Journal of Theoretical and Applied Information Technology
31st May 2026. Vol.104. No.10
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A MEDOID-BASED SEMI-SUPERVISED CLUSTERING APPROACH FOR WORD SENSE DISAMBIGUATION IN TELUGU

DURGAPRASAD PALANATI [1], SUNITHA K.V.N [2], PADMAJA RANI B [3]

[1] Research scholar, Computer Science and Engineering, JNTUH, Hyderabad, India
[2] Professor, Computer Science and Engineering, BVRIT HYDERABAD College of Engineering for Women, Hyderabad, India
[3] Professor, Computer Science and Engineering, JNTUCEH, Hyderabad, India
E-mail: [1]dp.cse5@gmail.com, [2] k.v.n.sunitha@gmail.com, [3]padmaja_jntuh@jntuh.ac.in

ABSTRACT

Word Sense Disambiguation (WSD) is critical component in all the Natural Language Processing related tasks. Due to limited availability of sense annotated corpus and morphologically richness of Telugu language its' very challenging to develop WSD systems. This paper proposes a semi-supervised clustering technique for Telugu WSD that minimizes the utilization of large annotated corpora. The proposed method uses IndicBERT-based sentence embeddings to find contextual semantics. A novel seed selection approach based on sum-of-squared error (SSE) is proposed to make sure the initialization of sense clusters, followed by a formula-based medoid selection mechanism and an adaptive similarity thresholding scheme for sense propagation. The approach is evaluated on a manually annotated corpus and also compared with baseline methods Most Frequent Sense, Most Common Sense and simple K-medoid. The performance metrics show that the proposed system surpasses the baseline approaches. This shows the effectiveness of proposed approach for Telugu WSD and applicability to Information retrieval tasks.

Keywords: *Contextual Sentence Embeddings, Lexical Ambiguity Resolution, Low-Resource Language Processing, Medoid-Based Clustering, Semi-Supervised Learning, Word Sense Disambiguation.*

1. INTRODUCTION

In natural language, meaning of a word varies depending on context. And this leads to ambiguity at the lexical level. As Word Sense Disambiguation (WSD) resolves lexical ambiguity, many Natural Language Processing (NLP) applications like machine translation, information retrieval, information extraction, text mining [2][3], etc., use it. For example, the Telugu word *'xvAramu'*[1] has two meanings[2] (senses) *'vAkili / praveSaM'* (entry) and *'upAyamu'* (a medium or way). Here, lexical sample WSD assigns the appropriate sense for the word present in particular context computationally. In the sentence, *'niReXiwa prAMwAllokis AmAnyulaki **xvAramu** niReXaM' (Civilians are not permitted to visit restricted areas)*, the WSD algorithm assigns *'praveSaM'* as the sense of the word ***'xvAramu'***. This shows the role of context around a polysemous word. Better context representation schemes are important for WSD systems [1][9]. Hence in [1], an efficient technique for context representation has been proposed. In this technique, 'sense-specific' [1] representation for words in semantic space is used to measure semantic similarity. The promising results reported in [1] motivated us to use context-based vector representation of the words. So, due to its vast application, many approaches based on supervised, unsupervised, semi-supervised, and knowledge-based have been developed (we suggest the reader refer [3-5] for an exclusive survey on various approaches as it is out of the scope of this paper). The primary motivation for us to develop a semi-supervised clustering approach for WSD in Telugu is as follows:1) paucity of resources (Ontology trees, Lexical knowledge bases to generate rules, and sense annotated corpora to train classifiers) in the Telugu language that are obligatory for knowledge-based and supervised approaches for WSD, 2) As Telugu is a free word-order (structure of the sentence is not rigid) language we need some new techniques for Telugu WSD apart from the existing English(sentence's structure is rigid) WSD systems.

[1] Hereafter we use WX-notation to represent Telugu words

[2] Meanings are taken from Telugu dictionary https://www.andhrabharati.com/dictionary/

The contributions of this work are as follows.

