International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

A Comprehensive Comparison of Knowledge-Based, Supervised, and Unsupervised Techniques for Word Sense Disambiguation

Shruti Garg

Guest Faculty, Data Science Department, PMCOE, Govt. Arts and Science College, Ratlam
Email: *shrutigarg878[at]gmail.com*

Abstract: ***Word Sense Disambiguation (WSD) is a cornerstone and persistent challenge in Natural Language Processing (NLP), critical for enabling machines to achieve a human-like understanding of language. The task involves computationally identifying the intended meaning of a polysemous word within a specific context. This paper presents a comprehensive survey and a rigorous comparative analysis of the three dominant paradigms in WSD: Knowledge-Based, Supervised, and Unsupervised methods. We provide an in-depth examination of the core methodologies, tracing their evolution from early heuristic and graph-based approaches to modern deep learning and sense embedding techniques. The comparison is structured across multiple dimensions, including performance, data dependency, computational efficiency, robustness, and interpretability. Our analysis confirms that while supervised deep learning models achieve state-of-the-art results on benchmark tasks, they are fundamentally constrained by the knowledge acquisition bottleneck—the scarcity of sense-annotated data. Knowledge-based methods offer greater domain independence and interpretability but often lag in accuracy. Unsupervised methods and, more recently, knowledge-informed neural models present a promising path forward by leveraging large, unlabeled corpora and structured lexical resources. The paper concludes that the optimal WSD technique is highly application-dependent, and future breakthroughs will likely stem from hybrid architectures that seamlessly integrate the robustness of knowledge bases with the representational power of contextualized language models.***

Keywords: Word Sense Disambiguation, Natural Language Processing, Computational Linguistics, Senseval, SemEval, Lesk Algorithm, Supervised Learning, Unsupervised Learning, Neural Networks, Word Embeddings, BERT

1. Introduction

Natural language is defined by its lexical ambiguity. A single word form can be associated with multiple meanings (senses), and the correct interpretation is almost exclusively determined by its linguistic context. For instance, the word "bass" can refer to a type of fish or low-frequency sound. While humans resolve this ambiguity subconsciously, automating this process, known as Word Sense Disambiguation (WSD), remains a central and challenging problem in NLP [1].

The significance of WSD is profound and directly impacts the performance of higher-level NLP applications. In Machine Translation, the correct sense dictates the lexical choice in the target language (e.g., "bass" fish translates to "loup" in French, while the sound translates to "basse"). In Information Retrieval, a query for "Python" should distinguish between the snake and the programming language to improve precision. Similarly, Question Answering, Text Summarization, and Semantic Search rely on a precise, unambiguous understanding of text, making WSD an indispensable component.

Decades of research have crystallized into three primary paradigms for tackling WSD:

1) **Knowledge-Based Methods:** Leverage structured information from lexical knowledge bases (e.g., WordNet, BabelNet) without requiring annotated data.
2) **Supervised Methods:** Frame WSD as a classification task, learning a mapping from contextual features to sense labels from manually annotated corpora.
3) **Unsupervised Methods:** Automatically cluster word usages into sense groups based on distributional properties, without relying on sense inventories or labeled data.

This paper provides a structured and detailed comparison of these techniques. Section 2 offers a deep dive into the methodologies of each paradigm. Section 3 presents a multi-faceted comparative analysis. Section 4 discusses the implications of our findings and outlines future research directions, and Section 5 concludes the paper.

2. Techniques for Word Sense Disambiguation

2.1 Knowledge-Based Methods

These methods exploit the rich semantic networks found in lexical knowledge bases (LKBs) like WordNet [2], Wiktionary, or BabelNet [3]. Their principal advantage is domain independence, as they do not require training on sense-tagged corpora.

1) **Lesk Algorithm and its Variants:** The original Lesk algorithm [4] operates on the premise that words in a given context will share a common topic. It disambiguates a target word by comparing its dictionary definition (gloss) with the glosses of every other word in its context. The sense whose gloss shares the largest number of overlapping words with the combined context glosses is selected. While intuitive, its performance is limited by the brevity of glosses.
   - **Simplified Lesk:** A more practical variant that compares the gloss of a target word sense only with the bag-of-words from the immediate context window, leading to improved efficiency and often better accuracy [5].

