QMorphVec: A Morphologically-Aware Embedding of Quranic Vocabulary

Doratossadat Dastgheib[💧🍃], **Alireza Sahebi**[≋], **Ehsan Khadangi**[♨], **Ehsaneddin Asgari**[🍃✲]
d_dastgheib@sbu.ac.ir, sahebi@sharif.edu, khadangi@shahed.ac.ir, easgari@hbku.edu.qa

[💧] Department of Computer Science, Shahid Beheshti University, Tehran
[≋] Computer Engineering Department, Sharif University of Technology, Tehran
[♨] Computer Engineering Department, Shahed University, Tehran
[🍃] Qatar Computing Research Institute, Qatar

Abstract

Developing effective word representations that incorporate linguistic features and capture contextual information is an essential step in natural language processing (NLP) tasks. When working with a text corpus from a specific domain with profound meanings, such as the Holy Quran, deriving word representations based on domain-specific textual contexts is particularly valuable. In this research, we employ a context-masking approach to generate separate embedding spaces for Quranic roots, lemmas, and surface forms, and then project them into a common space through linear mapping. We demonstrate that our in-domain embeddings, trained solely on Quranic text and it morphological contexts, perform comparably to—and, in some cases, better than—OpenAI's large embeddings while surpassing the multilingual XLM-R embeddings. Additionally, through qualitative analysis, we illustrate their utility in Quranic word analogy tasks. The code and the embeddings are available at: `https://github.com/language-ml/QMorphVec`.

1 Introduction

In natural language processing, language model–based embeddings offer compact representations of words that capture their lexical, syntactic, and semantic relationships. Unlike high-dimensional methods such as one-hot encoding, embeddings encode rich semantic information derived from word contexts, making them valuable for various text processing tasks. These tasks include information retrieval [11], text classification [14], and sequence-to-sequence applications like translation [23] and summarization [18, 2]. More recently, with the advent of large language models (LLMs), embeddings have become integral to nearly all language-related tasks [9].

Word embeddings have been employed in Quranic studies by enabling tasks such as verse similarity detection [7], binary classification [5], semantic search [17], story alignment [4], and categorizing verses into categories like prayer or social interaction [19]. These applications underscore the utility of embeddings in analyzing Quranic texts.

Despite these advancements, many previous approaches rely on translations from Arabic to English or on external sources, which can obscure the inherent meanings of the words/phrases. For example, [1] generated embeddings based on English translations; however, translations often fail to capture the semantic depth of Quranic words. The word الدّنيا, translated as the "nearest" in Surah 41:12 and the "world" in Surah 4:134, inherently encompasses both meanings, which are lost in translation [22].

[✲] Corresponding author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Although embeddings derived from Islamic texts, such as Hadiths [3], and modern Arabic language methods [6, 13] are valuable, they may not fully capture the unique contexts of the Quran.

To address these limitations, (i) we develop Quran-specific embeddings by utilizing only the Quranic text and its morphological features using a context-masking embedding approach. We create separate embedding spaces for Quranic roots, lemmas, and surface forms, which are then projected into a unified space through linear mapping. (ii) We show that our embeddings perform comparably to—and in some cases surpass—OpenAI's large embeddings while exceeding the multilingual XLM-R embeddings. (iii) Furthermore, qualitative analysis demonstrates their effectiveness in Quranic word analogy tasks.

2 Methodology

2.1 Dataset description

Quranic Text: The Quran, the foremost religious text for all Muslims, was revealed in Arabic and comprises 6,236 verses (ayahs) across 114 chapters (surahs). Arabic, like other Semitic languages, employs a root-based morphology where words are often derived from common linguistic roots consisting of three or four letters. These roots form the foundation for semantically related words. For example, the Quranic terms أَكْبَرُ (akbar: bigger), كبرياء (kibriyaa: pride), and أستكبر (istakbar: being arrogant) all originate from the root كبر (kbr). In addition to roots, lemmas provide finer classification by accounting for inflectional variations. For example, تستكبر (tastakbiru), يستكبر (yastakbiru), and استكبر (astakbar) all derive from the lemma استكبر (astakbara). The Quran comprises 1,642 unique roots and 4,832 lemmas. The Quranic dataset is relatively small for training language model-based embeddings. To efficiently augment the training data, we employ two strategies: (1) analyzing different morphological layers and (2) incorporating various notions of context.