1. **A medoid-based semi-supervised clustering approach** for WSD in Telugu is proposed which minimizes the dependency on large sense annotated corpora. This proposed technique is explained in detail in section 3.
2. **A novel seed data selection strategy based on sum-of-squared error (SSE)**, which ensures initial sense annotated instances are well diversified and effectively represent clustering semantically.
3. **A formula-based medoid selection approach**, guaranteeing that most representatives identified within clusters are semantically related thus eliminating the randomness in medoid initialization.
4. **For label propagation an adaptive similarity threshold mechanism** is designed to assign sense to unlabelled instances based on context similarity.
5. Experimental validation stating improvements over baseline methods.

The remainder of this paper is organized into Sections 2–7. Existing research on WSD discussed in Section 2. Section 3 outlines system overview followed by proposed methodology in Section 4. Experimental setup mentioned in Section 5 and result analysis in Section 6. Section 7 concludes and mentions future directions of research in WSD.

2. RELATED WORK

The approaches to WSD categorized based on the resources utilized [2]. These are Supervised approaches use large annotated corpora, unsupervised use unlabeled corpora, semi-supervised starts with small seed dataset, and knowledge-based utilize lexical reources. The state-of-the-art supervised WSD system [6] leverages word embeddings besides the standard WSD features. This system has shown better results by integrating word embeddings using the exponential decay strategy [6]. The WSD system used Bi-LSTM network to generate word embeddings [7]. Support Vector Machine (SVM) based classifier for WSD developed in [8] relies on features such as parts-of-speech tags, and pre-trained word embeddings. The WSD system [9] utilized BERT [10] based word embeddings and reported improved WSD performance. In [11], distance-based and word frequency-based coefficients were used to generate word embeddings for voting schemes. The supervised WSD systems discussed so far are suitable for resource-rich language. In recent years, WSD for Telugu has gained more popularity because most people prefer searching in their local languages. But, matured resources in local languages are unavailable, so we need to develop unsupervised and semi-supervised approaches to WSD for resource-poor languages (in our work, we concentrate on the Telugu language). Some semi-supervised WSD systems use graph-based and rule-based techniques. A small sense-labeled data set was used by the semi-supervised WSD system [12]. The Word sense induction (WSI) technique created this seed data set. Single-system and ensemble WSI developed in [12] for WSD. The ensemble WSI-based disambiguation reported higher accuracies because of linear kernel SVM (Support Vector Machine). The label-propagation (LP) graph transfers the senses from the labeled seed nodes to the large set of unlabeled sentences. Ensure that each vertex is connected to at least 10 other vertices to avoid sparse connectivity problems [13]. After enforcing this constraint LP algorithm with LSTM language, this LP+LSTM model-based WSD system reported improved performance [13]. The semi-supervised WSD system proposed in [14] also uses a small sense-annotated corpus to generate the context-based list. This system had given the best results for Hindi and Marathi languages. In [15], Bi-LSTM based context representation was proposed. Semi-supervised WSD system [15] used Bi-LSTM language model and LP technique. The semi-supervised graph-based WSD system [16] exploits all the semantic information in the context. Graph centrality measures the semantic similarity between the nodes. The Telugu WSD system developed in [20] used a decision list algorithm using the rules of the form <feature value, sense, score>. This system used a Maximum likelihood estimate Score based on the frequency of sense. In [21], two word and three-word disambiguation rules were used. These rules were generated using POS(Parts-of-Speech) tags. These two WSD systems [20-21] are rule-based and the drawback of these is they need lot of rules which are difficult to create. There was argument structure-based WSD systems for Telugu verb. These WSD systems proposed in [22-23] used a dictionary of features like human, animate, concrete, and edible. But these systems didn't utilize all the semantic features of the surrounding words. The context-based WSD developed in [24] measures the similarity between context and word bag. Collection of surrounding words within a specific window stored in context bag. Word bag is a collection of

words in the example sentences of the words semantically (hypernym, hyponymy, synonym, etc.,) related to the target ambiguous word. After creating these two bags, the WSD algorithm calculates the cosine similarity between these two bags to predict the correct sense. The unsupervised graph based WSD approach proposed for Hindi language in [25] utilized Hindi Wordnet. In this work sense selection is done ensuring graph centrality measures and they claimed PageRank based model achieved high performance. A supervised WSD approach proposed in [29] addressed the lexical ambiguity in machine translation. This method utilized Naïve Bayes classifier with collocation-based contextual features to assign correct sense to the ambiguous word. Another study [30] proposed semantic similarity-based WSD for Telugu, this measures cosine similarity between context and sense vectors. These vectors are computed using IndicBERT word embeddings. Thus, the highest similarity scored sense will be most appropriate sense of ambiguous word. Despite these efforts, most existing approaches heavily rely on large labeled corpora and lexical resources. To overcome these limitations, we have designed WSD system which minimizes the dependency on these resources.

