Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 257 (2025) 785–792

Procedia Computer Science

www.elsevier.com/locate/procedia

The 8th International Conference on Emerging Data and Industry (EDI40)
April 22-24, 2025, Patras, Greece

GlossGPT: GPT for Word Sense Disambiguation using Few-shot Chain-of-Thought Prompting

Deshan Sumanathilaka[a,*], Nicholas Micallef[a], Julian Hough[a]

[a] *Swansea University, Wales, United Kingdom*

Abstract

Lexical ambiguity is a major challenge in computational linguistic tasks, as limitations in proper sense identification lead to inefficient translation and question answering. General-purpose Large Language Models (LLMs) are commonly utilized for Natural Language Processing (NLP) tasks. However, utilizing general-purpose LLMs for specific tasks has been challenging, and fine-tuning has become a critical requirement for task specification. In this work, we craft advanced prompts with different contextual parameters to guide the model's inference towards accurate sense prediction to handle Word Sense Disambiguation (WSD). We present a few-shot Chain of Thought (COT) prompt-based technique using GPT-4-Turbo with knowledgebase as a retriever that does not require fine-tuning the model for WSD tasks and sense definitions are supported by synonyms to broaden the lexical meaning. Our approach achieves comparable performance on the SemEval and Senseval datasets. More importantly, we set a new state-of-the-art performance with the few-shot FEWS dataset, breaking through the 90% F1 score barrier.

Keywords: Word Sense Disambiguation, Knowledge Base Retrieval, Large Language Models, GPT-4-Turbo, Chain of Thought Prompting

1. Introduction

In Natural Language Understanding (NLU), accurately determining the meanings of words within sentences is crucial. Misinterpretations of word senses can lead to incorrect information and result in misinformation. Dealing with words that have multiple meanings (polysemy) poses a significant challenge for Natural Language Processing (NLP), particularly in tackling lexical ambiguity. Despite extensive research on Word Sense Disambiguation (WSD) across various languages, current methods still exhibit some weaknesses. Previous WSD studies have struggled with complex cases, often due to models' inadequate semantic understanding [20]. Ambiguous words, for instance, "bat", "bank", "run", "point", "light", and "set" have numerous senses, some of which are commonly used while others are less frequent. This variety poses a significant challenge in identifying the less common meanings, especially with

[*] Corresponding author. Tel.: +44 7706810130.
E-mail address: t.g.d.sumanathilaka@swansea.ac.uk

10.1016/j.procs.2025.03.101

Table 1. Glosses of *'bat'* and its sense tags

Sentence: A <WSD>bat</WSD> has a narrow handle and a broad flat end for hitting.		**Ambiguous word**	bat
Any of the small, nocturnal, flying mammals of the order Chiroptera, which navigate by echolocation.	bat.noun.0.	A club made of wood or aluminium, used for striking the ball in sports such as baseball, softball and cricket.	bat.noun.2
An old woman.	bat.noun.1	Packsaddle.	bat.noun.13
Possible senses	N: 13 V: 6	**Correct Sense**	bat.noun.2

limited resource datasets. Table 1 demonstrates an example of the word "bat", which contains 13 nouns and 6 verb interpretations. For demonstrative purposes, we have shared a few instances of glosses. However, research indicates that a word's meaning is heavily influenced by its surrounding context, demonstrating that analyzing words in isolation is inadequate for precise sense identification [14]. Consequently, incorporating positional value, Part-of-speech (POS) tags, and the broader sentence context is crucial for increasing the accuracy of the models. The advent of LLMs and Transformer-based generative AI have shown promising advancements in contextual understanding. These models exhibit a remarkable capability to manage complex language tasks. Fine-tuning models for downstream tasks such as question answering and domain-specific knowledge generation has yielded promising results [18]. In our study, we present a well-defined approach using GPT for the task of WSD by evaluating its capability to identify the correct meaning of words. We conducted research with different LLMs to determine the most suitable one for this study. The GPT-4 Turbo model provided clear evidence of its capabilities for WSD tasks, surpassing all other open-source and commercial LLMs [25]. Specifically, we aimed to explore how GPT models can be used to match words (with multiple meanings) to their correct sense. To guide the inference correctly, we carefully crafted the prompts using a few-shot Chain of Thought (COT) prompt-based approach with the assistance of synonyms of ambiguous words and knowledge base as a retriever. Synonyms have been considered an effective factor for WSD in previous studies [13].

