

# Is Word Sense Disambiguation Dead in the LLM Era?

Roberto Navigli

Sapienza University of Rome & Babelscape, Rome, Italy  
navigli@{diag.uniroma1.it, babelscape.com}

## Abstract

Word Sense Disambiguation (WSD) has been a central challenge since the earliest proposals for Machine Translation (MT), most famously Weaver’s 1949 memorandum. Classical systems treated WSD as an explicit task, grounded in lexical resources and annotated data. Recently, however, Large Language Models (LLMs) have blurred the boundary between disambiguation and general language understanding, leading some to suggest that WSD might be obsolete. This paper surveys the role of WSD in the LLM era, drawing on recent studies of encoder-based sense separation and disambiguation, and decoder-based definition selection and generation, as well as multilingual evaluation. Closed-source instruction-tuned LLMs now achieve performance comparable to specialized WSD systems, yet systematic weaknesses remain: non-predominant senses are often misclassified and disambiguation biases in MT persist. We argue that WSD is not “dead” but redefined as a diagnostic lens for assessing lexical-semantic competence, robustness, and interpretability in LLMs.

## Introduction

Word Sense Disambiguation (WSD) – the task of determining the intended meaning of a word in context – is one of the most long-standing problems in Natural Language Processing (NLP). Its roots trace back to Warren Weaver’s 1949 memorandum on translation, which observed that naive word-for-word substitution would fail “for problems of ‘literary’ translation” (Weaver 1949). Weaver argued that true translation required machines to capture the *meaning* of a sentence, implying that lexical ambiguity resolution is central to any computational understanding of language.

For instance, the following sentences use the word *shot* in five distinct senses:

1. The doctor gave her a flu *shot*.
2. He took a *shot* at the basket.
3. He drank a *shot* of whisky.
4. The arrow made a remarkable *shot* of 200 meters.
5. The cottage lay within *shot* of the lighthouse beacon.

While we limit ourselves here to five instances, the word can assume many more senses in English. In the literature, WSD has become a paradigmatic challenge in addressing the difficulty of selecting the appropriate sense of a word in context. Historically, sense selection has relied primarily on the dominant English sense inventory, WordNet (Miller 1995), and many different approaches have been proposed over the decades to tackle the task.

Until just a few years ago, knowledge-based methods relied on the structured information contained within lexical-semantic resources and graphs such as WordNet and, later, BabelNet (Navigli and Ponzetto 2012) to identify the most suitable sense, while supervised machine-learning approaches exploited annotated corpora such as SemCor (Miller et al. 1993) for training. Nonetheless, overall performance remained disappointing, with systems struggling to reach even 70% accuracy.

It was only with the advent of neural models that performance began to improve. These included gloss-informed bi-encoders (Blevins and Zettlemoyer 2020) and generative gloss-based systems (Bevilacqua, Maru, and Navigli 2020). In parallel, encoder-based models such as EWISER (Bevilacqua and Navigli 2020), ESCHER (Barba, Pasini, and Navigli 2021), ConSeC (Barba, Procopio, and Navigli 2021), and SANDWiCH (Guzman-Olivares, Quijano-Sánchez, and Liberatore 2025) pushed performance even further, showing that sense representations could be tightly integrated with contextual encoders. This allowed models to overcome the 80% performance ceiling, still with little information about the nature of the remaining 20% of “errors”.

The rise of Large Language Models (LLMs) has reignited the debate. Transformer-based models capture rich contextual information and, to some extent, appear to resolve senses implicitly. Recent studies suggest that instruction-tuned LLMs achieve performance on par with or even surpassing classical WSD systems (Meconi et al. 2025). Yet evidence remains mixed. Translation bias benchmarks such as DiBiMT (Campolungo et al. 2022; Martelli et al. 2025) show that non-predominant senses are systematically mistranslated, and multilingual evaluations – such as those enabled by XL-WSD (Pasini, Raganato, and Navigli 2021) – reveal uneven coverage and prompt sensitivity (Basile, Musacchio, and Siciliani 2024; Basile et al. 2025). Moreover, ambiguity probing studies highlight shallow heuristics

and overgeneralizations (Liu et al. 2023; Kamath et al. 2024; Capone et al. 2024).

The question, therefore, is not whether WSD is still relevant, but how it has transformed. Explicit WSD modules may no longer be necessary in modern NLP systems, but WSD continues to serve as a stress test for sense-aware generation, translation robustness, and semantic plausibility. In this paper, we survey and discuss findings from recent years to argue that WSD is not “dead” but evolving, i.e. transitioning to a diagnostic lens for evaluating whether LLMs truly *understand* word meaning in all its nuances and complexity.

## A Brief History of WSD

WSD research has progressed through waves: rule- and knowledge-driven methods, supervised models, unsupervised induction, neural architectures, and today’s LLM-driven approaches. Here we briefly overview the field, and highlight the key breakthroughs of each line of approach (see Bevilacqua et al. (2021) for a pre-LLM review of the literature; Navigli (2009) for a historical overview of the field; and Navigli (2018) for a wider discussion of the caveats and the need for explicit Natural Language Understanding).

**Early Approaches.** As mentioned above, WSD was identified as a computational challenge as early as Weaver’s memorandum (Weaver 1949). In the 1970s-1980s, early systems used hand-crafted rules, such as Wilks’s preference semantics, encoding selectional tendencies and semantic templates. These methods faced the so-called *knowledge acquisition bottleneck* (Navigli 2009): rules were costly to create, brittle and incomplete. The use of machine-readable dictionaries in the late 1980s partially alleviated this issue, but approaches remained knowledge-driven, providing foundations without scalability.