3. SYSTEM OVERVIEW

This section formalizes the core components of the proposed WSD system. The system is the integration of context representation and clustering-based learning. The semi-supervised framework of this WSD system operates well for low-resource languages. First, the system processes the context around the ambiguous word and represents the contextual semantics in a dense vector format i.e., sentence embeddings. IndicBERT model [19] is used to generate these sentence embeddings which helps in identifying the semantic relationships. The overall architecture consists of 4 phases: generation of sentence embeddings, initialization of seed dataset, clustering based on sense, and label propagation. The learning process starts with a small manually annotated sentences. These initial seed instances provide basic semantics of the target ambiguous word. For effective cluster formations we ensured the selected seed dataset is well-representative and diverse.

After initializing seed dataset, the system with the help of medoid-based clustering forms sense-specific clusters. For interpretability and robustness, the actual sentences are treated as medoids. The unlabeled sentence is assigned to a cluster based contextual similarity between sentence embedding and cluster representative.

To make sure the reliable sense assignments, the system employs similarity-based approach for label propagation. Only after satisfying a threshold value of semantic similarity the unlabeled sentences are assigned to a specific-sense cluster. In this way the without manual annotation the system creates a sense-labeled corpus.

Finally, the system selects a correct sense for target word. The architecture of proposed system can be applied any other low-resource languages. The semi-supervised clustering with the help of contextual embeddings offers an effective and scalable solution for Telugu Word Sense Disambiguation.

4. PROPOSED METHODOLOGY

This section presents a framework for WSD which is based on semi-supervised clustering. The proposed system hardly relies on large sense-annotated corpora but assigns correct sense to the target ambiguous word. The combination of contextual embeddings, selective medoid-based clustering and adaptive similarity threshold strategy makes the proposed system for effective label propagation. The complete process flow of the proposed system illustrated in figure 1.

4.1 Sentence embeddings as context

To resolve lexical ambiguity especially in free word order languages like Telugu effective representation of context is very important. In the proposed approach, pre-trained IndicBERT model is employed to generate sentence level embeddings. This model encodes the contextual semantics of polysemous word in these embeddings. IndicBERT model effectively represents context especially for Indian languages. The configuration settings of this model as per the work in [19,30] and mentioned in Apendix A.

To generate sentence embeddings, we have considered the ***last2avg*** strategy proposed in [28], which averages the last two layers of BERT architecture (in our case IndicBERT). This strategy generates semantically meaningful embeddings. For clustering and similarity computations the sentence is represented as 768-dimensional vector. The initial seed data set must be selected carefully such that it should be diverse to get effective clustering results [17]. Improper initialization of seed dataset leads to unreliable label assignments.

To overcome this issue, a novel strategy for seed selection based on the Sum-of-Squared Error. The strategy is as follows:

- In this, labeled sentence embeddings are the members of the seed data set.
- Measure the cosine similarity of vectors (sentence embeddings in 768-dimension) A,B

using eq (1). The cosine similarity range is [0,1].

$$\textbf{Similarity}\ (\mathbf{A},\mathbf{B}) = \frac{\mathrm{A}.\mathrm{B}}{\|\mathrm{A}\|\|\mathrm{B}\|} \qquad (1)$$

- The sum-of-squared error E between the members and their respective is measured using eq (2).

$$\boldsymbol{E} = \sum_{i=1}^{n} \sum_{sl \in C_i} sim(sl, m_i) \qquad (2)$$

Where, 'n' is the degree of polysemy of a particular word also the number of clusters. 'sl' be the labeled sentence embedding and 'm_i' is the medoid of cluster 'C_i'. And *sim()* is calculated from eq (1).

