IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Comprehensive Survey on Word Sense Disambiguation for Low-Resource Languages

[1]Kajal P. Visrani, [2]Dr K. P. Adhiya

[11]Research Scholar, Professor :::0009-0008-9746-2725
Department of Computer Engineering, SSBT's COET, Bambhori, Jalgaon, India
Email: kajal.visrani@gmail.com

[2] Professor, Professor ::0000-0002-3925-0213
Department of Computer Engineering, SSBT's COET, Bambhori,
Jalgaon, India
Email: adhiyakp@gmail.com

Abstract *– Word Sense Disambiguation (WSD) is a critical task in Natural Language Processing (NLP) that seeks to resolve lexical ambiguity by determining the appropriate meaning of a word within a specific context. While considerable progress has been made in WSD for high-resource languages, low-resource languages have received limited focus due to a lack of annotated datasets, lexical resources, and computational tools. This paper provides a thorough survey of current WSD techniques, including knowledge-based, supervised, unsupervised, and deep learning methods, and assesses their applicability in low-resource language environments. The study also addresses the challenges faced by low-resource languages, such as data scarcity, morphological complexity, the absence of standardized tools, and multilingual influences. Furthermore, it reviews WSD research in Indian languages, including Manipuri, Malayalam, Punjabi, Bengali, and Sindhi, to highlight both current advancements and limitations. Notably, Sindhi is recognized as one of the least investigated languages concerning WSD research. The findings of this survey underscore the necessity for developing resource-efficient and adaptable approaches, particularly leveraging machine learning and deep learning techniques, to enhance WSD performance in low-resource languages. The paper also outlines future research directions aimed at overcoming existing challenges and closing the research gap in this area.*

Keywords- Word Sense Disambiguation, Low-Resource Languages, NLP, Machine Learning, Deep Learning, Sindhi

INTRODUCTION

Word Sense Disambiguation (WSD) is the process of determining the correct meaning of a word in a given context when the word has multiple meanings. Many natural language words are ambiguous, meaning the words represent different concepts depending on their use in a sentence. For example, the word "bank" can refer to a financial institution or the side of a river, depending on the context. Resolving such ambiguities is essential for improving the performance of various Natural Language Processing (NLP) applications.

Many researchers have proposed various approaches to address the WSD problem. These approaches are broadly categorised into three main types: knowledge-based methods, supervised learning methods, and unsupervised learning methods. Knowledge-based methods rely on lexical resources such as dictionaries and thesauri, while supervised methods use machine learning algorithms trained on annotated corpora. Unsupervised methods, on the other hand, attempt to identify word senses without relying on labelled data. [1]

Despite significant progress in this field, most research has focused on high-resource languages such as English. Low-resource languages like Sindhi, Kashmiri, Manipuri, Bodo, Dogri, and Konkani still face several challenges, including the lack of annotated datasets, lexical databases, and computational tools. This creates a gap in research and highlights the need for developing efficient WSD techniques tailored to such languages. [2]

This paper presents a comprehensive survey of existing WSD approaches and analyses their effectiveness in low-resource language settings, highlighting key challenges and identifying potential directions for future research.

LITERATURE REVIEW

Agirre et al. (2021) introduced a multilingual and cross-lingual approach to Word Sense Disambiguation through the SemEval shared task. Their study shows that transfer learning and multilingual models can improve WSD performance in low-resource languages by leveraging knowledge from high-resource languages. However, the effectiveness of these approaches still depends on the availability of aligned multilingual datasets. [3]

Gul et al. (2021) focused on the development of a WordNet for the Saraiki language, highlighting the importance of lexical resources in enabling effective WSD. Their research highlights that the lack of organised semantic databases constitutes a considerable constraint in low-resource languages, and the development of these resources may significantly boost disambiguation performance. [4]

Gogoi and Baruah (2022) worked on Word Sense Disambiguation in the Assamese language using a combination of knowledge-based and machine learning approaches. Their study says hybrid techniques can improve disambiguation accuracy in low-resource settings; however, the lack of annotated datasets and standardised tools is still a major challenge. [5]

Rajan and Salgaonkar (2022) conducted a survey of Natural Language Processing resources for low-resource languages such as Sindhi, Nepali, and Konkani. Their work highlights the scarcity of linguistic resources, including corpora and lexical databases, which limits the development of effective NLP applications, including WSD, in these languages. [6]

OVERVIEW OF WORD SENSE DISAMBIGUATION TECHNIQUES

In Natural Language Processing, WSD has worked a lot to remove lexical ambiguity. These methods vary in terms of linguistic resources, annotated data, and computational models. Broadly, WSD techniques can be categorised into three main groups: knowledge-based, supervised, and unsupervised methods.

To determine a word's correct meaning, **knowledge-based approaches** rely on external lexical resources like dictionaries, thesauri, and semantic networks. Techniques such as the Lesk algorithm and semantic similarity measures use definitions and relationships between words to identify the most relevant sense in a given context. These methods are especially helpful when there is a lack of annotated data, but their efficacy is constrained by the quality and quantity of lexical resources [1].

