Investigating Lexical Ambiguity using Computational Linguistics

Deshan Sumanathilaka Nicholas Micallef Julian Hough

Swansea University, Wales, United Kingdom

Introduction

Lexical Ambiguity refers to situations in written and verbal communication where the meaning of a word or text is **unclear** or can be **interpreted in multiple ways**.

Figure 1. The usage of word "bank" in similar context

Consider the sentence, "**The Bank is closed for maintenance**". The word 'Bank' can be a 'Financial institution, Sloping land beside a river or an institute which stores data or information. Here, the correct meaning can be determined only by the context. Word Sense Disambiguation (WSD), a computational linguistic task that aims to determine the correct meaning or sense of a word in a given context, can accurately identify the intended sense of ambiguous words, enhancing the performance of NLP tasks, such as Machine Translation, Information Retrieval, and QA systems.

GLOSSGPT : GPT for WSD using Few-shot COT

In this work, we present a dynamic few-shot Chain-of-Thought (COT) prompt-based technique using **GPT-4-Turbo** with a knowledge base as a retriever, which eliminates the need for fine-tuning the model for WSD tasks. 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 achieved a new state-of-the-art performance on the few-shot FEWS dataset, surpassing the 90% F1 score barrier (Sumanathilaka et al., 2025c). Major outcomes of this work are:

- Proposing and evaluating a novel approach for WSD with SOTA performance for "FEWS and FOOL ME IF YOU CAN " datasets.
- Introducing two iterative models to tackle corner cases with high ambiguity.

Models	Dev	Unified Eval Framework				POS Tag based UEF					FEWS	
	SE07	SE2	SE3	SE13	SE15	N	V	A	R	ALL	Dev	Test
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 $_B$	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 $_L$	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 $_B$	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 $_L$	**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 $_B$	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***

Table 1. F1 score of on FEWS, SemEval and Senseval. Our approach is in italics.

Exploring the Impact of Temperature on LLMs: A Case Study for Classification Task based on WSD **(Best Paper Award at ICNLP 2025)**

In this study, we investigated the effect of the model's temperature on sense classification tasks for WSD. A carefully crafted COT prompt was used to conduct the study, and FEWS lexical knowledge was shared for the gloss identification task. GPT-3.5 and 4, LlaMa-3-70B and 3.1-70B, and Mixtral 8x22B have been used as the base models for the study, with evaluations conducted at 0.2 intervals between 0 and 1 (Sumanathilaka et al., 2025a). Outcomes of the study are :

- Demonstrating that temperature significantly affects the performance of LLMs in classification tasks, emphasizing the importance of a preliminary study to select the optimal temperature for a task.
- Demonstrating that temperatures above 1 disrupt the precision needed for multi-class classification tasks of WSD.

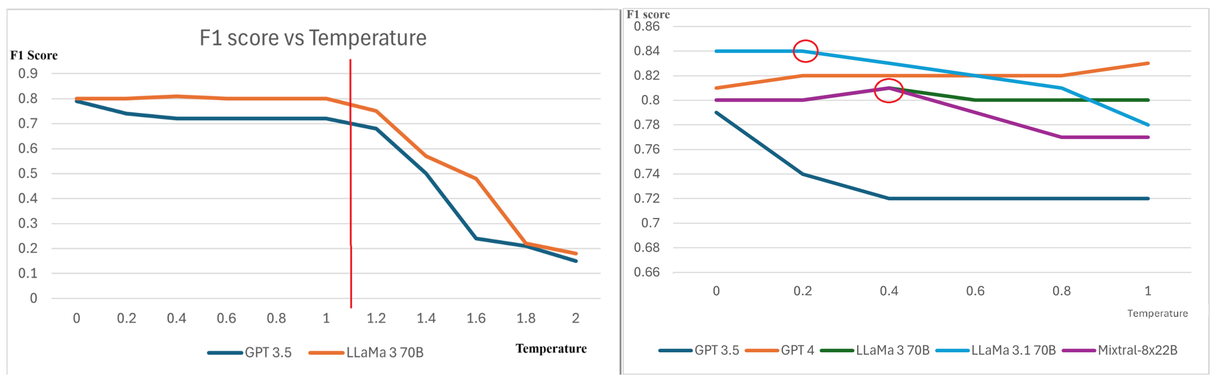

Figure 2. F1 score vs temperature distribution for initial study and extended study.

An EAD WSD Reasoning Framework with Low-Parameter LLMs

This study investigates whether low-parameter LLMs (<4B parameters) can achieve comparable results through fine-tuning strategies that emphasize reasoning-driven sense identification. Using the FEWS dataset augmented with semi-automated, rationale-rich annotations, we fine-tune eight small-scale open-source LLMs (e.g. LLama and Qwen). Our results reveal that COT-based reasoning combined with neighbour-word analysis achieves performance comparable to GPT-4-Turbo in zero-shot settings. Importantly, Gemma-3-4B and Qwen-3-4B models consistently outperform all medium-parameter baselines and state-of-the-art models on FEWS, with robust generalization to unseen senses. Furthermore, evaluation on the unseen "FOOL ME IF YOU CAN" dataset confirms strong cross-domain adaptability without task-specific fine-tuning (Sumanathilaka et al., 2026). Outcomes of the study are :

- Building three reasoning datasets using a semi-automated approach and releasing them as open-source resources for future research.
- Proposing and implementing lightweight adapters following a novel EAD (Exploration, Analysis and Disambiguation) framework that can be efficiently used for WSD with small-scale models.

