Frontiers in Emerging Computer Science and Information Technology (Open Access)

Volume 02, Issue 06, June 2025,
Publish Date: 20-06-2025
PageNo.16-21

Refining Word Sense Disambiguation via Multilevel Clustering of Lexical Representations

Prof. Omar F. Abdelrahman
Department of Computer Engineering, Cairo University, Cairo, Egypt

Laila K. Mostafa
Department of Computer Engineering, Cairo University, Cairo, Egypt

ABSTRACT

Word Sense Disambiguation (WSD), the task of identifying the correct meaning of a word in a given context, remains a pivotal challenge in Natural Language Processing (NLP). This article explores the application of multilevel clustering techniques to enhance WSD accuracy and provide clearer contextual understanding. By leveraging hierarchical analysis across various linguistic features, including word embeddings, lexical networks, and syntactic patterns, this approach aims to capture intricate semantic relationships. We discuss the methodological framework for integrating clustering at different analytical levels and synthesize the potential benefits, particularly in addressing polysemy and homonymy. The proposed multilevel clustering paradigm offers a robust pathway for refining sense assignments, leading to improved performance in downstream NLP applications.

KEYWORDS: Word sense disambiguation, multilevel clustering, lexical representations, natural language processing, semantic analysis, machine learning, computational linguistics, vector embeddings, language modeling, disambiguation algorithms.

INTRODUCTION

The ambiguity inherent in natural language is a fundamental hurdle for computational linguistics. Words, especially polysemous ones, often carry multiple meanings, and the precise interpretation depends heavily on the surrounding context [1]. Word Sense Disambiguation (WSD) is the computational task of assigning the most appropriate meaning (or "sense") to a word within a specific sentence or discourse [2]. The accurate resolution of word ambiguity is critical for a wide array of NLP applications, including machine translation, information retrieval, question answering, and text summarization, where a misunderstanding of a single word's intended meaning can propagate errors throughout the system [1, 9, 10, 19].

Historically, WSD approaches have broadly fallen into three categories: knowledge-based, supervised, and unsupervised methods [11]. Knowledge-based methods rely on lexical resources such as WordNet [14] or other thesauri to determine word senses, often employing measures of semantic similarity or relatedness between the ambiguous word's context and the definitions of its senses [1, 2]. Supervised methods treat WSD as a classification problem, training models on manually sense-tagged corpora. While often achieving high accuracy, their main limitation is the significant cost and effort associated with creating large, sense-annotated datasets [10]. Unsupervised methods, conversely, do not require pre-tagged data; instead, they infer word senses by clustering contexts in which an ambiguous word appears. These methods aim to group similar contexts together, with each cluster ideally corresponding to a distinct sense of the word [13, 5].

Clustering has emerged as a particularly promising technique within unsupervised and semi-supervised WSD paradigms. The core idea is that contexts in which a word shares the same meaning will exhibit higher similarity compared to contexts where the word carries different meanings. By grouping these similar contexts, distinct sense clusters can be identified [6, 13]. Advances in distributional semantics, particularly word embeddings, have further propelled clustering-based WSD, enabling the representation of words and their contexts in high-dimensional vector spaces, where semantic similarity can be quantified as vector proximity [18, 5].

Despite the advancements, WSD remains an open problem. Many existing clustering-based approaches operate at a single level of linguistic analysis, often focusing solely on word embeddings or lexical co-occurrence patterns. However, word senses are not merely defined by local context but are influenced by a complex interplay of lexical,

syntactic, and semantic factors that span multiple levels of linguistic representation. This paper proposes that a multilevel analytical approach, leveraging clustering techniques at various granularities, can provide a more comprehensive and accurate framework for WSD. By integrating information from different linguistic strata, we aim to capture richer contextual clues and disambiguate word senses with greater precision.
The objective of this article is to explore how multilevel clustering can enhance WSD by examining different analytical layers and their synergistic effects. We will delve into the methodological aspects of applying clustering at these distinct levels and discuss the potential improvements in WSD performance.

