Grenze International Journal of Engineering and Technology, June Issue

Contextual Embedding for Word Sense Disambiguation

Gauri Dhopavkar[1], Anant Gobade[2], Ayush Bhusari[3] and Samruddhi Joshi[4]
[1-4]Yeshwantrao Chavan College of Engineering, Nagpur, India
Email: gmdhopavkar@ycce.edu, anantgobade2003@gmail.com, bhusari.ayush22@gmail.com, samruddhijoshi2004@gmail.com

***Abstract*— Word Sense Disambiguation (WSD) requires users to identify the correct meaning of polysemous words through their surrounding text. The process of WSD becomes difficult for Marathi, which serves as a morphologically rich language with low resource availability, because there are few annotated corpora and its complex inflectional system, and lack of strong lexical resources. The paper shows a Marathi WSD system which uses linguistic processing to create transformer-based semantic models for its operation. The system first performs normalization and morphological analysis and lemma extraction before it creates contextual embeddings through the use of IndicBERT. The system fetches candidate gloss meanings from IndoWordNet, which it then transforms into semantic vectors to compare with sentence-level context vectors through cosine similarity for sense determination. The experimental results show that transfer learning effectively disambiguates both basic and advanced sentence constructions. This proves its usefulness for Indian language semantic analysis. The method demonstrates that NLP systems for low-resource domains achieve better results when they combine linguistic expertise with deep contextual embedding techniques.**

***Index Terms*— Marathi NLP, Word Sense Disambiguation, IndicBERT, Linguistic Preprocessing, Contextual Embeddings, Low-Resource Languages.**

I. INTRODUCTION

Natural Language Processing systems require precise word meaning understanding to function correctly because they work with languages which have multiple meanings depending on context. The Marathi language serves as a major Indo-Aryan tongue which more than 85 million people use. It remains underrepresented in computational resources. The Marathi language does not have large annotated datasets and domain-specific corpora, and high-quality pretrained language models which are available for English and other high-resource languages. The system operates with limited capabilities, which create obstacles for developing language understanding and generation tools. The main goal of Word Sense Disambiguation (WSD) in NLP involves selecting the right meaning for words which have multiple meanings based on their surrounding text. The process of Word Sense Disambiguation (WSD) needs to be accurate because it directly affects the performance of machine translation, and information retrieval, and speech recognition, and opinion mining, and conversational AI systems. The process of Word Sense Disambiguation (WSD) in Marathi and similar Indian languages faces multiple obstacles because these languages use inflectional morphology, and allow free word order, and create compound words, and multiple word senses share similar meanings. The traditional methods for Word Sense Disambiguation which use dictionary lookup, and rule-based heuristics, and co-occurrence statistics fail to identify the detailed contextual information [1].

Grenze ID: 01.GIJET.12.2.358_22

The development of transformer-based architectures together with contextual embeddings has enabled models to develop dynamic vector representations which change based on context. Thus, presenting an effective solution for semantic disambiguation in environments with limited resources. The multilingual transformer model IndicBERT enables learning of shared semantic spaces between Indian scripts, which enables models to understand context without needing large amounts of language-specific training data. The process of semantic alignment between sentence contextual embeddings and gloss embeddings from lexical resources becomes possible through the use of these models [2], [3].

Unlike existing Marathi Word Sense Disambiguation approaches that rely on static word embeddings, supervised classifiers, or unconstrained transformer representations, the proposed system introduces a linguistically constrained contextual embedding framework. The methodological novelty lies in the integration of morphological analysis and part-of-speech filtering prior to contextual embedding generation, followed by sentence-level alignment between IndicBERT-derived context embeddings and IndoWordNet gloss embeddings. This design enables precise sense discrimination without requiring supervised sense-labeled training data or task-specific fine-tuning, distinguishing the approach from earlier IndicBERT-based or gloss similarity methods used in Marathi WSD [4].

The study delivers three research results:

- Our research presents a new system which merges linguistic analysis with neural networks to perform Marathi word sense disambiguation.
- Our method uses gloss-based semantic similarity together with transformer embeddings to generate results.
- Our assessment reveals that contextual models serve as a solution for the restricted data access. This impacts languages with minimal resources.
- The proposed approach becomes clear through the following sections, which also show results from experimental testing and display system performance metrics.

