Assessing GPT's Potential for Word Sense Disambiguation: A Quantitative Evaluation on Prompt Engineering Techniques

Deshan Sumanathilaka
Department of Computer Science
Swansea University
Swansea, Wales, United Kingdom
deshankoshala@gmail.com

Nicholas Micallef
Department of Computer Science
Swansea University
Swansea, Wales, United Kingdom
nicholas.micallef@swansea.ac.uk

Julian Hough
Department of Computer Science
Swansea University
Swansea, Wales, United Kingdom
julian.hough@swansea.ac.uk

***Abstract*—Modern digital communications (including social media content) often contain ambiguous words due to their potential for multiple related interpretations (polysemy). This ambiguity poses challenges for traditional Word Sense Disambiguation (WSD) methods, which struggle with limited data and lack of contextual understanding. These limitations hinder efficient translation, information retrieval, and question-answering systems, thereby restricting the benefits of computational linguistics techniques when applied to digital communication technologies. Our research investigates the use of Large Language Models (LLMs) to improve WSD using various prompt engineering techniques. We propose and evaluate a novel method that combines a knowledge graph, together with Part-of-Speech (POS) tagging and few-shot prompting to guide LLMs. By utilizing prompt augmentation with human-in-loop on few-shot prompt approaches, this work demonstrates a substantial improvement in WSD. This research advances accurate word interpretation in digital communications, leading to important implications for improved translation systems, better search results, and more intelligent question-answering technology.**

Keywords— Large Language Models, Word Sense Disambiguation, FEWS sense tags, Few Shot Prompting, Knowledge Graph

I. Introduction

Understanding the contextual meaning of words and phrases within sentences is crucial in many spoken and written tasks, as incorrect interpretation can lead to ambiguity, which can result in misinformation [1]. Lexical ambiguity pose significant challenges that need to be addressed by using various computational techniques. To solve this problem, recent research has focused on WSD in both English and non-English languages, by employing diverse algorithms, achieving some success [2]. However, current WSD systems still struggle with particularly challenging instances, especially for words with a large number of senses [3]. For a instance consider the following sentence: "*I went to the bank to give some money to my friend.*" According to the FEWS sense tag, the word "*bank*" is ambiguous and can have 25 different meanings as a noun. In this context, "bank" could refer to "*an institution where one can place and borrow money and manage financial affairs*" or "*the edge of a river, lake, or other watercourse".* To accurately identify the correct sense of the word "bank", it is essential to analyze the surrounding words and context within the sentence. In these cases, disambiguating the intended meaning can prove exceptionally difficult, even for the most complex and optimized algorithms.

Recent research has shown the importance of contextual analysis for WSD demonstrating the limitations of isolated word analysis, particularly in cases of high polysemy [4]. While substantial research has been conducted on supervised WSD, often leveraging pragmatic relationships (synonyms, hyponyms, hypernyms), this study takes a different approach by utilizing a Retrieval-Augmented Generation (RAG)-inspired technique with a knowledge graph, together with POS tagging and few-shot prompting to guide LLMs. In previous work in WSD, Vial *et al.* reduced sense tags with a compressed sense vocabulary of Princeton WordNet to study disambiguation behaviour [5]. This strategy successfully decreased the size of neural WSD models. Recent work frequently employs transformer-based models to address word ambiguity. Architectures like BERT-large with feed-forward neural networks have been combined with knowledge graphs to enhance WSD, including the prediction of sense sets not present in training data [6]. GLOSSBERT fine-tuned a pre-trained BERT model with the SemCor dataset by constructing context gloss pairs, effectively transforming WSD into a sentence-pair classification problem. ConSEC reframed WSD as a text extraction problem using DeBERTA [7], surpassing previous state-of-the-art performance. ECS leverages definitional contexts for analysis and meaning extraction, framed as a span extraction problem [8]. A major limitation in these studies is the lack of datasets with appropriate sense annotations. SemCor and FEWS are among the few commonly used manually annotated corpora [9]. To address this, LLMs and generative AI demonstrate a remarkable capacity for solving complex linguistic tasks as they are trained on vast text corpora. Their ability to reason with language and generate error-free responses offers a promising solution for accurate WSD [10]. This study investigates the potential of LLMs for WSD using various prompt engineering techniques. We will explore the capabilities of LLMs in performing WSD tasks and examine techniques to optimize their accuracy and overall performance. The study is based on GPT 3.5 while following three distinct pipelines to assess LLM capabilities: WSD with prompting alone, WSD utilizing sentence augmentation to improve sense interpretation of ambiguous words, and a hybrid RAG-inspired model incorporating a knowledge base (KB) alongside the LLM.