**Knowledge-based Approaches.** From the late 1980s onward, dictionary- and graph-based approaches became prominent. Lesk’s algorithm (Lesk 1986) disambiguated words by detecting overlaps among dictionary glosses, inspiring extensions like conceptual density (Agirre and Rigau 1996) and semantic similarity measures (Resnik 1995; Leacock, Chodorow, and Miller 1998); see Navigli and Martelli (2019) for an overview. WordNet was starting to become a standard inventory, and its taxonomical structure fueled methods such as the Adapted Lesk (Banerjee and Pedersen 2002) and graph-based algorithms using PageRank (Mihalcea 2005; Agirre and Soroa 2009; Scozzafava et al. 2020). Multilingual graph resources like BabelNet (Navigli and Ponzetto 2012) – which integrates WordNet with Wikipedia, Wikidata, Wiktionary and other resources – and graph-based algorithms such as Babelfy (Moro, Raganato, and Navigli 2014) – which unifies WSD with entity linking – enabled the two tasks to be performed jointly across many languages.

Knowledge-based approaches remain attractive because they do not require annotated data, rely on structured knowledge that can be curated, and can complement and enhance supervised systems (Bevilacqua and Navigli 2020).

**Statistical and Machine-Learning Supervised Approaches.** In the 1990s, supervised learning reframed

WSD as classification, enabled by sense-tagged corpora such as SemCor. Early systems used Bayesian models (Gale, Church, and Yarowsky 1992), decision lists (Yarowsky 1994), and instance-based classifiers (Ng and Lee 1996), *inter alia*. Senseval evaluations, which started in 1998, established supervised methods as state-of-the-art, though performance plateaued around 70–73% F1 (Navigli 2009), depending on the setting. A few years later, the IMS system (Zhong and Ng 2010) became a standard baseline. Despite limited training data, supervised WSD dominated for over a decade, especially in lexical-sample tasks, where the focus is on a few words occurring many times across sentences, while all-words WSD, where all open-class words need to be tagged, remained more difficult.

**Unsupervised and Semi-supervised Learning.** Parallel work explored unsupervised and semi-supervised methods, often referred to as Word Sense Induction (WSI). Yarowsky (1995) introduced the principles of one sense per collocation and per discourse, bootstrapping classifiers with minimal seed data. Starting from raw text and context clustering, Schütze (1998) showed how to induce senses in a vector space. Subsequent methods included Pantel and Lin’s Clustering by Committee (Pantel and Lin 2002) and graph-based algorithms (Di Marco and Navigli 2013). While unsupervised WSD lagged behind in accuracy and posed evaluation challenges, it remained valuable for low-resource settings. More recently, contextual embeddings have enabled new WSI approaches (Amrami and Goldberg 2018; Kokosinski and Arefyev 2024).

**Neural and Deep Learning Methods.** The mid-2010s brought neural methods that broke long-standing ceilings. Embeddings improved feature representations (Iacobacci, Pilehvar, and Navigli 2016), followed by BiLSTM sequence models (Raganato, Camacho-Collados, and Navigli 2017). Contextualized embeddings, obtained with encoder-based models, transformed WSD: GlossBERT (Huang et al. 2019) and fine-tuned BERT models (Hadiwinoto, Ng, and Gan 2019) surpassed prior approaches. Bi-encoder models (Blevins and Zettlemoyer 2020) and hybrid approaches integrating WordNet (Bevilacqua and Navigli 2020) pushed accuracy beyond 80%. New formulations – e.g., QA-style (Barba, Pasini, and Navigli 2021) and definition generation (Bevilacqua, Maru, and Navigli 2020) – diversified the task.

The rise of LLMs renewed interest in WSD, thanks to the possibility of performing the task in a zero-shot fashion. But while LLMs excel at fluency, some studies found that they may overfit to more common interpretations (Liu et al. 2023) and that they fail to surpass specialized systems (Basile et al. 2025; Meconi et al. 2025). Prompting, fine-tuning, and instruction tuning improve results (Yae et al. 2024; Capone et al. 2024), but integrating explicit sense knowledge remains necessary (Kibria, Dipta, and Adnan 2024).

## Encoder-Based Models and Word Sense Differentiation

Transformer-based encoder models like BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), and DeBERTa (He et al.

2021) produce *contextualized* word embeddings that depend on a word’s usage in a sentence. This addresses the long-standing *meaning conflation deficiency* of static word embeddings (Wiedemann et al. 2019): a single vector can now represent different senses of a word based on context. For example, the word *shot* is associated with distinct embeddings when expressing different senses – and even within the same sense, though expectedly closer in the space – whereas static word embeddings represent it with a single vector. As a result, contextual models may implicitly perform disambiguation by encoding meaning in context (Ethayarajh 2019). A central question, however, is *how well* these embeddings *distinguish fine-grained senses* of a word. Below, we survey findings on this question, covering intrinsic evaluations of embedding spaces and extrinsic evaluations on sense-focused tasks.

### Intrinsic Evaluations: Sense Separation in the Embedding Space

One line of work probes whether the raw embeddings produced by a given model cluster by word sense. Early analyses of Transformer encoders found promising evidence that they do form distinct regions for different senses of the same word. For example, Reif et al. (2019) observed that BERT’s embedding space exhibits a fine-grained geometric representation of word senses: in a qualitative visualization, embeddings of a polysemous word were often found to form well-separated clusters corresponding to its senses. Wiedemann et al. (2019) directly compared encoder models and reported that pre-trained BERT clusters polysemous words into distinct sense regions in the embedding space, whereas previous pre-Transformer approaches, such as ELMo, do not.

More systematic studies examined variance and similarity of embeddings across contexts. Ethayarajh (2019) found that although the same word’s embeddings in different contexts are more similar to each other than to embeddings of unrelated words, this self-similarity drops substantially in upper network layers. In fact, in all layers of BERT, on average less than 5% of the variance in a word’s contextual embeddings can be explained by a static vector for that word. This means that the vast majority of the information in BERT’s word representations is context-specific rather than constant – a strong indication that fine-grained sense information is encoded.

