Grenze International Journal of Engineering and Technology, June Issue

A Survey of Word Sense Disambiguation: Applications, Approaches and Future Directions

Ms. Kajal P. Visrani[1] and K. P. Adhiya[2]
[1]Research Scholar, SSBT's COET, Jalgaon, India /Computer Department, Jalgaon
Email: kajal.visrani@gmail.com
[2]Professor, SSBT's COET, Jalgaon, India/ Computer Department, Jalgaon
Email: adhiyakp@gmail.com

***Abstract*—Word Sense Disambiguation (WSD) is a foundational problem in Natural Language Processing (NLP) that involves determining the correct meaning of a word based on its context. As language is inherently ambiguous, resolving lexical ambiguities is critical for improving the performance of various NLP applications, including machine translation, information retrieval, sentiment analysis, and question-answering. This paper comprehensively surveys WSD, tracing its historical evolution from rule-based approaches to contemporary deep learning techniques. It discusses key methodologies such as knowledge-based, supervised, and unsupervised learning, highlighting their strengths, limitations, and applications. The importance of WSD is explored in diverse contexts, including bioinformatics, semantic web development, and lexicography. Despite significant advancements, challenges such as handling polysemous words, adapting to domain-specific contexts, and addressing resource limitations in low-resource languages persist. The paper also examines the integration of linguistic resources and computational techniques to enhance WSD systems and underscores future research directions to develop robust, scalable solutions. By addressing these challenges, WSD holds the potential to drive innovation across NLP applications, ensuring accurate semantic understanding and interpretation of human language.**

***Index Terms*—Word Sense Disambiguation (WSD), Natural Language Processing (NLP), Lexical Ambiguity, Machine Learning, Deep Learning, Contextual Embedding, Low-Resource Languages, Machine Translation (MT), Information Retrieval (IR), Sentiment Analysis.**

I. INTRODUCTION

WSD is a highly respected, significant barrier in the field of NLP and one of the initial problems in computational linguistics. WSD is the process of determining the correct meaning of a word given its unique context. This task is particularly significant in languages with polysemous words, words that carry multiple meanings where interpretation can vary considerably based on the surrounding text, grammatical structure, and overall context of the passage [1, 2].

Effective methods for WSD are critical to increase linguistic clarity and allow for proper computational comprehension of text, which is a prerequisite for applications such as Machine Translation (MT) [3], information retrieval(IR), and question-answering [4]. But while substantial progress has made a lot of effort within the area of WSD for resource-rich languages, like English, under-resourced languages face massive obstacles in finding reliable disambiguation [5].

Grenze ID: 01.GIJET.11.2.239_20

The paper is structured as follows: It begins with an Introduction that underscores the significance of Word Sense Disambiguation (WSD) in Natural Language Processing (NLP). The History of WSD section provides a chronological overview of methodologies, from early rule-based approaches to contemporary deep learning models. Subsequent sections delve into the Importance of WSD in NLP Tasks, outline a General Working Model of WSD, discuss various Approaches to WSD, and highlight its Applications across different domains. The paper concludes by summarizing key insights and suggesting directions for future research.

II. HISTORY OF WSD

TABLE I

Period	Approach	Key features	challenges
1950s-1980s	Rule-based methods	Used handcrafted linguistic rules and dictionaries to infer word meanings	Labour-intensive, lacked scalability, rigid rules
1980s-1990s	Knowledge-based methods	Utilized lexical resources like WordNet to determine semantic similarity and lexical relations	Limited by resource coverage and quality
1990s-2010s	Supervised machine learning (e.g., Naïve Bayes, SVMs)	Trained on annotated corpora using contextual features for sense prediction	Required large labeled datasets, struggled with data sparsity
2010s-Present	Deep learning models (e.g., BERT, GPT)	Used contextual embedding's and transformers to dynamically interpret word meanings	High computational cost, challenges with rare senses and domain adaptation

III. IMPORTANCE OF WSD IN NATURAL LANGUAGE PROCESSING

A word's definition can vary depending on its language. WSD can detect the appropriate meaning of a word in any given situation automatically. For example, in the statement "A boy is playing cricket on the ground," the word "cricket" can mean both "a game" and "an insect." Similarly, in the sentence "Chirping noise is created by cricket in the night," the word "cricket" can mean both phrases. Uncertainty regarding terminology is a major problem in NLP. Using WSD alone has been deemed insufficient for MT from the beginning [8]. The process of MT entails translating from one language to another. A term that is ambiguous in the original language can have multiple translations in the target language depending on its context, whereas many words that are independent in the original language can only have one translation in the targeted word. It follows that WSD plays a pivotal role in resolving ambiguity between terms in the source and destination languages.