The contributions of the study are:

1) Proposing a novel approach to WSD using an LLM with prompt augmentation.
2) Presenting the effectiveness of hybrid models which use a knowledge base and the LLM-based model for identifying the sense tags of the ambiguous words.
3) Showing the capabilities of LLM to handle lexical ambiguity across different POS tags.

In the following sections we present related work, details on our methodology, the experimental results, draw conclusions, and outline potential future work.

II. Related Work

Ambiguity in spoken and written natural language poses a significant challenge for various automated natural language processing (NLP) tasks, with WSD being a fundamental problem. WSD has been one of the continuing research areas in different languages as the proper word sense directly impacts many NLP tasks like Machine Translation, Question Answering, Text Summarization, Text Classification, and Word Sense Induction [4]. Several advanced new neural architectures have been suggested by many of the researchers for the WSD task by integrating knowledge base models. Various NLP techniques are used to perform WSD effectively, and below we give an overview of some of the main approaches.

A. Supervised WSD using Lexical Knowledge

Supervised WSD relies on labelled datasets (e.g., Semcor, FEWS, Wordnet) to train models, with sense-annotated datasets available for multiple languages [8]. To improve accuracy, techniques like stacked bidirectional LSTM networks with attention mechanisms [11], data augmentation with SMSMix [12], and deep learning models such as BiLSTM (particularly successful in low-resource languages) [13] have been used. Systems like EWISER integrate knowledge bases and synset embeddings for better predictions [6] while GLOSSBERT leverages gloss knowledge [14]. Additionally, context-dependent methods [15] identification of multiple senses [16], and the use of synonyms and example phrases [17] are further employed to improve WSD performance. These studies demonstrate the crucial role of sense-annotated data in improving WSD results. We utilize the wealth of annotated data by feeding the model using few-shot prompting technique. The sense annotated data with respective sense id helps to drive the inference process smoothly to identify the intended sense id of the ambiguity word.

B. Knowledge base WSD

KB approaches to WSD leveraging external resources like lexical databases and ontologies have been widely used , often employing semantic similarity measures and graph-based algorithms. For instance, unsupervised graph-based algorithms utilize resources like Hindi WordNet to represent word senses and their relationships [18]. Bootstrapping techniques incorporating WordNet synsets have also been explored, demonstrating the effectiveness of integrating linguistic knowledge [19]. One graph-based approach is to use a complex network approach to represent ambiguous sentences as vertices in a network built on semantic similarities. This technique is particularly beneficial for contexts with limited information [20]. Additionally, context-aware semantic similarity-based research explores incorporating contextual information to improve disambiguation, even with limited annotated data [21]. KB strategies are expanding WSD toolkits, with frameworks like SREF augmenting sense embeddings with relation information and employing a refinement mechanism [22]. Other studies explore alternative approaches, such as Trie structure-based methods for Romanized Sinhala WSD [23], [24]. Additionally, studies explore the effectiveness of semi-supervised WSD using graph-based SSL algorithms and various word embeddings combined with parts-of-speech tags and word context [25]. KB approaches are well-suited to addressing limited annotated data issues, and our system incorporates advancements in KB techniques for improved results.

C. Hybrid Approach for WSD

Hybrid methodologies are emerging as a promising avenue for Word Sense Disambiguation (WSD). These approaches use techniques from different paradigms. For example, the TWE-WSD method integrates Latent Dirichlet Allocation (LDA) with word embedding techniques to achieve disambiguation [26]. However, TWE-WSD struggles with complex linguistic phenomena like homonymy and polysemy. Additionally, research on English word translation using a hybrid strategy incorporating translation aids and WordNet highlights the importance of hybrid model systems for effective WSD [27]. These studies show the importance of careful investigation of the strengths and weaknesses of different paradigms when designing hybrid models. Our work builds on this body of research by incorporating GPT in conjunction with a knowledge base approach to enhance the accuracy of WSD predictions.