Supervised approaches classify word meanings based on contextual factors using machine learning algorithms trained on corpora annotated with sense information. Neural networks, Support Vector Machines, and Naïve Bayes are common methods. Because these techniques can learn from labelled data, they typically achieve excellent accuracy; yet they require large-scale annotated datasets, which are frequently unavailable for low-resource languages [1], [2].

Unsupervised approaches, on the other hand, seek patterns and similarities in unannotated text to determine word meanings without relying on labelled datasets. To group comparable word usages, methods including co-occurrence analysis, word embeddings, and context clustering are frequently employed. These techniques are generally less accurate than supervised ones, despite being more adaptable and scalable [1].

The area of WSD has improved dramatically in recent years thanks to deep learning-based methods that allow models to capture intricate semantic linkages and contextual dependencies. Strong performance across a range of jobs and languages has been shown by methods based on transformer topologies and contextual embeddings, which eliminates the requirement for human feature engineering [2].

The effectiveness of existing WSD techniques is highly dependent on the availability of linguistic resources and annotated corpora. When it comes to low-resource languages, where these resources are hard to get by, this restriction becomes very important. Therefore, it is essential to examine the applicability and limitations of these techniques in low-resource settings.

WSD IN LOW-RESOURCE LANGUAGE

Word Sense Disambiguation (WSD) in low-resource languages poses difficulties that have a substantial impact on the effectiveness and suitability of current methods. Low-resource languages lack the computational tools, annotated datasets, and linguistic resources necessary for efficient disambiguation, in contrast to high-resource languages like English**.** That's why many WSD approaches cannot be directly applied or generalized to these low-resource languages.

These languages typically have limited data resources, lack standardized lexical databases like WordNet, and receive less research focus. This group includes languages like Sindhi, Nepali, Konkani, Kashmiri, and Assamese, where the lack of digital resources limits the creation of reliable NLP systems. Implementing supervised and deep learning-based WSD techniques, which mostly rely on sizable annotated datasets, is severely hampered by these restrictions [1], [2].

WSD FOR INDIAN LANGUAGES

WSD has been well-researched for many European and Asian languages, especially English. However, research on WSD for Indian languages has been relatively limited. These languages are complex, involving a variety of morphological inflexions. There is also a lack of adequate resources like sense inventories, machine-readable dictionaries, and annotated corpora to support WSD. Some efforts at WSD for the major Indian languages, such as Manipuri, Malayalam, Punjabi, and Bengali, are tabulated below. [7]

Manipuri

Manipuri, spoken in the northeastern region of India, has a unique syntactic and semantic structure that sets it apart from many other Indian languages. The first significant effort to develop a WSD system for Manipuri was proposed in 2013 [7]. This system works in two phases: training and testing. This process involves several important steps, including preprocessing, feature selection, training, testing, and performance evaluation. The system works with specific features like the use of the keyword, its position in the sentence, and neighbouring words. Thus, it emphasises positional features because Manipuri lacks other relevant morphological features that might become accessible during disambiguation. By understanding the surroundings in which the target word operates, the system can unravel ambiguities and assign an appropriate sense of the word itself.

Malayalam

The Malayalam language, spoken in Kerala, is again a Dravidian language with complexities in syntax and vocabulary. The first automation attempt for WSD in Malayalam was knowledge-based; it used two main strategies: one using a hand-devised knowledge source and the other using Malayalam WordNet as a lexical resource. The LESK and Walker algorithms were used to compute the overlap between contextual words and predefined sense definitions from the knowledge source. A Conceptual Density-based algorithm was also applied to measure semantic relatedness between words, using WordNet depth as a key metric [7]. The sense with the lowest depth (or highest conceptual density) was chosen as the correct sense. These methods combined contextual word overlap with semantic relationships, so they assisted in improving the accuracy of WSD in the Malayalam language.

Punjabi

Punjabi, being rich in morphology and ample inflexion, poses several difficulties for WSD. A WSD algorithm based on the Modified LESK Algorithm has been proposed to remove ambiguity from Punjabi text. The system functions on the basis that co-occurring words within a sentence should be assigned closely related senses, and definitions of the related senses should have maximum overlap. This approach uses co-occurrence information at the sentence level to select the most relevant sense for an ambiguous word. The algorithm can effectively identify the correct sense by measuring overlap in definitions at the sense level, and it thus becomes a very useful approach for WSD in Punjabi. [8]

Bengali

Bengali is another major language spoken in India and Bangladesh. Research on WSD in the context of machine translation has also been conducted for Bengali and Bengali-Hindi. In the former case, a graph-based clustering approach for unsupervised sense clustering is used by a Bengali-Hindi machine translation system. Here, words with similar senses are grouped based on the co-occurrence patterns of the words within the text. Graph-based clustering provides a way to organise word senses by their contextual use and thus facilitates easier choice of the correct sense for ambiguous words [78]. The unsupervised approach becomes particularly effective for Bengali because the word senses can differ widely according to context, and such a system can process vast amounts of data without having to use manually annotated resources. [8]

Sindhi

Sindhi is an Indo-Aryan language with rich morphology and a complex Arabic-based script, which makes Word Sense Disambiguation (WSD) a challenging task. Due to the lack of annotated corpora, lexical resources such as WordNet, and computational tools, research in Sindhi WSD remains limited. Existing efforts mainly focus on basic resource development and adapting knowledge-based or machine learning approaches using contextual features. However, the absence of sufficient data restricts the use of supervised techniques, highlighting the need for resource-efficient and language-specific WSD methods.