- **Corpus-Lesk:** Enhances the sense representation by incorporating example sentences, related terms, and other information from a large corpus, providing a richer feature set for comparison and mitigating data sparsity [6].
- **Concept - Based Methods (Graph -Based):** These methods model the context as a graph of interconnected concepts (senses) from an LKB and use graph algorithms to identify the most coherent set of senses.
- **Personalized PageRank (PPR):** This is a seminal approach [7]. It constructs a graph where nodes are synsets (e.g., from WordNet) and edges are semantic relations (e.g., hypernymy, meronymy). A random walk is initiated from the synsets of all content words in the context. The PPR algorithm biases the walk towards these start nodes, and the sense of the target word with the highest stationary probability is selected. This effectively identifies the most "central" or "relevant" sense within the local semantic graph.

Strengths: High portability across domains, no need for annotated data, decisions are interpretable.

Weaknesses: Performance is capped by the coverage and quality of the LKB; struggles with fine-grained sense distinctions and domain-specific senses not recorded in the LKB.

2.2 Supervised Methods: A Deep Dive

Supervised methods treat WSD as a supervised classification task, where each instance of a target word is a data point represented by a feature vector, and the goal is to learn a classifier that maps this vector to a discrete sense label.

2.2.1 Feature Engineering and Classical Models

The performance of early supervised systems hinged on sophisticated feature engineering. Key features included:

- **Local Collocations:** The words and their positions in a fixed window (e.g., ±3 words) around the target word. These are highly discriminative features.
- **Syntactic Features:** Part-of-speech tags of the target and surrounding words, syntactic dependencies (e.g., the subject or object of the target word).
- **Bag-of-Words:** All content words in a larger context window, often weighted by TF-IDF, to capture the broader topic.
- **Semantic Features:** Features derived from subject codes (e.g., from LDOCE) or
- topical representations from Latent Dirichlet Allocation (LDA).

Prominent classical models included:

- **Support Vector Machines (SVM):** Became the workhorse for WSD due to their effectiveness in high-dimensional feature spaces [8]. They were particularly successful in Senseval/SemEval competitions.
- **Decision Trees and Naïve Bayes:** Provided strong baselines and were valued for their relative interpretability.

2.2.2 The Deep Learning Revolution

Neural network models have superseded feature-based models by automatically learning relevant feature representations from raw text.

a) Neural Sequence Models:

Bi-directional LSTMs (Bi-LSTMs): These models process the sentence sequentially in both directions, creating a context-aware representation for each word. The hidden state corresponding to the target word is then fed into a softmax classifier for sense prediction [9]. Bi-LSTMs effectively capture long-range dependencies that are crucial for disambiguation.

b) Contextualized Word Embeddings and Fine-Tuning:

BERT and Transformers: Models like BERT (Bidirectional Encoder Representations from Transformers) [10] represent a paradigm shift. They generate dynamic, context-sensitive embeddings for each word token. For WSD, the standard approach is to take the contextualized embedding of the target word (e.g., the word piece embeddings for "bank") and feed it into a classification layer. The entire model is then fine-tuned on a sense-annotated corpus like SemCor. This approach has achieved state-of-the-art results on standard English all-words WSD benchmarks [11].

Strengths: State-of-the-art accuracy on benchmark datasets when sufficient training data is available; ability to capture complex, non-linear contextual patterns.

Weaknesses: The Knowledge Acquisition Bottleneck: Heavy reliance on large, high-quality, manually annotated datasets, which are expensive and scarce. Models are often word-specific (one classifier per word) and generalize poorly to new words or domains not seen during training.

2.3 Unsupervised and Knowledge-Lean Methods: A Detailed View

These methods aim to circumvent the need for sense-annotated data by leveraging the distributional hypothesis—that words with similar meanings occur in similar contexts.

2.3.1 Pure Unsupervised Methods (Sense Induction)

These methods do not assume a pre-defined sense inventory. Their goal is to automatically discover a word's senses from raw text.

