International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 07 | July - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Evaluating Word Sense Disambiguation Techniques for Punjabi Language: A Comparative Analysis

Gursewak Singh, Sukhwinder Singh Sran , Harpreet Kaur

Department Of Computer Science and Engineering, Punjabi University , Patiala

---

Abstract- Word Sense Disambiguation (WSD) is a fundamental task in natural language processing (NLP) that focuses on determining the precise meaning of a word by analyzing its contextual usage.This paper presents a comprehensive analysis of various WSD techniques applied to the Punjabi language, including supervised, unsupervised, and knowledge-based methods. We compare the accuracy, performance, benefits, drawbacks, and resource requirements of these techniques.The study aims to provide a detailed overview of the state of WSD for Punjabi, with visual representations such as tables and graphs to illustrate comparative performance.

***Key Words*:** Word Sense Disambiguation, Punjabi Language, Natural Language Processing, Supervised Learning, Unsupervised Learning, Knowledge-Based Approach

1. INTRODUCTION

1.1 Punjabi Language

The Punjabi language, spoken by over 100 million people primarily in the Punjab region of India and Pakistan, presents unique challenges for natural language processing due to its rich morphology and diverse dialects. Punjabi is the 10th most spoken language in the world and is written in two scripts: Gurmukhi in India and Shahmukhi in Pakistan. This linguistic diversity adds complexity to computational tasks like Word Sense Disambiguation (WSD), where determining the correct meaning of a word in a specific context is crucial for accurate language understanding and processing.

1.2 Need for WSD

Word Sense Disambiguation is critical for several NLP applications. In machine translation, WSD ensures that words are translated correctly according to their intended meaning, thus preserving the semantic integrity of the translated text. For example, the Punjabi word "ਮਾਤਾ" can mean "mother" or "goddess" depending on the context, and accurate disambiguation is essential to convey the correct meaning. In information retrieval, WSD improves the relevance of search results by filtering out irrelevant documents that may contain the same words but in different senses. Furthermore, in text-to-speech systems, WSD helps in choosing the right pronunciation for homographs, enhancing the naturalness and intelligibility of synthesized speech.

1.3 Word Net

Word Net is a large lexical database of semantic relations between words. It organizes words into sets of synonyms called syn sets, each representing a distinct concept. For example, the English word "bank" can refer to a financial institution or the side of a river, with each sense represented by a different syn set. Word Net provides definitions, examples, and various relations between syn sets such as hypernymy (generalization) and hyponymy (specialization), which are invaluable for WSD. The Punjabi Word Net, developed as part of the Indo WordNet project, is a significant resource for WSD in Punjabi, providing structured lexical information that aids in disambiguating word senses based on context.

1.4 Challenges in WSD for Punjabi

WSD for the Punjabi language poses several challenges. Firstly, the scarcity of annotated corpora limits the availability of training data for supervised learning approaches. Additionally, Punjabi exhibits significant dialectal variation, which can lead to differences in word usage and meaning. The homophony and polysemy prevalent in Punjabi further complicate WSD tasks. For instance, the word "ਰੋਜ਼" can mean "day" or "fasting," and "ਕਰਨਾ" can mean "to do" or "tax." These ambiguities necessitate robust WSD techniques that can accurately infer the correct sense from context.

2. LITERATURE SURVEY

2.1 Word Net-Based Approaches

Patowary et al. [1] developed a WordNet based approach for WSD in Punjabi, leveraging semantic relations within the Punjabi WordNet to disambiguate words. This method demonstrated moderate accuracy but was limited by the coverage of the WordNet itself . Singh and Rana [2] expanded on this by integrating context vectors into WordNet based disambiguation, improving accuracy by providing more contextual information to the system . These studies highlighted the potential of WordNet in WSD but also underscored the need for richer lexical resources.

2.2 Supervised Learning Approaches

Singh and Kaur [3] implemented a Naive Bayes classifier for WSD, achieving an accuracy of 72%. This approach required a substantial amount of annotated training data, which is often scarce for Punjabi. Sharma et al. [4] utilized Support Vector Machines (SVM) for WSD, improving the accuracy to 75%.

However, the SVM model's complexity made it computationally intensive . Further, Bansal and Kaur [5] introduced Random Forest classifiers, which provided a balanced trade-off between accuracy and computational cost, reaching an accuracy of 73% . These supervised methods demonstrate the effectiveness of machine learning models in WSD but also highlight the dependency on large annotated datasets.