Other intrinsic probes have trained classifiers on frozen embeddings to predict a word’s specific sense label or to decide if two instances share a sense. Successful probing results indicate that information to disambiguate senses is present in the representations. For instance, linear classifiers trained on BERT embeddings can distinguish different senses for many words, implying a linear separability of sense categories in the embedding space (Wiedemann et al. 2019). However, limitations remain: contextual embeddings do not always produce perfectly separated clusters for very fine-grained senses, and embeddings tend to over-represent frequent senses. Recently, Teglia, Tedeschi, and Navigli (2025) showed that encoder models like DeBERTa-v3 differentiate word senses most effectively in their middle

layers (e.g., layers 7–8), significantly outperforming deeper and output layers in sense-separation accuracy, further highlighting both the potential and the nuances of intrinsic sense encodings.

### Task-Based Evaluations: WSD and Word-in-Context

The most direct test of a model’s ability to produce contextualized word embeddings that effectively separate senses is indeed WSD. Loureiro and Jorge (2019) showed that simply using pre-trained BERT embeddings with a nearest-neighbor classifier yielded state-of-the-art WSD performance, surpassing prior supervised systems. They constructed sense embeddings for each WordNet sense and disambiguated unseen instances by choosing the nearest sense embedding to the context-specific embedding. This demonstrated that BERT’s embedding space is rich enough to distinguish senses with minimal additional training.

Subsequent work extended this idea: SensEmBERT (Scarlini, Pasini, and Navigli 2020) combined BERT embeddings with lexical-semantic knowledge to create context-enhanced sense representations, achieving further improvements and making the representations multilingual, while ARES (Scarlini, Pasini, and Navigli 2020) leveraged semi-supervised learning to improve recall for less common senses. In parallel, fine-tuning strategies such as GlossBERT (Huang et al. 2019) treated WSD as a sentence-pair classification task between context and gloss, achieving strong results. Models like KnowBERT (Peters et al. 2019) and EWISER (Bevilacqua and Navigli 2020) injected lexical knowledge during pre-training or ensemble inference to further boost accuracy. ConSeC (Barba, Procopio, and Navigli 2021), instead, reformulates WSD as extractive question answering, using sequence-to-sequence architectures to select the correct sense gloss. Overall, contextual encoder models have dramatically improved WSD, with results rising from the 65–70  $F_1$  range to 83% on all-words WSD benchmarks, in some cases surpassing the historical inter-annotator agreement based on WordNet sense distinctions.

To address the limits of traditional WSD task framing, Pilehvar and Camacho-Collados (2019) introduced the Word-in-Context (WiC) task, which provides a direct evaluation of sense representation. Given a pair of sentences containing the same word, WiC prompts the model to judge whether the word is used with the same sense. BERT fine-tuned on WiC achieved reasonable accuracy (below 70%), above context-insensitive baselines but below human performance (80%), while RoBERTa and SenseBERT (Levine et al. 2020) improved results, with the latter showing the value of injecting lexical knowledge into pre-training.

Knowledge-enriched models trained with sense annotations or glosses achieved stronger performance: SemEq-Large (Yao et al. 2021) attained 75.9%, outperforming RoBERTa and even large models like T5. Other approaches, such as BERT<sub>ARES</sub> and KnowBERT, similarly demonstrated that lexical-semantic information reinforces sense distinctions. Beyond knowledge integration, contrastive learning has proven effective. Lo et al. (2023) introduced

CLWiC, combining contrastive and supervised objectives, reaching 77.5%, a 4-point gain over fine-tuned BERT. Overall, current WiC systems reach the mid-to-high 70s, narrowing the gap to the 80–90% human level and underscoring the value of sense-informed modeling. However, despite being useful for probing sense distinctions in contextualized embeddings, WiC does not solve the limits that affect WSD.

### Takeaways from Encoder Models

Recent results show that contextualized embeddings from encoder-based models capture a certain amount of word sense information. Intrinsic evaluations demonstrate that embeddings tend to cluster by sense, while extrinsic evaluations show that they achieve state-of-the-art results, however such results are still far from being optimal. Improvements in architecture and pre-training correlate with stronger sense distinctions, and integrating lexical knowledge (e.g., SensEmBERT, SenseBERT, KnowBERT) further improves robustness. However, models still struggle with very fine-grained and rare senses, often preferring most frequent ones.

### Large Language Models for WSD

With the advent of decoder-based Transformer models, i.e. LLMs, a natural question has emerged: do these models truly *understand* word senses, or do they simply approximate distributional regularities? Recent studies have therefore begun to evaluate GPT-style models on standard WSD benchmarks, including all-words disambiguation datasets from Senseval and SemEval (Kilgariff and Palmer 2000; Preiss and Yarowsky 2001; Snyder and Palmer 2004; Pradhan et al. 2007; Navigli, Jurgens, and Vannella 2013; Moro and Navigli 2015) and the Word-in-Context (WiC) task (Yae et al. 2024; Basile et al. 2025). Decoder-only LLMs have demonstrated strong zero-shot and few-shot performance on WSD, though early results indicated they were not yet surpassing the best dedicated WSD systems. For instance, Kocoń et al. (2023) reported that ChatGPT (GPT-3.5) achieved roughly 70%  $F_1$  on an English all-words WSD evaluation – a respectable result but still below the state of the art attained by prior supervised encoder-based models (see the previous section). This finding aligned with the view of ChatGPT as a “jack of all trades, master of none”, performing well across many NLP tasks but generally lagging behind specialized models in 2023.

