Pirates at ArabicNLU2024: Enhancing Arabic Word Sense Disambiguation using Transformer-Based Approaches

Tasneem Wael[1], Eman Elrefai[2], Mohamed Makram[3], Sahar Selim[1] and Ghada Khoriba[1]

1 NLP Research Group, Center for Informatics Science,
School of Information Technology and Computer Science (ITCS),
Nile University, Egypt
2 Widebot, Egypt, 3 NeuroTech, Egypt
T.mohamed, ghadakhoriba, sselim @nu.edu.eg
eman.elrefai@widebot.net, m.makram@neurotech.com

Abstract

This paper presents a novel approach to Arabic Word Sense Disambiguation (WSD) leveraging transformer-based models to tackle the complexities of the Arabic language. Utilizing the SALMA dataset, we applied several techniques, including Sentence Transformers with Siamese networks and the SetFit framework optimized for few-shot learning. Our experiments, structured around a robust evaluation framework, achieved a promising F1-score of up to 71%, securing second place in the ArabicNLU 2024: The First Arabic Natural Language Understanding Shared Task competition. These results demonstrate the efficacy of our approach, especially in dealing with the challenges posed by homophones, homographs, and the lack of diacritics in Arabic texts. The proposed methods significantly outperformed traditional WSD techniques, highlighting their potential to enhance the accuracy of Arabic natural language processing applications.

1 Introduction

Arabic is recognized for its morphological richness—from diacritics to the extensive use of prefixes and suffixes. This presents significant challenges for computational linguistics. These linguistic features introduce complexities. Particularly when homophones, which may differ in meaning based on diacritical marks, are standardized in academic texts (Kaddoura and Nassar, 2024). These challenges increase uncertainty, challenging the Word Sense Disambiguation (WSD) task.

WSD in Arabic involves determining the contextual meaning of a word, a task complicated by homophones, homographs, contextual ambiguity, and the frequent absence of diacritics in written texts. Arabic's rich inflectional and derivational morphology further complicates this task, necessitating robust and innovative computational approaches (Kaddoura and Nassar, 2024; El-Razzaz et al., 2021).

Recent literature on Arabic WSD is abundant, and methodologies range from traditional knowledge-based to advanced deep-learning techniques. Our study focuses on harnessing the power of transformer-based models, which have shown promise in other domains.Specifically, we explore the efficacy of contextualized word embeddings, mainly through sentence transformers, enhanced by techniques such as verb sense disambiguation and Arabic Word Sense Induction based on verb preposition classes (Djaidri et al., 2023). These techniques have demonstrated promising results, particularly in contexts where traditional methods fail.

Drawing inspiration from contrastive learning techniques utilized in computer vision tasks such as face recognition (Ou et al., 2023; Kumar et al., 2023), we adapt these approaches to Arabic WSD. Our methodology includes using the SetFit framework (Tunstall et al., 2022), optimized for few-shot learning, which is particularly suited to the constraints of Arabic language datasets in terms of size and quality. SetFit's application in this context has yielded remarkable improvements with minimal data input. We were opening the possibility of using it for WSD tasks.

Furthermore, we evaluate a classification approach, which has proven effective against traditional WSD methods like the word expert supervised method (Huang et al., 2019). This approach and exploring few-shot learning techniques form the cornerstone of our comparative analysis.

In summary, our study goes beyond comparing transformer-based approaches by incorporating insights from computer vision. We utilize cutting-edge, pre-trained models to adapt these advanced techniques to the specific challenges of Arabic WSD. By merging these methodologies, our research aims to substantially advance Arabic natural language processing.

Proceedings of The Second Arabic Natural Language Processing Conference, pages 372–376

2 Proposed Dataset Construction Approach

We used two dataset approaches to tackle the Arabic WSD task.The first approach involved collecting data from resources similar to the SALMA dataset, and the second approach utilized the SALMA development set, which was provided by the Shared Task competition (Khalilia et al., 2024).