METHODS

The proposed approach for refining Word Sense Disambiguation (WSD) leverages a multilevel clustering framework, designed to capture the nuanced semantic distinctions of polysemous words by analyzing contextual information at varying levels of abstraction. This section outlines the theoretical underpinnings and methodological components of this multilevel analysis, emphasizing how clustering techniques are applied at each stage.

Multilevel Analytical Framework

The concept of multilevel analysis in WSD stems from the understanding that word senses are not solely determined by immediate neighbors but by a hierarchy of linguistic cues. Our framework identifies three primary levels of analysis:

1. Lexical-Semantic Level: This level focuses on the intrinsic semantic properties of words and their direct relationships within a lexical network. Resources like WordNet play a crucial role here, providing a structured hierarchy of senses and their definitions, synonyms, antonyms, and hypernyms/hyponyms [14, 2].
2. Contextual Embedding Level: This level captures the distributional semantics of words within their immediate and broader textual environments. Word embeddings, such as Word2Vec, GloVe, or context-aware models like BERT, represent words as dense vectors where semantically similar words are mapped to proximate points in a high-dimensional space [18, 9].
3. Syntactic and Relational Level: This level considers the grammatical roles and relational dependencies between words in a sentence. While often implicitly captured by advanced contextual embeddings, explicit analysis of syntactic structures can provide additional constraints for disambiguation, especially for distinguishing between different usages of a verb or noun [7].

The integration of these levels allows for a more robust understanding of word meaning, moving beyond simple co-occurrence counts to richer semantic and structural representations [12].

2.2. Clustering Algorithms and Their Application

Clustering forms the core of this multilevel approach, employed at each analytical layer to identify distinct sense groups. Various clustering algorithms are suitable, chosen based on the nature of the data representation at each level:

2.2.1. Sense Clustering on Lexical Networks

At the lexical-semantic level, clustering can be applied directly to the WordNet graph or similar knowledge graphs. Nodes in such graphs represent word senses, and edges represent semantic relations. Graph-based clustering algorithms are particularly effective here [8]. For instance, approaches like TKB-UO utilize sense clustering within WordNet to group fine-grained senses into broader, more manageable clusters, thereby mitigating the data sparseness problem often encountered with very specific sense distinctions [3]. This involves creating a similarity graph where nodes are WordNet senses and edge weights reflect their semantic relatedness. Spectral clustering [6] or community detection algorithms [4] can then be applied to partition this graph into sense clusters.

2.2.2. Contextual Clustering with Word Embeddings

This is a prominent approach in unsupervised WSD. For each occurrence of an ambiguous word in a corpus, its context is represented as a vector. This can be achieved by averaging the embeddings of surrounding words, using specialized context vectors from models like BERT, or even by combining word and context embeddings [5, 16]. Once these context vectors are generated, traditional clustering algorithms such as K-means, DBSCAN, or hierarchical clustering are applied. Each resulting cluster is hypothesized to correspond to a distinct sense of the ambiguous word [13]. For example, "bank" appearing in financial contexts would cluster separately from "bank" appearing in riverine contexts. Unsupervised most frequent sense detection has been explored using word embeddings, where the largest cluster is often designated as the most frequent sense [5]. CluBERT, for instance, uses a cluster-based approach for learning sense distributions by leveraging BERT embeddings [16].

2.2.3. Hybrid and Semi-supervised Clustering

To bridge the gap between purely unsupervised methods and the resource-intensive supervised approaches, semi-supervised clustering methods can be incorporated [10, 15]. These methods can leverage a small amount of labeled data (e.g., a few sense-tagged examples) to guide the clustering process, improving the quality of the resulting clusters. This often involves incorporating constraints (must-link/cannot-link) into the clustering objective function based on the labeled data. Kernel methods have also been explored to handle non-linear relationships in data for WSD, which can be integrated into clustering or classification frameworks [11]. Furthermore, a combination of clustering and classification can be used, where clusters are first formed, and then a classifier is trained on a small subset of labeled cluster members [20].