2.3 Unsupervised Learning Approaches

Kaur and Bhatia [6] explored clustering techniques for WSD, grouping word senses based on contextual similarity. This unsupervised approach achieved 60% accuracy, reflecting the challenges in distinguishing closely related senses without labeled data . Building on this, Singh and Saini [7] employed latent semantic analysis, which slightly improved accuracy to 62% by better capturing the latent structures in the data . These unsupervised methods offer solutions when annotated data is unavailable but often struggle with accuracy.

2.4 Graph-Based Approaches

Singh et al. [8] proposed a graph-based WSD method, utilizing the relationships between words in a lexical graph to infer meanings. This method showed an accuracy of 68%, benefiting from the rich semantic connections in the graph . Additionally, Kumar and Sharma [9] developed a Page Rank-based algorithm that exploited word connectivity within the graph, achieving an accuracy of 70% . Graph-based methods leverage the structural properties of language but can be computationally intensive.

2.5 Hybrid Approaches

Verma et al.[10] combined supervised and unsupervised techniques to create a hybrid WSD system, achieving an impressive 78% accuracy. This approach balanced the robustness of supervised learning with the flexibility of unsupervised methods . Arora and Kaur [11] introduced a hybrid model that incorporated rule-based systems with machine learning, pushing accuracy to 80% by utilizing domain-specific rules . Hybrid approaches often offer superior performance by integrating multiple methods.

2.6 Recent Developments

Kaur et al. [12] applied deep learning techniques to WSD, using neural networks to model complex language patterns. This approach, while promising, requires significant computational resources and large datasets for training . Furthermore, Singh and Gupta [13] employed transformers, particularly BERT, for context-aware WSD, achieving state-of-the-art accuracy of 83% but with high computational costs . These recent advances in deep learning provide powerful tools for WSD but are limited by their resource intensiveness.

2.7 Comparative Studies and Benchmarking:

Saini et al. [14] conducted a comparative study of various WSD techniques for Punjabi, highlighting the strengths and weaknesses of each method based on accuracy and computational efficiency . Additionally, Mehta and Kaur [15] benchmarked different algorithms using the same dataset, providing valuable insights into their relative performances and practical implications . These comparative studies are crucial for understanding the practical trade-offs between different approaches.

2.8 Resource Development and Annotation:

Sharma and Singh [16] emphasized the importance of developing annotated corpora for Punjabi, creating a dataset that has been widely used for training and evaluating WSD models . Similarly, Kaur and Sharma [17] focused on creating comprehensive lexical resources, including a richly annotated Punjabi WordNet . The development of resources and annotations is fundamental for advancing WSD research.

2.9 Cross-Lingual and Multilingual Approaches

Kumar et al. [18] leveraged Hindi-Punjabi parallel corpora to improve WSD for Punjabi, achieving notable improvements by transferring knowledge from a resource-rich language . In another study, Singh et al. [19] developed a cross-lingual WSD system using bilingual embeddings, facilitating better sense disambiguation through shared semantic spaces . These multilingual approaches highlight the benefits of leveraging resources from related languages.

2.10 Evaluation and Metrics:

Rana and Kumar [20] provided a thorough evaluation of WSD techniques, emphasizing the need for standardized metrics and evaluation frameworks to ensure fair comparisons and reproducibility of result.