2.1 WSD dataset

Using data resembling the test set distribution helps the model bridge the gap between the training data and test data distribution, thereby reducing covariate shift (Pan and Yang, 2010). This dataset was constructed using two resources. The first resource was the Arabic Dictionary El Gani (Abul-Azm, 2014), which was also used in the construction of the SALMA dataset (Jarrar et al., 2023). Hence, our training data would resemble the test set, a sample of the SALMA dataset. The second resource was Arabic Context Gloss pairs (El-Razzaz et al., 2021).

2.1.1 El Gani

The dataset constructed using El Gani consists of 527,66 samples, 6006 words, and 142,22 sentences.

- **50%** are **positive samples** where the sense matches the word.
- **25%** are **negative samples** which were made by shuffling the senses.
- **25%** are **hard negative samples** made through grouping the data by word and shuffling the senses while reserving the sentence.

2.1.2 Arabic context Gloss pairs

The Arabic context gloss data consists of 310,98 samples, 5347 words, and 151,30 sentences.

- **50%** are **positive samples**.
- **50%** are **hard negative samples**.

2.1.3 Combined data

We combined the two datasets, El Gani and Arabic context Gloss, into one balanced dataset with 83864 samples—the dataset distribution as shown in Figure 1.

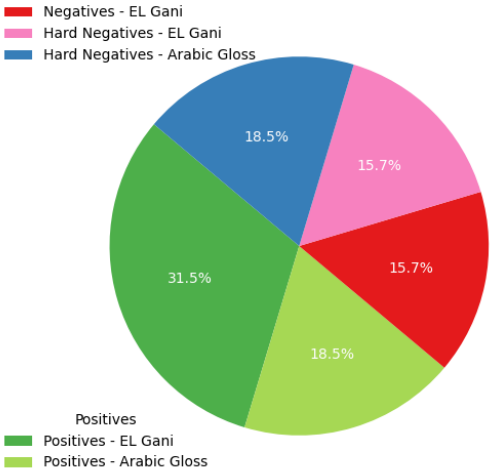

Figure 1: Combined dataset distribution

2.2 SALMA DATASET

We used the development data shared with us for training (Jarrar et al., 2023). This dataset follows the same distribution as the test set, which is expected to yield the highest F1-score. The development data was shared with us as ground truth, dictionary JSON files, and a prepared test file on the development data.We converted this format into columns for sentences, words, and meanings, and we made a label column using the Sense IDs ground truth. This allowed us to create 2,436 samples from 18 sentences, with 586 words and 2,436 different senses.

We created a feature column following the same structure mentioned in (Jarrar et al., 2023). We included the sentence, the CLS token, and the meaning. The selected word in the sentence was placed between two brackets. We split this dataset into training and validation sets with a ratio of 0.8 and 0.2, respectively. This dataset was used in experiments detailed in sections 3.2 and 3.3

3 Proposed transformer-based Pipeline for Arabic WSD

We conducted three experiments. The first experiment used the combined dataset. We used the SALMA-Dev set as training for the second and third experiments.All models were ultimately tested on the SALMA test set. We used the same base model, AraBERTv2 (Antoun et al., 2020), with three different approaches: Sentence Transformer with a Siamese network, the SetFit framework, which utilizes Sentence Transformer in a more optimized pipeline, and lastly, the classifica-

tion approach.

3.1 Experiment 1: Sentence Transformer

In this experiment, we fine-tuned AraBERTv2 as a sentence transformer (Reimers and Gurevych, 2019) with contrastive loss using the combined dataset, as shown in Figure 2.

The combined dataset was prepared for training by placing the target word between two brackets [ ] and adding it to the sentence. The model received two inputs to extract their embeddings. The first input was the sentence with the [word], and the second input was the sense of the word in the sentence context. We were using contrastive loss. We then calculated the Euclidean distance between the two embeddings to choose a threshold for the two classes: matching sense and mismatch. Additionally, we used a logistic regression model to find the best distance threshold that separated the two classes, which was 0.63.