(1) Morphological layers of data: Given that Quranic words can share meanings but differ in form and diacritics, incorporating roots and lemmas alongside surface forms can enhance embedding quality. We leverage morphological data from the Quranic Corpus website to analyze the Quranic text at three levels: **(i) surface forms**, the original words as they appear; **(ii) lemmas**, groups based on inflectional variations; and **(iii) roots**, the linguistic roots consisting of three or four letters. To further refine the analysis, we prepared two versions of the Quranic text: (i) a unified version with intact surface forms and (ii) a disaggregated version that segments surface forms into their morphological components. For example, رزَقْنَاهُم (razaqnaahum: "we have provided them") is decomposed into رَزَقْ (razaq: "to provide"), نا (na: "we"), and هُم (hum: "them").

(2) Notions of context: we augmented the Quranic text by considering different levels of contexts: (i) verses, (ii) paragraphs (ruku'at) [24], and (iii) chapters. This approach provided diverse input sequences, offered varying contexts for the words, and enhanced weight learning in the embedding process. Further details on these methods are discussed in Section 2.2.

2.2 Embedding Approach

1. Base model architecture: For our base model, we employ the skip-gram approach [15] and its subword variant, fastText [8]. FastText extends skip-gram by incorporating subword information, enabling the model to better handle rare words and capture morphological variations. Both models utilize a shallow feed-forward neural network to predict surrounding context words of window size 'c' for a given target word, effectively capturing semantic relationships based on the distributional semantic hypothesis [12].

2. Morphology and context strategies: To capture the linguistic richness of the Quranic text, we generate embeddings at three levels of morphological analysis: (1) roots, (2) lemmas, and (3) surface forms. These embeddings were trained using various hyperparameters and settings, including embedding dimensions of 50, 100, 200, and 500, and sliding window sizes of 3, 5, 10, and 20. We experimented with different notions of context in the Quranic text: **(1) ayatx3**: The Quran corpus at the verse (ayah) level. **(2) surahx3**: The Quran corpus at the chapter (surah) level. **(3) ayatSurahRuku3at**: A sequential repetition of the Quran, combining verse, chapter, and section (ruku) levels. **(4) comx3**: A randomized combination of lemmas, roots, and surface forms. **(5)**

https://corpus.quran.com/

ayatRuku3atx2: A sequential repetition of the Quran, combining verse and section (ruku) levels. These variations allow us to explore how different morphology- and context-based strategies at the input influence the quality of the final embeddings.

3. Embedding projection into the joint space: Embedding Quranic roots and lemmas alongside surface forms in a shared space clusters words with diverse morphological forms around central semantic vectors, facilitating the identification of similarities across variants. By projecting surface, root, and lemma embeddings into this common joint space, we leverages the strengths of each embedding level, resulting in a more nuanced and semantically informed representation of Quranic vocabulary. Inspired by cross-lingual embedding techniques [21], we map average word embeddings to their corresponding root and lemma embeddings and align average embeddings of the same lemma using a linear projection. This approach assumes a linear relationship between corresponding embeddings.

If we consider ℓ_i as a lemma embedding and $w_{i_1}, \cdots, w_{i_k}$ as the corresponding surface word embeddings aligned with ℓ_i, we aim to find a transformation matrix T such that the relationship $(\frac{w_{i_1}+\cdots+w_{i_k}}{k})T = \ell_i$ holds. Implementing this method with neural networks involves solving the optimization problem $\min_T \sum_{i=1}^{n} (T - (\frac{w_{i_1}+\cdots+w_{i_k}}{k}) - \ell_i)^2$.

In addition to the neural network approach, we explored the direct matrix pseudo-inverse multiplication method to compute the transformation T. Let L be a matrix whose rows are the lemma embeddings, and let W be a matrix where each row represents the average of the surface word embeddings corresponding to the lemma embedding in the same row of L. The transformation matrix T computed as: $T = W^{+}L$. where W^{+} denotes the pseudo-inverse of W. This method is particularly useful when W is not invertible. Consequently, we utilized this method for all embedding levels to construct the optimal common space.

3 Results

3.1 Mean Reciprocal Rank

We used the Mean Reciprocal Rank (MRR) to assess how well embeddings from different levels—root, lemma, and surface form—align with their corresponding counterparts in other levels. After projecting embeddings from one level to another, words in the target space were ranked by cosine similarity to their aligned counterparts. If a word's match ranked $i - th$, it was assigned a value of $\frac{1}{i}$.

Figure 1 presents the average MRR scores across all mappings between root, lemma, and surface form levels across various corpora. The results demonstrate that using a shared embedding space effectively aligns embeddings across levels, enhancing semantic coherence compared to models like XLM-R [10] and OpenAI's embedding-large [20].

The embeddings of 500 dimensions and a window size of 3, based on the "*ayatSurahRuku3at*" corpus produced the best performing MRR on average. Table 1 shows projections from root to lemma and surface form for this setting. The results demonstrate that projected lemmas and surface forms align closely with their respective roots. The full results and code are available on GitHub.