Building on these foundations, previous research in WSD has attempted to determine the correct gloss or sense tag by approaching the problem from various perspectives [17]. However, these studies face a significant challenge in identifying the sense of complex instances with varied meaning distributions. A study by Pasini [21] highlights that current architectures lack the confidence to accurately predict the sense of highly ambiguous words with multiple interpretations. To address this issue, we utilize the pre-trained knowledge of language models, which are trained on extensive text corpora and thus hold the potential for mitigating data scarcity problems in supervised learning. We chose prompt augmentation approaches to leverage the extensive knowledge embedded in LLMs. This process follows a human-in-the-loop approach to identify the optimal prompting technique. The identified advanced prompt was used to evaluate the LLM's capabilities for WSD, and the outcomes are presented in the results section. Notably, we did not use fine-tuning for the WSD task; instead, we reframed the prompts to guide the inference process toward WSD. The study's contributions include

- Proposing and evaluating a novel approach for Word Sense Disambiguation using few-shot chain of thought prompting, supported with GPT and synonyms to enhance lexical knowledge.
- Introducing iterative models to tackle corner cases with high ambiguity.

In summary, this introduction has given a thorough look at the field of research and the main points we will be exploring. Moreover, we achieved state-of-the-art performance on English-WSD without fine-tuning the GPT models for the downstream task. Henceforth, The subsequent sections will explore similar studies, explain our chosen research methods, present our findings, and discuss the limitations of the proposed methods. We will also suggest areas for potential future research.

2. Related Works

Ambiguity in natural language presents a significant challenge for various natural language processing tasks, making WSD a fundamental issue. WSD remains a persistent research area across different languages, as accurate word sense identification directly impacts numerous NLP applications. Researchers have proposed several advanced neural architectures for WSD by integrating knowledge-based models [1]. Various NLP techniques have been combined to perform effective WSD, and an overview of these works is provided below.

2.1. *Supervised WSD*

Supervised WSD is a well-researched area in literature. Properly labelled datasets, such as Semcor and FEWS, along with WordNet, are commonly used for WSD studies [23]. Various computational techniques have been evaluated in the context of supervised WSD, with researchers attempting to reframe the WSD problem into different computational challenges. One notable study, ConSec, introduced a novel approach to WSD by reframing the task as a text extraction problem [4]. It incorporated a feedback loop strategy to target the ambiguous word, considering the context and the explicit sense assigned to nearby words. EWISER explored the possibility of leveraging Lexical Knowledge Bases (LKB) in their study, where synset embeddings and relations were used to train neural architecture [6]. SpareLLM utilized sparse contextualized word representations [5], while the Bi-Encoder model integrated target words with their surrounding context and glosses [8]. Various BERT variations, including fine-tuning pre-trained BERT models, were also explored in the context of WSD [11]. It is evident from previous studies that using synonyms and contextual meaning of the surrounding words (paradigmatic relations) heavily impacts the decision-making process of sense id. Therefore, our study incorporated synonyms to enrich word sense interpretations during inferencing.

