Comparison of Cross-encoder and Bi-encoder Approaches for Arabic question answering task

Anonymous ACL submission

Abstract

With the recent advancement in Transformer networks and large language models, various encoder-based approaches have been proposed as solutions. When textual data for questions and answers are available, cross-encoder approaches encode them jointly, while bi-encoder approaches encode them separately. In this research, the performance of these approaches for question-answer pairs using an Arabic medical dataset is compared. Five variants of the Transformer model were utilized for this study. These models differ in design but share the objective of leveraging large amounts of text data to build a general language understanding model. Then, fine-tuned on an answer selection task and evaluated for performance using accuracy and execution time metrics. The results indicate that the AraBERT model with a cross-encoder architecture achieved the highest accuracy of 0.96.

1 Introduction

The Arabic language poses many challenges in Natural Language Processing (NLP), including the question-answering (QA) task. One of the most prominent recent NLP techniques applied to the QA task in Arabic is pre-trained transformer-based models, which can achieve state-of-the-art performance. These models are capable of learning universal representations of language that can be fine-tuned for specific tasks without the need to train each model from scratch (Ortiz-Barajas et al., 2022).

Cross-encoders (Reimers and Gurevych, 2019) are transformer-based models designed to capture the relationship between input pairs. Cross-encoders take two inputs, usually a pair of sentences or sequences, and encode them together into a shared representation. Cross encoders can effectively model interactions and dependencies between the input elements by jointly considering both inputs, enhancing performance on various downstream tasks (MS et al., 2024). Another approach is the bi-encoder model, which use separate encoders for each input sentence by a Siamese network. Each sentence is encoded independently, producing two separate representations. These representations are then compared using a similarity metric to determine the relationship between the sentences (Ortiz-Barajas et al., 2022). Both cross-encoders and bi-encoders have their advantages and are suitable for different scenarios. Cross-encoders excel at capturing the interaction between sentences, while bi-encoders are computationally efficient (Ortiz-Barajas et al., 2022). In this study, an empirical analysis of state-of-the-art models using these approaches for the task of question answering in the medical domain is conducted. The goal is to find and compare the approaches that best fit the Arabic QA task. The structure of the paper is as follows. In Section 2, the related work is described, focusing on previous approaches applied to the QA task. In Section 3, the proposed architecture and the experimental setup are explained. Finally, in Sections 4 and 5, the results and conclusions are presented, respectively.

2 Related Work

The Transformer network (Vaswani et al., 2017) is an architecture that encodes texts in parallel using attention mechanisms instead of the sequential mechanisms found in Recurrent Neural Networks. In the Transformer-based encoder models, there are two primary architectures: bi-encoders and cross-encoders. For QA tasks, some research has focused on developing models based on these architectures.

In addition, other models, like the one introduced by (Risch et al., 2021), have introduced a new evaluation metric for question-answering (QA) models called SAS (Semantic Answer Similarity). SAS is designed to evaluate the semantic similarity between model predictions and ground-truth answers rather than relying solely on lexical overlap. The SAS metric uses a cross-encoder architecture based on transformer models and shows a better correlation with human judgment compared to traditional lexical metrics. Moreover, a bi-encoder based on the Sentence-BERT model has been proposed by (Nie, 2022) to handle answer selection tasks. They fine-tuned a pre-trained Sentence Transformer model for the insurance QA task using the InsuranceQA Corpus, resulting in a significant improvement in accuracy from 0.26 to 0.48. Additionally, (Alfarizy and Mandala, 2022) proposed a bi-encoder based on the Sentence-BERT (SBERT) model for the verification of unanswerable questions in QA systems. They focused on reading comprehension by proposing a modification to ELECTRA, incorporating similarity parameters using SBERT, and then using cosine similarity for comparison. The similarity value is a decimal number between 0 and 1, with the greatest similarity value taken to represent the similarity of context with a declarative sentence. After obtaining the value of sentence similarity, they determined a limit value for labeling two sentences with the labels "similar" and "not similar". Values that are "similar" are considered "answerable" while those that are "not similar" are considered "unanswerable".

3 Methodology

This section presents the comparison of cross-encoder and bi-encoder approaches for the question-answer pairs task. To achieve this goal, five Transformer models were fine-tuned on the same training sets and evaluated with the same validation and test sets.

3.1 Dataset

The dataset used in this research, namely the Arabic Medical Community QA dataset (AM-CQA), was collected from the Altibbi platform. The Altibbi platform (a medical consultation platform)[1] provides reliable medical diagnoses by the best doctors in the Arab world. It has been developed to enhance doctor-patient consultations. The AM-CQA dataset was created from Arabic medical forums that contain a mix of informal and formal language and different Arabic dialects. This dataset consists of 107,268 women's healthcare question-answer pairs, three columns are used: question description, one correct answer, and one incorrect answer.