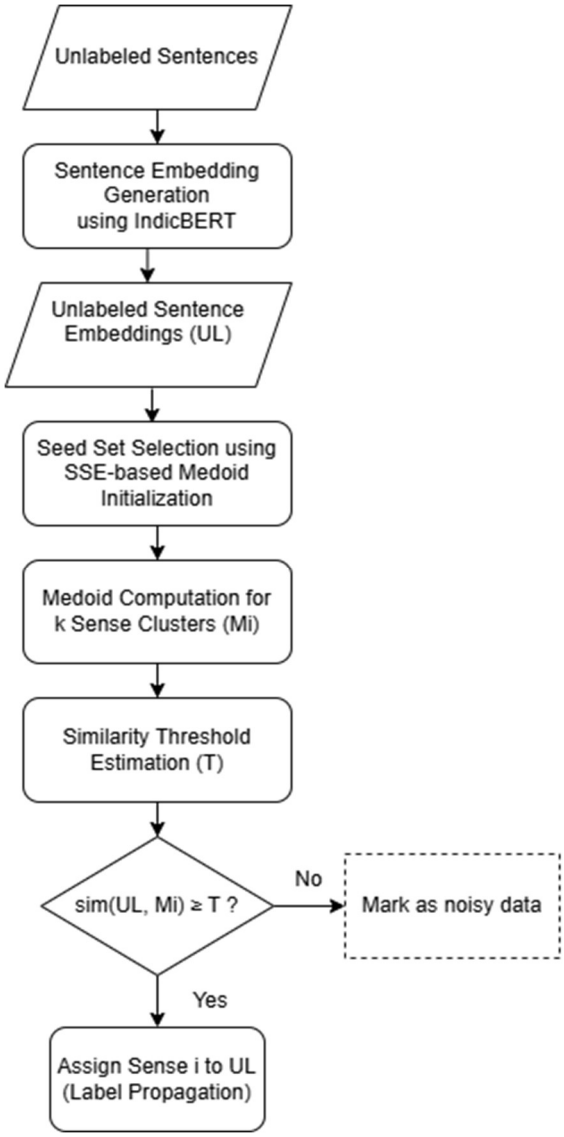

Figure 1 Process flow of proposed system

- Let us consider an ambiguous word ***'ila'*** with a frequency of 6308 and degree of polysemy=3. Figure2 shows the sentence embeddings in 2-D space (used PCA[3] to convert 768-D vector into 2-D).
- Now, we select a subset(cluster) of size not less than ten from sentence embeddings such that the value of 'E' (sum-of-squared error calculated using eq 2) is minimum. And manually annotate the sentences in these clusters. Figure 3 shows an example of seed clusters selected from the sentence embeddings shown in figure 2.

[3] Principal Component Analysis

- Use this seed data set to propagate the sense labels to the remaining un-annotated sentences, and the final clusters look like figure 4.
-