2.3. Multilevel Integration and Disambiguation Strategy

The strength of the multilevel approach lies in the integration of information derived from clustering at each level. The disambiguation process then becomes a decision-making task that considers the sense assignments from all levels.

- Initial Sense Candidates: For an ambiguous word occurrence, initial sense candidates can be generated by querying a lexical resource like WordNet [2].
- Contextual Refinement: The contextual embedding of the target word instance is then calculated. This instance is assigned to the nearest cluster in the contextual embedding space. This provides a data-driven, usage-based sense prediction.
- Lexical-Semantic Validation: The sense suggested by the contextual clustering is then validated against the sense clusters derived from the lexical network. If the contextual cluster's predicted sense aligns well with a predefined sense cluster from WordNet, the confidence in the assignment increases.
- Syntactic Augmentation (if applicable): For highly ambiguous cases or specific linguistic phenomena, syntactic parse trees or dependency relations can be used to further refine the sense selection. For example, a word might have different senses depending on whether it acts as a verb or a noun, or its specific arguments. Approaches like those focusing on multilevel center embedding for sentence similarity [7] can inform how structural context influences meaning.
- Ensemble or Weighted Voting: Finally, a mechanism (e.g., weighted voting, ensemble learning, or a rule-based system) combines the "votes" or probabilities from each level's clustering output to arrive at the final, most probable sense. This ensures that the final decision benefits from the strengths of each analytical layer.

This comprehensive, multilevel approach aims to address the limitations of single-level WSD systems by creating a richer and more robust representation of word meaning through integrated clustering strategies.

RESULTS (Synthesized Findings)

While this article presents a conceptual framework for a multilevel clustering approach to Word Sense Disambiguation (WSD), the "results" section synthesizes expected outcomes and general findings observed in related research that employs similar principles of hierarchical or integrated analysis. The efficacy of a multilevel clustering paradigm in WSD is supported by the improved ability to resolve semantic ambiguities through a more comprehensive contextual understanding.

3.1. Enhanced Disambiguation Accuracy

Multilevel clustering approaches are anticipated to yield higher disambiguation accuracy compared to single-level methods. By combining insights from lexical networks (e.g., WordNet's structured knowledge), contextual embeddings (capturing distributional semantics), and potentially syntactic patterns, the system gains a richer set of features for sense discrimination.

- Improved Handling of Fine-Grained Senses: The integration of knowledge-based sense clustering [3] helps in grouping overly fine-grained dictionary senses into more practically distinguishable clusters. This reduces the problem of data sparsity often encountered with very specific sense definitions, leading to more robust sense assignments.
- Robustness to Ambiguous Contexts: When a word appears in a highly ambiguous or sparse context, relying solely on local co-occurrence or basic embeddings can be insufficient. A multilevel approach allows for fallback or supplementary information from broader semantic relationships within lexical graphs, making the disambiguation process more resilient [17]. For example, the automatic word sense disambiguation and construction identification based on corpus multilevel annotation demonstrates the utility of integrating varied linguistic information [12].

3.2. Better Capture of Semantic Nuances

The hierarchical nature of multilevel clustering allows for the capture of subtle semantic differences that might be overlooked by flatter models.

- Contextual Richness: By leveraging advanced contextual embeddings (e.g., from BERT-like models), the system can distinguish senses based on nuanced semantic and syntactic patterns in the surrounding text [9, 16]. The clustering of these rich contextual vectors leads to more semantically coherent sense clusters.
- Relational Insights: The incorporation of relational information, either implicitly through contextual embeddings or explicitly through graph-based analysis, helps differentiate between senses based on the typical arguments or grammatical roles a word assumes. For instance, distinguishing between "drive" as a verb meaning to operate a vehicle versus "drive" as a noun referring to a hard disk relies on the different syntactic

contexts and semantic roles the word plays. Some approaches, such as Multilevel Center Embedding for Sentence Similarity, hint at how complex structures can be analyzed for deeper semantic understanding [7].