II. Literature Review

The research field of Marathi Word Sense Disambiguation (WSD) investigates its development through three main approaches, which include linguistic resource development, machine learning techniques, and contextual embedding methods. The first research study focused on resource-based methods which employed Marathi WordNet as their main lexical database to extract sense definitions and glossary information. Ransing and Gulati discovered two primary obstacles which prevented lexical methods from achieving their desired outcomes. The researchers did not have enough annotated data sets because their study language contained sophisticated morphological elements [5]. Distributed word representations became popular which led to the development of unsupervised methods that used embedding techniques. The research team used FastText and Word2Vec embeddings to compare context–sense similarity which improved their results without requiring any manual annotation. The models faced two main problems because they could not handle extended complex sentences and their semantic understanding power was not detailed enough [6]. The supervised learning methods used classical classifiers, which included Naïve Bayes, SVM, Decision Tree, and Random Forest, to perform sense prediction tasks. The models achieved good accuracy results during testing yet their ability to apply learned information to new data stayed limited because the training set contained only a small number of labelled examples. The researchers applied semi-supervised bootstrapping techniques to reduce annotation costs by generating additional training examples through multiple applications of their starting dataset [7]. The research applied morphological preprocessing techniques which included lemma extraction and root-word mapping to decrease vocabulary size and improve embedding quality.

The research team combined lexical resources with clustering algorithms, including K-means, but the system faced two main obstacles. It struggled to identify clear sense boundaries, and no established Marathi WSD evaluation datasets existed. The development of WSD systems runs parallel to Marathi Named Entity Recognition (NER) research because both fields encounter identical language processing obstacles, which include agglutination and restricted data resources, and multiple meanings of words. The researchers found Marathi NLP resources to be extremely limited because of these language processing challenges [8]. The field of instruction tuning together with transformer-based modelling now enables large language models to develop contextual reasoning abilities through their training with instruction–response datasets. The system achieves its purpose when it receives high-quality data. This data contains broad semantic information [9]. The research shows that lexical methods together with statistical techniques solve only some problems, but transformer-based contextual modelling appears to be an effective solution for improving WSD performance with low-resource Marathi language data [10]. Although recent studies have explored contextual embeddings and lexical resources

independently for Marathi WSD, existing systems do not explicitly integrate linguistic constraint-based filtering with transformer-driven sentence-to-gloss semantic alignment. This gap motivates the proposed hybrid framework, which systematically combines morphological filtering with IndicBERT-based contextual similarity for unsupervised sense resolution.

III. Proposed Methodology

The proposed Marathi Word Sense Disambiguation system adopts a hybrid computational-linguistic design that integrates rule-based preprocessing with transformer-based contextual embeddings. The system functions through five main components which start with dataset formation and continue through text standardization, linguistic attribute extraction, sense candidate selection, and meaning resolution. The system needs to identify the specific definition. This approach differs from prior Marathi WSD systems by enforcing linguistic constraints before contextual embedding generation and by performing sentence-level context–gloss alignment rather than isolated word-level similarity. Fig. 1 shows the WSD core processing steps.

A. Dataset Construction

The lack of accessible standardized Marathi Word Sense Disambiguation datasets led to the creation of a specialized corpus for testing purposes. The research team collected ambiguous Marathi words from multiple sources which included news articles, Wikipedia pages, digital books, blogs, and public Marathi text corpora to achieve linguistic variety. The researchers chose ambiguous target words through three main selection criteria: (i) IndoWordNet needed to show at least two distinct meanings for each word, (ii) the words had to belong to open-class categories which mainly consist of nouns and verbs, and (iii) the team removed senses that appeared in archaic language or extremely rare usage or domain-specific contexts to make annotation more straightforward. Each dataset instance consists of a tuple which includes the input sentence together with the ambiguous target word and its corresponding part-of-speech tag.

Sense annotation involved human annotators who matched each sentence's target word meaning with the suitable IndoWordNet gloss definition. The annotation process followed a strict set of rules, which required annotators to base their decisions on the complete sentence meaning instead of depending on separate word definitions. The dataset maintained its reliability because it removed all ambiguous examples. These did not have definite sense category assignments. The final dataset contains sense assignments, which show high confidence levels, so they can effectively assess the performance of unsupervised word sense disambiguation (WSD) [11].

B. Text Preprocessing

The preprocessing stage converts unprocessed Marathi text into a uniform format. The system performs:

Tokenization: The hybrid regex tokenizer breaks down Devanagari text into individual segments which process both punctuation marks, numbers, and text that includes different writing systems. The system maintains token positional indices to enable correct alignment between POS tags and morphological attributes.

