Learning Word Embeddings from Glosses: A Multi-Loss Framework for Arabic Reverse Dictionary Tasks

Engy Ibrahim, Farhah Adel, Marwan Torki, Nagwa El-Makky
Computer and Systems Engineering Department
Alexandria University, Egypt
{es-Engy.Ibrahim2024, es-farhah.adel1823, mtorki, nagwamakky}@alexu.edu.eg

Abstract

We address the task of reverse dictionary modeling in Arabic, where the goal is to retrieve a target word given its definition. The task comprises two subtasks: (1) generating embeddings for Arabic words based on Arabic glosses, and (2) a cross-lingual setting where the gloss is in English and the target embedding is for the corresponding Arabic word. Prior approaches have largely relied on BERT models such as CAMeLBERT or MARBERT trained with mean squared error loss. In contrast, we propose a novel ensemble architecture that combines MARBERTv2 with the encoder of AraBART, and we demonstrate that the choice of loss function has a significant impact on performance. We apply contrastive loss to improve representational alignment, and introduce structural and center losses to better capture the semantic distribution of the dataset. This multi-loss framework enhances the quality of the learned embeddings and leads to consistent improvements in both monolingual and cross-lingual settings. Our system achieved the best rank metric in both subtasks compared to the previous approaches. These results highlight the effectiveness of combining architectural diversity with task-specific loss functions in representational tasks for morphologically rich languages like Arabic.

1 Introduction

The reverse dictionary task (Hill et al., 2016) aims to retrieve a target word based on its definition or description. Unlike traditional dictionary lookup, which maps words to their meanings, reverse dictionary systems assist users in finding the right word when they can only recall its definition. This task has practical applications in writing assistance, vocabulary learning, and aiding users experiencing the tip-of-the-tongue phenomenon (Brown and McNeill, 1966)—when a person knows the meaning of a word but cannot recall the word itself. It is especially valuable for second-language learners and multilingual users who might grasp a concept in one language but struggle to retrieve the corresponding word in another.

This work presents our solution to the Arabic Reverse Dictionary Shared Task (Al-Matham et al., 2023), which involves predicting word embeddings from glosses in either Arabic or English. The dataset includes Arabic words paired with their glosses and corresponding word embeddings based on SGNS (Mikolov et al., 2013) and ELECTRA (Clark et al., 2020). Subtask 1 focuses on Arabic glosses, while Subtask 2 uses English glosses to predict the same Arabic word embeddings. In this work, we focus on the ELECTRA embeddings, which provide stronger semantic representations due to their transformer-based pretraining.

Prior approaches in Arabic reverse dictionary modeling have typically relied on BERT-based models (Devlin et al., 2019) trained using mean squared error (MSE) objective. While these models can capture contextual information, they often fail to structure the embedding space in a way that facilitates discriminative retrieval. In particular, MSE-based training encourages numerical closeness to the target embedding but does not explicitly enforce semantic clustering, separation between unrelated words, or alignment between gloss and word embeddings (Gao et al., 2021). As a result, the predicted embedding may be close to the correct target, but not necessarily closer to it than to other distractor words, which can harm rank performance.

In this work, we propose a novel ensemble-based model for Arabic reverse dictionary modeling that combines the encoder of AraBART (Eddine et al., 2022), a sequence-to-sequence model trained on large Arabic corpora, with MARBERTv2 (Abdul-Mageed et al., 2020), a BERT-based model specialized for Arabic. To improve the quality and discriminability of the generated embeddings, we

Proceedings of The Third Arabic Natural Language Processing Conference, pages 384–388

design a multi-loss training objective that integrates contrastive (Chen et al., 2020), structural, and center alignment losses. Our method achieves state-of-the-art performance on both the monolingual and cross-lingual subtasks of the 2023 Arabic Reverse Dictionary Shared Task.

Our contributions can be summarized as follows:

1. We present a new ensemble architecture for Arabic reverse dictionary modeling, combining AraBART and MARBERTv2 to leverage complementary semantic representations learned from generative and masked language modeling objectives.

2. We introduce a multi-loss training objective that combines contrastive, structural alignment, and center alignment losses to improve the structure and quality of the learned embedding space.

3. We evaluate our method on both monolingual and cross-lingual settings and show that it achieves state-of-the-art performance on rank metric.