3.3. Adaptability and Generalization

Multilevel clustering offers enhanced adaptability and generalization capabilities, particularly in unsupervised or semi-supervised settings.

- Reduced Reliance on Labeled Data: While supervised WSD methods often achieve high accuracy, they suffer from the bottleneck of requiring extensive manually tagged corpora. Multilevel clustering, especially when employed within an unsupervised or semi-supervised framework, significantly reduces this dependency. By effectively clustering unlabeled contextual data and validating against knowledge bases, it can generalize well to unseen words or domains with minimal supervision [10, 15].
- Cross-Lingual Potential: Although not explicitly detailed in the methods, the principles of multilevel clustering, especially those leveraging distributional semantics, can be extended to cross-lingual WSD. Techniques like CluBERT, which focus on learning sense distributions in multiple languages, show the potential for such approaches to generalize across linguistic boundaries [16].

In summary, the synthesis of findings from related research strongly suggests that a multilevel clustering approach provides a more comprehensive, accurate, and adaptable solution for Word Sense Disambiguation. By integrating information from diverse linguistic levels, it addresses the limitations of single-level systems, leading to a clearer and more precise understanding of word meanings in context.

DISCUSSION

The exploration of a multilevel clustering framework for Word Sense Disambiguation (WSD) presents a compelling direction for advancing the field. This approach builds upon the strengths of unsupervised and semi-supervised clustering techniques by systematically integrating various levels of linguistic information, from lexical-semantic networks to contextual embeddings and syntactic patterns.

4.1. Advantages of Multilevel Clustering for WSD

The primary advantage of this multilevel paradigm lies in its ability to harness a broader spectrum of linguistic cues, leading to a more robust and accurate disambiguation process.

- Comprehensive Contextualization: By considering both local contextual embeddings and global semantic relationships from knowledge graphs (like WordNet), the system gains a holistic view of word meaning. This allows for better discrimination between closely related senses that might otherwise be conflated by single-level models. Knowledge-based approaches using WordNet have long been foundational [2, 14], and combining their structural richness with the empirical power of contextual embeddings offers a potent synergy.
- Addressing Data Sparsity: Unsupervised clustering methods inherently address the data sparsity problem associated with supervised WSD, as they do not require large amounts of sense-tagged data for training [13]. Furthermore, by clustering senses within knowledge graphs, very fine-grained distinctions can be grouped into coarser, more robust sense inventories, making the task more tractable [3].
- Leveraging Implicit and Explicit Knowledge: The approach effectively combines implicit knowledge captured by word embeddings (derived from vast text corpora) with explicit, human-curated knowledge encoded in lexical databases. This dual approach helps in both discovering novel sense distinctions and validating them against established semantic structures.
- Flexibility in Integration: The modular nature of analyzing different levels allows for flexibility. New types of linguistic features (e.g., discourse-level information, sentiment, or domain-specific terminology) can be incorporated as additional levels of analysis, further enriching the contextual understanding.

4.2. Challenges and Limitations

Despite its potential, the multilevel clustering approach faces several challenges:

- Feature Engineering and Representation: Deciding on the optimal representation for each level (e.g., choice of word embedding model, method for context vector generation, graph construction from lexical resources) is crucial. The quality of these representations directly impacts the clustering results [18].
- Integration Complexity: Combining the outputs from multiple clustering processes is non-trivial. Designing an effective ensemble or decision-making mechanism that optimally weighs the contributions from each level requires careful consideration. Simple voting mechanisms might not always capture the nuanced interactions between different linguistic features.
- Computational Cost: Analyzing and clustering data at multiple levels, especially with large corpora and high-dimensional embeddings, can be computationally intensive. Optimizing algorithms for efficiency is paramount.
- Evaluation Metrics: Traditional WSD evaluation metrics might not fully capture the benefits of a multilevel

approach, particularly if it leads to more nuanced sense distinctions or better generalization on out-of-domain data. Developing metrics that assess the richness of sense representations and their utility in downstream applications could be beneficial.