2.2. *Knowledge-Base (KB) WSD*

Knowledge-based methods for WSD employ external resources using semantic similarity metrics and graph-based algorithms. Wang et al. [29] have utilized semantic space and semantic paths hidden within sentences to improve WSD using WordNet. Techniques including Latent Semantic Analysis (LSA) and PageRank are employed in these methods. A study by Chaplot and Salakhutdinov [9] used topic modelling to scale words in context linearly, proposing a variant of Latent Dirichlet Allocation (LDA) where synset proportions are used instead of topic proportions. The Babelfy study, which employs Entity Linking (EL) connected to named entities, introduces a unified graph-based method for EL and WSD. This approach relies on a broad identification of potential meanings and uses the densest subgraph heuristic to choose the most coherent semantic interpretations, showing promising results in multilingual setups [19]. Early WSD solutions proposed random walks over large knowledge bases such as Extended WordNet [2]. Jha et al. [12] used Hindi WordNet with weighted graphs representing word senses and their relations, while Duarte et al. [10] combined graph-based approaches with word embeddings and contextual information for semi-supervised WSD. The study by Martinez-Gil [16] incorporated similarity measurements, proving the importance of contextual information. The Synset Relation-Enhanced Framework (SREF) for enriched sense embeddings expanded the WSD toolkit by augmenting basic sense embeddings with sense relations and incorporating a try-again mechanism [28]. Furthermore, some studies have proposed suggestion-level modules, demonstrating the importance of knowledgebases [27]. These studies have illustrated the promising usage of KB for the disambiguation process of ambiguous words. This motivates us to incorporate a KB for an effective few-shot retrieval process.

3. Methodology

Previous studies are a strong indication that the usage of LLM-based approaches for WSD tasks can be effective. Prior work has explored the capacity of different LLMs for WSD incorporating different computational approaches, namely parameter tuning and prompt augmentation [25]. Inspired by the insights gained from these previous studies, the optimal prompt for this study has been constructed using an iterative human-in-loop approach. The synonyms of ambiguous words are introduced to the pipeline as a novel addition to drive accurate sense selection.

3.1. *Dataset Selection*

3.1.1. *FEWS*

This work focused on the recently introduced FEWS dataset, designed explicitly for few-shot WSD and its evaluation [7]. The dataset included sense tag lists, training data, and test data. It uniquely covered definitions and example

Table 2. Sense tag definition used and expected sense tag for input sentence

***Senseid:**	dictionary.noun.0	**Tags**	en	***Word**	dictionary	**Depth**	1
***Gloss**	A reference work with a list of words from one or more languages, normally ordered alphabetically, explaining each word's meaning, and sometimes containing information on its etymology, pronunciation, usage, translations, and other data.						
***Synonyms**	wordbook						
Sentence	The first author meticulously cross-checked the manuscript against various </WSD>dictionaries </WSD>, striving to ensure both word choice and proper usage.				**Output**	dictionary. noun.0	

sentences provided in Wiktionary, addressing words with uncommon ambiguities. This study evaluated the models' ability to correctly assign defined sense tags to polysemous terms positioned between <WSD>tokens within sentences. The sense tag distribution from the FEWS dataset is shown in Table 2.

3.1.2. Semcor, Semeval and Senseval

We selected SemCor 3.0 for KB creation because it is the largest manually annotated corpus with WordNet senses for WSD. Our method was further evaluated using several English all-words WSD datasets. To ensure a fair comparison, we utilized the benchmark datasets proposed by [22], which encompass five standard all-words fine-grained WSD datasets from the Senseval and SemEval competitions: Senseval-2 (SE2), Senseval-3 (SE3), SemEval-2007 (SE07), SemEval-2013 (SE13), and SemEval-2015 (SE15). Both FEWS and the unified evaluation framework (UEF) evaluation set has been formatted to maintain the input sentence and the expected output, as shown in Table 2. The training data from FEWS and SemCor have been organized into two knowledge bases based on the POS tag and the words to supply the necessary few-shot examples for the pipeline. The word serves as the root nodes, while the POS tags were assigned as the first-level parent nodes. All related instances, sense tags and synonyms from the dataset were stored in the leaf nodes accordingly. Sentences and their senses were arranged in a unique list for each word-POS tag pair. This structure facilitated the extraction of relevant examples from the KB in constant time, regardless of the size of the training set. The extracted data from the KBs were utilized to facilitate the few-shots required for the GPT model inferencing. Few-shots were required for FEWS Dev, and test set evaluations were computed through FEWS training data, while Senseval and SemEval were based on SemCor.