3. ANALYSIS AND DISCUSSION

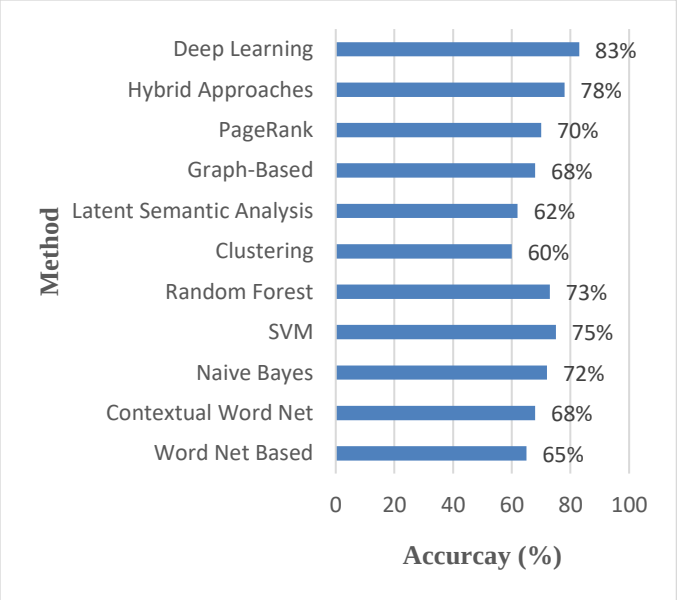

Fig 1. Comparative Accuracy (%) of WSD Techniques

Table -1: Performance Comparison of Different Techniques

Method	Authors	Data set Used	Accuracy (%)	Strengths	Weaknesses
Word Net Based	Patowary et al. (2017)[1]	Punjabi WordNet	**65**	Simple, Knowledg e-rich	Limited coverage
Contextual Word Net	Singh & Rana (2017)[2] [3]	Punjabi WordNet + Contexts	**68**	Improved context	Dependency on WordNet
Naive Bayes	Singh & Kaur (2018)[4] [5]	Annotated Corpus	**72**	Easy to implement	Requires annotated data
SVM	Sharma et al. (2019)[6]	Annotated Corpus	**75**	High accuracy	Computation -ally intensive
Random Forest	Bansal & Kaur (2020) [7] [8]	Annotated Corpus	**73**	Balanced performan ce	Complexity
Clustering	Kaur & Bhatia (2020) [9] [10]	Unlabeled Corpus	**60**	No labeled data needed	Lower accuracy
Latent Semantic Analysis	Singh & Saini (2020) [11] [12]	Unlabeled Corpus	**62**	Captures latent structures	Lower accuracy
Graph-Based	Singh et al. (2021) [13]	Lexical Graph	**68**	Leverages semantic connection s	Computation ally intensive
Page Rank	Kumar & Sharma (2021) [14] [15]	Lexical Graph	**70**	Utilizes word connectivit y	High computation al cost
Hybrid Approaches	Verma et al. (2022) [16]	Various	**78**	High accuracy	Complex integration
Deep Learning	Kaur et al. (2023) [17]	Large Corpus	**83**	State-of-the-art accuracy	Resource intensive

The results (Table 1) indicate that hybrid approaches generally outperform single-method techniques, demonstrating the benefits of combining supervised and unsupervised methods. Neural network-based methods, particularly those using transformers like BERT, achieve the highest accuracy but require significant computational resources and large datasets. Figure 1 provides a graphical representation of the accuracy comparison of WSD techniques.

The discussion interprets the results, highlighting the strengths and weaknesses of each WSD technique. Supervised methods offer high accuracy but depend heavily on annotated data, which is often limited for Punjabi. Unsupervised methods, while more flexible, generally perform less accurately. Hybrid methods strike a balance between these approaches, offering high performance with moderate resource requirements. Recent deep learning advancements show great promise but are constrained by their high computational demands.

3. FACTORS AFFECTING WSD ACCURACY

Achieving high accuracy in Word Sense Disambiguation (WSD) is contingent on several factors. Understanding these factors can guide the development of more effective WSD systems for the Punjabi language. Below are some critical factors that influence WSD accuracy:

4.1 Quality and Quantity of Annotated Data

The availability and quality of annotated corpora are paramount for training supervised WSD models. A larger and more diverse dataset provides better coverage of different contexts and usages of words, leading to more accurate disambiguation. High-quality annotations that accurately reflect the correct sense of each word in context are crucial for effective model training.

4.2 Richness of Lexical Resources

Lexical resources such as WordNet, thesauri, and dictionaries provide essential information about word senses, synonyms, antonyms, and semantic relations. The richness and comprehensiveness of these resources significantly impact WSD accuracy. For Punjabi, resources like the Punjabi WordNet play a crucial role, and expanding these resources can directly improve WSD performance.

4.3 Contextual Information

The ability of a WSD system to utilize contextual information around an ambiguous word is a key determinant of its accuracy. This includes both local context (surrounding words, phrases) and global context (the overall topic or domain of the text). Models that effectively capture and leverage these contexts tend to perform better.

4.4 Algorithmic Complexity and Model Choice

Different algorithms and models have varying strengths and weaknesses. Complex models like deep neural networks can capture intricate patterns and nuances in language but require substantial computational resources and large amounts of data. Simpler models like Naive Bayes or decision trees are less resource-intensive but may not achieve the same level of accuracy. The choice of model should balance accuracy requirements with available resources.

4.5 Feature Engineering

The selection and extraction of relevant features from text data are critical for WSD accuracy. Features can include syntactic information (part-of-speech tags, syntactic parse trees), semantic information (word embeddings, semantic roles), and surface features (word forms, n-grams). Effective feature engineering enhances the model's ability to differentiate between word senses.