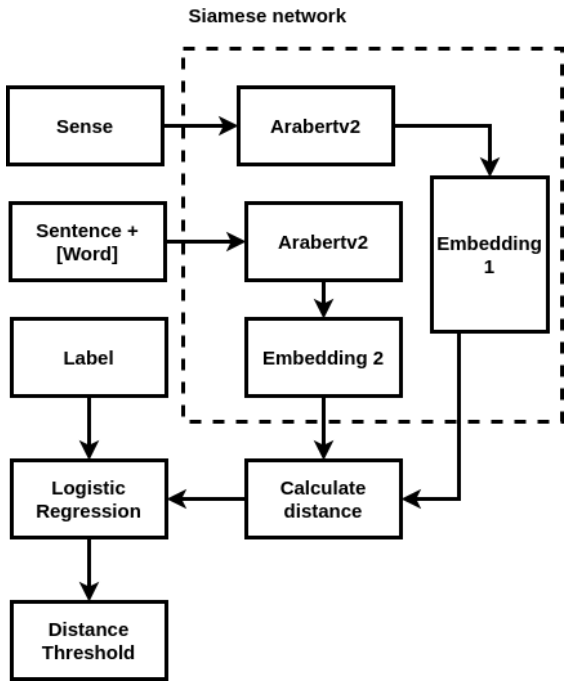

Figure 2: Arabertv2 sentence transformer pipeline

3.2 Experiment 2: SetFit

In this approach, we used a similar approach to experiment one. The main difference is that we applied a few-shot classification technique using SetFit, as shown in Figure 3. SetFit optimizes model performance by learning to generalize from several samples to new, unseen data. We used the AraBERTv2 model with 128 random samples extracted from the second dataset mentioned in Section 2.2.

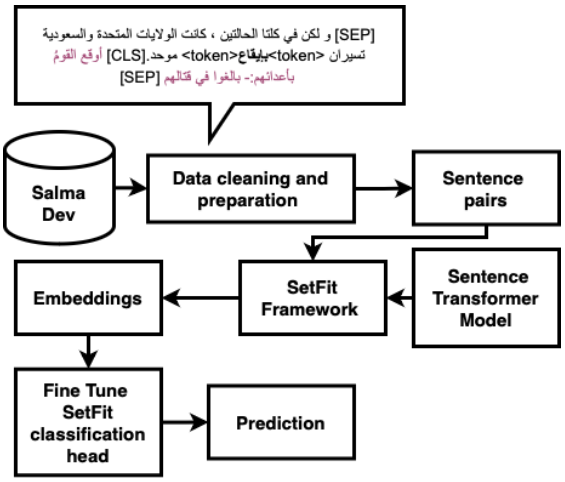

Figure 3: SetFit pipeline

The input data was the feature column, consisting of the sentence, word, and meaning, all separated by unique tokens. The output was the binary label. The loss function was cosine similarity.

3.3 Experiment 3: Classification Approach

This approach involved using transformer models for classification tasks. It provided the best results and secured second place in the competition (Khalilia et al., 2024). This commonly used and stable approach utilized the BertForSequenceClassification model and the pre-trained BERT model with a single linear classification layer. We fine-tuned the pre-trained AraBERTv2 model on 2,192 samples from the second dataset mentioned in Section 2.2.

The input and output data were the same as in the second approach (SetFit), and We used the AdamW optimizer.

4 Results

For a comparative study, all approaches were tested on the SALMA test set and dev set. We Used F1-Score for evaluation. Experiments Results are shown in table 1; in the first experiment, we evaluated our combined dataset test set, which gave 89%.The second and third experiments had the same input, and the classification layer of the SetFit approach F1-Score was 50%. In contrast, the Classification approach achieved an F1-score of 94% on the development data and 71% on the test data.

Table 1: Summary of Experiments Results

Experiments	Trained on	Tested on	F1 Score
Exp 1 sentence transformer	Combined Train	Combined Test	89%
		SALMA Dev	50%
		SALMA Test	48%
Exp 2: SetFit	SALMA Dev	SALMA Dev	72%
		SALMA Test	50%
Exp 3: Classification	SALMA Dev	SALMA Dev	94%
		SALMA Test	71%

5 Discussion

Our journey through the three approaches was guided by the scores we acquired during this research. Starting with the ArabertV2 Sentence Transformer, this contrastive learning approach using our combined dataset achieved a final F1-score of 48% on the SALMA-Test set despite an 89% score on the combined dataset test set. This discrepancy indicates that the model struggled to generalize effectively. The next step involved applying the few-shot classification technique using SetFit, which allowed us to utilize the SALMA-Dev set. SetFit, known for high performance with small datasets, resulted in a 50% F1-score. While this was promising, it highlighted the need for further optimization. Finally, the classification approach using the ArabertV2 transformer model and the SALMA-Dev set yielded our best result with a 71% F1-score. However, this performance indicated overfitting, suggesting that the model's generalization ability could be improved. Another observation is that if we compared experiment 1 and experiment 2, where the models were tested with the SALMA test set, we could find they had a very close performance. This indicates that the setfit few-shot approach is promising.