3.2. Few-shot COT based Prompting

This phase aimed to identify the optimal prompt for extracting the correct sense ID from the sense tags associated with ambiguous words within a given sentence. According to previous studies, GPT-3.5-Turbo and GPT-4-Turbo outperformed all other LLMs for WSD tasks without fine-tuning [24]. This study employed these GPT versions to evaluate the outcomes of FEWS and unified evaluation test cases. A human-in-the-loop approach was employed, where the lead researcher used prompt engineering techniques to develop the most suitable prompt for extracting the sense ID. An iterative approach was adopted, with careful refinement of the prompt based on the results of each iteration. Incorrect predictions were systematically analyzed to improve the prompt and generate optimal results. This phase explored various prompting techniques, including Zero-Shot prompting, few-shot prompting, and Zero and few-shot COT prompting [31], to identify the most effective approach. However, the previous results revealed that some challenging, ambiguous words could not be identified without a proper understanding of each sense tag [25]. To address this limitation, a KB approach using few-shot COT prompting was proposed. Synonyms were utilized to give a better lexical understanding of ambiguous words during the prompting process. The model was prompted on instances of each sense tag along with their corresponding glosses and instructed to learn about the meaning of each tag based on the provided instances and the word's usage within a sentence. The models used in our study were not trained for the WSD task; instead, we designed the prompts to operate the general task language model for the WSD task. The absence of fine-tuning or training of base models in our approach is deliberate. We have demonstrated comparable results by utilizing carefully crafted prompts with few-shot examples.

In the prompt, definitions extracted from the lexical knowledge and sense IDs are shared. Let's consider the word "bank" from WordNet 3.0. The below format is used to share the appropriate sense after the POS tag-based filtration process. Two instances are used to illustrate the execution process. However, in the exact execution, all ten senses are passed for the inference.

- bank.n.06:01: a container (usually with a slot in the top) for keeping money at home Synonyms: savings_bank, coin_bank, money_box.
- bank.n.14:00: a financial institution that accepts deposits and channels the money into lending activities Synonyms: depository_financial_institution, banking_concern, banking_company

The example section in the prompt offers relevant instances from KB to enhance comprehension of each ambiguous word. We have ensured that the utmost three instances are shared for each sense. These examples enrich the inference process, providing additional context for the few-shot prompting process.

3.3. Predict and Verify approach for corner cases

This phase of the study primarily focused on improving the performance of sense prediction, particularly in corner cases. Incorrect predictions from phase 2 of the FEWS test set were further analyzed here. Due to the sense distribution nature of the FEWS test set, corner cases of FEWS from the initial study were selected for the subsequent studies. The previous prompt usage has been redesigned into an iterative predict-and-verify methodology, where each subsection is responsible for sense prediction and verification of the identified sense. This approach re-framed the WSD to self-refinement, and ReAct approaches proposed in previous studies [15, 30]. During the iterative process, the predicted sense from the lexical space was validated by a "Verify prompt". If the predicted sense was identified as incorrect during the verification, the lexical space was updated by removing the incorrect sense from the initial run. The model was then re-run with the updated lexical sense space. This process continued until the verification process confirmed the correct sense. If a correct prediction is not identified, it returns "None" at the end. In the best-case scenario, the model produces the answer in constant time, while in the worst-case scenario, it operates in linear time.

3.4. Iterative Binary Classification with few-shot COT prompting for corner cases

The predict-and-verify approach has shown promising results. However, there is still room for improvement in handling corner cases. To address this limitation, we have proposed an iterative binary classification-driven approach inspired by the binary classification problem commonly used for classifying two classes. This method aimed to enhance attention on two senses simultaneously. In this sense prediction approach, two senses were considered at a time, and the most confident answer from the selection was utilized in the subsequent iteration. Each iteration included the gloss of two sense IDs, a few-shot examples of each sense, and any available synonyms. This approach has been implemented and evaluated to give utmost attention to two senses simultaneously. It is important to note that, as this is an iterative process, the time complexity and token usage were comparatively higher than in phases 1 and 2. However, the answers were produced linearly in both best-case and worst-case scenarios. Codes and the prompts can be accessed here[1].