D. Large Language Models for WSD

LLMs demonstrate an inherent understanding of word senses, even without explicit WSD training [28]. By framing WSD as a textual entailment problem, researchers can evaluate whether LLMs determine the sense label accurately in a sentence containing an ambiguous word. Zero-shot and few-shot approaches show promise, surpassing random guesses and sometimes rivalling supervised WSD systems [29]. Further research explores cross-lingual WSD using pre-trained language models and zero-shot approaches informed by cross-lingual knowledge [30]. Beyond prediction, LLMs like GPT-2 have been used for contextual data augmentation, showcasing their broad utility in WSD [31]. Our work explores the use of the inherited knowledge embedded in LLMs for enhanced prediction of the correct sense of ambiguous words in spoken and written linguistic data.

III. Methodology

According to the discussion of the previous studies, it is evident that available algorithms and systems need to be further finetuned and examined for the proper identification of ambiguous words [32]. In our work, we evaluate the semantic understanding of LLMs by using different techniques, namely Prompt augmentation and knowledge-based few-shot prompting. We attempted to use well-crafted prompts using different prompt engineering techniques to drive the inference process to the downstream task of WSD.

A. Dataset

This work was conducted using the publicly available FEWS dataset [33] which contains the sense tag list, training data and the test data. The selection of FEWs dataset was influenced by its nature of sense arrangement and diversity of the test cases. In all the proposed methods, the models are evaluated for their ability to correctly assign sense tags to ambiguous words positioned between <WSD> tokens.

TABLE I. FORMAT OF THE SENSE TAG INPUT/OUTPUT

*Sentence: The lead author meticulously cross-checked the manuscript against various <WSD> dictionaries </WSD>, striving to ensure both word choice and proper usage.			
*sense_id:	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		

* Model parameters used for the Study

The Inputs, outputs and the sense tag format are highlighted in Table I. This dataset contains a large training set which consists of different interpretations of ambiguous words. The sense tags arrangement of FEWS is depicted in Table I and the selected parameters for the study is shown using the asterisk (*) mark.

The FEWS training data distribution used for the KB creation based on POS tag is listed in Table II. The training data has been pre-processed into the format of Table I, consisting of ambiguous words and their sense tag representations. During pipeline 2, the relevant training data along with the sense ID and gloss is extracted and shared with the GPT-3.5-turbo model for few-shot prompting purposes.

TABLE II. TRAINING DATA DISTRIBUTION

POS Tag	No of Records	POS Tag	No of Records
Nouns	55442	Adjectives	19269
Verbs	24396	Adverbs	2324
Total			101458

B. Study Procedure

This study is divided into two phases. In the first phase, we evaluate how general zero-shot prompting and Chain of thought (COT) prompts can enhance WSD performance. Our methodology uses a human-in-the-loop approach in the prompt engineering process, with the aim of refining the prompts to achieve improved model behaviour. Specifically, the lead researcher used a human-in-the-loop approach to iteratively refine the prompts by carefully analysing the results of the incorrect model predictions. The first phase follows the data flow shown in Fig 2 mainly consists of 2 steps: preprocessing the sentence with ambiguity word and prompting the LLM. In the initial preprocessing step, the FEWS sense tag is stored, and the sentence is pre-processed to remove irrelevant tokens or symbols. NLTK Universal tag set is used to identify the POS tag of the target word for WSD, which is used in the following step to filter the sensed tag. Filtered sense tags and the sentence are fed to GPT-3.5-turbo for response generation using the prompting listed in Table III. In both general and enhanced prompting mode, {meaning} contains the filtered definitions from the FEWS sense tag (Lexical knowledge) based on the POS tag identified. It contains the gloss and the sense ID only.

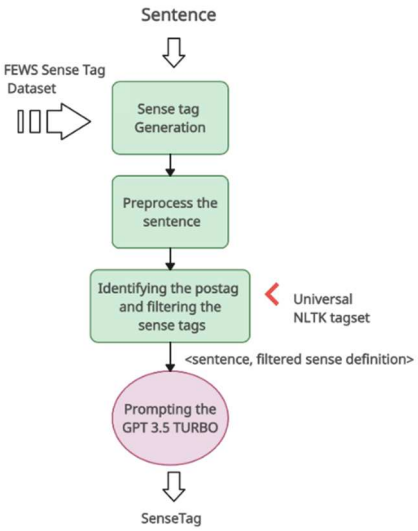

Fig. 1. Pipeline for Phase 1 with Zero Shot Prompting.