We identified several issues that we intend to address in our future work. Despite achieving the best performance with the classification approach, the scores indicate overfitting. By exploring the distribution of the datasets and its effect on model generalization, we aim to achieve better scores in the future. Understanding the impact of dataset distribution on model performance is crucial, and future work will involve deeper investigations into this aspect to enhance model robustness. Although SetFit is optimized for intent classification rather than WSD, its application has raised several questions. For example, can we achieve better scores by using different loss functions and more tuning?

6 Conclusion

This paper presented our system through three experiments to address Arabic WSD, a task complicated by the language's complexities and limited data resources. Our investigation leveraged transformer-based models, particularly the pretrained AraBERTv2, capable of generating contextual embeddings. We used these embeddings to tackle the Arabic WSD problem. The experiments on the SALMA dataset yielded promising results, achieving an F1-Score of up to 71%.

Despite the notable success of our classification approach, which secured second place in the ArabicNLP 2024 competition, we identified areas requiring further exploration and improvement. These areas include optimizing model generalization to handle diverse datasets effectively and exploring more tailored loss functions. Additionally, the varied results across different datasets have prompted us to consider deeper investigations into dataset distributions and their impact on model performance.

Future work will enhance data augmentation to generate more diverse train data to improve model generalization. Additionally, we would experiment with different loss functions and hyperparameters to optimize SetFit for WSD. Furthermore, exploring hybrid approaches by combining multiple techniques could leverage the strengths of each approach. We also aim to incorporate strategies such as verb preposition classes (Djaidri et al., 2023).

References

Abdul-Ghani Abul-Azm. 2014. *Al-ghani al-zaher dictionary*. Al-Ghani Publishing Institution, Rabat.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Process-*

ing Tools, with a Shared Task on Offensive Language Detection, pages 9–15, Marseille, France. European Language Resource Association.

Asma Djaidri, Hassina Aliane, and Hamid Azzoune. 2023. The contribution of selected linguistic markers for unsupervised arabic verb sense disambiguation. 22(8).

Mohammed El-Razzaz, Mohamed Waleed Fakhr, and Fahima A. Maghraby. 2021. Arabic gloss wsd using bert. *Applied Sciences*, 11(6).

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514, Hong Kong, China. Association for Computational Linguistics.

Mustafa Jarrar, Sanad Malaysha, Tymaa Hammouda, and Mohammed Khalilia. 2023. SALMA: Arabic sense-annotated corpus and WSD benchmarks. In *Proceedings of ArabicNLP 2023*, pages 359–369, Singapore (Hybrid). Association for Computational Linguistics.

Sanaa Kaddoura and Reem Nassar. 2024. Enhanced-bert: 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.

Mohammed Khalilia, Sanad Malaysha, Reem Suwaileh, Mustafa Jarrar, Alaa Aljabari, Tamer Elsayed, and Imed Zitouni. 2024. Arabicnlu 2024: The first arabic natural language understanding shared task. In *Proceedings of the 2nd Arabic Natural Language Processing Conference (Arabic-NLP), Part of the ACL 2024*. Association for Computational Linguistics.

C. Ranjeeth Kumar, Saranya N, M. Priyadharshini, Derrick Gilchrist E, and Kaleel Rahman M. 2023. Face recognition using cnn and siamese network. *Measurement: Sensors*, 27:100800.

Lizhen Ou, Yiping Yao, Xueshan Luo, Xinmeng Li, and Kai Chen. 2023. Contextad: Context-aware acronym disambiguation with siamese bert network. *International Journal of Intelligent Systems*, 2023:1–14.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *arXiv preprint*.