Yae et al. (2024) conducted a comprehensive evaluation of seven generative LLMs (both closed-source and open-weight) on WSD tasks formulated as prompt-based queries. They considered multiple evaluation formats – from selecting the correct sense from a list of candidates to answering true/false if a given sense is correct – using standard WSD corpora (SemCor and Senseval/SemEval datasets) with senses from BabelNet. Interestingly, and today we can say expectedly, a consistent pattern emerged: larger models yielded higher disambiguation accuracy. For example, in a multiple-choice WSD setting, the 70B-parameter Llama-2 model reached around 61% accuracy, whereas OpenAI’s GPT-4 achieved about 70.4%, the top performance in that group. Smaller 7B models obtained lower performance, underscoring the strong correlation between model

scale and WSD competence. Notably, bigger LLMs outperformed smaller open-weight models even without any fine-tuning, indicating that pre-training endows them with substantial knowledge for disambiguation. Similar trends were observed for a binary WiC-style task: GPT-4 attained nearly 77% accuracy on true/false sense discrimination, compared to mid-60% for GPT-3.5 and Llama-70B.

While these results show that GPT-based LLMs are competitive with conventional WSD systems, they also highlight some limitations. One key challenge is handling rare or novel word senses that the model has not encountered during training. Yae et al. (2024) introduced a particularly difficult zero-shot WSD test in which the target word in a sentence was replaced by an *unseen* nonce word (with the original senses as options). All models struggled with this task: even GPT-4’s accuracy dropped to about 56%, and smaller models fell below 40%, barely above random guessing. This gap suggests that LLMs rely partly on prior lexical knowledge and frequency biases; when forced to infer a sense purely from context with no familiar lexical cues, their performance degrades considerably. Inconsistent behavior was also noted across different benchmarks – e.g. an LLM might do better on WiC (which only requires detecting a sense change or not) than on a multi-class all-words WSD test – but overall, the ordering of models by performance remained consistent. These findings reinforce that larger, more domain-informed models have an advantage, yet truly solving WSD in the general case (especially for low-frequency senses) remains challenging for current LLMs.

Recent work has also shown that with appropriate prompting or fine-tuning, GPT-style models can not only match but even surpass prior state-of-the-art WSD systems. Sumanathilaka, Micallef, and Hough (2025) proposed a prompt-engineering approach called GlossGPT, which leverages GPT-4 with few-shot examples, chain-of-thought reasoning, and incorporated sense glosses to guide disambiguation. This method achieved new state-of-the-art WSD accuracy without any task-specific model training, attaining all-words WSD performance on standard benchmarks in the mid-80s, and exceeding the best previous results reported for supervised encoder-based models. Similarly, Basile et al. (2025) found that a medium-sized open-weight LLM, when fine-tuned on a multilingual WSD dataset, reached an average 84.7% WSD accuracy across languages and about 86.5% on English – outperforming all other evaluated models, including the prior state-of-the-art systems. Notably, their zero-shot experiments confirmed that untuned LLMs perform well on WSD but still fall short of dedicated models, whereas a fine-tuned LLM is able to close that gap and even surpass past approaches.

Meconi et al. (2025) provided the most extensive recent evaluation. They test a wide range of instruction-tuned LLMs, both open-weight – from 1B to 70B models, including the Llama, Phi, Gemma and Qwen and Mistral families – and closed-source – including GPT-4o and DeepSeek-V3 – across four benchmarks (the Senseval-SemEval collection refined by Maru et al. (2022) and three sense-skewed datasets: 42D, focused on as many domains, HARDEN, on difficult instances, and a newly curated dataset of 5,500

items). The task is framed as *definition selection*: given a word in context and candidate dictionary definitions (from WordNet), the model must pick the correct definition for that word (i.e. its sense).

In a zero-shot setting, GPT-4o achieved an  $F_1$  score that is statistically comparable to that of ConSeC, the top supervised encoder-based system. Crucially, Meconi et al. (2025) compared GPT-4o with a human annotator, finding that humans reached 91.3  $F_1$ , well above the model, thus demonstrating that state-of-the-art LLMs still fall short of human-level WSD. However, a manual analysis of systematic error patterns – including over-generalization, metonymy, gross misinterpretations, and sensitivity to prompt design – showed that many of GPT-4o’s “wrong” sense choices were not serious. If we could eliminate serious mistakes – that is, errors unlikely to be made by a professional human – then GPT-4o would nearly match human performance on the all-words WSD task.

However, when evaluated on domain-specific or highly challenging items, performance drops sharply. On 42D, LLMs show a decline in  $F_1$  (GPT-4o at 77% and DeepSeek-V3 at 75%). Even worse, on HARDEN, GPT-4o attained only 45.6  $F_1$ , underscoring the difficulty of deeply challenging cases. These results resonate with findings from the Disambiguation Bias in Machine Translation (DiBiMT) benchmark (Campolungo et al. 2022; Martelli et al. 2025), a fully manually curated test set designed to probe the translation of lexically ambiguous words across multiple language pairs (from English to languages as different as Chinese, German, Italian, Russian, Slovene, and others). DiBiMT contains sentences each featuring a target ambiguous word, along with human-verified “good” and “bad” translations of that word. Indeed, DiBiMT evaluations reveal that state-of-the-art Neural Machine Translation and LLM systems often mistranslate ambiguous words used in non-predominant senses, despite high BLEU scores. The bad translation performance of these cases (GPT-4 on average translates suitably 70% of the occurrences, around 60% in some languages) indicates that implicit WSD is insufficient for accurate multilingual applications.

These results mirror our observations with WiC-style tasks: LLMs can perform well on general items, where senses echo Zipfian distributions, but when semantics is pushed – through polysemy, domain focus, or rare senses – the performance gap becomes clear: Martelli et al. (2025) and Meconi et al. (2025) show that models tend to default to more frequent meanings, failing to detect infrequent or subtle context-dependent senses. In both settings – WSD and MT of ambiguous words – semantic biases inherent in pre-training or task framing emerge as a key limitation, confirming that mainstream metrics often mask serious deficits in sense-sensitive performance.