4. We provide a detailed analysis of the contribution of each loss function, illustrating how each component—contrastive, structural alignment, and center alignment loss—contributes to learning more discriminative and semantically aligned embeddings.

2 Dataset

We use the dataset from the Arabic Reverse Dictionary Shared Task, designed for both monolingual and cross-lingual modeling. It consists of three subsets:

Subset 1: Arabic Dictionary. Contains Arabic glosses, their corresponding Arabic words, and two target embeddings (SGNS and ELECTRA). This subset is used in Subtask 1, which involves predicting Arabic word embeddings from Arabic definitions.

Subset 2: English Dictionary. Each entry includes an English gloss, the corresponding English word, and its SGNS and ELECTRA embeddings. It mirrors Subset 1 in structure.

Subset 3: Cross-lingual Mapping. Provides alignment data, including Arabic and English glosses, their corresponding words, and the Arabic embeddings. It supports Subtask 2, which predicts Arabic embeddings from English definitions.

All subsets are split into training, development, and test sets, as summarized in Table 1.

Subset	Train	Dev	Test
Arabic Dict	45,200	6,400	6,410
English Dict	50,877	12,719	N/A
Cross-lingual Mapping	2,862	301	1,213

Table 1: Summary of the three dataset subsets provided by the Arabic Reverse Dictionary Shared Task.

In our work, we focus specifically on predicting ELECTRA embeddings, leveraging their transformer-based structure to obtain richer semantic representations of Arabic words.

3 Method

Our system finetunes two pretrained Arabic language models independently—MARBERTv2 and the encoder of AraBART—on the Arabic Reverse Dictionary dataset. For the monolingual task (Subtask 1), we train both MARBERTv2 and AraBART encoders using the first subset. For the cross-lingual task (Subtask 2), we follow the strategy proposed by (ElBakry et al., 2023) inspired from (Artetxe et al., 2023) such that instead of processing the original English glosses directly, we use their Arabic translations as input to our finetuned Arabic models. This approach allows us to maintain a unified Arabic modeling pipeline across both subtasks, reducing system complexity while leveraging cross-lingual alignment.

Both models are trained to map input glosses to the corresponding target ELECTRA embeddings using a multi-loss training framework. This framework includes three objectives, each contributing to a different aspect of embedding quality.

Contrastive Loss. We use an NT-Xent contrastive loss to ensure that each predicted embedding is closest to its correct target embedding. This loss pulls the prediction toward its corresponding ground truth and pushes it away from all other targets in the batch. Given normalized predicted embeddings $\hat{y}_i$ and target embeddings y_i for a batch of size B, the loss is defined as:

$$\mathcal{L}_{contrast} = \frac{1}{B}\sum_{i=1}^{B} CrossEntropy\left(\frac{\hat{y}_i^\top Y}{\tau},\, i\right),$$

where Y is the matrix of all target embeddings in the batch, τ is a temperature hyperparameter, and i is the index of the correct target for $\hat{y}_i$.

Structural Alignment Loss. This loss enforces that the similarity structure among predictions mirrors that of the ground truth embeddings. That is, if two target embeddings are similar, their predicted

embeddings should also be similar. Using cosine similarity, the structural alignment loss is given by:

$$\mathcal{L}_{struct} = \left\| \hat{Y}\hat{Y}^\top - YY^\top \right\|_F^2,$$

where $\hat{Y}$ and Y are the matrices of normalized predicted and ground truth embeddings, respectively, and $\| \cdot \|_F^2$ denotes the squared Frobenius norm.

Center Alignment Loss. To ensure that the global distributions of predictions and targets are aligned, we minimize the distance between their mean vectors:

$$\mathcal{L}_{center} = \left\| \frac{1}{B} \sum_{i=1}^{B} \hat{y}_i - \frac{1}{B} \sum_{i=1}^{B} y_i \right\|_2^2.$$

Overall Objective. The final training objective is a weighted sum of the three losses:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{contrast} + \lambda_2 \mathcal{L}_{struct} + \lambda_3 \mathcal{L}_{center},$$

where λ_1, λ_2, and λ_3 are hyperparameters that control the contribution of each term.

