ICRAMCS 2026
THE EIGHTH EDITION OF THE INTERNATIONAL CONFERENCE ON RESEARCH IN APPLIED MATHEMATICS AND COMPUTER SCIENCE
April 23-24-25, 2026 | Marrakech, Morocco

Evaluating Large Language Models for Word Sense Disambiguation in Modern Standard Arabic: Sense Selection and Sense Ranking

Communication Info

Authors:
Said Belbachir[1]
Ouafae Nahli[2]
Mohammed El Mohajir[1]
Mohamed Chahhou[1]
[1] New Technology Trends for Innovation Laboratory, Faculty of Sciences, Abdelmalek Essaadi University, Tetouan, Morocco
[2] Instituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche, Pisa, Italy

Keywords:
(1) Word Sense Disambiguation (WSD)
(2) Modern Standard Arabic (MSA)
(3) Lexical resource construction
(4) Large Language Models (LLMs)
(5) Natural Language Processing (NLP)
(6) Low-resource NLP

Abstract

Word Sense Disambiguation (WSD) remains a difficult problem in Natural Language Processing (NLP), particularly in Modern Standard Arabic (MSA), where there are few standard evaluation methods or benchmarks. Before the advent of large language models (LLMs), Arabic WSD relied on traditional methods such as knowledge-based approaches, supervised models with contextual embeddings, and similarity-based scoring methods [1,2,3]. While prompting-based WSD methods are now more common with LLMs, their effectiveness for MSA has not been tested consistently.

In this study, we introduce an evaluation framework to benchmark LLMs on MSA WSD across two tasks. The first task is sense selection, where the model chooses the correct sense from a list of candidate senses for a target word in context. The second task is sense ranking, in which the model ranks all candidate senses by their likelihood in the same context. The evaluation dataset is drawn from an open MSA WSD dataset [4,5], where each target word is paired with its context and candidate senses. This approach allows for fair comparison of different models and prompts. We test several LLMs using controlled prompts, and report results as accuracy for sense selection and ranking-based metrics for sense ranking. LLMs perform well at sense selection when given candidate lists and context. However, they struggle with sense re-ranking and do not always outperform existing WSD systems. Their performance drops for rare senses, specialized terms, and complex word forms. This suggests that Arabic LLMs perform best when they have encountered common-sense patterns and have strong lexical supervision. Our framework helps standardize LLM evaluation for MSA sense identification, identifies current limits in detailed sense ranking for Arabic, and addresses the lack of widely used Arabic WSD benchmarks [4] by

References

[1] El-Razzaz, M., Fakhr, M. W., & Maghraby, F. A. (2021). Arabic Gloss WSD Using BERT. *Applied Sciences*, *11*(6), 2567. https://doi.org/10.3390/app11062567

[2] Laatar, R., Aloulou, C., & Belghuith, L.H. (2018). Word Embedding for Arabic Word Sense Disambiguation to create a Historical Dictionary for Arabic Language. 2018 8th International Conference on Computer Science and Information Technology (CSIT), 131-135.

[3] Alian, M., & Awajan, A. (2023). Arabic word sense disambiguation using sense inventories. *International Journal of Information Technology*, *15*(2), 735-744.

[4] Kaddoura, S., & Nassar, R. (2024). A comprehensive dataset for Arabic word sense disambiguation. *Data in Brief*, *55*, 110591.

[5] Kaddoura, S., & Nassar, R. (2024). EnhancedBERT: A feature-rich ensemble model for Arabic word sense disambiguation with statistical analysis and optimized data collection. Journal of King Saud University-Computer and Information Sciences, 36(1), 101911.

[6] Jarrar, M., Malaysha, S., Hammouda, T., & Khalilia, M. (2023, December). Salma: Arabic sense-annotated corpus and wsd benchmarks. In Proceedings of ArabicNLP 2023 (pp. 359-369).