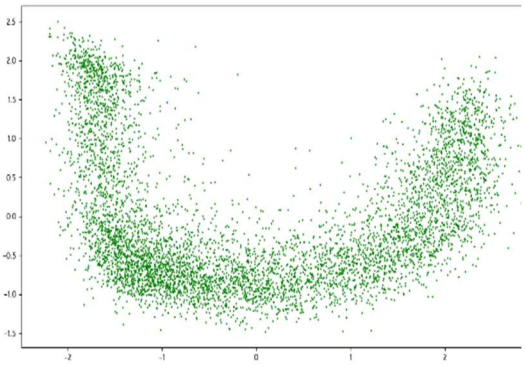

Figure 2. Sentence embeddings in 2-D space

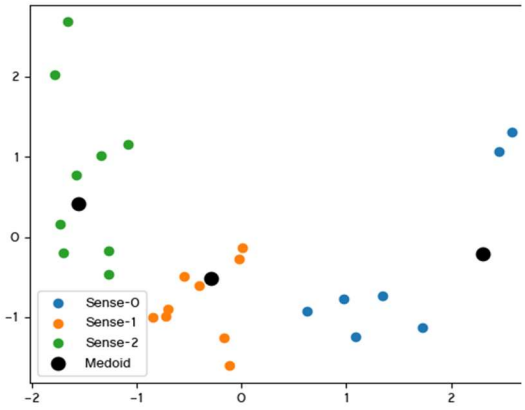

Figure 3. Seed set example-1

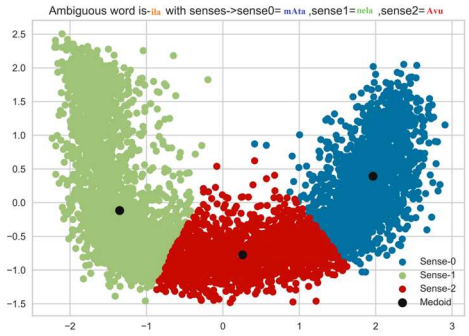

Figure 4. Sense clusters after completion of annotation

4.3 Formula-based Medoid Selection

In high dimensional semantic spaces, random initialization of medoids in k-medoid clustering algorithm may lead to low clustering quality. To eliminate randomness, we have proposed a formula-based medoid selection strategy. In this strategy, we have to select an object more similar to others in that cluster instead of randomly selecting a cluster member as a medoid. The object (sentence embedding) with the maximum Sum of the similarities to the remaining objects in the respective cluster will be the medoid of that cluster, and we find this object 'O_i' using eq (3).

$$i = \text{argmax}_i \sum_{j=1}^{n} \textbf{Similarity}\,(O_i, O_j) \qquad (3)$$

For i≠j, i=1,…n

4.4 Adaptive Similarity Threshold Estimation

After formation of initial clusters and selecting initial medoids, a similarity threshold value must be defined. This value is crucial in assigning the unlabeled sentence to a sense-specific cluster. For reliable assignment, instead of fixed threshold value an adaptive similarity threshold mechanism is proposed.

If the similarity between cluster medoid 'm' and the new object 'O' meets or exceeds the threshold value(Φ_{sim}), then 'O' will be assigned to the respective cluster. In our work, we find the similarity threshold value (Φ_{sim}) using the process proposed in [18] with some modifications. We preferred the cosine similarity of objects over distance measures (Minkowski, Euclidean, Manhattan distance) [18].

Our proposed process for selecting similarity threshold is as follows:

- Most clusters with a minimum of five objects are significant [27]. So, we initially consider the dense region of a minimum of ten from the annotated seed data set.
- Find the similarities from medoid to the remaining members of the cluster using eq (1).

Next, we propose two methods to determine threshold value:

a. Then find the average of similarities using eq (4), i.e., ***the similarity threshold value (Φ_{sim_AVG})***.

$$\Phi_{sim\_AVG} = \frac{\sum_{i=1}^{n} sim(O_i, O_m)}{n} \qquad (4)$$

Where 'O_m' is medoid of the cluster and 'n' number of sentence embeddings ('O_i') in the cluster

b. Consider top five similarity values as threshold values ***similarity threshold values.***

This adaptive similarity threshold mechanism enables the system to track sense-specific semantic variation and reduces noisy assignments during label propagation.

4.5 Sense Label Propagation

The sense labels are assigned to unlabeled sentences using selective medoids and adaptive threshold similarity. By comparing the unlabeled sentence embeddings with the cluster medoid, if the similarity exceeds the threshold value, then only the unlabeled sentence is assigned to that cluster.

To reduce the error propagation and improve the WSD accuracy similarity-based constraints are used. These constraints ensure only most semantically related sentences are clustered together.

5. EXPERIMENTAL SETUP

This section describes the dataset preparation, sense inventory creation, annotation strategy and experimental configuration to evaluate the proposed medoid-based semi-supervised WSD system for the Telugu language.

5.1 Sense Inventory Construction

We have created a sense inventory by collecting ambiguous words from various Telugu dictionaries[4]. The average degree of polysemy of the total 8200 words is approximately 3.24. This indicates that most words have more than one sense which represents complex bottleneck for Word Sense Disambiguation.

5.2 Corpus Description

The raw corpus is collected from Wikipedia[5], IITH Hyderabad[6], Kaggle[7] , and some news websites[8]. This text corpus has the collection about various domains such as news, stories, general knowledge and formal writings.