Ensembling. After training, we obtain the final predicted embedding by averaging the outputs of MARBERTv2 and AraBART:

$$\hat{y}_{final} = \frac{1}{2}(\hat{y}_{marbert} + \hat{y}_{arabart}).$$

Hyperparameters. We train both models using AdamW (Loshchilov and Hutter, 2017) with a learning rate of 5×10^{-5} and batch size of 100. For contrastive learning, we use a temperature of 0.07. Both models are trained for 10 epochs with early stopping on the development set. We also used a weight decay of 1×10^{-4}.

4 Results

Following the official shared task protocol, we report results using three evaluation metrics in the prescribed order: **rank**, **mean squared error (MSE)**, and **cosine similarity**. The *rank* metric, used as the primary evaluation criterion, computes the proportion of target embeddings that are more similar to the predicted embedding than the correct target. Lower values indicate better performance. MSE quantifies the squared distance between predicted and target embeddings, while cosine similarity measures their angular alignment.

Model	Rank ↓	MSE ↓	CosSim ↑
MARBERTv2	0.0557	0.233	0.352
AraBART	0.0663	0.244	0.301
Ensemble	**0.0496**	**0.232**	**0.355**

Table 2: Development set performance of Subtask 1 on each component using rank, mean squared error (MSE), and cosine similarity.

Model	Rank ↓	MSE ↓	CosSim ↑
MARBERTv2	0.0400	0.249	0.382
AraBART	0.0537	0.261	0.324
Ensemble	**0.0372**	**0.248**	**0.384**

Table 3: Development set performance of Subtask 2 on each component using rank, mean squared error (MSE), and cosine similarity.

4.1 Subtask 1

Table 2 presents the development set performance of our system and its individual components, while Table 4 compares our final ensemble approach to prior work on the test set.

Our system achieves substantial improvements over prior work in the rank metric. Specifically, our ensemble reduces the rank error from 0.242 to 0.0508 compared to the best baseline on the test set, reflecting a significant performance gain.

4.2 Subtask 2

Table 3 shows how our individual models and ensemble perform on the development set, while Table 5 highlights our ensemble's performance against prior systems on the test set.

As in Subtask 1, our ensemble achieves superior performance in the primary rank metric, further demonstrating the robustness and generalizability of our method across settings.

5 Analysis

To understand the contribution of each loss component, we first trained the model using only contrastive loss. This resulted in a noticeable drop in cosine similarity (0.248) and poor structural organization. As shown in Table 6, the predicted embeddings exhibited significantly lower pairwise similarity than the target embeddings, indicating that semantically similar concepts were mapped to distant points. This is expected, as contrastive loss pushes all non-matching pairs apart—even if they are semantically related.

To mitigate this, we introduced a structural alignment loss to preserve the relational structure within the embedding space. This led to a substantial in-

	Rank ↓	MSE ↓	CosSim ↑
Rosetta Stone (ElBakry et al., 2023)	0.242	0.152	0.645
Abed Team (Qaddoumi, 2023)	0.285	0.157	0.625
Qamosy (Sibaee et al., 2023)	0.281	0.236	0.519
Proposed Approach	**0.0508**	0.218	0.370

Table 4: Test set performance Comparison of Subtask 1.

	Rank ↓	MSE ↓	CosSim ↑
Rosetta Stone (ElBakry et al., 2023)	0.127	0.17	0.659
Abed Team (Qaddoumi, 2023)	0.281	0.206	0.565
Proposed Approach	**0.0278**	0.253	0.394

Table 5: Test set performance Comparison of Subtask 2.

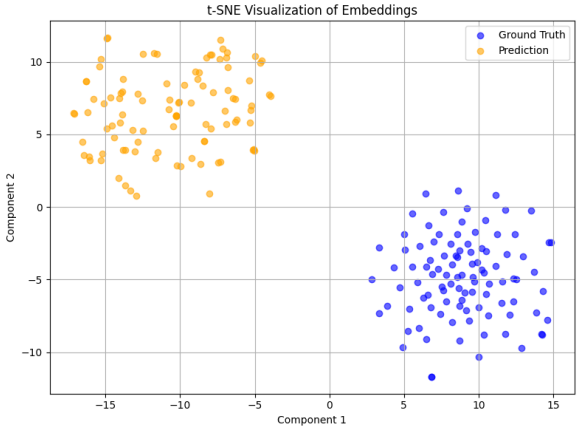

Figure 1: t-SNE visualization of predicted and target embeddings after applying structural alignment loss. The two distributions form distinct clusters.