TABLE III. PROMPT USED FOR ZERO SHOT COT PROMPTING

General Prompting (Zero Shot)
Examine the sentence. {sentence}. Identify the most suitable Meaning associated with the word enclosed within the <WSD>. Return most suitable sense id associated with from below. it contains sense ID and its definition {meanings}.
Enhanced Prompting (Zero Shot COT)
You are going to identify the sense tag of an ambiguity word in English. Do the following tasks. 1. Examine the sentence below. "{sentence}". 2. Identify the meaning of the word enclosed within the <WSD> tags. You need to consider the total sentence before you get the exact meaning of the word. 3. Based on the identified meaning, try to find the most appropriate senseIDs from the below. "{meanings}". 4. If you have more than one senseIDs identified, you can return the senseIDs in order of confidence level. 5. Return a proper JSON object that contains the ambiguity word and the finalized senseIDs. Use the following format for the output. <JSON object that contains ambiguity word and the finalized senseIDs>

{sentence} contains the sentence with ambiguous words enclosed with <WSD> tags. The concatenated sense ids and the sentence with ambiguous word is passed to the model to reason and generate the response. From the results of phase 1, it was clear that some of the challenging ambiguous words could not be identified by general prompting alone. The applicability of some sense tags need to be further eloborated to the models to enhance the comprehension the proper usability. Therefore, in the next phase, we proposed an enhancement to the pipeline by sharing some examples related to ambiguous words from the training data. The model was given different instances of the corresponding gloss and the sense tag to pre-learn, together with the sentence for tagging the corresponding tag. The flow of the proposed pipeline is shown in Fig 3. In the KB based pipeline, FEWS training data is structured in a Tree, maintaining the word as the root and first level parent node as the POS tag. Each instance is kept as child nodes to first-level parent.

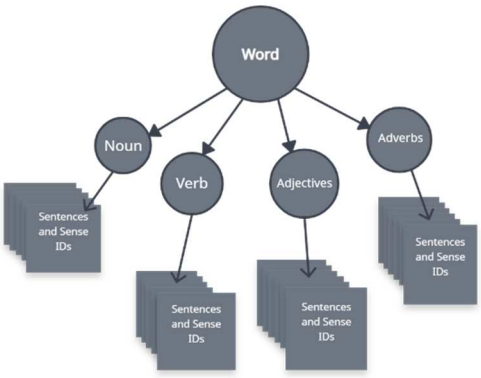

Fig. 2. Knowledge Tree structure used for the study.

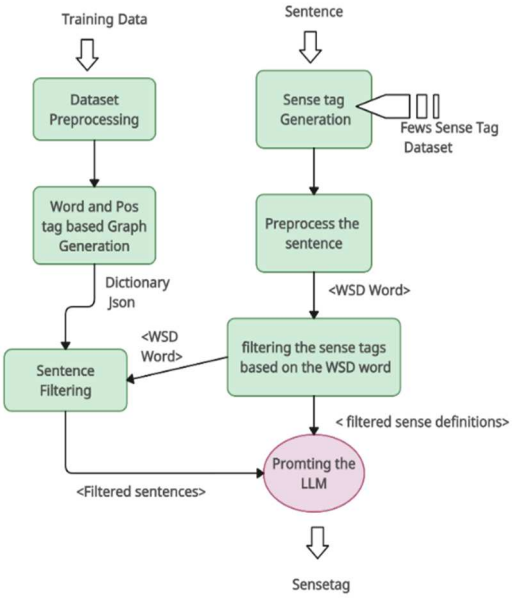

Fig. 3. Pipeline for Phase 2 with few-shot Prompting.

Arranged data is stored in a JSON file, and based on the ambiguous word, the required information is retrieved in a constant time and shared with the model for few-shot prompting. Notably the tree-based knowledge graph has improved the efficiency in the retrieval process due to its fast and efficient searching techniques. As an enhancement to pipeline 1, prompt augmentation and knowledge sharing have been incorporated. The prompt in Table IV is used as input to the GPT-3.5-turbo model for inference.