Stopword Filtering: The researchers created a list of Marathi stopwords, which they used to eliminate function words that do not add semantic value. The system kept Negation terms such as “नाही” and “नको” because they change the meaning of the text [12].

Unicode Normalization: The token normalization process uses NFC format to standardize Devanagari characters between decomposed and composed forms while it removes emojis and URLs, and replaces numbers with generic token representations.

Lemmatization: The system converts inflected words into their base form lemmas through a process which combines dictionary references with rules that remove word endings. The system needed this feature to process Marathi language rules. These create different forms based on gender, number, tense, and case.

C. Linguistic Feature Extraction

The input representation receives additional linguistic information through two separate modules, which add syntactic, and morpho-semantic attributes to the data.

Morphological Analysis: The morphological analyser which uses rules to enhance its analysis determines grammatical features including gender, and number, and tense by analysing suffix patterns and performing dictionary lookups. The system uses these features to remove gloss candidates which have contradictory semantic meanings.

Part-of-Speech Tagging: The statistical unigram POS tagger uses its tagging process to identify the syntactic categories which tokens belong to. The system uses POS information to limit sense retrieval operations, which focus on content words (nouns, and verbs). This avoids performing unnecessary calculations [13].

D. Candidate Sense Retrieval

The system performs ambiguity detection before it starts searching Marathi WordNet to find all possible meanings of the target word. Each sense entry contains three essential elements, which include gloss definitions, synonyms, and unique sense identifiers. The disambiguation process uses these gloss descriptions to create the semantic hypothesis space. This serves as foundation.

E. Contextual Embedding Generation

The semantic interpretation system functions through IndicBERT which serves as a multilingual transformer encoder that trained on multiple Indian language datasets. The model produces a large amount of contextual data from its operation. The system creates two sets of high-dimensional embeddings through its process, which combines the entire input sentence to generate the context vector, and each candidate gloss definition to produce sense vectors. The system allows word definitions to change when words show up in different textual contexts. The system operates differently from traditional static embedding models because it uses this approach.

F. Similarity-Based Sense Selection

The system calculates cosine similarity by comparing the context embedding with each sense embedding. The sense which achieves the highest similarity score becomes the most likely meaning of the text. The mathematical formula defines the winning sense S* as the sense which achieves the highest cosine similarity between the context vector A and each candidate sense vector B_n:

$$\text{Cosine}(\theta) = \frac{A.B}{\|A\| \times \|B\|}$$

$$S^* = \arg\max_n \text{Cosine}(A, B_n) \quad (1)$$

The score-based inference system applies mathematical principles to perform semantic matching. It operates independently from any need for supervised sense annotation.

G. Handling Sentence Complexity

The system tracks two different processing methods which it uses for sentences that have either one ambiguous token or more than one ambiguous token. The system solves basic cases through direct methods, yet it processes complex cases by analysing each token individually while using the entire sentence for context during each disambiguation step. The system prevents error spread through its design. This also keeps the original context intact during processing.

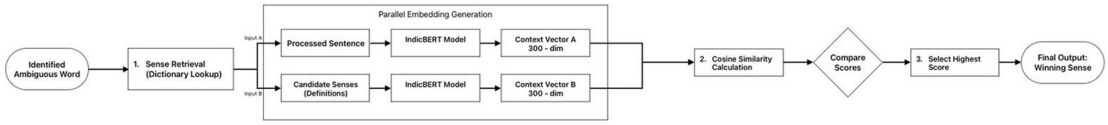

Figure 1. WSD Core Processing Steps

IV. Results and Discussion

The evaluation process for the Marathi Word Sense Disambiguation system involved testing it against a new evaluation dataset, which researchers developed by selecting data from their supervised dataset. The system needed evaluation to determine its capability of understanding semantic meanings, which show up in different linguistic structures and vocabulary variations.

A. Evaluation Setup

The dataset included 2,000 Marathi sentences which contained multiple meaning words that appeared in news articles, literary texts, and spoken communication. The dataset was divided into the following categories:
The dataset contains 600 basic sentences which consist of a single word that can mean different things.
The dataset contains 1400 complex sentences, which include multiple ambiguous words and complex nested clause structures.
The system evaluation process for performance measurement applied standard NLP assessment measures, which included accuracy, together with precision, recall, and F1-score. The system reached accurate predictions through its ability to link assigned senses with the exact meanings found in reference gloss interpretations.