5.3 Annotation Process

The sentences containing ambiguous word are extracted from the raw corpus. For each sense of ambiguous word, minimum 10 sentences were manually annotated.

5.4 Experimental Configuration

The representation of sentence is generated by the IndicBERT model [19]. The sentence embeddings generated by this model are vectors of 768-dimension. We measure cosine similarity between the sentence embeddings. This tells the contextual similarity. The average threshold (Φ_{sim_AVG}) is set to 0.72 which is empirically determined. The values of Φ_{sim1}, Φ_{sim2}, Φ_{sim3}, Φ_{sim4}, Φ_{sim5} varies for each ambiguous word depending on semantic distribution. Finally, set degree of polysemy as the cluster count.

6. RESULTS AND ANALYSIS

Experimental results of proposed system and their comparisons with standard baseline approaches are discussed in this section.

6.1 Baseline Methods

We have evaluated the proposed system against the following widely used baseline approaches [26]:

- **Most Frequent Sense (MFS):** Sense with highest frequency is assigned to an ambiguous word.
- **Most Common Sense (MCS):** An unsupervised MFS which predicts most likely sense from unlabelled data.
- **Simple K-Medoid:** A clustering-based baseline without the proposed seed dataset selection, formula-based medoid formulation, and adaptive thresholding mechanisms.

6.2 Evaluation Metrics

The following metrics are used to evaluate performance of WSD system:

- **Accuracy (ACC):** Measures overall correctness of sense assignments.
- **Macro-averaged F1-score (MACROAVG):** Treats all senses equally and highlights performance on minority senses.
- **Weighted F1-score (WHTAVG):** Accounts for class imbalance by weighting each sense according to its frequency.

6.3 Quantitative Results

The comparative performance of the proposed approach against baseline methods shown in Table 1. The figure 5 visually shows that the proposed k-medoids system consistently achieves high performance, especially in case of accuracy and macro-F-1 score. Thus, it indicates improvement in WSD accuracy compared to baseline.

Table 1. Performance evaluation against baseline techniques

[4] https://andhrabharati.com/dictionary/- (searched for telugu-to-telugu meanings)

[5] https://te.wikipedia.org/wiki/

[6] https://ltrc.iiit.ac.in/showfile.php?filename=downloads/sentiraama/

[7] https://www.kaggle.com/sudalairajkumar/telugu-nlp

[8] https://www.andhrajyothy.com/, https://www.bbc.com/telugu

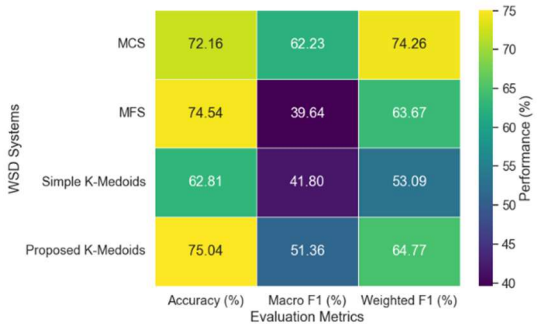

Figure 5. Performance comparison of WSD systems

6.4 Discussion

The experimental results show that the proposed medoid-based semi-supervised clustering approach is effective for Telugu WSD. **SSE-based seed selection strategy**, ensures diversity of initial seed dataset selection. And the cluster stability is maintained with the help of **formula-based medoid selection,** which reduces the randomness in medoid selection. The improvement in metrics of proposed KMED system over simple KMED has proven the importance of proposed seed selection and medoid initialization strategies. Even though the proposed system utilizes small labelled data set it performs better than baseline methods. Based on these results we can say the proposed system is suitable for resource-poor languages and applicable for tasks such as information retrieval [31-32].

7. CONCLUSION AND FUTURE WORK

In this work, we presented an improved version of simple medoid-based clustering approach for WSD system for data-scarce environments and complex morphological structured languages such as Telugu. This system's main feature is that it utilizes small annotated corpora. IndicBERT model is used to generate sentence embeddings. A novel strategy based on sum-of-squared error (SSE) is proposed for initialization of seed dataset. Formula-based medoid selection is done to ensure effective initial clusters. Robust performance for WSD is achieved even in poor resource-constrained set up because sense propagation started from carefully selected seed dataset.