- Dynamic Nature of Language: Language is dynamic, with new word usages and senses emerging over time. While distributional semantics can capture this to some extent, integrating dynamic updates into static knowledge graphs or requiring re-clustering can be challenging.

4.3. Future Directions

Several avenues for future research can further enhance the multilevel clustering approach for WSD:

- Deep Learning Integration: Explore more sophisticated deep learning architectures that can inherently learn multilevel representations and perform joint clustering and disambiguation. For instance, extending models like GlossBERT [9] or CluBERT [16] to explicitly incorporate multiple, distinct clustering layers rather than just end-to-end classification.
- Reinforcement Learning for Integration: Investigate the use of reinforcement learning to dynamically learn the optimal weighting or combination strategy for integrating outputs from different clustering levels, rather than relying on heuristic-based ensembles.
- Explainability: Focus on developing methods to make the decision-making process of multilevel WSD more transparent and explainable. Understanding why a particular sense is chosen based on features from different levels would be valuable for debugging and improving trust in the system.
- Domain Adaptation: Research how multilevel clustering can be effectively adapted for specific domains where word senses might differ significantly from general language. This could involve domain-specific fine-tuning of embeddings and integration with domain ontologies.
- Cross-Lingual Multilevel WSD: Extend the framework to handle cross-lingual WSD, leveraging multilingual embeddings and parallel corpora to align word senses across languages. This could involve clustering aligned contexts or sense definitions.
- Interactive WSD: Develop interactive systems where human annotators can provide feedback to refine clusters or correct sense assignments, creating a human-in-the-loop system that continuously improves the WSD model. The principles of semi-supervised learning integrated with classifier combination [10] and learning model order from labeled and unlabeled data [15] could be further explored in this interactive context.

In conclusion, the adoption of a multilevel clustering approach for WSD offers a powerful methodology to tackle the complexities of lexical ambiguity. By synthesizing information from diverse linguistic levels, it moves towards a more comprehensive and accurate understanding of word meanings, paving the way for more intelligent and robust NLP systems. Continued research in this area, particularly focusing on integration strategies, computational efficiency, and explainability, holds significant promise for unlocking the full potential of WSD. The application of graph-based algorithms [17] and approaches considering genetic algorithms for specific languages [21] further underscore the diverse avenues for innovation within this framework.

REFERENCES

[1] Agirre E., L´opez de Lacalle O., Soroa A.: Random Walks for Knowledge-Based Word Sense Disambiguation, Computational Linguistics, vol. 40(1), pp. 57–84, 2014. doi: 10.1162/COLIa00164.

[2] AlMousa M., Benlamri R., Khoury R.: A novel word sense disambiguation approach using WordNet knowledge graph, Computer Speech & Language, vol. 74, 101337, 2022. doi: 10.1016/j.csl.2021.101337.

[3] Anaya-S´anchez H., Pons-Porrata A., Berlanga-Llavori R.: TKB-UO: Using Sense Clustering for WSD. In: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, June 2007, pp. 322–325, 2007. doi: 10.3115/1621474.1621544.

[4] Berahmand K., Li Y., Xu Y.: A Deep Semi-Supervised Community Detection Based on Point-Wise Mutual Information, IEEE Transactions on Computational Social Systems, vol. 11(3), pp. 3444–3456, 2023. doi: 10.1109/TCSS.2023.3327810.

[5] Bhingardive S., Singh D., Rudramurthy V., Redkar H., Bhattacharyya P.: Unsupervised most frequent sense detection using word embeddings. In: Proceedings of the 2015 conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies, pp. 1238–1243, 2015. doi: 10.3115/v1/n15-1132.

[6] Chifu A.G., Hristea F., Mothe J., Popescu M.: Word sense discrimination in information retrieval: A spectral clustering-based approach, Information Processing & Management, vol. 51(2), pp. 16–31, 2015. doi: 10.1016/j.ipm.2014.10.007.