B. Quantitative Outcomes

The proposed system correctly disambiguated 1,847 out of 2,000 ambiguous instances, yielding an overall accuracy of 92.35 percent. The system processed various sentence types, which resulted in different performance outcomes during its operation:

The system reached an accuracy level of 95.33 percent when it processed basic sentences.

Table I shows the system performance metrics of various approach. The system achieved 91.07 percent accuracy when it processed complex sentences. The system achieved an F1-score of 93.64% because precision and recall values both exceeded 92%.

To contextualize the effectiveness of the proposed approach, performance is compared with previously reported Marathi WSD systems that employ lexical, embedding-based, and supervised learning techniques. Since no standardized Marathi WSD benchmark exists, reported results from prior studies are used for reference rather than direct experimental replication.

TABLE I. SYSTEM PERFORMANCE METRICS

Approach	Methodology	Supervision	Reported Accuracy
Word2Vec-based WSD [6]	Static embeddings + cosine similarity	No	~78–82%
FastText-based WSD [6]	Subword embeddings	No	~80–85%
Supervised ML classifiers [2], [4]	NB, SVM, RF	Yes	~85–88%
Semi-supervised bootstrapping [5]	Iterative labeling	Partial	~88–90%
Proposed Method	Linguistic filtering + IndicBERT + gloss similarity	No	92.35%

The proposed system achieves higher accuracy than previously reported unsupervised and semi-supervised Marathi WSD approaches, while avoiding the dependency on annotated sense-labeled training data.

C. Error Analysis

The research team separated all misclassified cases into two main categories after they studied the different failure patterns which occurred in the system.

Detection and Identification Errors: The system failed to identify ambiguous tokens because they appeared in uncommon morphological patterns, and their inflected forms were not common. The system encountered these situations most often when processing conversational text and informal writing, which used unconventional suffix patterns.

Semantic Resolution Errors: Semantic resolution errors primarily occurred when multiple IndoWordNet glosses exhibited highly overlapping semantic descriptions, resulting in minimal cosine similarity margins. For example, in the sentence “तो बँकेत काम करतो”, the ambiguous word “बँक” can refer to a financial institution or a riverbank. While the broader sentence context indicates the financial sense, the gloss definitions for both senses contain overlapping occupational and location-related terminology, leading to incorrect sense selection in a small number of cases.

The restriction appeared in metaphorical expressions which included the phrase "त्याच्या बोलण्यात धार आहे" because the word "धार" expresses both physical edges and metaphorical sharpness. The glosses in IndoWordNet provide only basic separation between concrete and abstract meanings, which makes it challenging to separate their meanings through embedding-based semantic methods.

D. Qualitative Observations

Three qualitative findings emerged during analysis:

Writers create complex sentences by adding multiple subordinate clauses, which results in decreased semantic information within their sentences. The multiple subordinate clauses which writers add to their sentences create less clear semantic information. This makes it tough to identify their candidate gloss vectors.

Impact of Morphology: Researchers can reduce the number of possible meanings which nouns and verbs can have through the process of word analysis using lemmatization and POS tagging. The research findings indicate that Marathi language users obtain better results from morphological preprocessing because their words include multiple inflectional variations.

Sense Inventory Limitations: The gloss definitions in IndoWordNet demonstrate restricted detail levels, which affect particular situations. The disambiguation process would become more effective through expanded gloss resources because the differences between literal and metaphorical meanings, and other semantically related terms, showed minimal cosine margin variations.

E. Discussion

From a computational perspective, the proposed framework remains scalable for practical use. For each sentence, the system performs a single forward pass through IndicBERT to obtain the contextual embedding, followed by similarity comparisons with a limited set of candidate gloss embeddings. Since the number of senses per ambiguous word in IndoWordNet is relatively small, the similarity computation scales linearly with respect to the number of candidate senses. Gloss embeddings can be precomputed and cached, significantly reducing inference time. As a result, the approach is suitable for batch processing of large corpora and can be adapted for near real-time NLP applications with moderate computational resources.

The research findings demonstrate that hybrid systems which combine linguistic elements with neural networks achieve better disambiguation results than rule-based systems, or statistical methods, when working with languages that have complex morphological structures.

V. Conclusion

The research introduces a hybrid framework for Marathi Word Sense Disambiguation which merges linguistic preprocessing techniques with contextual transformer embeddings to identify the correct meaning of words with multiple senses. The system resolves problems which emerge from complex word inflections and insufficient linguistic data through its combination of normalization techniques, morphological analysis, lemma extraction, and IndicBERT-based semantic matching. The proposed approach demonstrated superior performance during testing with a selected dataset which included both basic and advanced sentence structures. The results showed that contextual embeddings help users distinguish word meanings when working with languages. These languages have limited resources.