Performing WSD requires two steps. The first step involves determining the number of senses an ambiguous word can have from the sense inventory; the second step involves assigning the correct sense to the ambiguous word. Clarifying the meanings of terms that are open to interpretation is an intricate and domain-specific undertaking. For terms with several definitions, the WSD process consults a standard sense inventory, such as WordNet or a dictionary, and then employs various machine learning algorithms to arrive at the correct definition. When building NLP systems, WSD is an essential building block for systems that aim to comprehend and interpret human language in its context and semantics. Because lexical ambiguity is a fundamental aspect of natural language, finding a solution to it is essential for enhancing the precision and reliability of NLP applications in domains including MT, sentiment analysis, text summarization, and question response systems. MT, information retrieval, or text analysis are all areas where mistakes in interpreting word meanings can have a major impact, making precise context comprehension of words critical for NLP [9].

IV. GENERAL WORKING MODEL OF WSD

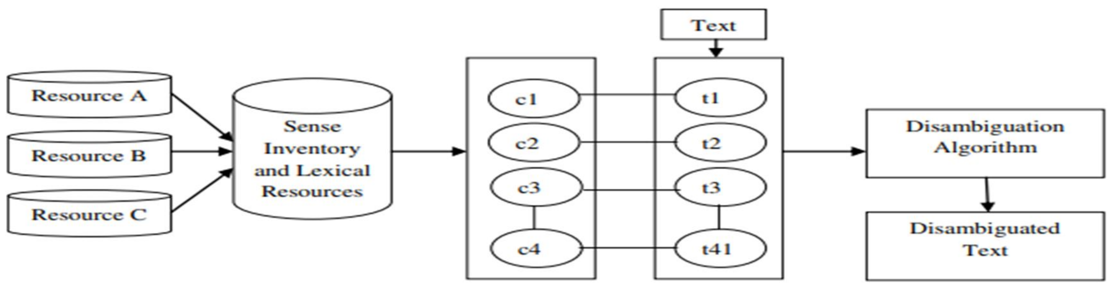

Figure 1 working of WSD

Figure 1. illustrates the workflow of a WSD system, showcasing its core components and their interconnections. The process begins with multiple external resources, labeled Resource A, Resource B, and Resource C, which provide essential lexical and semantic data regarding word meanings. These resources are consolidated into a Sense Inventory and Lexical Resources repository, which serves as a structured database containing word senses and their relationships. This repository allows access and ordering of relevant word meanings when disambiguation occurs. The system reads the input text with ambiguous words by relating it to the sense inventory. In the image, elements from the sense inventory (c1, c2, c3, c4) are mapped to corresponding elements in the text (t1, t2, t3, t41). These mappings indicate the relevance between the possible word meanings and contextual cues derived from the input text. The associations illustrate how contextual and lexical knowledge is exploited to discern the possible senses of ambiguous words. This information is provided to a Disambiguation Algorithm, which employs contextual relationships and sense inventory information to recognize the most suitable sense of each ambiguous word. Through the observation of patterns, semantic relationships, and textual clues, the algorithm resolves the ambiguity and produces Disambiguated Text as output wherein meanings of words are deduced depending on their given context. Figure successfully captures the structured and systematic nature of a WSD system by emphasizing the adaptation of external lexical resources, sense inventories, and contextual analysis toward achieving accurate word sense disambiguation. This framework shows how blending linguistic resources with computational techniques can be used to address the problems of polysemy in NLP [9]

V. APPROACHES

A. Knowledge-Based Methods

In order to resolve ambiguities, knowledge-based techniques rely on external lexical resources like WordNet, thesauri, and ontologies. These techniques use contextual information and semantic associations (such as synonyms, hypernyms, and semantic similarity metrics) to determine word senses. [9]. Despite being resource-efficient when labelled data is not available, these methods are constrained by the calibre and scope of the lexical resources that are employed. The approach can be further divided into several methods, with four primary kinds often applied: LESK Algorithm, Semantic Similarity, Selection, Selection Preferences, Heuristic, and Walker algorithm.