3.2 Analogy

To evaluate embeddings, we examined semantic relationships between Quranic words using analogy tests, inspired by [16]. For example, in the analogy **Allah - Nur (Lightness) + Zulmat (Darkness)**, the closest result was **"Allat"**, an idol mentioned in verse 53:19. Another analogy, **Jahannam (Hell) - Kafiroon (Unbelievers) + Mumin (Believers),**" yielded **"Jannah (Paradise)"** as a close result, suggesting deep semantic connections within the embeddings. These findings highlight the nuanced meanings captured by the embeddings.

https://github.com/language-ml/QMorphVec

Table 1: Examples of the top three most similar projected lemmas and surface forms (words) to the roots *fxr*, *rsl*, *Elm*, and *Alh* are provided for embeddings generated using the **ayatSurahRuku3at** corpus, configured with 500 dimensions and a window size of 3.

Root							
اله -(A,l,h)		علم -(E ,l,m)		رسل -(r,s,l)		فخر -(f,x,r)	
Lemma	**similarity**	**Lemma**	**Similarity**	**Lemma**	**Similarity**	**Lemma**	**Similarity**
اللّه	0.6388	عَلّم	0.7261	مُرْسل	0.6725	فَخُور	0.7554
إلَه	0.6168	مُعَلّم	0.6703	أرْسَل	0.6629	فَخّار	0.7452
اللّهُم	0.6119	عَلَمَت	0.6143	مٌرسل	0.6376	تَفَاخُر	0.6886
Word	**Similarity**	**Word**	**Similarity**	**Word**	**Similarity**	**Word**	**Similarity**
ءَالِهَةَ	0.6947	عَلّمُ	0.6842	رّسُولَا	0.7312	فَخُورٍ	0.7633
أَلِهَت	0.6799	أعْلَمَ	0.6580	أرْسَلَ	0.7009	فَخّارِ	0.7249
ءَالِهَةً	0.6702	نَعْلَمُ	0.6542	أرْسَلُ	0.6918	فَخُورٌ	0.7082

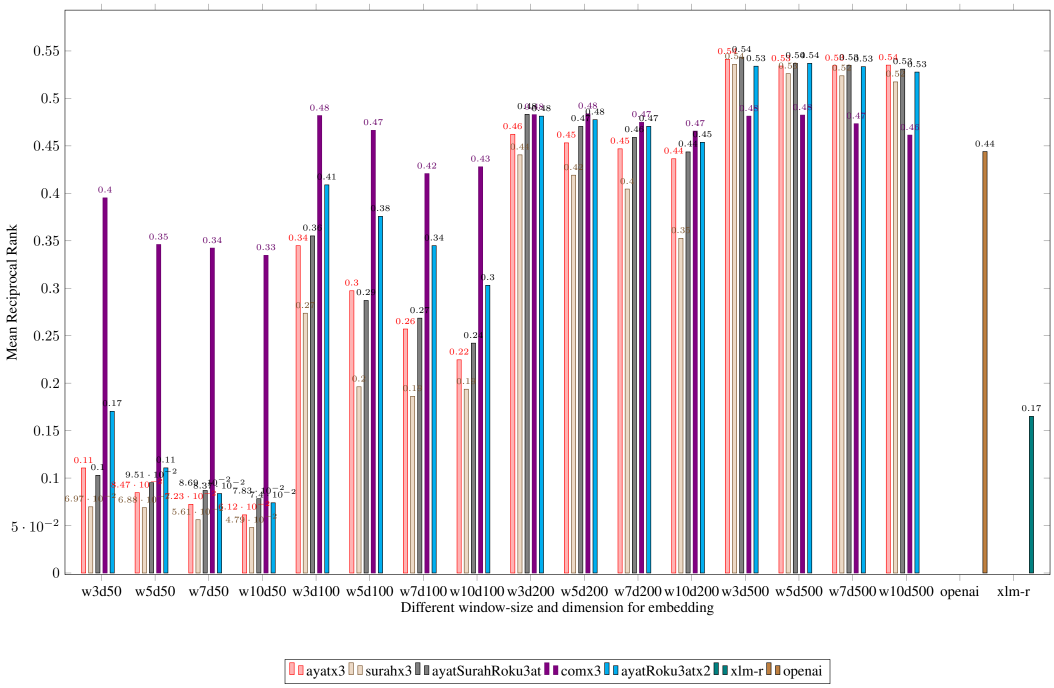

Figure 1: MRR score for different window sizes and embedding dimensions. w**A**d**B** is stands for embedding common space with window-size **A** and embedding dimension **B**. MRR is calculated to measure how different levels—root, lemma, and surface form—align with their corresponding counterparts in other levels.