### Takeaways from Decoder Models

In summary, decoder-only LLMs have rapidly become competitive in WSD. Larger models like GPT-4 now approach expert-level disambiguation performance and, with carefully designed prompting or fine-tuning, they can even exceed the accuracy of traditional WSD systems. There are

clear trends of improved WSD capability with increasing model scale, reflecting the benefits of extensive pre-training on semantic content. At the same time, challenges remain: LLMs tend to default to common senses and struggle with low-resource senses. While differences and inconsistencies across benchmarks and languages have been observed, indicating that evaluation setting and language-specific knowledge can influence outcomes, overall the literature shows a convergence – modern GPT-style LLMs are becoming effective WSD solvers, narrowing the long-standing gap between general language understanding models and specialized disambiguation systems. Ongoing research is now examining how to further exploit LLMs’ capabilities for WSD (e.g. via better prompts, integration of knowledge bases, or hybrid ensemble methods) and to address the rare or domain-oriented failure cases and ambiguities in a principled and effective way.

## So Is WSD Dead?

### The Case for Obsolescence

LLMs *internalize* sense distinctions in their contextual representations, making explicit sense labeling unnecessary for many end tasks. In open-ended QA, abstractive summarization, and a large fraction of contemporary MT pipelines, strong decoder-only neural architectures resolve lexical ambiguity implicitly at inference time, often rivaling or surpassing classic supervised WSD baselines without a dedicated WSD head (Kocoń et al. 2023; Yae et al. 2024; Sumanathilaka, Micallef, and Hough 2025; Basile et al. 2025). From a pipeline perspective, therefore, now that LLMs provide fluent and effective outputs in many cases, a modular WSD component no longer appears *instrumental* to competitive performance in mainstream applications; disambiguation emerges as a byproduct of the model’s broader semantic competence, aided by scale, context, instruction-tuning, and effective prompting.

### The Case for Persistence

Yet a growing body of evidence shows that WSD is far from solved, and that explicit, sense-centric evaluation continues to reveal systematic gaps, namely:

- **Non-predominant senses remain under-served.** Even very large models exhibit frequency bias: performance drops on rare or long-tail senses (Blevins, Joshi, and Zettlemoyer 2021; Kocoń et al. 2023; Yae et al. 2024; Martelli et al. 2025; Meconi et al. 2025). Prompt engineering narrows but does not eliminate this effect (Sumanathilaka, Micallef, and Hough 2025).
- **Multilingual transfer exposes hidden failures.** Encoder-based and decoder-only systems that look robust in English often degrade in cross-lingual or low-resource settings; sense inventories and domain coverage interact with pretraining data in ways that surface errors especially outside English (Basile et al. 2025; Martelli et al. 2025).
- **Domain robustness is brittle.** Targeted stress tests such as HARDEN and 42D show marked drops for all models,

including state-of-the-art GPT-style systems. On these cases, supervised WSD systems may fail catastrophically, while LLMs retain modest but still limited competence (Meconi et al. 2025).

Taken together, and considering the enduring importance of grounding text in Knowledge Graphs and KBs (Agrawal et al. 2024), these findings argue against declaring WSD dead or obsolete: WSD remains crucial not only to make understanding interpretable but also as an *evaluation paradigm* to expose blind spots that task-level metrics can mask.

## WSD as a Diagnostic Paradigm

Based on the above discussion, we propose reframing WSD as a well-grounded diagnostic for the lexical-semantic competence of current (and future) NLP technology:

1. **Stress-testing lexical competence.** Carefully controlled all-words evaluations and WiC-style pair judgments (Pilehvar and Camacho-Collados 2019; Raganato, Camacho-Collados, and Navigli 2017) assess the ability of models to handle complex lexical semantics, testing not only their sensitivity to subtle compositional cues but also their capacity to avoid relying on superficial heuristics (e.g., topical similarity). Error categories – e.g., over-generalization, metonymy, and gross misinterpretation – help identify specific failure modes at scale (Meconi et al. 2025). Testing lexical-semantic robustness under diverse conditions would offer an effective way to track equity and generalizability of models across domains and languages, and in biased scenarios.
2. **Assessing sense-aware generation.** Recent studies show that, when freed from rigid inventories, top LLMs can generate high-quality definitions and explanations, yet still exhibit sensitivity to prompt phrasing and residual bias toward prototypical meanings (Meconi et al. 2025; Sumanathilaka, Micallef, and Hough 2025). WSD-style setups (i.e., context-gloss matching, sense-conditioned generation, etc.) provide auditable hooks to evaluate *which* sense a model is committing to in explanations, paraphrases, and translations, offering a valuable tool for LLM interpretability. This requires robust (Zhang et al. 2025b) and adversarial benchmarks (Zhang et al. 2025a).
3. **Linking to cognitive plausibility.** As Tedeschi et al. (2023) caution, claims of “superhuman” Natural Language Understanding warrant grounded comparisons to expert human performance. Sense-disambiguation benchmarks with expert baselines provide precisely such anchors: recent work shows significant headroom for improvement in the performance of top LLMs compared to expert annotators on challenging subsets (Martelli et al. 2025; Meconi et al. 2025), enabling calibrated statements about progress and remaining gaps.
4. **Exploring low-resource languages and multilinguality.** Word sense inventories and annotated corpora vary widely across languages and domains. WSD-based probes can thus reveal how models leverage (or fail to leverage) resource-rich vs. resource-poor settings, and whether sense distinctions transfer across languages. We

expect performance gaps to be especially pronounced in under-resourced languages, echoing findings from recent work on low-resource African languages (Alhanai et al. 2025). Results from the DiBiMT benchmark are particularly revealing: they provide a gold standard for measuring how models handle lexical ambiguity in multilingual MT settings, and confirm that sense-sensitive evaluation remains crucial even when system performance appears high according to standard metrics such as BLEU.

