Abed at KSAA-RD Shared Task: Enhancing Arabic Word Embedding with Modified BERT Multilingual

Abed Qaddoumi
Independent, NYC, NY, 11216
amq259@nyu.edu

Abstract

This paper presents a novel approach to the Arabic Reverse Dictionary Shared Task at ArabicNLP 2023 by leveraging the Bidirectional Encoder Representations from Transformers (BERT) Multilingual model and introducing modifications augmentation and using a multi attention head. The proposed method aims to enhance the performance of the model in understanding and generating word embeddings for Arabic definitions, both in monolingual and cross-lingual contexts. It achieved good results compared to benchmark and other models in the shared task 1 and 2.

1 Introduction

The Arabic Reverse Dictionary Shared Task at ArabicNLP 2023 poses unique challenges in generating word embeddings from definitions, especially in a cross-lingual setting. While traditional models have shown promise, the complexity of the Arabic language and its rich morphological structure necessitates advanced techniques. This paper introduces a modified BERT Multilingual model, incorporating changes augmentation and using a multi attention head, to address these challenges.

This paper describes the system used for the Arabic Reverse Dictionary Shared task at ArabicNLP 2023. The task was released for the the first based on the SemEval 2022 Shared Task #1: Comparing Dictionaries and Word Embeddings (CODWOE) (Mickus et al., 2022) but for the Arabic language. Competition results highlight two main trends:

1. Baseline architectures still perform competitively against new participant solutions.

2. The overall scores, especially in the definition modeling track, are unsatisfactory.

Participants identified challenges such as subpar data quality, small training corpora, and mainstream natural language generation (NLG) metrics' limited relevance. Teams have experimented with Transformer, Recurrent Neural Networks (RNN), and Convolutional neural networks (CNN) models and found success with multi-task training. There's no single architecture that stands out as the best, with some evidence suggesting that Transformers may not be ideal for this task (Mickus et al., 2022). For future research, the focus should be on enhancing dataset size and quality and re-evaluating metrics. The competition has spotlighted a variety of natural language processing (NLP) models and approaches, underscoring the field's dynamic nature.

For the Arabic Reverse Dictionary Shared task at ArabicNLP 2023, our primary objective was to assess the competitiveness of our current BERT Multilingual Cased implementation in the context of comparing dictionaries and embeddings. Additionally, we endeavored to incorporate a data augmentation strategy to enhance our results. Remarkably, our experiments with data augmentation yielded significant improvements in the development set results. The augmentation techniques employed were relatively basic, involving operations such as word addition, deletion, and swapping within sentences. Due to the limited size of the available data, even with augmentation, our training was restricted to just two epochs. Despite these constraints, our approach demonstrated competitive outcomes.

2 Literature Review

2.1 BERT:

BERT (Bidirectional Encoder Representations from Transformers) has revolutionized the field of natural language processing with its transformer architecture and pre-trained embeddings. The BERT Multilingual model, in particular, is trained on multiple languages, making it a suitable candidate for cross-lingual tasks (Devlin et al., 2018).

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 472–476

2.2 Data Augmentation:

In a detailed survey (Feng et al., 2021) addressing data augmentation (DA) in the realm of Natural Language Processing (NLP), researchers spotlighted the increasing significance of DA, especially with the rise of low-resource domains, new NLP tasks, and expansive neural networks. Although DA has been pivotal in machine learning and computer vision, its adoption in NLP remains tentative due to the challenges arising from the discrete nature of language (Feng et al., 2021). The paper underscores the necessity of DA in NLP given the expansion of large pre-trained models and the proliferation of domains with scarce training data. The authors categorize DA techniques into rule-based, example interpolation-based, and model-based strategies, emphasizing their application across various NLP tasks, from bias mitigation to few-shot learning. They further provide a continually updated GitHub repository as a resource for researchers delving into DA in NLP .

2.3 Reverse Dictionary

Reverse dictionaries, also known as retrograde dictionaries, represent a paradigm shift from traditional dictionary structures, enabling users to locate words based on anticipated definitions. A significant challenge in this domain revolves around generating definition glosses that align with user expectations. Subsequently, a growing trend in NLP has centered on the development of dynamic reverse dictionaries capable of interpreting user-input definitions and mapping them back to corresponding words. Pioneering works in this field emphasized the augmentation of definitions using semantically linked words, including synonyms, hypernyms, or hyponyms, a strategy explored across languages like English, Turkish, and Japanese. Successive research has integrated comprehensive lexical resources, including WordNet (Fellbaum, 2010) and the Oxford dictionary, among others, to further refine this approach.