3.5. Experimental Setup

In this study, our objective was to assess the effectiveness of prompt engineering along with different computational techniques. Due to their established capabilities in numerous semantics-related tasks, we focused on GPT-3.5 Turbo and GPT-4 models. We obtained an OpenAI API key from a tier-one OpenAI account, maintaining a temperature of 0 and a maximum token limit of 500 for each output. Additionally, we performed a separate study on temperature settings, evaluating results across a range of values (0–1) with 0.2 intervals for the WSD task. We identified that temperature 0 produced deterministic results, while a higher temperature made incorrect sense selections [26]. However,

[1] https://github.com/Sumanathilaka/GlossGPT-GPT-4-WSD-with-COT

Table 3. F1 score of the different models on FEWS, SemEval and Senseval. Our approach is in italics.

Models	**Dev**	**Unified Eval Framework**				**POS Tag based UEF**					**FEWS**	
	SE07	**SE2**	**SE3**	**SE13**	**SE15**	**N**	**V**	**A**	**R**	**ALL**	**Dev**	**Test**
MFS	54.5	65.6	66.0	63.8	67.1	67.7	49.8	73.1	80.5	65.5	52.8	51.5
Lesk	51.6	63.0	63.7	66.2	64.6	69.8	51.2	51.7	80.6	63.7	42.5	40.9
EWISE	67.3	73.8	71.1	69.4	74.5	74.0	60.2	78.0	82.1	71.8	-	-
GlossBERT	72.5	77.7	75.2	76.1	80.4	79.8	67.1	79.6	87.4	77.0	-	-
BEM	74.5	79.4	77.7	79.7	81.7	81.4	68.5	83.0	87.9	79.0	79.3	79.0
ARES	71.0	78.0	77.1	77.3	83.2	80.6	68.3	80.5	83.5	77.9	-	-
SemEq $_{\text{Base Expert}}$	74.1	81.0	78.5	79.9	82.6	82.5	69.9	82.5	88.4	79.9	80.4	80.1
SemEq $_{\text{Large Expert}}$	74.9	81.8	79.6	81.2	81.8	83.2	71.1	83.2	87.9	80.7	81.8	82.3
ESR $_{\text{base}}$	77.4	81.4	78.0	81.5	83.9	83.1	71.1	83.6	87.5	80.7	77.9	77.8
ESR $_{\text{Large}}$	**78.5**	82.5	80.2	82.3	**85.3**	84.4	73.0	74.4	88.0	82.0	83.8	83.4
RTWE $_{\text{Base}}$	74.5	82.3	80.9	81.8	83.7	83.3	72.2	87.4	87.6	81.6	78.0	78.4
CoNSEC	77.4	82.3	79.9	**83.2**	85.2	**85.4**	70.8	84.0	87.3	**82.0**	-	-
GlossGPT	*76.2*	***86.1***	***82.9***	*75.4*	*83.0*	*82.6*	***73.1***	***91.9***	***88.6***	*81.8*	***90.2***	***90.7***

to ensure reproducibility, we set the temperature to 0 for this study. The temperature determines the randomness of the model's output, with higher values leading to more diverse and creative responses, while lower values result in more deterministic and focused outputs. The primary task assigned to all the LLMs was word sense identification, defining their role as "helpful assistant for identifying word senses". Subsequently, we analyzed the number of correct predictions, execution time, and token distribution to assess the model's performance.