[1] https://altibbi.com/

3.1.1 Dataset Pre-processing

The following pre-processing steps are applied to the AM-CQA dataset.

- Remove diacritics using Pyarabic, an Arabic plugin tool for Python.
- Remove questions with attachment files.
- Remove HTTP links, special characters, English alphabet, English numbers, Arabic numbers, and extra spaces using regular expressions, a built-in Python package.
- Normalize text that replaces the letters “ أ", “ إ", “ ا", “ آ" and with “!".
- Replace English question marks “?" with Arabic question marks “ ؟" to unify.

3.2 Models Architecture

This section presents the approach to train bi-encoders and cross-encoders, based on semantic similarity in the Arabic medical domain. First, the data is prepared for training the models, and divided into three subsets: Train, Development, and Test. The training set, containing 85,812 QA pairs, is used to create data from the corpus for model fine-tuning. The development set, containing 10,728 QA pairs, is used to prepare evaluation data for assessing the accuracy of the fine-tuned model's QA task. The test set, also containing 10,728 QA pairs, is used for evaluating the QA system and obtaining the final performance of the system.

3.2.1 Cross-Encoders Model

The first model in the proposed architecture in the research is Cross-embeddings, which incorporates detailed question-answer interactions and is derived from the Cross-Encoders (Reimers and Gurevych, 2019). Illustrated in Figure 1, Cross-Encoders take input from both the question and answer sentences. To accurately capture the interaction between questions and answers, the matched correct answer and question is employed to guide the encoding of the question with correct answer. Then the output predicts a label for this question-answer pair, with 0 indicating a wrong answer and 1 indicating a correct answer.

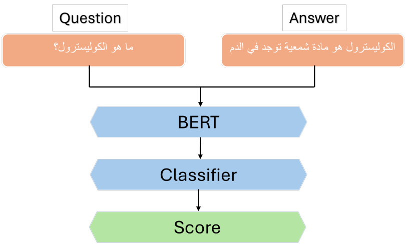

Figure 1: Cross-encoder Architecture of BERT Model

3.2.2 Bi-encoder Model

The second model in the proposed architecture in the research is the bi-encoder model, which is based on the Sentence-BERT model (Reimers and Gurevych, 2019). Sentence-BERT uses a modification of a Siamese neural network capable of obtaining individual vectors of fixed size from each text (Reimers and Gurevych, 2019). In a bi-encoder, as illustrated in Figure 2, both the input question and the answer are encoded into vectors. Then, a pooling operation is applied to the last hidden state of the BERT model to obtain a sentence vector for each question and answer. These sentence vectors are represented as u and v, respectively. Then, concatenate the sentence embedding u and v with their element-wise absolute difference $|u - v|$,this concatenated vector is multiplied by a trainable weight matrix $W_t \in R^{3n \times k}$, as shown in Eq. 1 (Reimers and Gurevych, 2019).

$$o = softmax(W_t[u, v, |u - v|]) \tag{1}$$

where u: is embedding of the first sentence, v: embedding of the second sentence. $|u-v|$: is element-wise absolute difference capturing how the embedding differ. And n represents the dimensional of the sentence embedding. The total dimensional of the concatenated vector is $3n$ and k denotes the number of labels, with k= 2:0 indicating a wrong answer and 1 indicating a correct answer. Where 3 represent the three embedding sentences u, v, and $|u - v|$. The model is trained by optimizing the cross-entropy loss. This loss was used in bi-encoder model to train the SBERT model on data. It adds a softmax classifier on top of the output of two transformer networks.

3.3 Fine-Tuning

Fine-tuning is a method of making precise adjustments to improve the performance and accuracy of a pre-trained network (Mustafid et al., 2020). In this research, five pre-trained models from the Hugging Face library are selected, the model's detail

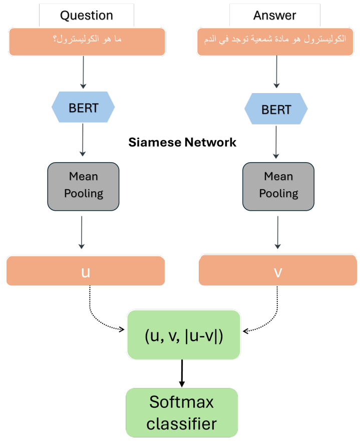

Figure 2: Bi-encoder Architecture of Sentence-BERT Model.

are illustrated in Table 1, for fine-tuning the QA dataset.