The trajectory of research also unveils a subset focused on utilizing dictionaries as benchmarks for compositional semantics, as seen in the works of (Zanzotto et al., 2010) and (Hill et al., 2016). They employed neural networks and LSTMs respectively to leverage dictionaries for training. Modern iterations of reverse dictionaries utilize neural language models, exemplified by the WantWords system, which is rooted in a Bidirectional Long Short-Term Memory (BiLSTM) architecture and embraces auxiliary tasks to enhance performance. (Yan et al., 2020) endeavored to integrate pre-trained models like BERT for cross-lingual capabilities. The most recent advancements, such as the Persian reverse dictionary by (Malekzadeh et al., 2021), maintain the momentum of NLP innovations in this realm. This evolution culminates in the CODWOE shared task's interest, which emphasizes the reconstruction of word embeddings from their definitions, a premise intimately linked to prior works.

3 Methods

This section explains the on Data Augmentation and Model Architecture part of the paper.

The Data Augmentation section talks about using this method in NLP to make the model stronger and more adaptable by exposing it to a wider variety of language. Different text changing techniques like swapping synonyms, adding or removing words, and switching word order are used to make the training data more varied, which helps the model perform better.

The Model subsection explains the design and training steps, the usage of BERT Multilingual model, andd highlights key parts like Multihead Attention, a Linear Layer, and the choice of Loss Function and Optimizer. This structured approach reflects a systematic endeavor to enhance model performance and adaptability in handling text regression tasks across varying linguistic scenarios.

3.1 Data Augmentation

In the realm of natural language processing, data augmentation is a crucial strategy to enhance the robustness and generalization capabilities of models. By introducing variations in the training data, we can simulate a broader range of linguistic structures and nuances, thereby preparing the model to handle diverse real-world scenarios more effectively.

To achieve this, we have incorporated the following text augmentation techniques:

- **Synonym Replacement:** This technique is designed to introduce variations in word choice while preserving the overall meaning of the sentence. It operates by randomly selecting words from a given sentence and substituting them with their synonyms. These synonyms are sourced from WordNet, a comprehensive lexical database. This was only used for English Task.

- **Random Insertion:** This method involves adding new words into the sentence at random positions. These additional words are synonyms of existing words in the sentence, introducing diversity and expanding the vocabulary. This was only used for English Task.

- **Random Deletion:** By probabilistically removing words from the sentence, this process mimics natural language noise and encourages the model to be more robust by learning to handle missing or incomplete input.

- **Random Swap:** The random word swap technique shuffles the positions of words within the sentence. Words are swapped randomly while ensuring that the sentence's overall structure remains intact. This operation encourages the model to understand word order more flexibly.

Through the integration of these augmentation techniques, we aim to enrich our training data, thereby enhancing the model's performance and adaptability across diverse linguistic scenarios.

3.2 Model

In the context of our text regression task, the architecture and training process of the model are of paramount importance. The BERT Multilingual model serves as the foundation of our approach. It is pre-trained on 106 languages, including Arabic and English, making it a robust choice for the task at hand. The input or the model is 256 for skip-gram with negative-sampling (SGNS) embeddings which is based on word2vec models (Mikolov et al., 2013) trained with gensim (Řehřek and Sojka, 2010). The input or the model is 300 for Electra embeddings (Clark et al., 2020). The following steps elucidate the core components of our approach:

1. **Multihead Attention Head:** Our model incorporates a Multihead Attention. This component is pivotal for text regression tasks, and a cornerstone of the Transformer architecture. It empowers the model to concentrate on various segments of the input sequence, capturing intricate patterns and relationships. The output is 256 for SGNS, and 300 for Electra.

2. **Linear Layer:** fully connected layer that transforms the attention mechanism's output to the desired dimension. It is seamlessly integrated into the model, enabling it to predict continuous values of embeddings from the input text.

3. **Loss Function and Optimizer:**

   - *Loss Function Selection:* The mean squared error (MSE) loss function is employed. This function is a standard choice for regression tasks, quantifying the squared discrepancies between the model's predictions and the actual values.
   - *Optimizer Initialization:* The AdamW optimizer is utilized for optimizing the model's parameters. This optimizer is a variant of the conventional Adam optimizer tailored for deep learning models.
   - *Learning Rate (lr):* The learning rate, a pivotal hyperparameter, is set to 2e-5. It dictates the optimization step size and plays a crucial role in model convergence.

4. **Epochs:** The training encompasses multiple iterations, referred to as epochs, over the entire dataset. For this model, only **two** epochs are executed. Limiting the training to two epochs was done because the validation loss began to increase afterward, which is likely due to the small size of the dataset.

4 Results

To validate the effectiveness of our proposed modifications, we conducted experiments on the provided dataset for the shared task.

4.1 Dataset

The dataset comprises Arabic word definitions and their corresponding word embeddings. It also includes English definitions for the cross-lingual task. The data augmentation for the Arabic => Arabic only used deletion and swapping methods from data augmentation. We generated five different variations of each sentences that was longer than two words. The punctuation was removed. For English => Arabic we used Natural Language Toolkit (NLTK) (Bird et al., 2009) word synonyms to replace words randomly.