1) **Context Clustering:** The classic approach involves collecting all context vectors (e.g., bag-of-words representations) for a target word from a large corpus and applying a clustering algorithm like K-means or Hierarchical Agglomerative Clustering. Each resulting cluster is presumed to represent a distinct sense of the word [12]. The number of clusters k is a critical and often difficult-to-set parameter.
2) **Word Sense Induction via Sense Embeddings:** Modern approaches learn multiple embeddings per word type, each representing a different sense.
   - **Neelakantan et al. (2014) [13]:** This seminal work proposed non-parametric models that extend the Skip-gram architecture to learn multiple sense-specific embeddings for a word. The model dynamically

assigns a context to one of the word's sense clusters during training.
- **AutoExtend:** This method projects pre-trained word embeddings onto the synsets of a lexical resource like WordNet, effectively inducing synset (sense) embeddings without direct supervision [14].

2.3.2 Semi- Supervised and Knowledge -Lean Methods

These methods use a very small amount of supervision (e.g., a few seed words or a sense inventory) to guide the disambiguation process.

Bootstrapping (Yarowsky's Algorithm) : A classic semi-supervised method [15]. It starts with a small set of seed examples or highly reliable rules for each sense (e.g., "river" strongly indicates
the geographical sense of "bank"). It then iteratively:

1) Labels new instances in a large unlabeled corpus using the current seeds/rules.
2) Learns new predictive features (collocations) from the newly labeled data.
3) Adds the most confident new examples to the seed set.

This process repeats, gradually expanding the training data.

Strengths: Do not require sense-annotated data; can adapt to new domains and discover novel, emerging word senses (e.g., "twitter" as a social media platform).

Weaknesses: Induced senses may not align with human-curated sense inventories (e.g., WordNet), making evaluation difficult; overall accuracy is generally lower than fully supervised methods; parameter tuning (like the number of clusters *k*) can be non-trivial.

3. Comparative Analysis

This section provides a structured comparison of the three paradigms. Performance metrics are generalized trends based on results from standard benchmarks like Senseval-2, Senseval-3, and SemEval tasks.

Table 1: Comparative Analysis of WSD Techniques

Dimension	Knowledge based	Supervised	Unsupervised
Performance (F1-Score)	Moderate (60-75% F1)	High (75- 80%+ F1)	Low to Moderate (50-70% F1)
Data Dependency	None (uses LKBs)	High (needs annotated data)	None / Low (uses raw text)
Computational Cost (Inference)	Low to Moderate	Low (after training)	Moderate to High (per-word clustering)
Robustness & Scalability	High (domain-independent)	Low (domain-dependent)	High
Interpretability	High (gloss overlap, graph centrality)	Low (black-box models)	Moderate (cluster centroids)
Handling New Words/ Senses	No (limited by LKB)	No	Yes

4. Discussion of Comparison

- **Performance:** Supervised methods, particularly those based on fine-tuned transformers like BERT, consistently achieve the highest F1-scores on standardized tests. They excel at modeling complex, non-linear contextual patterns. Knowledge-based methods are competitive for words with highly distinct senses but struggle with fine-grained distinctions. Unsupervised methods' performance is highly variable and depends on the corpus and clustering parameters.
- **Data Dependency:** This is the most critical trade-off. The superior performance of supervised methods is contingent upon the existence of large, high-quality labeled datasets for thousands of words, which are a major bottleneck. This makes them impractical for many real-world, multi-domain scenarios. Knowledge-based and unsupervised methods are far more practical and scalable in this regard.
- **Robustness and Scalability:** Supervised models are prone to domain shift; a model trained on news text will perform poorly on biomedical abstracts. Knowledge-based methods, being rooted in general-purpose LKBs, and unsupervised methods, which can learn from any domain-specific corpus, are inherently more robust and scalable.

The Interpretability Spectrum: Knowledge-based methods offer clear reasoning—a decision can be traced back to gloss overlaps or PageRank scores. In contrast, the decisions of a deep neural network are largely inscrutable. Unsupervised methods offer some interpretability by examining the characteristic words in each cluster centroid.