The main finding of this research shows that linguistically constrained contextual embedding alignment methods solve Word Sense Disambiguation problems for Marathi, which is a low-resource language, without needing any supervised sense-labeled data. The system achieves strong performance through its combination of morphological filtering with part-of-speech constraints and IndicBERT-based sentence-to-gloss similarity for both basic and complex sentence structures. The system encounters two main challenges because it cannot process lengthy sentences with complex syntax and it struggles to identify subtle sense differences because IndoWordNet glosses do not provide enough detail. The upcoming research will concentrate on creating better sense inventories while testing two different contextual modelling techniques, which use hierarchical structures and discourse-based information to improve semantic discrimination.

Acknowledgment

The authors want to express their deep appreciation to Gauri Dhopavkar because she provided essential guidance. The research team found guidance through this process which helped them complete their research tasks. The author found ongoing inspiration through field experts who provided essential feedback. This helped them develop their work. The team members reached better results because they enhanced their expertise in technical areas and their ability to organize structure. The team members obtained crucial direction through their discussions which enabled them to create their fundamental ideas while they worked through various problems which emerged during the design process.

The authors express their gratitude to their academic environment because it provided essential support which enabled their research development. The project reached its deadline through correct application of research facilities, computational resources, and by finding suitable academic resources. The authors deeply appreciate the teamwork, which produced valuable feedback that guided their entire research process.

References

[1] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, et al., “Instruction Tuning for Large Language Models: A Survey,” arXiv preprint arXiv:2308.10792, 2023.

[2] R. Ransing and A. Gulati, “Word Sense Disambiguation for Marathi Language Using Supervised Learning,” in Proc. ICON, 2023.

[3] R. Ransing and A. Gulati, “Unsupervised Word Sense Disambiguation for Marathi Using Word Embeddings,” Int. J. Intell. Syst. Appl. Eng. (IJISAE), vol. 12, no. 3, pp. 1374–1380, 2024.

[4] R. Ransing, A. Gulati, and P. Kulkarni, “Supervised Marathi WSD Using Multiple Classifiers: Naïve Bayes, SVM, Decision Tree, Random Forest, and Logistic Regression,” Int. J. Sci. Res. Sci. Technol. (IJSRST), vol. 10, no. 10, pp. 345–351, 2023.

[5] G. Gahankari, M. Jadhav, and V. Patil, “Semi-Supervised Bootstrapping Approach for Marathi Word Sense Disambiguation,” Int. J. Sci. Res., 2023.
[6] Choudhari, P. Deshmukh, and R. Sharma, “Embedding and Morphology-Based Methods for Marathi WSD Using Word2Vec and Root-Word Preprocessing,” in Proc. Int. Conf. Comput. Linguistics for Indian Languages, 2022.
[7] P. Khot and V. Bhosale, “Hybrid Approaches for Word Sense Disambiguation Using Machine Learning and Lexical Knowledge,” Int. J. Comput. Appl., vol. 175, no. 18, pp. 12–18, 2020.
[8] J. Wei, M. Bosma, V. Zhao, K. Guu, A. Yu, B. Lester, N. Du, A. Dai, and Q. Le, “Finetuned Language Models Are Zero-Shot Learners,” arXiv preprint arXiv:2112.01364, 2021.
[9] S. Patil, R. Kulkarni, and A. Deshpande, “Challenges in Marathi Named Entity Recognition and Implications for WSD,” in Proc. Workshop on Indian Language NLP, 2021.
[10] S. Mishra, D. Khashabi, C. Baral, and H. Hajishirzi, “Cross-Task Generalization via Natural Language Crowdsourcing Instructions,” in Proc. NAACL, 2021, pp. 122–133.
[11] K. Scaria, H. Gupta, S. Goyal, S. Sawant, and S. Mishra, “InstructABSA: Instruction Learning for Aspect-Based Sentiment Analysis,” in Proc. Int. Conf. Natural Language Processing, 2022.
[12] Y. Liu, J. Wu, and T. Chen, “Instruction Tuning for Few-Shot Aspect-Based Sentiment Analysis,” arXiv preprint arXiv:2210.06629, 2022.
[13] Kumar and S. Singh, “Instruction Tuning with Retrieval-Based Example Ranking for Sentiment and WSD Tasks,” in Findings of the Association for Computational Linguistics (ACL), 2024, pp. 3562–3573.