Table 2: Sample relationships between Quranic words evaluated using analogy tests.

3th score	3st asnwer	2th score	2st asnwer	1th score	1st asnwer	Index	Most Related	Analogy
0.4547	مُمْسكَت	0.4819	عَلَمَت	0.5565	اللّت	1	اللّت	اللّه - نُور + ظُلُمَت =
0.3910	مُهَيْمِن	0.4037	جَنّة	0.4119	مٌؤْمِنَة	2	جَنّة	جَهَنّم - كَفِرُون + مُؤْمِن =
0.4568	خِزْى	0.4716	اخذ	0.5517	حَيَوة	1	حَيَوة	مُسْلِم - دّنْيَا + اخِر =
0.461	رَجْم	0.4739	عَجَل	0.4747	يَجْهَلُ	2	عَجَل	عَبَد - عَقَل + يَجْهَل =

4 Conclusion

In this study, we developed domain-specific word representations for Quranic text by generating separate embedding spaces for roots, lemmas, and surface forms using a context-masking approach. These embeddings were unified into a common space through linear mapping, utilizing neural transformations and matrix operations. By training exclusively on Quranic text and its morphological contexts, while augmenting data with varying context segments (verses, paragraphs, and chapters), we demonstrated that our embeddings perform comparably to OpenAI's large embeddings and surpass multilingual embeddings like XLM-R, particularly in tasks involving semantic similarity and analogy. Qualitative evaluations further validated the reliability of these embeddings for Quranic studies, underscoring their utility in applications requiring deep semantic understanding.

References

[1] Z. Aghahadi and A. Talebpour. Word emebedding in small corpora: A case study in quran. In *ICCCKE*, 2018.

[2] M. Al-Maleh and S. Desouki. Arabic text summarization using deep learning approach. *Journal of Big Data*, 7(1):109, 2020.

[3] A. M. Alargrami and M. M. Eljazzar. Imam: Word embedding model for islamic arabic nlp. In *NILES*, 2020.

[4] M. Aldawsari, E. Asgari, and M. A. Finlayson. Story fragment stitching: The case of the story of moses. In *1st Workshop on Artificial Intelligence for Narratives (AI4N 2020)*, 2021.

[5] A. N. Alsaleh, E. Atwell, and A. Altahhan. Quranic verses semantic relatedness using arabert. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 185–190. Association for Computational Linguistics, 2021.

[6] W. Antoum, F. Baly, and H. Hajj. Arabert: Transformer-based model for language understanding. In *Workshop Language Resources and Evaluation Conference*, 2020.

[7] E. Asgari. Ayat: Detecting similar quranic verses using embeddings, 2016.

[8] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[9] T. B. Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[10] A. Conneau. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[11] L. Galke, A. Saleh, and A. Scherp. Evaluating the impact of word embeddings on similarity scoring in practical information retrieval. In M. Eibl and M. Gaedke, editors, *INFORMATIK*, pages 2155–2167, Bonn, 2017. Gesellschaft für Informatik.

[12] Z. Harris. Distributional structure. *Word*, 10(2–3):146–162, 1954.

[13] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, and N. Habash. The interplay of variant, size, and task type in arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Association for Computational Linguistics, 2021.

[14] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: V. 2, Short Papers*, pages 427–431. Association for Computational Linguistics, 2017.

[15] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representation of words and phrases and their compossibility. In *NIPS 13*, volume 2, pages 3111–3119, 2013.

[16] T. Mikolov, W. Yih, and G. Zweig. Linguistic regularities in continuous space word representation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751. Association for Computational Linguistics, 2013.

[17] E. H. Mohamed and E. M. Shokry. Qsst: A quranic semantic search tool based on word embedding. *Journal of King Saud University - Computer Science and Information Sciences*, 34(3):934–945, 2022.

[18] R. Nallapati, B. Zhou, C. Santos, C. Gulcehre, and B. Xiang. Abstractive text summarization using sequence-to-sequence rnns and beyond. 2016.

[19] A. Noor and A. Ali. Multiclass imbalanced classification of quranic verses using deep learning approach. In *ICCIS*, 2021.

[20] OpenAI. New embedding models and api updates, 2024. Accessed: 2024-11-16.

[21] S. Ruder, I. Vulić, and A. Søgaard. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631, 2019.

[22] Saheeh International. *The Qur'an: Saheeh International Translation*. Saheeh International, 2013.

[23] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *NIPS 14*, volume 2, pages 3104–3112. MIT Press, 2014.

[24] A. Zarezardini, M. A. Lesani Fisharaki, and M. Khalili. Discourse genres in rokuat (case study of surah al-baqarah). *Linguistic Research in the Holy Quran*, 8(2):73–94, 2019.