4. Results and Discussion

The results in Table 3 have been evaluated using both the few-shot dev set and a test set of the FEWS dataset and UEF. It provides the proper benchmark for the WSD compared to the current existing approach. Notably presented results are F1 scores. The result presents a comprehensive analysis of disambiguation techniques used by different research works for WSD. Our approach (GlossGPT) has noted a promising result of 90.2 F1 for the dev set and 90.7 F1 for the few-shot test set, creating a new state-of-art performance in the few-shot word sense disambiguation in the FEWS dataset. Further, we have evaluated the zero-shot setting with COT prompting approach with zero-shot dev and test set where 0.81 and 0.79 F1 were achieved, showing the effectiveness of the proposed few-shot COT prompting approach in GlossGPT. GlossGPT has outperformed all the other models in SE2 and SE3 while showing significant results in verb, adjective and adverb-based sense identification. The combination of synonyms and the few-shot COT prompting has enabled the new arena in the WSD without fine-tuning the models for the downstream task. Synonyms for lexical knowledge and COT prompts have enriched the sense interpretations, resulting in high performance. This approach utilizes an average of 695 tokens per request. Though FEWS and UEF evaluations show promising results, UEF dataset underperforms due to multiple ambiguous words in a single sentence, where one incorrect disambiguation can lead to the incorrect sense identification of other words. This is not the case in FEWS, as each record contains only one noted ambiguous word. Although our approach tested in phase 1 shows promising results, handling less frequent ambiguous words is a limitation. Therefore, we have re-designed the prompts to tackle the edge cases of FEWS in phase 1 using an iterative manner. The predict and verify approach and iterative binary classification approaches present promising results with corner cases. The predict and verify approach handled 62 edge cases using the GPT-4-Turbo as the base model. Overall, a 13% overall improvement is demonstrated in the above work. The iterative binary classification approach shows a promising result in handling corner cases with a 28.81% improvement on corner cases. Although the proposed architectures are capable of handling edge cases, these approaches are not well suited for synchronizing with real-time applications since the process takes a high amount of time and token count (Average Token Count per one sense identification: Predict & verify-1306, Iterative Binary Classification-1126). However, observed improvements highlight a promising direction for future research to refine WSD methodologies. The prompts and the knowledge base can be accessed here: https://github.com/Sumanathilaka/GlossGPT-GPT-4-WSD-with-COT.

4.1. Ablation Study

To further demonstrate the generalizability and robustness of the proposed model, we conducted a validation study using set 3 of the FOOL Me if You Can dataset, which contains ambiguous contexts with adjectives that do not align with the intended meaning of the homonyms in those sentences [3]. The selection of set 3 was motivated by the comparatively low performance observed with the zero-shot prompting approach proposed in the original study. For this validation, we mapped the dataset to FEWS sense tags to fit within the proposed framework. The study setup followed the same hyperparameters discussed in the experiment setup section to ensure consistency in evaluation. The knowledge base was built using the training data from the FOOL dataset, where the retriever employed a semantic similarity-based score to select the most relevant few-shot examples needed for the inference process. Each sense was supported by up to three few-shot instances. The results of the study are promising, with the GPT-4o model achieving an F1 score of 0.89, setting a new state-of-the-art performance on this dataset. This further confirms that the proposed architecture is well-suited for WSD. Codes and the prompts can be accessed here[2].

5. Conclusion and Future Directions

This research illustrates the efficacy of integrating synonyms with few-shot COT prompting using the knowledge-base as an extractor along with the GPT-4-turbo model. GlossGPT is equipped with well-designed prompts enhanced by prompt augmentation. This methodology has introduced a new state-of-art performance for WSD without incorporating training or fine-tuning. The corner cases are handled by a novel approach which utilizes an iterative binary classification-based mechanism, achieving a significantly improved result. Future endeavours should prioritize assessing these techniques through additional parameters that can improve the performance of nouns. Subsequent investigations must be performed on various languages to validate the approach's generalizability and explore its applicability in real-world scenarios.

References

[1] Abeysiriwardana, M., Sumanathilaka, D., 2024. A survey on lexical ambiguity detection and word sense disambiguation. arXiv preprint arXiv:2403.16129 .

[2] Agirre, E., López De Lacalle, O., Soroa, A., 2014. Random Walks for Knowledge-Based Word Sense Disambiguation. Computational Linguistics 40, 57–84. URL: https://direct.mit.edu/coli/article/40/1/57-84/1454, doi:10.1162/COLI_a_00164.