The arrangement of the knowledge base is shown in Fig 2. Compared to the prompt in Table III, {examples} has been introduced. Retrieved instances from the knowledge are used as few-shots required for the prompt. These examples consist of instances related to the ambiguity of the word and their sense interpretation. The meaning of the sense interpretation is shared along with the same prompt. This helps the GPT model to broaden the lexical knowledge of a word sense and its usage in a proper context. This task has helped the model to learn the context usage of ambiguous words where necessary. This approach has shown significant improvements in performance of less frequently used cases during the evaluation process.

TABLE IV. PROMPT USED FOR FEW-SHOT COT PROMPTING

Improved Prompting with Knowledge Base (Augmented few-shot COT)
You are going to identify the corresponding sense tag of an ambiguity word in English sentences. Do the following tasks. 1. {word} has different meanings. Below are possible meanings. Comprehend the sensetags and meanings. {filtered_definitions} 2. You can learn more on the usage of each word and the meaning through below Examples. Examples are "{examples}". 3. Now Examine the sentence below. You are going to identify the most suitable meaning for ambiguity word. "{sentence}" 4. Try to identify the meaning of the word in the above sentence which is enclosed with the <WSD>. You can think of the real meaning of sentence and decide the most suitable meaning for the word. 5. Based on the identified meaning, try to find the most appropriate senseIDs from the below. You are given definition of each sense tag too."{filtered_definitions}". 6. If you have more than one senseIDs identified after above steps, you can return the senseIDs in order of confidence level. 7. Return JSON object that contains the ambiguity word and the finalized senseIDs. Use the following format for the output. <JSON Object with ambiguity word and the finalized senseIDs >

C. Experimental Setup

We aimed to test the efficacy of different prompting strategies with an off-the-shelf and widely used LLM which shows the promising results in the Natural language tasks. The GPT 3.5 turbo model was selected for this study as this model has shown to be capable of performing many semantics-related tasks [34]. An OpenAI API key was generated from a tier-one OPEN AI account. The model evaluation process maintains a temperature of 0 and an output token limit of 500 tokens. GPT-3.5-Turbo was used as the primary model, tasked with word sense identification. To evaluate the above pipelines, the test data from the FEWS dataset was used and the data was selected according to a 4:3:3 ratio for Nouns, Verbs, and Adjectives, a ratio approximately based on the distribution of POS tags in the training data. Additionally, 50 instances of adverbs were evaluated. In total, the evaluation set consisted of 1050 instances. For each testing instance, a disambiguation is deemed correct if the predicted sense tag matches the target sense tag for that word in its context. Accuracy is calculated as the percentage of correct predictions out of all test cases. The number of correct predictions, execution time and token distribution are analysed.

IV. RESULTS AND DISCUSSION

Table V shows the accuracy and statistical significance of each approach using the McNemar test based on the p values. Our evaluation demonstrates a clear hierarchy of performance among the prompting techniques: using improved prompting as shown in Table IV exhibits the most accurate results, while general prompting performs the worst. Notably, general prompting consistently underperformed compared to enhanced prompting, even when incorporating a KB. The results show that enhanced prompting itself is capable of handling WSD even without the KB. However, the effectiveness of the KB is dependent upon its proper integration into the prompt, as evidenced by Approach 4's underperformance compared to Approach 2.

TABLE V. EXPERIMENTAL RESULTS OF PROMPTING TECHNIQUES

Approach	POS TAG			
	Noun	Verb	Adj	Adv
	400	300	300	50
Baseline LLM (Zero-Shot)	0.65	0.45	0.58	0.40
LLM with Enhanced prompting (Zero-Shot COT)	0.77	0.62	0.75	0.74
General prompting with a knowledge base (few-Shot)	0.70	0.60	0.64	0.58
Enhanced prompting with knowledge base (few-Shot COT)	0.76	0.65	0.67	0.74
Improved prompting with knowledge Base (Augmented few-Shot COT)	**0.85***	**0.78***	**0.80***	**0.86***

* Indicate statistically significant differences ($p < 0.05$) using a McNemar test