5. Discussion and Future Directions

The choice of a WSD technique is not a question of which is universally "best," but which is most suitable for a given application's constraints and objectives. For high-stakes applications where maximum accuracy is required and annotated data exists for the domain, a supervised neural model is the optimal choice. For broad-coverage, domain-independent applications (e.g., a general-purpose semantic search engine), a robust knowledge-based method like PPR is more appropriate. For exploring new domains or tracking semantic change over time, unsupervised sense induction is the only viable path.

Promising future research directions are focused on hybrid models and leveraging new paradigms:

1) **Knowledge-Informed Neural Models:** The most promising direction involves integrating structured knowledge from LKBs directly into neural network architectures. Using Graph Neural Networks (GNNs) to propagate information through WordNet or BabelNet during training can enhance a model's semantic reasoning and provide a bridge between data-driven and knowledge-driven approaches [16].
2) **Fully Unsupervised WSD with Pre-trained LMs:** Instead of fine-tuning, new methods probe the intrinsic capabilities of large pre-trained language models (LLMs) like BERT and GPT for WSD, using techniques like sentence-pair classification or analyzing attention patterns to perform "zero-shot" or "few-shot" WSD [17].

3) **Cross-Lingual and Multilingual WSD:** Leveraging annotations in resource-rich languages (like English) to perform WSD in low-resource languages, using aligned multilingual knowledge bases like BabelNet [3] and cross-lingual word embeddings.
4) **Evaluation Beyond Fine-Grained WSD:** Developing benchmarks and metrics for tasks like domain-specific WSD, sense discovery, and lexical semantic change detection, which better reflect real-world challenges.

6. Conclusion

This paper has presented a comprehensive and detailed comparison of the primary techniques for Word Sense Disambiguation. We have delineated the core principles, key algorithms, and evolutionary trajectory of knowledge-based, supervised, and unsupervised paradigms, providing a clear exposition of their respective strengths and weaknesses. The analysis confirms a fundamental trade-off: supervised methods deliver superior accuracy but are critically hampered by their dependence on annotated data, while knowledge-based and unsupervised methods offer greater flexibility, robustness, and scalability at the cost of some precision. The field is now converging on hybrid and knowledge-informed models that seek to combine the data-driven power of neural networks with the structured, interpretable knowledge of lexical resources. As NLP systems continue to evolve towards a more profound and nuanced understanding of human language, advancing the state-of-the-art in WSD will remain a critical and dynamic research frontier.

References

[1] Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2), 1-69.
[2] Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11), 39-41.
[3] Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193, 217-250.
[4] Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation* (pp. 24-26).
[5] Kilgarriff, A., & Rosenzweig, J. (2000). Framework and results for English SENSEVAL. *Computers and the Humanities*, 34(1-2), 15-48.
[6] Banerjee, S., & Pedersen, T. (2002). An adapted Lesk algorithm for word sense disambiguation using WordNet. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 136-145).
[7] Agirre, E., & Soroa, A. (2009). Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the ACL* (pp. 33-41).
[8] Lee, Y. K., & Ng, H. T. (2002). An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 41-48).
[9] Kågebäck, M., & Salomonsson, H. (2016). Word sense disambiguation using a bidirectional LSTM. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)* (pp. 51-56).
[10] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
[11] Huang, L., Sun, C., Qiu, X., & Huang, X. (2019). GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3509-3514).
[12] Pedersen, T., & Bruce, R. (1997). Distinguishing word senses in untagged text. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing* (pp. 197-207).
[13] Neelakantan, A., Shankar, J., Passos, A., & McCallum, A. (2014). Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1059-1069).
[14] Rothe, S., & Schütze, H. (2015). AutoExtend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1793-1803).
[15] Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics* (pp. 189-196).
[16] Wang, X., Wang, D., Xu, C., He, X., Cao, Y., & Chua, T. S. (2019). Explainable reasoning over knowledge graphs for recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, No. 01, pp. 5329-5336).
[17] Wiedemann, G., Remus, S., Chawla, A., & Biemann, C. (2020). Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2020)*.