5. **Going beyond sense inventories.** Meconi et al. (2025) showed that when freed from predefined sense inventories, GPT-4o could convey suitable meanings with up to 96% accuracy in unconstrained definition-generation tasks, suggesting that rigid WSD benchmarks may underestimate LLMs’ semantic competence<sup>1</sup>. This points to the need for evaluation protocols that balance the audibility of inventory-based tests with the flexibility of open-ended meaning representation. Future benchmarks could therefore combine WSD with free-form definition, paraphrase, or explanation tasks. These would offer a more holistic view of how models capture, articulate, and generalize lexical meaning beyond fixed ontologies.

In contemporary systems, explicit WSD *modules* may recede, but explicit WSD *tests* should not. We advocate maintaining sense-centric evaluations alongside task metrics, extending them with:

1. rarity- and domain-focused reporting,
2. multilingual coverage beyond English,
3. generation-oriented probes tied to sense commitments.

This repositioning treats WSD not as an end in itself, but as a necessary diagnostic lens on lexical meaning in LLM-era NLP. It is also worth noting that encoder-based architectures should not be considered obsolete: recent evidence suggests that contrastive and sense-aware encoders may still play a role as complementary components in hybrid systems.

## Conclusion

Is WSD dead? Not quite. Despite the remarkable progress of LLMs, WSD as a research problem endures. Implicit disambiguation in LLMs has absorbed much of the burden: LLMs excel at surface-level sense discrimination, often rivaling specialized WSD systems, but they do not model ambiguity in a well-grounded way. This is evident both intrinsically, when WSD is directly evaluated, and extrinsically, e.g. in measuring disambiguation bias in Machine Translation: the apparent semantic fluency of LLM masks persistent weaknesses in deeper lexical understanding, with systematic failures in domain and long-tail senses and gaps across languages.

While earlier the impact of WSD was less evident, we posit it now reemerges as a diagnostic paradigm and conceptual lens for evaluating lexical-semantic competence in current and future NLP models, underscoring the value of meaning-centric evaluation for more equitable, reliable, and interpretable language technology.

<sup>1</sup>This, however, was obtained in an all-words scenario, so the caveats noted above for rare and domain-oriented cases still apply.

## Acknowledgements

The author acknowledges the support of the IT4LIA AI Factory Grant. He also wishes to thank all his past and current co-authors and colleagues for an exciting, never-ending journey in the field of semantics.

## References

- Agirre, E.; and Rigau, G. 1996. Word Sense Disambiguation using Conceptual Density. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, 16–22.
- Agirre, E.; and Soroa, A. 2009. Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, 33–41. Athens, Greece: Association for Computational Linguistics.
- Agrawal, G.; Kumara, T.; Alghamdi, Z.; and Liu, H. 2024. Can Knowledge Graphs Reduce Hallucinations in LLMs? : A Survey. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3947–3960. Mexico City, Mexico: Association for Computational Linguistics.
- Alhanai, T.; Kasumovic, A.; Ghassemi, M.; Zitzelberger, A.; Lundin, J.; and Chabot-Couture, G. 2025. Bridging the Gap: Enhancing LLM Performance for Low-Resource African Languages with New Benchmarks, Fine-Tuning, and Cultural Adjustments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 27802–27812.
- Amrami, A.; and Goldberg, Y. 2018. Word Sense Induction with Neural biLM and Symmetric Patterns. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4860–4867. Brussels, Belgium: Association for Computational Linguistics.
- Banerjee, S.; and Pedersen, T. 2002. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*.
- Barba, E.; Pasini, T.; and Navigli, R. 2021. ESC: Redesigning Word Sense Disambiguation with Extractive Sense Comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4661–4672. Online: Association for Computational Linguistics.
- Barba, E.; Procopio, L.; and Navigli, R. 2021. ConSeC: Word Sense Disambiguation as Continuous Sense Comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1492–1503. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.
- Basile, P.; Musacchio, E.; and Siciliani, L. 2024. Ita-Sense: Evaluating LLMs’ Ability for Italian Word Sense Disambiguation. In *Proceedings of CLiC-it*.
- Basile, P.; Siciliani, L.; Musacchio, E.; and Semeraro, G. 2025. Exploring the Word Sense Disambiguation Capabilities of Large Language Models. *arXiv preprint arXiv:2503.08662*.
- Bevilacqua, M.; Maru, M.; and Navigli, R. 2020. Generational or “How We Went Beyond Word Sense Inventories and Learned to Gloss”. In *Proceedings of EMNLP*, 7207–7221.
- Bevilacqua, M.; and Navigli, R. 2020. Breaking Through the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2854–2864. Online: Association for Computational Linguistics.
- Bevilacqua, M.; Pasini, T.; Raganato, A.; and Navigli, R. 2021. Recent Trends in Word Sense Disambiguation: A Survey. In Zhou, Z., ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, 4330–4338. ijcai.org.
- Blevins, T.; Joshi, M.; and Zettlemoyer, L. 2021. FEWS: Large-Scale, Low-Shot Word Sense Disambiguation with the Dictionary. In Merlo, P.; Tiedemann, J.; and Tsarfaty, R., eds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 455–465. Online: Association for Computational Linguistics.
- Blevins, T.; and Zettlemoyer, L. 2020. Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-Encoders. In *Proceedings of ACL*, 1006–1017.
- Campolungo, N.; Martelli, F.; Saina, F.; and Navigli, R. 2022. DiBiMT: A Novel Benchmark for Measuring Word Sense Disambiguation Biases in Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4331–4352. Dublin, Ireland: Association for Computational Linguistics.
- Capone, L.; Auriemma, S.; Miliani, M.; Bondielli, A.; and Lenci, A. 2024. Lost in Disambiguation: How Instruction-Tuned LLMs Master Lexical Ambiguity. In *Proceedings of \*ACL*.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*, 4171–4186.
- Di Marco, A.; and Navigli, R. 2013. Clustering and Diversifying Web Search Results with Graph-Based Word Sense Induction. *Computational Linguistics*, 39(3): 709–754.
- Ethayarajh, K. 2019. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMO, and GPT-2 Embeddings. In *Proceedings of EMNLP-IJCNLP*, 55–65.
- Gale, W. A.; Church, K. W.; and Yarowsky, D. 1992. Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the 4th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*.