[3] Ballout, M., Dedert, A., Abdelmoneim, N.M., Krumnack, U., Heidemann, G., Kühnberger, K.U., 2024. FOOL ME IF YOU CAN! an adversarial dataset to investigate the robustness of LMs in word sense disambiguation, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA. pp. 5042–5059. URL: https://aclanthology.org/2024.emnlp-main.290/, doi:10.18653/v1/2024.emnlp-main.290.

[4] Barba, E., Procopio, L., Navigli, R., 2021. ConSeC: Word Sense Disambiguation as Continuous Sense Comprehension, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic. pp. 1492–1503. URL: https://aclanthology.org/2021.emnlp-main.112, doi:10.18653/v1/2021.emnlp-main.112.

[5] Berend, G., 2020. Sparsity Makes Sense: Word Sense Disambiguation Using Sparse Contextualized Word Representations, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online. pp. 8498–8508. URL: https://www.aclweb.org/anthology/2020.emnlp-main.683, doi:10.18653/v1/2020.emnlp-main.683.

[6] Bevilacqua, M., 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, Association for Computational Linguistics, Online. pp. 2854–2864. URL: https://www.aclweb.org/anthology/2020.acl-main.255, doi:10.18653/v1/2020.acl-main.255.

[7] Blevins, T., Joshi, M., Zettlemoyer, L., 2021. Fews: Large-scale, low-shot word sense disambiguation with the dictionary, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. URL: https://blvns.github.io/papers/eacl2021.pdf.

[8] Blevins, T., Zettlemoyer, L., 2020. Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online. pp. 1006–1017. URL: https://www.aclweb.org/anthology/2020.acl-main.95, doi:10.18653/v1/2020.acl-main.95.

[2] https://github.com/Sumanathilaka/FOOL-ME-IF-YOU-CAN-dataset-Meets-FEWS-sense-Tags

[9] Chaplot, D.S., Salakhutdinov, R., 2018. Knowledge-based Word Sense Disambiguation using Topic Models. URL: http://arxiv.org/abs/1801.01900. arXiv:1801.01900 [cs].

[10] Duarte, J.M., Sousa, S., Milios, E., Berton, L., 2021. Deep analysis of word sense disambiguation via semi-supervised learning and neural word representations. Information Sciences 570, 278–297. URL: https://linkinghub.elsevier.com/retrieve/pii/S0020025521003273, doi:10.1016/j.ins.2021.04.006.

[11] Huang, L., Sun, C., Qiu, X., Huang, X., 2020. GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge. URL: http://arxiv.org/abs/1908.07245. arXiv:1908.07245 [cs].

[12] Jha, P., Agarwal, S., Abbas, A., Siddiqui, T.J., 2023. A Novel Unsupervısed Graph-Based Algorıthm for Hindi Word Sense Disambiguation. SN Computer Science 4, 675. URL: https://link.springer.com/10.1007/s42979-023-02116-1, doi:10.1007/s42979-023-02116-1.

[13] Li, Y., Chen, J., Li, Y., Yu, T., Chen, X., Zheng, H.T., 2023. Embracing ambiguity: Improving similarity-oriented tasks with contextual synonym knowledge. Neurocomputing 555, 126583. URL: https://linkinghub.elsevier.com/retrieve/pii/S0925231223007063, doi:10.1016/j.neucom.2023.126583.

[14] Luo, F., Liu, T., Xia, Q., Chang, B., Sui, Z., 2018. Incorporating Glosses into Neural Word Sense Disambiguation. URL: http://arxiv.org/abs/1805.08028. arXiv:1805.08028 [cs].

[15] Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B.P., Hermann, K., Welleck, S., Yazdanbakhsh, A., Clark, P., 2023. Self-Refine: Iterative Refinement with Self-Feedback. URL: http://arxiv.org/abs/2303.17651. arXiv:2303.17651 [cs].

[16] Martinez-Gil, J., 2023. Context-Aware Semantic Similarity Measurement for Unsupervised Word Sense Disambiguation. URL: http://arxiv.org/abs/2305.03520. arXiv:2305.03520 [cs].