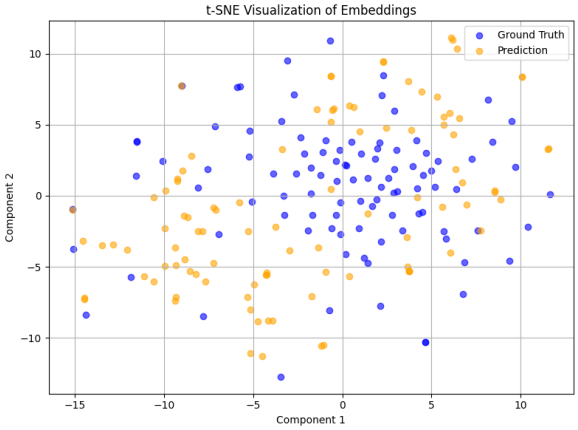

Figure 2: t-SNE visualization of predicted and target embeddings after applying center alignment loss. The two distributions are now overlapping.

crease in pairwise cosine similarity, aligning the internal structure of predictions more closely with that of the targets (Table 6).

However, despite the improved structure, evaluation metrics degraded into 0.219, 0.408 and 0.121 for cosine similarity, MSE and rank respectively. As visualized in Figure 1, the predicted and target embeddings formed separate clusters, suggesting that structural alignment alone was insufficient for proper distributional alignment.

To resolve this, we added a center alignment loss, encouraging the predicted distribution to align with the center of the target embeddings. As shown in Figure 2, this led to a more overlapping and well-aligned distribution. Also, pairwise similarity remained close to the target's value as shown in Table 6, indicating that this loss combination successfully balances spatial alignment with internal structure. All metrics improved as a result of that combination loss as well.

While our method improves the primary metric (rank), it leads to a drop in cosine similarity. This is due to the contrastive loss forcing predictions to be the closest to their specific targets and farther from all others, even when multiple targets form a semantically coherent cluster.

Loss	Preds CosSim
contrastive loss	0.0097
contrastive + structural loss	0.292
contrastive + structural + center loss	0.282

Table 6: Pairwise cosine similarity among predicted embeddings under different loss settings. The target embeddings have an internal similarity of 0.327.

6 Conclusion

We proposed an ensemble approach for Arabic reverse dictionary modeling, combining AraBART and MARBERTv2 with a multi-loss objective that includes contrastive, structural, and center alignment losses. Our method achieved state-of-the-art rank performance on both monolingual and cross-lingual subtasks of the 2023 shared task, highlighting the value of model diversity and semantically informed training.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020. Arbert & marbert: Deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

Rawan Al-Matham, Waad Alshammari, Abdulrahman AlOsaimy, Sarah Alhumoud, Asma Wazrah, Afrah Altamimi, Halah Alharbi, and Abdullah Alaifi. 2023. KSAA-RD shared task: Arabic reverse dictionary. In *Proceedings of ArabicNLP 2023*, pages 450–460, Singapore (Hybrid). Association for Computational Linguistics.

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting machine translation for cross-lingual classification. *arXiv preprint arXiv:2305.14240*.

Roger Brown and David McNeill. 1966. The "tip of the tongue" phenomenon. *Journal of verbal learning and verbal behavior*, 5(4):325–337.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Moussa Kamal Eddine, Nadi Tomeh, Nizar Habash, Joseph Le Roux, and Michalis Vazirgiannis. 2022. Arabart: a pretrained arabic sequence-to-sequence model for abstractive summarization. *arXiv preprint arXiv:2203.10945*.

Ahmed ElBakry, Mohamed Gabr, Muhammad ElNokrashy, and Badr AlKhamissi. 2023. Rosetta stone at ksaa-rd shared task: A hop from language modeling to word–definition alignment. *arXiv preprint arXiv:2310.15823*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Abdelrahim Qaddoumi. 2023. Abed at ksaa-rd shared task: Enhancing arabic word embedding with modified bert multilingual. In *Proceedings of ArabicNLP 2023*, pages 472–476.

Serry Sibaee, Samar Ahmad, Ibrahim Khurfan, Vian Sabeeh, Ahmed Bahaaulddin, Hanan Belhaj, and Abdullah Alharbi. 2023. Qamosy at arabic reverse dictionary shared task: Semi decoder architecture for reverse dictionary with sbert encoder. In *Proceedings of ArabicNLP 2023*, pages 467–471.