Tuning the prompt in approach 5 successfully addresses this issue, integrating the KB and lexical knowledge appropriately and leading to the highest performance across all approaches. The model has performed best on noun-related tasks for all evaluated approaches (see Column 2 in Table V). This is likely because the nouns often represent concrete objects or concepts with less ambiguous meanings, allowing the model to identify their intended sense more easily within a given context. This indicates that the base GPT model seems more suitable for WSD for nouns compared to other parts of speech. A reasonable accuracy was achieved for adjectives and verbs, but more research on the field can improve performance in the future. We observed that the initial approaches struggled to identify sense tags with similar meanings due to a lack of definitional understanding. To address this, we directly provided the model with relevant definitions before querying it with test data. This human-in-loop approach yielded significant improvements, proving the effectiveness of this approach for prompt tuning. The enhanced prompts used more tokens and high execution time to generate the responses, but the increased specificity led to superior model performance.

TABLE VI. F1 SCORE BENCHMARKING WITH EXISTING APPROACHES

Model	Few-shot Dev	Few-shot Test
MFS	52.8	51.5
Lesk [35]	42.5	40.9
Probe	72.3	72.1
BEM [36]	79.3	79.0
RTWE [37]	78.0	78.4
ESR base [17]	77.9	77.8
Our Model KB + few-shot COT+ GPT-3.5-turbo	76.0	75.9

We evaluated our best resulting approach with few-shot dev (5000 instances) and few-shot test (5000 instances) from FEWS dataset. While our results with GPT-3.5-turbo model do not outperform some existing models, it is worth mentioning that the models used in our study were not trained/finetuned for the WSD task. We designed the prompts to leverage 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. These results indicate that exploring high-parameter versions of GPT models and fine-tuning GPT 3.5 turbo models using lexical knowledge along with training data could improve model accuracies and performances.

V. CONCLUSION AND FUTURE DIRECTIONS

This study demonstrates the effectiveness of combining GPT-3.5-based prompt tuning with a knowledge-based approach for WSD. The study proves the effectiveness of using the general purpose LLMs for downstream tasks like WSD with properly crafted prompts using different computational techniques. Few-shot COT prompting has demonstrated promising results revealing positive directions for the researchers to explore in semantic-related computational linguistics tasks. Future work should focus on evaluating the proposed methods on open source and commercial models like GPT-4 with SenseEval and SemEval datasets, potentially refining it further with SemCor training data with additional parameters like synonyms. While the core research remains sound, the lead researcher prioritizes improving how prompts are created. This suggests that expertise in crafting prompts can significantly impact choosing the best one. To ensure fair evaluation, it's important to further explore techniques that minimize bias in prompt selection.

This work establishes a new approach for WSD achieving promising results which demonstrates the potential for real-world applicability. In future work, we plan to carry out further investigations across a diverse set of commercial and open-sourced models, to solidify the generalizability of the approach and unlock its potential for various real-world applications.

ACKNOWLEDGEMENT

We would like to express our gratitude to the OpenAI Researcher Access Program for providing credits to support the development of this project.

REFERENCES

[1] R. Mente, S. Aland, and B. Chendage, "Review of Word Sense Disambiguation and It'S Approaches," *SSRN Journal*, 2022, doi: 10.2139/ssrn.4097221.

[2] Q.-P. Nguyen, A.-D. Vo, J.-C. Shin, and C.-Y. Ock, "Effect of Word Sense Disambiguation on Neural Machine Translation: A Case Study in Korean," *IEEE Access*, vol. 6, pp. 38512–38523, 2018, doi: 10.1109/ACCESS.2018.2851281.

[3] A. Raganato, J. Camacho-Collados, and R. Navigli, "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*, Valencia, Spain: Association for Computational Linguistics, 2017, pp. 99–110. doi: 10.18653/v1/E17-1010.

[4] M. Bevilacqua, T. Pasini, A. Raganato, and R. Navigli, "Recent Trends in Word Sense Disambiguation: A Survey," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, Montreal, Canada: International Joint Conferences on Artificial Intelligence Organization, Aug. 2021, pp. 4330–4338. doi: 10.24963/ijcai.2021/593.