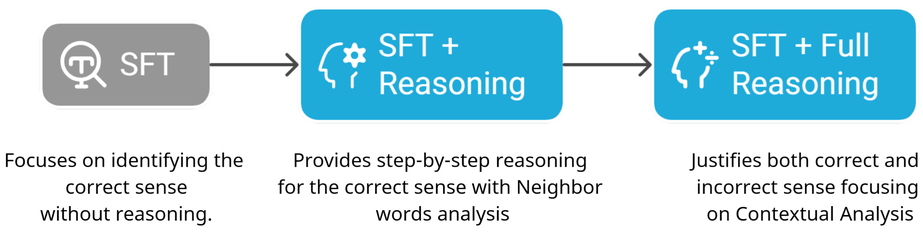

Figure 3. Finetuning strategies employed.

Models	Noun	Verb	Adjective	Adverb	Overall
Fine Tuned with COT based Neighbour words analysis Approach (Zero shot)					
Qwen 3 4B	0.79	0.67	**0.75**	0.68	0.74**
Gemma 3 4B	**0.81**	**0.71**	0.72	**0.76**	**0.75****
Fine Tuned with COT based Neighbour words analysis Approach (Few shot)					
Qwen 3 4B	0.74	0.62	0.68	0.64	0.68
Gemma 3 4B	**0.78**	**0.68**	**0.70**	**0.78**	**0.72**
Advanced Reasoning with Correct and Incorrect Sense Analysis					
Gemma 3 4B	0.76	0.60	0.66	0.64	0.68
Qwen 2.5 3B	0.74	0.61	0.68	0.66	0.68
Qwen 3 4B	**0.76**	**0.67**	**0.70**	**0.80**	**0.72**
Current Baseline					
Gemma 7B	0.49	0.41	0.51	0.46	0.47
Mixtral 7B	0.43	0.32	0.46	0.42	0.41
Yi - 34B	0.65	0.51	0.57	0.52	0.58
GPT 4o-mini	0.37	0.30	0.31	0.32	0.33

Table 2. Benchmark against Different Reasoning Strategies and Inferencing. All models are finetuned for 2 epochs. Statistical significance: * $p < 0.05$, ** $p < 0.01$

Related other Outcomes

- Assessing GPT's Potential for Word Sense Disambiguation: A Quantitative Evaluation on Prompt Engineering Techniques (Sumanathilaka et al., 2024a).
- Can LLMs assist with Ambiguity? A Quantitative Evaluation of various Large Language Models on Word Sense Disambiguation (Sumanathilaka et al., 2024b).
- Prompt Balance Matters: Understanding How Imbalanced Few-Shot Learning Affects Multilingual Sense Disambiguation in LLMs (Sumanathilaka et al., 2025d).
- From Rules to Large Language Models: A Systematic Review of Word Sense Disambiguation in English, Multilingual, and Low-Resource Languages Sumanathilaka et al. (2025b).

Publications

Sumanathilaka, D., Micallef, N., and Hough, J. (2024a). Assessing gpt's potential for word sense disambiguation: A quantitative evaluation on prompt engineering techniques. In *2024 IEEE 15th Control and System Graduate Research Colloquium (ICSGRC)*, pages 204–209.

Sumanathilaka, D., Micallef, N., and Hough, J. (2025a). 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)*, pages 178–182.

Sumanathilaka, D., Micallef, N., and Hough, J. (2025b). From rules to large language models: A systematic review of word sense disambiguation in english, multilingual, and low-resource languages. *Natural Language Processing*. Under review.

Sumanathilaka, D., Micallef, N., and Hough, J. (2025c). Glossgpt: Gpt for word sense disambiguation using few-shot chain-of-thought prompting. *Procedia Computer Science*, 257:785–792.

Sumanathilaka, D., Micallef, N., and Hough, J. (2025d). Prompt balance matters: Understanding how imbalanced few-shot learning affects multilingual sense disambiguation in llms. In *Proceedings of the Workshop on Beyond English: Natural Language Processing for All Languages in an Era of Large Language Models (GlobalNLP 2025)*. ACL.

Sumanathilaka, D., Micallef, N., and Hough, J. (2026). An exploration-analysis-disambiguation reasoning framework for word sense disambiguation with low-parameter llms. In *Proceedings of the 2026 Conference on Language Resources and Evaluation (LREC 2026)*. Under review.

Sumanathilaka, D. K., Micallef, N., and Hough, J. (2024b). Can LLMs assist with ambiguity? a quantitative evaluation of various large language models on word sense disambiguation. In Mitkov, R., Ezzini, S., Ranasinghe, T., Ezeani, I., Khallaf, N., Acarturk, C., Bradbury, M., El-Haj, M., and Rayson, P., editors, *Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security*, pages 97–108, Lancaster, UK. International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security.