonyms, and instances associated with each meaning of the ambiguous word, in conjunction with the contextual information provided by the phrase. This improvement is seen in comparison to considering simply the gloss of each sense of the ambiguous word. The suggested methodology enables the disambiguation of nouns, verbs, adjectives, and adverbs. Nevertheless, the algorithm’s effectiveness is compromised in instances when a phrase has two consecutive words that possess many interpretations. To enhance the precision of Word Sense Disambiguation for the Marathi language, a viable technique may be to use of supervised learning. However, this method necessitates the availability of a sense-annotated corpus.

References

[1] Arora, H.S., Bhingardive, S., Bhattacharyya, P.: Detecting most frequent sense using word embeddings and babelnet. In: Proceedings of the 8th Global WordNet Conference (GWC), pp. 21–25 (2016)

[2] Lahoti, P., Mittal, N., Singh, G.: A survey on nlp resources, tools, and techniques for marathi language processing. ACM Transactions on Asian and Low-Resource Language Information Processing 22(2), 1–34 (2022)

[3] Eranpurwala, F., Ramane, P., Bolla, B.K.: Comparative study of marathi text classification using monolingual and multilingual embeddings. In: International Conference on Advanced Network Technologies and Intelligent Computing, pp. 441–452 (2021). Springer

[4] Bevilacqua, M., Pasini, T., Raganato, A., Navigli, R.: Recent trends in word sense disambiguation: A survey. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21 (2021). International Joint Conference on Artificial Intelligence, Inc

[5] Ransing, R., Gulati, A.: A survey of different approaches for word sense disambiguation. In: ICT Analysis and Applications: Proceedings of ICT4SD 2022, pp. 435–445. Springer (2022)

[6] Bhingardive, S., Singh, D., Rudramurthy, V., Bhattacharyya, P.: Using word embeddings for bilingual unsupervised wsd. In: Proceedings of the 12th International Conference on Natural Language Processing, pp. 59–64 (2015)

[7] Orkphol, K., Yang, W.: Word sense disambiguation using cosine similarity collaborates with word2vec and wordnet. Future Internet 11(5), 114 (2019)

[8] Popale, L., Bhattacharyya, P.: Creating marathi wordnet. The WordNet in Indian Languages, 147–166 (2017)

[9] Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM 38(11), 39–41 (1995)

[10] Petrolito, T.: A language-independent lesk based approach to word sense disambiguation. In: Proceedings of the 8th Global WordNet Conference (GWC), pp. 275–281 (2016)

[11] Lesk, M.: 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, pp. 24–26 (1986)

[12] Banerjee, S., Pedersen, T.: An adapted lesk algorithm for word sense disambiguation using wordnet. In: International Conference on Intelligent Text Processing and Computational Linguistics, pp. 136–145 (2002). Springer

[13] Bhingardive, S., Puduppully, R., Singh, D., Bhattacharyya, P.: Merging verb senses of hindi wordnet using word embeddings. In: Proceedings of the 11th International Conference on Natural Language Processing, pp. 344–352 (2014)

[14] Scarlini, B., Pasini, T., Navigli, R., et al.: With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3528–3539 (2020). The Association for Computational Linguistics

[15] Kumar, S., Kumar, S., Kanojia, D., Bhattacharyya, P.: “a passage to india”: Pretrained word embeddings for indian languages. In: Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), pp. 352–357 (2020)

[16] Gaikwad, V., Haribhakta, Y.: Adaptive glove and fasttext model for hindi word embeddings. In: Proceedings of the 7th ACM IKDD CoDS and 25th COMAD, pp. 175–179 (2020)

[17] Khanuja, S., Bansal, D., Mehtani, S., Khosla, S., Dey, A., Gopalan, B., Margam, D.K., Aggarwal, P., Nagipogu, R.T., Dave, S., et al.: Muril: Multilingual representations for indian languages. arXiv preprint arXiv:2103.10730 (2021)

[18] Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N., Bhattacharyya, A., Khapra, M.M., Kumar, P.: Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 4948–4961 (2020)

[19] Kharate, N.G., Patil, V.H.: Word sense disambiguation for marathi language using wordnet and the lesk approach. In: Proceeding of First Doctoral Symposium on Natural Computing Research: DSNCR 2020, pp. 45–54 (2021). Springer

[20] Kumari, A., Lobiyal, D.: Word2vec’s distributed word representation for hindi word sense disambiguation. In: International Conference on Distributed Computing and Internet Technology, pp. 325–335 (2019). Springer

[21] Bhingardive, S., Singh, D., Rudramurthy, V., Redkar, H., Bhattacharyya, P.: Unsupervised most frequent sense detection using word embeddings. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1238–1243 (2015)

[22] Laatar, R., Aloulou, C., Belguith, L.H.: Word sense disambiguation of arabic language with word embeddings as part of the creation of a historical dictionary. In: LPKM (2017)

[23] Laatar, R., Aloulou, C., Belguith, L.H.: Evaluation of stacked embeddings for arabic word sense

disambiguation. Computaci´on y Sistemas 27(2) (2023)

[24] Panjwani, R., Kanojia, D., Bhattacharyya, P.: pyiwn: a python based api to access indian language wordnets. In: Proceedings of the 9th Global Wordnet Conference, pp. 378–383 (2018)