KEY CHALLENGES IN LOW-RESOURCE WSD

Main c**hallenges with** Word Sense Disambiguation in low-resource languages:

Data Scarcity: **The biggest problem in** training supervised and deep learning models is the absence of large-scale annotated corpora. Developing sense-annotated datasets is a time-consuming process and is rarely available

Limited Lexical Resources: Knowledge-based techniques are less effective since resources like WordNet, dictionaries, and semantic networks are either unavailable or insufficient.

Morphological Complexity: These languages often have rich morphology, which includes inflexions and variants that complicate sense identification and create ambiguity.

Lack of Standard Tools: NLP tools that are necessary for preprocessing in WSD systems, such as token, POS taggers, and parsers, are scarce or rarely available

Multilingual Influence and Code-Mixing: Speakers often mix multiple languages (e.g., Sindhi with Hindi or English), which creates additional ambiguity in identifying word meanings.

Domain and Context Variability: Word meanings can vary depending on the context or domain, and it is more difficult to capture these differences when there is little data.

Lack of Evaluation Benchmarks: For low-resource languages, there aren't many standard datasets or benchmarks available to assess and contrast WSD systems.

CONCLUSION

Word Sense Disambiguation (WSD) is a crucial task in Natural Language Processing, aimed at resolving lexical ambiguity and enhancing language comprehension. This paper provides a thorough survey of various WSD techniques, including knowledge-based, supervised, unsupervised, and deep learning methods. It also assesses their applicability to low-resource languages. The study indicates that, despite significant advancements for high-resource languages, low-resource languages continue to encounter substantial challenges due to the scarcity of annotated datasets, lexical resources, and computational tools. Additionally, linguistic complexities such as rich morphology and code-mixing exacerbate the difficulties of effective disambiguation in these languages. A review of Indian languages, including Manipuri, Malayalam, Punjabi, Bengali, and Sindhi, reveals that research efforts remain limited and fragmented. Notably, Sindhi is one of the least explored languages, with minimal resources and few studies specifically focused on WSD. This highlights a considerable research gap and underscores the necessity for targeted initiatives in this domain. In summary, the findings of this survey underscore the importance of developing resource-efficient and adaptable WSD approaches specifically designed for low-resource languages.

FUTURE WORK

Future research in Word Sense Disambiguation for low-resource languages should priorities the development of effective Natural Language Processing techniques utilising machine learning and deep learning approaches. It is essential to enhance the availability of linguistic resources and to leverage cross-lingual methods to improve performance. Furthermore, designing hybrid and context-aware models will be crucial in addressing challenges such as data scarcity and code-mixed language usage, especially for languages like Sindhi.

ACKNOWLEDGMENT

I express my sincere gratitude to my academic mentors (guide) for their insightful guidance and encouragement throughout this study. Special thanks to the research community and authors whose work laid the foundation for this study, serving as a valuable source of knowledge and inspiration. Finally, I am deeply grateful to my family, colleagues, and friends for their unwavering support and motivation during this research journey. Their belief in my work has been a constant source of strength.

REFERENCES

[1] Ms. Kajal P. Visrani & K. P. Adhiya, "A Survey of Word Sense Disambiguation: Applications,," *Grenze International Journal of Engineering and Technology, June Issue,* vol. 11, no. 2, pp. 2370-2374, 2025.

[2] D. K. P. A. Kajal P. Visrani, "CONTEXTUAL LEARNING APPROACHES FOR WORD SENSE DISAMBIGUATION," *ACCENT JOURNAL OF ECONOMICS ECOLOGY & ENGINEERING,* vol. 11, no. 2, 2026.

[3] A. S. e. a. Eneko Agirre, "SemEval-2021 Task 2: Multilingual and Cross-lingual Word Sense Disambiguation," 2021.

[4] e. a. Sadia Gul, "Developing WordNet for Saraiki Language," *Linguistic and Literary Review,* p. https://doi.org/10.32350/llr.72.04, 2021.

[5] N. B. Amit Gogoi, "Word Sense Disambiguation in Assamese Language Using Knowledge-Based and Machine Learning Approaches," 2022.

[6] A. S. A. Rajan, "A Survey of NLP Resources for Low-Resource Languages: Sindhi, Nepali and Konkani," *Advances in Intelligent Systems and Computing,* 2022.