3.4 Evaluation Metrics

This research used popular metrics for evaluation, namely accuracy and running time for each model. Accuracy is widely used for measuring QA task performance (Shaheen and Ezzeldin, 2014). The running time measuring in second. Accuracy (Acc) is defined as the percentage of correctly answered questions over the total number of questions, as shown in Eq. 2. Let K be the number of correctly answered questions, and Q is the total number of questions (Shaheen and Ezzeldin, 2014).

$$Acc = (\frac{K}{Q}) \tag{2}$$

3.5 Configuration

The experiments have been conducted completely in Google Colaboratory Pro Plus. The virtual machine associated with the GPU in Colab Pro+ has 166.1 GB of disk space and provides up to 52 GB of RAM. All models are trained with a batch size of 8. The learning rate is set to 1e-5 using the Adam optimizer, with a linear learning rate warm-up over 10% of the training data. All models are trained for four epochs. The maximum sequence length is set to 128.

4 Results and Discussion

The experiment in this study assested state-of-the-art transformer models for the Arabic QA task. The focus was on identifying the best architectures that performed well on the AM-CQA corpus. Table 1 presents the performance of different pre-trained

No.	Model Name	bi-encoder		cross-encoder	
		Accuracy	Running Time	Accuracy	Running Time
1	disistiluse-base-multilingual-cased-v2	0.79	4840 s	0.83	12633 s
2	bert-base-arabertv2	0.86	**4523 s**	**0.96**	16409 s
3	paraphrase-TinyBERT-L6-v2	0.81	7387 s	0.85	5069 s
4	bert-base-arabic-camelbert-mix	0.85	7069 s	0.87	12023 s
5	stsb-roberta-base-v2	0.84	17055 s	0.88	6456 s

Table 1: Performance of different pre-trained Transformer models on the medical QA task.

Transformer models on a medical QA task, comparing both bi-encoder and cross-encoder architectures. Table 1 presents the evaluation results of various pre-trained Sentence Transformer models on the medical QA task. For both cross-encoder and bi-encoder architectures, five different models were evaluated. Each model corresponds to an independent run using different random seeds. All models were fine-tuned on the AM-CQA corpus specifically for the QA task. Models 1 through 5 use different models of BERT. We observed that bi-encoder models generally offer lower accuracy by 0.86 with AraBERT compared to cross-encoder models due to the lack of joint context between the question and answer sentences in bi-encoders. While cross-encoders are slower and more memory-intensive by 16409 seconds, they provide significantly higher accuracy by 0.96 with AraBERT.

5 Conclusion

In this research, a comparative analysis of cross-encoder and bi-encoder architectures for question-answering tasks using an Arabic medical dataset is presented. Five different the Transformer models are fine-tuned on a QA task and their performance is evaluated using accuracy and execution time metrics. The findings showed that the AraBERT model with a cross-encoder architecture achieved the highest accuracy of 0.96, indicating that cross-encoders are more effective for this specific task. However, they come at a higher computational cost.

Acknowledgments

The authors extend their appreciation to the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University for funding and supporting this work through Graduate Students Research Support Program and the Artificial Intelligence Center for Arabic Language Processing at King Salman Global Academy for partially funding the work.

References

Giffari Alfarizy and Rila Mandala. 2022. Verification of unanswerable questions in the question answering system using sentence-bert and cosine similarity. In *2022 9th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–6. IEEE.

Ankith MS, Arindam Bhattacharya, Ankit Gandhi, Vijay Huddar, Atul Saroop, and Rahul Bhagat. 2024. Diskco: Disentangling knowledge from cross-encoder to bi-encoder. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 346–354.

Ahmad Mustafid, Muhammad Murah Pamuji, and Siti Helmiyah. 2020. A comparative study of transfer learning and fine-tuning method on deep learning models for wayang dataset classification. *IJID (International Journal on Informatics for Development)*, 9(2):100–110.

Ercong Nie. 2022. Fine-tuned sentence transformer model for question answering task.

Jesus-German Ortiz-Barajas, Gemma Bel-Enguix, and Helena Gómez-Adorno. 2022. Sentence-crobi: A simple cross-bi-encoder-based neural network architecture for paraphrase identification. *Mathematics*, 10(19):3578.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Julian Risch, Timo Möller, Julian Gutsch, and Malte Pietsch. 2021. Semantic answer similarity for evaluating question answering models. *arXiv preprint arXiv:2108.06130*.

Mohamed Shaheen and Ahmed Magdy Ezzeldin. 2014. Arabic question answering: systems, resources, tools, and future trends. *Arabian Journal for Science and Engineering*, 39:4541–4564.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.