[5] L. Vial, B. Lecouteux, and D. Schwab, "Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation." arXiv, Aug. 27, 2019. Accessed: Feb. 08, 2024. [Online]. Available: http://arxiv.org/abs/1905.05677

[6] M. Bevilacqua and R. Navigli, "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*, Online: Association for Computational Linguistics, 2020, pp. 2854–2864. doi: 10.18653/v1/2020.acl-main.255.

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

[8] E. Barba, T. Pasini, and R. Navigli, "ESC: Redesigning WSD with Extractive Sense Comprehension," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online: Association for Computational Linguistics, 2021, pp. 4661–4672. doi: 10.18653/v1/2021.naacl-main.371.

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

[10] S. Minaee *et al.*, "Large Language Models: A Survey." arXiv, Feb. 20, 2024. Accessed: May 22, 2024. [Online]. Available: http://arxiv.org/abs/2402.06196

[11] Y. Sun and J. Platoš, "Attention-based Stacked Bidirectional Long Short-term Memory Model for Word Sense Disambiguation," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, p. 3594780, May 2023, doi: 10.1145/3594780.

[12] H. S. Yoon, E. Yoon, J. Harvill, S. Yoon, M. Hasegawa-Johnson, and C. D. Yoo, "SMSMix: Sense-Maintained Sentence Mixup for Word Sense Disambiguation." arXiv, Dec. 21, 2022. Accessed: Feb. 08, 2024. [Online]. Available: http://arxiv.org/abs/2212.07072

[13] M. Le, M. Postma, J. Urbani, and P. Vossen, "A Deep Dive into Word Sense Disambiguation with LSTM," 2018.

[14] L. Huang, C. Sun, X. Qiu, and X. Huang, "GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge." arXiv, Jan. 05, 2020. Accessed: Feb. 08, 2024. [Online]. Available: http://arxiv.org/abs/1908.07245

[15] N. Koppula, K. S. Rao, and B. VeeraSekharReddy, "Word Sense Disambiguation Using Context Dependent Methods," in *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India: IEEE, Jun. 2021, pp. 1582–1590. doi: 10.1109/ICOEI51242.2021.9452823.

[16] R. Orlando, S. Conia, F. Brignone, F. Cecconi, and R. Navigli, "AMuSE-WSD: An All-in-one Multilingual System for Easy Word Sense Disambiguation," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 298–307. doi: 10.18653/v1/2021.emnlp-demo.34.

[17] Y. Song, X. C. Ong, H. T. Ng, and Q. Lin, "Improved Word Sense Disambiguation with Enhanced Sense Representations," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 4311–4320. doi: 10.18653/v1/2021.findings-emnlp.365.

[18] P. Jha, S. Agarwal, A. Abbas, and T. J. Siddiqui, "A Novel Unsupervısed Graph-Based Algorıthm for Hindi Word Sense Disambiguation," *SN COMPUT. SCI.*, vol. 4, no. 5, p. 675, Sep. 2023, doi: 10.1007/s42979-023-02116-1.

[19] A. Gahankari, D. A. S. Kapse, P. K. Biyani, and D. A. S. Kapse, "Marathi Word Sense Disambiguation using Bootstrapping Method," vol. 10, no. 10, 2023.

[20] C. D. Kokane, S. D. Babar, P. N. Mahalle, and S. P. Patil, "Word Sense Disambiguation: Adaptive Word Embedding with Adaptive-Lexical Resource," in *Proceedings of International Conference on Data Analytics and Insights, ICDAI 2023*, vol. 727, N. Chaki, N. D. Roy, P. Debnath, and K. Saeed, Eds., in Lecture Notes in Networks and Systems, vol. 727. , Singapore: Springer Nature Singapore, 2023, pp. 421–429. doi: 10.1007/978-981-99-3878-0_36.

[21] J. Martinez-Gil, "Context-Aware Semantic Similarity Measurement for Unsupervised Word Sense Disambiguation." arXiv, Dec. 13, 2023. Accessed: Feb. 08, 2024. [Online]. Available: http://arxiv.org/abs/2305.03520

[22] M. Wang and Y. Wang, "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)*, Online: Association for Computational Linguistics, 2020, pp. 6229–6240. doi: 10.18653/v1/2020.emnlp-main.504.