We have conducted experiments on a large Telugu corpus containing 8200 ambiguous words with average polysemy degree of approximately 3.24. The results demonstrate the effectiveness of proposed method. The experimental results show that proposed approach is scalable and can be adapted to any resource poor languages.

Despite the promising results of proposed system, there are some open challenges to be addressed. First, the current work focused on nouns which can be adapted to other parts-of-speech. Second, the performance for highly polysemous words can be improved by considering dynamic context size. Third, larger multi-domain corpora may be used to evaluate the proposed system for domain robustness. Finally, to further increase the WSD accuracy some basic lexical knowledge from resources such as IndoWordNet can be incorporated.

SYSTEM	ACCURACY (%)	MACRO AVG F-1 (%)	WHT AVG F-1 (%)
MCS	72.16	62.23	74.26
MFS	74.54	39.64	63.67
Simple KMED	62.81	41.80	53.09
Proposed KMED	75.04	51.36	64.77

REFERENCES:

[1] Reisinger, Joseph, and Raymond Mooney. "Multi-prototype vector-space models of word meaning." In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 109-117. 2010.

[2] "Word sense disambiguation: The state of the art." *Computational Linguistics* 24, no. 1 (1998): 1-40.

[3] Bevilacqua, Michele, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. "Recent trends in word sense disambiguation: A survey." In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conference on Artificial Intelligence, Inc, 2021.

[4] Singh, Himanshu, and Pushpak Bhattacharyya. "A survey on word sense disambiguation." *ACM Comput. Surv.(CSUR)* (2019).

[5] Iacobacci, Ignacio, Mohammad Taher Pilehvar, and Roberto Navigli. "Embeddings for word sense disambiguation: An evaluation study." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 897-907. 2016.

[6] Kågebäck, Mikael, and Hans Salomonsson. "Word sense disambiguation using a

bidirectional lstm." *arXiv preprint arXiv:1606.03568* (2016).

[7] Papandrea, Simone, Alessandro Raganato, and Claudio Delli Bovi. "Supwsd: A flexible toolkit for supervised word sense disambiguation." In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 103-108. 2017.

[8] Hadiwinoto, Christian, Hwee Tou Ng, and Wee Chung Gan. "Improved word sense disambiguation using pre-trained contextualized word representations." *arXiv preprint arXiv:1910.00194* (2019).

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4171–4186.

[10] Fahandezi Sadi, Majid, Ebrahim Ansari, and Mohsen Afsharchi. "Supervised word sense disambiguation using new features based on word embeddings." *Journal of Intelligent & Fuzzy Systems* 37, no. 1 (2019): 1467-1476.

[11] Başkaya, Osman, and David Jurgens. "Semi-supervised learning with induced word senses for state of the art word sense disambiguation." *Journal of Artificial Intelligence Research* 55 (2016): 1025-1058.

[12] Yuan, Dayu, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. "Semi-supervised word sense disambiguation with neural models." *arXiv preprint arXiv:1603.07012* (2016).

[13] Rani, Pratibha, Vikram Pudi, and Dipti Misra Sharma. "Semisupervied Data Driven Word Sense Disambiguation for Resource-poor Languages." In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pp. 503-512. 2017.

[14] Li, Z. H. I., F. A. N. Yang, and Yaoru Luo. "Context embedding based on Bi-LSTM in semi-supervised biomedical word sense disambiguation." *IEEE Access* 7 (2019): 72928-72935.

[15] Abdalgader, Khaled, and Aysha Al Shibli. "Context expansion approach for graph-based word sense disambiguation." *Expert Systems with Applications* 168 (2021): 114313.

[16] Abdalgader, Khaled, and Aysha Al Shibli. "Context expansion approach for graph-based word sense disambiguation." *Expert Systems with Applications* 168 (2021): 114313.

[17] Sajidha, S. A., Siddha Prabhu Chodnekar, and Kalyani Desikan. "Initial seed selection for K-modes clustering–a distance and density based approach." *Journal of King Saud University-Computer and Information Sciences* 33, no. 6 (2021): 693-701.