4.2 Experimental Setup

We fine-tuned the modified BERT Multilingual model on the training dataset and evaluated its performance on the test set. The evaluation dataset was only used for inference.

4.3 Results

Table 1: Reverse Dictionary Track (RD)

Dataset	Metric	SGNS	Electra
Benchmark	Cosine	35.61%	48.85%
Benchmark	MSE	35.61%	24.94%
Benchmark	Ranking	38.52%	31.28%
Dev	Cosine	49.45%	61.69%
Dev	MSE	3.48%	16.75%
Dev	Ranking	31.45%	24.97%
Test	Cosine	53.8%	62.5%
Test	MSE	3.1%	15.7%
Test	Ranking	29.1%	28.5%

Table 2: Cross-lingual Reverse Dictionary Track (CLRD)

Dataset	Metric	SGNS	Electra
Benchmark	Cosine	26.23%	54.09%
Benchmark	MSE	4.92%	22.11%
Benchmark	Ranking	50.17%	36.22%
Dev	Cosine	27.72%	58.06%
Dev	MSE	5.07%	19.55%
Dev	Ranking	45.77%	25.88%
Test	Cosine	27.0%	56.5%
Test	MSE	5.0%	20.6%
Test	Ranking	45.2%	28.1%

The tables illustrate the model's performance on Reverse Dictionary (RD) and Cross-lingual Reverse Dictionary (CLRD) tasks, comparing the Benchmark results with the Development (Dev) results generated by the Multilingual BERT with data augmentation.

In the RD track, the development and test datasets show a notable improvement in Cosine Similarity compared to the Benchmark dataset, indicating better vector space alignment. The MSE metric in the Dev dataset is significantly lower, suggesting a reduction in error rates. The Ranking metric also shows a decrease, which might indicate an improved model performance in ranking the dictionary entries correctly. The main difference between dev and test datasets for RD is that the Electra ranking was worse in test compared to dev unlike SGNS.

Similarly, in the CLRD track, the development and test datasets show an improvement in Cosine Similarity, indicative of better alignment in the vector space. The MSE is slightly higher in the Dev dataset, suggesting a slight increase in the error rate. The Ranking metric shows a decrease in implies a better performance in ranking tasks. Similar to the previous task the ranking was worse in Electra test unlike SGNS.

The variations in performance metrics between the Benchmark and Dev datasets could be attributed to the utilization of a Multilingual BERT model coupled with data augmentation techniques, which might have contributed to enhancing the model's generalization capabilities and performance in both RD and CLRD tasks.

5 Future research:

1. Augmenting our training dataset by introducing nuanced variations to the glosses, potentially employing paraphrasing techniques or deliberately infusing noise such as typos and word order alterations.

2. Adapting of multi-task learning; alongside our primary regression task, training the model to concurrently predict attributes like part of speech (POS) might bolster its gloss representation capabilities.

3. Integrating additional features, such as the gloss length or its associated POS.

6 Discussion and Conclusion

This paper presented a novel approach to the Arabic Reverse Dictionary Shared Task using a modified BERT Multilingual model. The introduced modifications augmentation and using a multi attention head, have shown promise in enhancing the model's performance, paving the way for future research in this domain. Our experiments demonstrate the potential of the BERT Multilingual model, even simple modifications such as data augmentation and using a multi attention head still provides good results but the improvements in SGNS embeddings is less impressive.

Limitations

While our proposed model demonstrates promise in the Arabic Reverse Dictionary task, there is still

room for major improvements mentioned in the discussion. The major limitation was the limited amount of training data.

Ethics Statement

We have ensured that our research adheres to the highest ethical standards. Our methodologies and data handling processes will be released as we are committed to transparency, fairness, and the responsible application of our findings in real-world scenarios.

Acknowledgements

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

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. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*.

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.

Arman Malekzadeh, Amin Gheibi, and Ali Mohades. 2021. Predict: persian reverse dictionary. *arXiv preprint arXiv:2105.00309*.

Timothee Mickus, Kees Van Deemter, Mathieu Constant, and Denis Paperno. 2022. Semeval-2022 task 1: CODWOE – comparing dictionaries and word embeddings. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1–14, Seattle, United States. Association for Computational Linguistics.

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

Radim Řehřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora.

Hang Yan, Xiaonan Li, and Xipeng Qiu. 2020. Bert for monolingual and cross-lingual reverse dictionary. *arXiv preprint arXiv:2009.14790*.

FM Zanzotto, Ioannis Korkontzelos, Francesca Fallucchi, Suresh Manandhar, et al. 2010. Estimating linear models for compositional distributional semantics. In *Proceedings of the 23rd international conference on computational linguistics (COLING)(GGS Conference Rating 2 A)*.

A Example Appendix

This is a section in the appendix.