- Guzman-Olivares, D.; Quijano-Sánchez, L.; and Liberatore, F. 2025. SANDWiCH: Semantical Analysis of Neighbours for Disambiguating Words in Context. *arXiv preprint arXiv:2503.05958*. Preprint; introduces cluster-based WSD using semantic neighbor information.
- Hadiwinoto, C.; Ng, H. T.; and Gan, W. C. 2019. Improved Word Sense Disambiguation Using Pre-Trained Contextualized Word Representations. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5297–5306. Hong Kong, China: Association for Computational Linguistics.
- He, P.; Liu, X.; Gao, J.; and Chen, W. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations (ICLR)*.
- Huang, L.; Sun, C.; Qiu, X.; and 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 (EMNLP)*, 3509–3514.
- Iacobacci, I.; Pilehvar, M. T.; and Navigli, R. 2016. Embeddings for Word Sense Disambiguation: An Evaluation Study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 897–907.
- Kamath, G.; Schuster, S.; Vajjala, S.; and Reddy, S. 2024. Scope Ambiguities in Large Language Models. *Transactions of the Association for Computational Linguistics*, 12: 738–754.
- Kibria, R.; Dipta, S. I. U.; and Adnan, M. A. 2024. On Functional Competence of LLMs for Linguistic Disambiguation. In *Proceedings of CoNLL*, 143–160.
- Kilgarriff, A.; and Palmer, M. 2000. Introduction to the Special Issue on SENSEVAL. *Computers and the Humanities*, 34(1/2): 1–13. SENSEVAL: Evaluating Word Sense Disambiguation Programs.
- Kocoń, J.; Cicheński, I.; Kaszyński, O.; Kochanek, M.; Szydło, D.; Baran, J.; Bielaniec, J.; Gruza, M.; Janz, A.; Kanclerz, K.; et al. 2023. ChatGPT: Jack of all trades, master of none. *Information Fusion*, 99: 101861.
- Kokosinskii, D.; and Arefyev, N. 2024. Multilingual Substitution-based Word Sense Induction. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 11859–11872. Torino, Italia: ELRA and ICCL.
- Leacock, C.; Chodorow, M.; and Miller, G. A. 1998. Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics*, 24(1): 147–165.
- 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 (SIGDOC)*, 24–26.
- Levine, Y.; Lenz, B.; Dagan, O.; Ram, O.; Padnos, D.; Sharir, O.; Shalev-Shwartz, S.; Shashua, A.; and Shoham, Y. 2020. SenseBERT: Driving Some Sense into BERT. In *Proceedings of ACL*, 4656–4667.
- Liu, A.; Wu, Z.; Michael, J.; Suhr, A.; West, P.; Koller, A.; Swayamdipta, S.; Smith, N. A.; and Choi, Y. 2023. We’re Afraid Language Models Aren’t Modeling Ambiguity. *Proceedings of EMNLP*, 790–807.
- Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.
- Lo, P.-C.; Lee, Y.-Y.; Chen, H.-H.; Kwee, A. T.; and Lim, E.-p. 2023. Contrastive Learning Approach to Word-in-Context Task for Low-Resource Languages. In *Proceedings of the 37th AAAI Workshop on Knowledge Augmented Methods for Natural Language Processing*, 1–8. Washington, DC.
- Loureiro, D.; and Jorge, A. 2019. Language Modelling Makes Sense: Propagating Representations through WordNet for Full-Coverage Word Sense Disambiguation. In *Proceedings of ACL*, 5682–5691.
- Martelli, F.; Perrella, S.; Campolungo, N.; Munda, T.; Kovaleva, S.; Tiberius, C.; and Navigli, R. 2025. DiBiMT: A Gold Evaluation Benchmark for Studying Lexical Ambiguity in Machine Translation. *Computational Linguistics*, 51(2): 343–413.
- Maru, M.; Conia, S.; Bevilacqua, M.; and Navigli, R. 2022. Nibbling at the Hard Core of Word Sense Disambiguation. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4724–4737. Dublin, Ireland: Association for Computational Linguistics.
- Meconi, D.; Stirpe, S.; Martelli, F.; Lavalle, L.; and Navigli, R. 2025. Do Large Language Models Understand Word Meanings? *Proceedings of EMNLP*, 33885–33904.
- Mihalcea, R. 2005. Unsupervised Large-Vocabulary Word Sense Disambiguation with Graph-based Algorithms for Sequence Data Labeling. In *Proceedings of the Joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, 411–418. Vancouver, British Columbia, Canada: Association for Computational Linguistics.
- Miller, G. A. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11): 39–41.
- Miller, G. A.; Leacock, C.; Teng, R.; and Bunker, R. T. 1993. A Semantic Concordance. In *Proceedings of the Workshop on Human Language Technology, HLT ’93*, 303–308. Stroudsburg, PA, USA: Association for Computational Linguistics.
- Moro, A.; and Navigli, R. 2015. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. In Nakov, P.; Zesch, T.; Cer, D.; and Jurgens, D., eds., *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 288–297. Denver, Colorado: Association for Computational Linguistics.