[17] Mente, R., Aland, S., Chendage, B., 2022. Review of Word Sense Disambiguation and It'S Approaches. SSRN Electronic Journal URL: https://www.ssrn.com/abstract=4097221, doi:10.2139/ssrn.4097221.

[18] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., Gao, J., 2024. Large Language Models: A Survey. URL: http://arxiv.org/abs/2402.06196. arXiv:2402.06196 [cs].

[19] Moro, A., Raganato, A., Navigli, R., 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. Transactions of the Association for Computational Linguistics 2, 231–244. URL: https://direct.mit.edu/tacl/article/43316, doi:10.1162/tacl_a_00179.

[20] Nguyen, Q.P., Vo, A.D., Shin, J.C., Ock, C.Y., 2018. Effect of Word Sense Disambiguation on Neural Machine Translation: A Case Study in Korean. IEEE Access 6, 38512–38523. URL: https://ieeexplore.ieee.org/document/8399736/, doi:10.1109/ACCESS.2018.2851281.

[21] Pasini, T., 2020. The Knowledge Acquisition Bottleneck Problem in Multilingual Word Sense Disambiguation, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, Yokohama, Japan. pp. 4936–4942. URL: https://www.ijcai.org/proceedings/2020/687, doi:10.24963/ijcai.2020/687.

[22] Raganato, A., Camacho-Collados, J., Navigli, R., 2017. Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Association for Computational Linguistics, Valencia, Spain. pp. 99–110. URL: http://aclweb.org/anthology/E17-1010, doi:10.18653/v1/E17-1010.

[23] Scarlini, B., Pasini, T., Navigli, R., 2020. Sense-Annotated Corpora for Word Sense Disambiguation in Multiple Languages and Domains .

[24] Sumanathilaka, T., Micallef, N., Hough, J., 2024a. Assessing GPT's Potential for Word Sense Disambiguation: A Quantitative Evaluation on Prompt Engineering Techniques, in: Proceedings of 15th Control and System Graduate Research Colloquium (ICSGRC), IEEE, Mardhiyyah Hotel & Suites, Shah Alam, Malaysia.

[25] Sumanathilaka, T., Micallef, N., Hough, J., 2024b. Can LLMs assist with Ambiguity? A Quantitative Evaluation of various Large Language Models on Word Sense Disambiguation, in: The First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security NLPAICS2024.

[26] Sumanathilaka, T., Micallef, N., Hough, J., 2025. Exploring the impact of temperature on large language models: A case study for classification task based on word sense disambiguation, in: 2025 7th International Conference on Natural Language Processing (ICNLP),Guangzhou, China.

[27] Sumanathilaka, T., Weerasinghe, R., Priyadarshana, Y., 2023. Swa-Bhasha: Romanized Sinhala to Sinhala Reverse Transliteration using a Hybrid Approach, in: 2023 3rd International Conference on Advanced Research in Computing (ICARC), IEEE, Belihuloya, Sri Lanka. pp. 136–141. URL: https://ieeexplore.ieee.org/document/10145648/, doi:10.1109/ICARC57651.2023.10145648.

[28] Wang, M., Wang, Y., 2020. A Synset Relation-enhanced Framework with a Try-again Mechanism for Word Sense Disambiguation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online. pp. 6229–6240. URL: https://www.aclweb.org/anthology/2020.emnlp-main.504, doi:10.18653/v1/2020.emnlp-main.504.

[29] Wang, Y., Wang, M., Fujita, H., 2020. Word Sense Disambiguation: A comprehensive knowledge exploitation framework. Knowledge-Based Systems 190, 105030. URL: https://linkinghub.elsevier.com/retrieve/pii/S0950705119304344, doi:10.1016/j.knosys.2019.105030.

[30] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y., 2023. ReAct: Synergizing Reasoning and Acting in Language Models. URL: http://arxiv.org/abs/2210.03629. arXiv:2210.03629 [cs].

[31] Zhang, Z., Zhang, A., Li, M., Smola, A., 2022. Automatic chain of thought prompting in large language models. arXiv preprint arXiv:2210.03493 .