[7] Dubey S., Kohli N.: A Multilevel Center Embedding approach for Sentence Similarity having Complex structures. In: 2023 World Conference on Communication & Computing (WCONF), pp. 1–8, 2023. doi: 10.1109/wconf58270.2023.10235102.

[8] Guerrieri A., Rahimian F., Girdzijauskas S., Montresor A.: Tovel: Distributed graph clustering for word sense disambiguation. In: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), pp. 623–630, IEEE, 2016. doi: 10.1109/icdmw.2016.0094.

[9] Huang L., Sun C., Qiu X., Huang X.: GlossBERT: BERT for word sense disambiguation with gloss knowledge, arXiv preprint arXiv:190807245, 2019. doi: 10.18653/v1/d19-1355.

[10] Le A.C., Shimazu A., Huynh V.N., Nguyen L.M.: Semi-supervised learning integrated with classifier combination for word sense disambiguation, Computer Speech & Language, vol. 22(4), pp. 330–345, 2008. doi: 10.1016/j.csl.2007.11.001.

[11] Li X., Qing S., Zhang H., Wang T., Yang H.: Kernel methods for word sense disambiguation, Artificial Intelligence Review, vol. 46, pp. 41–58, 2016. doi: 10.1007/s10462-015-9455-5.

[12] Lyashevskaya O., Mitrofanova O., Grachkova M., Romanov S., Shimorina A., Shurygina A.: Automatic Word Sense Disambiguation and Construction Identification Based on Corpus Multilevel Annotation. In: Text, Speech and Dialogue: 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1–5, 2011. Proceedings 14, pp. 80–90, Springer, 2011. doi: 10.1007/978-3-642-23538-2_11.

[13] Mart´ın T., Berlanga-Llavori R.: A clustering-based approach for unsupervised word sense disambiguation, Procesamiento del Lenguaje Natural, vol. 49, pp. 49–56, 2012.

[14] Miller G.A., Chodorow M., Landes S., Leacock C., Thomas R.G.: Using a semantic concordance for sense identification. In: Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994, 1994. doi: 10.3115/1075812.1075866.

[15] Niu Z.Y., Ji D.H., Tan C.L.: Learning model order from labeled and unlabeled data for partially supervised classification, with application to word sense disambiguation, Computer Speech & Language, vol. 21(4), pp. 609–619, 2007. doi: 10.1016/j.csl.2007.02.001.

[16] Pasini T., Scozzafava F., Scarlini B.: CluBERT: A cluster-based approach for learning sense distributions in multiple languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4008–4018, 2020. doi: 10.18653/v1/2020.acl-main.369.

[17] Patil A.P., Ramteke R., Bhavsar R., Darbari H.: Graph-Based Algorithm for Word Sense Disambiguation: A Performance and Comparison, Sambodhi, vol. 44(03), pp. 77–79, 2021.

[18] Pelevina M., Arefyev N., Biemann C., Panchenko A.: Making sense of word embeddings, arXiv preprint arXiv:170803390, 2017. doi: 10.48550/arXiv.1708.03390.

[19] Seo H.C., Chung H., Rim H.C., Myaeng S.H., Kim S.H.: Unsupervised word sense disambiguation using WordNet relatives, Computer Speech & Language, vol. 18(3), pp. 253–273, 2004. doi: 10.1016/j.csl.2004.05.004.

[20] Shirai K., Nakamura M.: JAIST: Clustering and classification based approaches for Japanese WSD. In: Proceedings of the 5th International Workshop on Semantic Evaluation, Uppsala, Sweden, 15–16 July 2010, pp. 379–382, 2010.

[21] Vaishnav Z.B., Sajja P.S.: Knowledge-Based Approach for Word Sense Disambiguation Using Genetic Algorithm for Gujarati. In: S. Satapathy, A. Joshi (eds.), Information and Communication Technology for Intelligent Systems. Smart Innovation, Systems and Technologies, vol. 1, pp. 485–494, Springer, Singapore, 2019. doi: 10.1007/978-981-13-1742-2_48.