[18] Kelkar, Bhagyashri Abhay, Sunil F. Rodd, and Umakant P. Kulkarni. "Estimating distance threshold for greedy subspace clustering." *Expert Systems with Applications* 135 (2019): 219-236.

[19] Kakwani, Divyanshu, Anoop Kunchukuttan, Satish Golla, N. C. Gokul, Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. "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.

[20] Palanati Durga Prasad, and Ramakrishna Kolikipogu. "Decision list algorithm for word sense disambiguation for TELUGU natural language processing." *Int. J. Electron. Commun. Comput. Eng* 4, no. 6 (2013): 176-180.

[21] Sreedhar, J., S. Viswanadha Raju, and A. Vinaya Babu. "Two-Word and Three-Word Disambiguation Rules for Telugu Language Sentences: A Practical Approach." *Global Journal of Computer Science and Technology* (2014).

[22] http://hdl.handle.net/10603/1699

[23] Ch.Mandakini, KVN Sunitha, "Disambiguating the sense of verb in Telugu sentence using the argument structure", International Journal of Computational Linguistics and Natural Language Processing IJCLNLP , Volume 1, Issue 5, December 2012, Special Issue on Word Sense disambiguation, ISSN: 2279 – 0756, pgnos:151-155.

[24] DurgaPrasad Palanati, K. V. N. Sunitha, and B. Padmaja Rani. "Context-based word sense disambiguation in Telugu using the statistical techniques." In *Proceedings of the Second International Conference on Computational Intelligence and Informatics*, pp. 271-280. Springer, Singapore, 2018.

[25] Jha, P., Agarwal, S., Abbas, A., Singh, S., & Siddiqui, T. J. (2024). Unsupervised hindi word sense disambiguation using graph-based

centrality measures. *Int J Artif Intell*, *13*(4), 4957-4964.

[26] Bhingardive, Sudha, Dhirendra Singh, Rudra Murthy, Hanumant Redkar, and Pushpak Bhattacharyya. "Unsupervised most frequent sense detection using word embeddings." In *DENVER*. 2015.

[27] Zhang, Huirong, Yan Tang, Ying He, Chunqian Mou, Pingan Xu, and Jiaokai Shi. "A novel subspace clustering method based on data cohesion model." *Optik* 127, no. 20 (2016): 8513-8519.

[28] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 9119–9130, November 16–20, 2020,Association for Computational Linguistics.

[29] Abraham, A., Gupta, B. K., Maurya, A. S., Verma, S. B., Husain, M., Ali, A., ... & Gupta, S. (2024). Naïve Bayes approach for word sense disambiguation system with a focus on parts-of-speech ambiguity resolution. *IEEE Access*.

[30] Durgaprasad, P., Sunitha, K. V. N., & Padmajarani, B. (2022). Resolving Lexical Level Ambiguity: Word Sense Disambiguation for Telugu Language by Exploiting IndicBERT Embeddings. In *Communication, Software and Networks: Proceedings of INDIA 2022* (pp. 357-368). Singapore: Springer Nature Singapore.

[31] Kolikipogu, R., Padmaja Rani, B., & Swapna, N. (2013). Pseudo relevance feedback by linking WordNet for expanding queries in information retrieval process. *Int. J. Model. Optim*, *3*(5), 462-467.

[32] Kolikipogu, R. (2014). Vector Space Model for Telugu Information Retrieval.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

APPENDIX A

IndicBERT$_{BASE}$ with the configuration settings [30]

of : { "model_type": "albert",
"attention_probs_dropout_prob": 0,
"hidden_act": "gelu",
"hidden_dropout_prob": 0,
"embedding_size": 128,
"hidden_size": 768,
"initializer_range": 0.02,
"intermediate_size": 3072,
"max_position_embeddings": 512,
"num_attention_heads": 12,
"num_hidden_layers": 12,
"num_hidden_groups": 1,
"net_structure_type": 0,
"gap_size": 0, "num_memory_blocks":
0, "inner_group_num": 1,
"down_scale_factor": 1,
"type_vocab_size": 2,
"vocab_size": 200000}.