- Moro, A.; Raganato, A.; and Navigli, R. 2014. Entity Linking meets Word Sense Disambiguation: A Unified Approach. *Transactions of the Association for Computational Linguistics*, 2: 231–244.
- Navigli, R. 2009. Word Sense Disambiguation: A Survey. *ACM Computing Surveys*, 41(2): 1–69.
- Navigli, R. 2018. Natural Language Understanding: Instructions for (Present and Future) Use. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, 5697–5702.
- Navigli, R.; Jurgens, D.; and Vannella, D. 2013. SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. In Manandhar, S.; and Yuret, D., eds., *Second Joint Conference on Lexical and Computational Semantics (\*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 222–231. Atlanta, Georgia, USA: Association for Computational Linguistics.
- Navigli, R.; and Martelli, F. 2019. An overview of word and sense similarity. *Nat. Lang. Eng.*, 25(6): 693–714.
- Navigli, R.; and Ponzetto, S. P. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193: 217–250.
- Ng, H. T.; and Lee, H. B. 1996. Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL)*, 40–47.
- Pantel, P.; and Lin, D. 2002. Discovering Word Senses from Text. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 613–619.
- Pasini, T.; Raganato, A.; and Navigli, R. 2021. XL-WSD: An Extra-Large and Cross-Lingual Evaluation Framework for Word Sense Disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 13648–13656. AAAI Press.
- Peters, M. E.; Neumann, M.; Logan, R.; Schwartz, R.; Joshi, V.; Singh, S.; and Smith, N. A. 2019. Knowledge Enhanced Contextual Word Representations. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 43–54. Hong Kong, China: Association for Computational Linguistics.
- Pilehvar, M. T.; and Camacho-Collados, J. 2019. WiC: The Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations. In *Proceedings of NAACL*, 1267–1273.
- Pradhan, S.; Loper, E.; Dligach, D.; and Palmer, M. 2007. SemEval-2007 Task-17: English Lexical Sample, SRL and All Words. In Agirre, E.; Màrquez, L.; and Wicentowski, R., eds., *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, 87–92. Prague, Czech Republic: Association for Computational Linguistics.
- Preiss, J.; and Yarowsky, D., eds. 2001. *Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*. Toulouse, France: Association for Computational Linguistics.
- Raganato, A.; Camacho-Collados, J.; and Navigli, R. 2017. Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. In *Proceedings of EACL*, 99–110.
- Reif, E.; Yuan, A.; Wattenberg, M.; Viegas, F. B.; Coenen, A.; Pearce, A.; and Kim, B. 2019. Visualizing and Measuring the Geometry of BERT. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Resnik, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*.
- Scarlino, B.; Pasini, T.; and Navigli, R. 2020. SensEmBERT: Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation. In *Proceedings of AAAI*, volume 34, 8758–8765.
- Schütze, H. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1): 97–123.
- Scozzafava, F.; Maru, M.; Brignone, F.; Torrisi, G.; and Navigli, R. 2020. Personalized PageRank with Syntagmatic Information for Multilingual Word Sense Disambiguation. In Celikyilmaz, A.; and Wen, T.-H., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 37–46. Online: Association for Computational Linguistics.
- Snyder, B.; and Palmer, M. 2004. The English all-words task. In *Proceedings of SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, 41–43. Barcelona, Spain: Association for Computational Linguistics.
- Sumanathilaka, T. G. D. K.; Micallef, N.; and Hough, J. 2025. GlossGPT: Prompt Engineering for Word Sense Disambiguation with GPT-4. In *Proceedings of the 2025 Conference on Computational Natural Language Learning*.
- Tedeschi, S.; Bos, J.; Declerck, T.; Hajič, J.; Hershcovich, D.; Hovy, E.; Koller, A.; Krek, S.; Schockaert, S.; Sennrich, R.; Shutova, E.; and Navigli, R. 2023. What's the Meaning of Superhuman Performance in Today's NLU? *Proceedings of ACL*, 12471–12491.
- Teglia, S.; Tedeschi, S.; and Navigli, R. 2025. How Much Do Encoder Models Know About Word Senses? In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2266–2277. Vienna, Austria: Association for Computational Linguistics.
- Weaver, W. 1949. Translation. Memorandum, Rockefeller Foundation.
- Wiedemann, G.; Remus, S.; Chawla, A.; and Biemann, C. 2019. Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. *CoRR*, abs/1909.10430.

- Yae, J. H.; Skelly, N. C.; Ranly, N. C.; and LaCasse, P. M. 2024. Leveraging Large Language Models for Word Sense Disambiguation. In *Proceedings of CoNLL*.
- Yao, W.; Pan, X.; Jin, L.; Chen, J.; Yu, D.; and Yu, D. 2021. Connect-the-Dots: Bridging Semantics between Words and Definitions via Aligning Word Sense Inventories. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7741–7751. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.
- Yarowsky, D. 1994. Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 88–95.
- Yarowsky, D. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 189–196.
- Zhang, K.; Liu, Q.; Zhang, L.; Zheng, C.; Li, S.; Xu, B.; Yang, M.; Qiao, X.; and Lu, W. 2025a. MADAWS: Multi-Agent Debate Framework for Adversarial Word Sense Disambiguation. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 22294–22313. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.
- Zhang, L.; Li, S.; Li, Y.; Kang, K.; Zhang, K.; Wang, C.; and Lu, W. 2025b. RoDEval: A Robust Word Sense Disambiguation Evaluation Framework for Large Language Models. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 17095–17126. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.
- Zhong, Z.; and Ng, H. T. 2010. It Makes Sense: A Wide-Coverage Word Sense Disambiguation System for Free Text. In *Proceedings of the ACL 2010 System Demonstrations*, 78–83.