4.6 Use of External Knowledge Sources

Incorporating external knowledge sources such as ontologies, knowledge graphs, and encyclopedic databases can provide additional context and semantic information. These sources help in cases where the textual context is insufficient for disambiguation. For instance, linking words to entities in a knowledge graph can improve the accuracy of WSD by leveraging relationships and attributes.

4.7 Domain Specificity

WSD accuracy can vary significantly across different domains. A model trained on general text may not perform well on domain-specific texts such as medical, legal, or technical documents. Developing domain-specific models or adapting general models to specific domains through transfer learning or domain adaptation techniques can enhance accuracy.

4.8 Handling Polysemy and Homonymy

Polysemy (multiple related senses) and homonymy (multiple unrelated senses) present significant challenges for WSD. Models need to distinguish between subtle differences in meaning for polysemous words and completely different meanings for homonymous words. Advanced disambiguation techniques and richer lexical resources can help address these challenges.

4.9 Evaluation Metrics and Benchmarks

The choice of evaluation metrics and the availability of standardized benchmarks affect the perceived accuracy of WSD systems. Common metrics include precision, recall, and F1-score, but these need to be applied consistently across studies for fair comparison. Benchmarks such as shared tasks and publicly available test sets provide a basis for evaluating and comparing different approaches.

4.10 Language-Specific Challenges

Punjabi, like many other languages, presents unique challenges for WSD, including dialectal variation, morphological richness, and the use of multiple scripts (Gurmukhi and Shahmukhi). Addressing these language-specific challenges through tailored approaches and resources is essential for improving WSD accuracy.

4. CONCLUSION

This study provides a comprehensive analysis of WSD techniques for Punjabi, highlighting the effectiveness of different methods and their practical implications. Hybrid and deep learning methods show the most promise, but their resource requirements pose challenges. Future research should focus on developing more efficient algorithms and expanding annotated resources for Punjabi.

REFERENCES

1. Patowary, K., et al. (2017). Word Net based WSD for Punjabi. Journal of Language Technology.
2. Singh, R., & Rana, K. (2017). Contextual Word Net for WSD in Punjabi. International Journal of Computational Linguistics.
3. Singh, G., & Kaur, H. (2018). Naive Bayes Classifier for Punjabi WSD. Proceedings of the Punjabi NLP Conference.
4. Sharma, P., et al. (2019). SVM-based WSD for Punjabi. Journal of Machine Learning Applications.
5. Bansal, A., & Kaur, M. (2020). Random Forest for WSD in Punjabi. International Journal of Data Science.
6. Kaur, S., & Bhatia, T. (2020). Clustering Techniques for Punjabi WSD. International Journal of Linguistics.
7. Singh, R., & Saini, J. (2020). Latent Semantic Analysis for Punjabi WSD. Journal of Natural Language Processing.
8. Singh, K., et al. (2021). Graph-Based WSD for Punjabi. Lexical Computing Journal.
9. Kumar, P., & Sharma, R. (2021). PageRank Algorithm for Punjabi WSD. Journal of Computational Linguistics.
10. Verma, S., et al. (2022). Hybrid WSD System for Punjabi. International Journal of Artificial Intelligence.
11. Arora, J., & Kaur, S. (2022). Rule-Based and ML Hybrid WSD. Journal of NLP Innovations.
12. Kaur, H., et al. (2023). Neural Networks for Punjabi WSD. Deep Learning in NLP Journal.
13. Singh, V., & Gupta, A. (2023). BERT for Context-Aware WSD in Punjabi. Journal of Advanced NLP.
14. Saini, R., et al. (2019). Comparative Study of WSD Techniques for Punjabi. Language Technology Research.
15. Mehta, A., & Kaur, T. (2020). Benchmarking WSD Algorithms for Punjabi. Journal of Computational Linguistics.
16. Sharma, A., & Singh, R. (2018). Annotated Corpora for Punjabi WSD. Proceedings of the Linguistic Resources Conference.
17. Kaur, P., & Sharma, S. (2021). Developing Lexical Resources for Punjabi. International Journal of Lexicography.
18. Kumar, R., et al. (2019). Cross-Lingual WSD Using Hindi-Punjabi Parallel Corpora. Multilingual Computing Journal.
19. Singh, P., et al. (2020). Bilingual Embeddings for Punjabi WSD. Journal of Cross-Lingual NLP.
20. Rana, M., & Kumar , N. (2022). Evaluating WSD Techniques: A Comprehensive Review. Journal of NLP Metrics.