Supervised techniques: Supervised techniques train machine learning models that anticipate a word's right meaning using labelled datasets. To categorise word senses, algorithms such as Naïve Bayes, decision trees, and neural networks analyse contextual cues. These approaches are difficult to use in low-resource languages or specialised fields since they are limited by the availability of huge annotated corpora, despite the fact that they frequently attain high accuracy. The approach can be further divided into several methods: Decision List, Decision Tree, Naïve Bayes, Neural Network, Exemplar Based, Support vector Machine

Unsupervised Methods: To find patterns and groups of related word usages, unsupervised techniques use unannotated text corpora. Word occurrences with comparable contexts are grouped using techniques like clustering, topic modelling, and word embedding’s to determine their senses. Although these techniques are adaptable and scalable, they usually lack the accuracy of supervised methods and may need a lot of fine-tuning to produce useful outcomes. The approach can be further divided into several methods context clustering, Word clustering, co-occurrence graph

VI. APPLICATIONS OF WSD

While MT was the first and most apparent use of WSD, disambiguation has been included in nearly all NLP applications and is gaining importance in fields like bioinformatics or the semantic web.

Machine Translation (MT): To prevent misconceptions in cross-linguistic communication, WSD in machine translation makes sure that polysemous words are translated according to their accurate meanings in each context. For instance, in the sentence "He went to the bank to fish" as opposed to "He went to the bank to deposit money," WSD enables the system to differentiate between the financial institution in the latter sentence and "bank," which refers to the riverbank.

The creation of translations that are both contextually accurate and semantically consistent depends on this distinction. Without WSD, translation systems may provide outputs that are unclear or completely inaccurate, which might result in translations that are not very usable or of high quality [14].

WSD enables MT systems to produce translations that are more accurate, fluid, and faithful to the original text's intended meaning.

Information Retrieval (IR): WSD is essential for enhancing information retrieval systems, like search engines, which make sure that the meaning of ambiguous terms used in searches is reflected in the search results. For example, when a person searches for "python," it might refer to either the computer language or the snake. To provide the user with the most pertinent information, WSD assists search engines in deciphering these phrases and ranking them appropriately. This skill is crucial in specialised fields where word meanings must be highly exact, such as retrieving legal or medical information. WSD allows for more accurate and pertinent results when search queries are properly interpreted, which lessens user annoyance and probably improves the search experience overall.[14]

Text Summarization: WSD plays a crucial role in text summarisation by maintaining important concepts, allowing for the creation of succinct but insightful summaries [15]. Summarising text can be difficult when polysemous terms are used, particularly in technical or specialized literature. For example, when a report uses the terms "virus" (biological) and "virus" (computer), WSD enables the summarisation system to maintain the word's meaning in its context, preventing any misunderstandings [16]. In applications where preserving the original content's correctness and clarity is crucial, such as news summarisation, legal document summarisation, and academic article summarisation, this disambiguation is crucial. Sentiment Analysis: WSD is vital for sentiment analysis, where the goal is to determine the emotional tone behind the text, whether positive, negative, or neutral. Words with multiple meanings can drastically affect sentiment scores if misinterpreted. For instance, the word “charged” in “he felt charged up” conveys excitement, while in “he was charged with theft,” it has a legal implication with a negative tone [17]. Without WSD, sentiment analysis might misinterpret these examples, leading to incorrect conclusions about the emotional tone. By correctly interpreting word meanings based on context, WSD enables more accurate sentiment categorization, which is especially valuable for businesses analysing customer feedback, social media sentiment, and product reviews.

VIII. CONCLUSIONS

Word Sense Disambiguation (WSD) remains a critical task in Natural Language Processing (NLP) due to its pivotal role in resolving lexical ambiguity and enhancing the performance of numerous downstream applications. This paper has explored the evolution of WSD methodologies, from early rule-based systems and knowledge-based approaches to advanced machine learning and deep learning techniques. While each approach has contributed significantly to the development of WSD, challenges such as polysemy, domain-specific adaptability, and the scarcity of annotated resources in low-resource languages persist.

WSD’s importance extends beyond traditional NLP tasks like machine translation and information retrieval to emerging fields such as bioinformatics, semantic web applications, and lexicography. The integration of linguistic resources with advanced computational techniques, including contextual embedding’s and deep neural networks, has substantially improved the accuracy and scalability of WSD systems.