[23] T. G. D. K. Sumanathilaka, R. Weerasinghe, and Y. H. P. P. Priyadarshana, "Swa-Bhasha: Romanized Sinhala to Sinhala Reverse Transliteration using a Hybrid Approach," in *2023 3rd International Conference on Advanced Research in Computing (ICARC)*, Belihuloya, Sri Lanka: IEEE, Feb. 2023, pp. 136–141. doi: 10.1109/ICARC57651.2023.10145648.

[24] D. Sumanathilaka, N. Micallef, and R. Weerasinghe, "Swa-Bhasha Dataset: Romanized Sinhala to Sinhala Adhoc Transliteration Corpus," in *2024 4th International Conference on Advanced Research in Computing (ICARC)*, Belihuloya, Sri Lanka: IEEE, Feb. 2024, pp. 189–194. doi: 10.1109/ICARC61713.2024.10499771.

[25] J. M. Duarte, S. Sousa, E. Milios, and L. Berton, "Deep analysis of word sense disambiguation via semi-supervised learning and neural word representations," *Information Sciences*, vol. 570, pp. 278–297, Sep. 2021, doi: 10.1016/j.ins.2021.04.006.

[26] [L. Jia, J. Tang, M. Li, J. You, J. Ding, and Y. Chen, "TWE-WSD: An effective topical word embedding based word sense disambiguation," *CAAI Trans on Intel Tech*, vol. 6, no. 1, pp. 72–79, Mar. 2021, doi: 10.1049/cit2.12006.

[27] D. Ji and G. Xiao, Eds., *Chinese Lexical Semantics: 13th Workshop, CLSW 2012, Wuhan, China, July 6-8, 2012, Revised Selected Papers*, vol. 7717. in Lecture Notes in Computer Science, vol. 7717. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. doi: 10.1007/978-3-642-36337-5.

[28] O. Sainz, O. L. de Lacalle, E. Agirre, and G. Rigau, "What do Language Models know about word senses? Zero-Shot WSD with Language Models and Domain Inventories," 2023.

[29] M. Ortega-Martín, Ó. García-Sierra, A. Ardoiz, J. Álvarez, J. C. Armenteros, and A. Alonso, "Linguistic ambiguity analysis in ChatGPT." arXiv, Feb. 20, 2023. Accessed: Feb. 08, 2024. [Online]. Available: http://arxiv.org/abs/2302.06426

[30] H. Kang, T. Blevins, and L. Zettlemoyer, "Translate to Disambiguate: Zero-shot Multilingual Word Sense Disambiguation with Pretrained Language Models." arXiv, Apr. 26, 2023. Accessed: Feb. 08, 2024. [Online]. Available: http://arxiv.org/abs/2304.13803

[31] D. Loureiro, K. Rezaee, M. T. Pilehvar, and J. Camacho-Collados, "Analysis and Evaluation of Language Models for Word Sense Disambiguation," *Computational Linguistics*, pp. 1–57, May 2021, doi: 10.1162/coli_a_00405.

[32] D. Loureiro, M. T. Pilehvar, K. Rezaee, and J. Camacho-Collados, "Language Models and Word Sense Disambiguation: An Overview and Analysis," *Computational Linguistics*, vol. 0, no. 1.

[33] T. Blevins, M. Joshi, and L. Zettlemoyer, "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: Main Volume*, Online: Association for Computational Linguistics, 2021, pp. 455–465. doi: 10.18653/v1/2021.eacl-main.36.

[34] T. B. Brown *et al.*, "Language Models are Few-Shot Learners." arXiv, Jul. 22, 2020. Accessed: May 22, 2024. [Online]. Available: http://arxiv.org/abs/2005.14165

[35] P. Basile, A. Caputo, and G. Semeraro, "An Enhanced Lesk Word Sense Disambiguation Algorithm through a Distributional Semantic Model".

[36] T. Blevins and L. Zettlemoyer, "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*, Online: Association for Computational Linguistics, 2020, pp. 1006–1017. doi: 10.18653/v1/2020.acl-main.95.

[37] X. Zhang *et al.*, "Word Sense Disambiguation by Refining Target Word Embedding," in *Proceedings of the ACM Web Conference 2023*, Austin TX USA: ACM, Apr. 2023, pp. 1405–1414. doi: 10.1145/3543507.3583191.