Despite these advancements, achieving robust, scalable solutions that address rare senses and specialized domains remains an ongoing research priority. Future innovations in WSD will require interdisciplinary efforts, combining advancements in computational linguistics, artificial intelligence, and domain-specific knowledge. By overcoming existing limitations, WSD has the potential to further enhance the understanding, interpretation, and processing of human language, driving innovation across a broad spectrum of applications in NLP and beyond.

ACKNOWLEDGMENT

I express my sincere gratitude to my academic mentors 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] Advaith, V., Shivkumar, A., & Lakshmi, B. S. (2022, July). Parts of Speech Tagging for Kannada and Hindi Languages using ML and DL models. In 2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONNECT) (pp. 1-5). IEEE.

[2] Pu, X., Pappas, N., Henderson, J., & Popescu-Belis, A. (2018). Integrating weakly supervised word sense disambiguation into neural machine translation. Transactions of the Association for Computational Linguistics, 6, 635-649.

[3] Asim, M. N., Isim, M., Khan, M. U. G., Mahmood, N., & Mahmood, W. (2019). The use of ontology in retrieval: a study on textual, multilingual, and multimedia retrieval. IEEE Access, 7, 21662-21686.

[4] Sumanth, C., & Inkpen, D. (2015, September). How much does word sense disambiguation help in sentiment analysis of micro post data? In Proceedings of the 6th workshop on computational approaches to subjectivity, sentiment and social media analysis (pp. 115-121).

[5] Xu, G., Yu, Z., Yao, H., Li, F., Meng, Y., & Wu, X. (2019). Chinese text sentiment analysis based on extended sentiment dictionary. IEEE Access, 7, 43749-43762.

[6] Asim, M. N., Isim, M., Khan, M. U. G., Mahmood, N., & Mahmood, W. (2019). The use of ontology in retrieval: a study on textual, multilingual, and multimedia retrieval. IEEE Access, 7, 21662-21686.

[7] Advaith, V., Shivkumar, A., & Lakshmi, B. S. (2022, July). Parts of Speech Tagging for Kannada and Hindi Languages using ML and DL models. In 2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONNECT) (pp. 1-5). IEEE.

[8] Menai, Mohamed El Bachir. "Word sense disambiguation using an evolutionary approach." Informatica 38, no. 3 (2014).

[9] Lake, Brenden M., and Gregory L. Murphy. "Word meaning in minds and machines." Psychological review 130, no. 2 (2023): 401.

[10] Gul, S., Azher, M., & Nawaz, S. (2021). Development of Saraiki wordnet by mapping of word senses: A corpus-based approach. Linguistics and Literature Review, 7(2), 46-66.

[11] Rajan, A., & Salgaonkar, A. (2022). Survey of NLP resources in low-resource languages Nepali, Sindhi and Konkani. In Information and Communication Technology for Competitive Strategies (ICTCS 2020) ICT: Applications and Social Interfaces (pp. 121-132). Springer Singapore.

[12] Gogoi, A., & Baruah, N. (2022). A Lemmatizer for Low-resource Languages: WSD and Its Role in the Assamese Language. Transactions on Asian and Low-Resource Language Information Processing, 21(4), 1-22.

[13] Sodhar, I. N., Sulaiman, S., Buller, A. H., & Sodhar, A. N. (2023). Hybrid Approach Used to Analyze the Sentiments of Romanized Text (Sindhi). International Journal of Advanced Computer Science and Applications, 14(3).

[14] Sharma, Dilip Kumar. "A comparative analysis of Hindi word sense disambiguation and its approaches." In International conference on computing, communication & automation, pp. 314-321. IEEE, 2015.

[15] Sutriawan, Sutriawan, Supriadi Rustad, Guruh Fajar Shidik, Pujiono Pujiono, and Muljono Muljono. "Review of Ambiguity Problem in Text Summarization Using Hybrid ACA and SLR." Intelligent Systems with Applications (2024): 200360.

[16] Li, Qing, Zhihang Hu, Yixuan Wang, Lei Li, Yimin Fan, Irwin King, Gengjie Jia, Sheng Wang, Le Song, and Yu Li. "Progress and opportunities of foundation models in bioinformatics." Briefings in Bioinformatics 25, no. 6 (2024): bbae548.

[17] Sultonov, Akmal, Sharof Musaev, Mavludahon Xajimatova, Shokhrukh Ustemirov, and Akbarali Sattorov. "Pollutant Standards for Mining Enterprises." (2021).