Quranic Verses Semantic Relatedness Using AraBERT

Abdullah N. Alsaleh
University of Leeds
School of Computing
scanaa@leeds.ac.uk

Eric Atwell
University of Leeds
School of Computing
e.s.atwell@leeds.ac.uk

Abdulrahman Altahhan
University of Leeds
School of Computing
a.altahhan@leeds.ac.uk

Abstract

Bidirectional Encoder Representations from Transformers (BERT) has gained popularity in recent years producing state-of-the-art performances across Natural Language Processing tasks. In this paper, we used AraBERT language model to binary classify pairs of verses provided by the QurSim dataset to either be semantically related or not. We have pre-processed The QurSim dataset and formed three datasets for comparisons. Also, we have used both versions of AraBERT, which are AraBERTv0.2 and AraBERTv2, to recognise which version performs the best with the given datasets. The best results was AraBERTv0.2 with 92% accuracy score using a dataset comprised of label '2' and label '-1', the latter was generated outside of QurSim dataset.

1 Introduction

In recent years, Natural Language Processing (NLP) has been evolved with the introduction of Transformer architecture by Vaswani et al. (2017). BERT, built on the transformer layer, has showed and produced state-of-the-art accuracy in a number of NLP tasks such as machine translation and text classification (Devlin et al., 2019). There are two stages of BERT: pre-training and fine-tuning. Pre-training is used for masked language modeling and next sentence prediction. For fine-tuning, is to add one or more layers designed for specific task on top of the final encoder layer (Rogers et al., 2020). Google provides pre-trained models for English and other languages including Arabic. Several studies provide their own language model based on BERT to perform better on specific tasks. AraBERT is recently published that contributes to Arabic language model that is pre-trained to suit a wide range of Arabic NLP related tasks. In this paper, we have used AraBERT language model to classify QurSim dataset in semantic relatedness task. Section 2 will outline the related work in semantic similarity and relatedness while section 3 will discuss AraBERT. Section 4 will discuss QurSim dataset and filtration process. Finally, section 5 will showcase the results of using QurSim datasets using AraBERT.

2 Related Work

2.1 Semantic Similarity/Relatedness in Arabic and Quranic Text

Several studies that involved Arabic and Quranic text using different methods to extract semantic similarity or relatedness. Mohamed et al. (2015) built a system Al-Bayan for evaluating semantic interpreter between Arabic questions and answers. The system used Morphological Analysis and Disambiguation for Arabic (Habash et al., 2009) as preprocessing tool and Decision Tree Classifier to predict the label. The proposed system achieved 74.5% accuracy score. Another study by Al-Bataineh et al. (2019) presented a system to identify similar Arabic questions in Quora using ELMo (Peters et al., 2018), Word2vec (Mikolov et al., 2013a) (Mikolov et al., 2013b) and Sent2vec (Pagliardini et al., 2018). The dataset used for training word and sentence embeddings includes Mawdoo3 question-to-question (Q2Q) (Seelawi et al., 2019b) and Madar Dialect Q2Q (Bouamor et al., 2018) datasets. The study found that ELMo performed better with 93% in Modern Standard Arabic and 82% in Arabic Dialects compared to other models such as Sent2vec and Word2vec.

2.2 Semantic Similarity using BERT

There are studies that use BERT models for identifying semantic similarity between text. A team competing at NSURL-2019 used BERT model with pre-trained multilingual to detect similar Arabic questions (Al-Theiabat and Al-Sadi, 2019). The dataset used for this competition mostly from Mawdoo3 (Seelawi et al., 2019a). The results showcased BERT model outperform other models with F1-Score of 95.92%. Another study by Peinelt et al.

Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 185–190
Kyiv, Ukraine (Virtual), April 19, 2021.

(2020) combines topic model and BERT (tBERT) to enhance semantic similarity detection and prediction between pairs of English sentences. Reimers and Gurevych (2019) presented Sentence-BERT (SBERT) that uses siamese and triplets network structure with modified pre-trained BERT to derive semantically meaningful text embeddings that can be compared using cosine-similarity. However, to the best of our knowledge there is no research on semantic similarity or relatedness using BERT or AraBERT on Holy Quran text.

3 AraBERT

AraBERT is an Arabic language model in which BERT was trained on a large Arabic corpus (Antoun et al., 2020). The dataset includes Arabic Wikipedia, 1.5 billion words from Arabic corpora (El-Khair, 2016) and the Open Source International Arabic News Corpus (Zeroual et al., 2019). The corpus covers news articles from several Arab news media with different topics and from different Arab countries. The size of the pre-training dataset is 70 million sentences that amounts to approximately 24GB of text (Antoun et al., 2020).

AraBERT has produced better results on various Arabic NLP tasks. In sentiment analysis, AraBERT performed better than mBERT, which is a multilingual BERT model developed by Google, on most tested datasets (Antoun et al., 2020). Also, AraBERT outperformed mBERT and TF-IDF on the new Twitter-based benchmark dataset for Arabic Sentiment Analysis (Alharbi et al., 2020). In Named Entity Recognition (NER), AraBERTv01 had better results over Bi-LSTM-CRF model with macro-F1 score of 84.2, in which AraBERT new state of the art for NER on ANERcorp (Antoun et al., 2020). AraBERT also outperformed other Arabic NER tools such MADAMIRA and FARASA in NER tasks using AQMAR and NEWS datasets (Helwe et al., 2020).

4 QurSim

QurSim is a work of Arabic text that pertains to the Holy Quran (Sharaf and Atwell, 2012). The dataset showcases 7679 pairs of verses that are similar or related verses according to comments of Ibn Kathir's Tafsir, which is highly respected for its interpretation of the Holy Quran. It also improves its dataset using lexical similarity approach such as Term Frequency-Inverse Document Frequency (TF-IDF). QurSim dataset classifies pairs of Quranic verses that are related into three classes.

Label '2' indicates that the two verses are strongly related and they share similar lexicon. Label '1' means that the two verses are related based on the main topics that were mentioned in these verses but they share less similar lexicon. Finally, label '0' is being classified as not obvious relation between verses as, for example, it draws analogies that are different from each other but they serve the same purpose. Label '2', '1' and '0' comprises 40.09%, 48.41% and 11.49% respectively of the total QurSim dataset.

4.1 Mapping QurSim dataset to Quranic Verses Text

The QurSim data contains numeric values that entail the location of chapters and verses. However, for this research, BERT needs text as inputs in order to perform its models including BERTforSequenceClassification. Therefore, We needed to map Quranic numeric verses to texts. A Quran dataset by Aloufi (2019) has been used to map the numeric verses to their texts. Then, the verses were manually checked to see if they are in correct order.

4.2 Dataset Filtration

4.2.1 Duplicated Pairs Elimination

We found there were duplicated pairs of verses in QurSim datasets with the total of 764 records of duplicated pairs of verses. When we examined the duplicates, there were 592 records of duplicated pairs were labelled the same but ordered differently. For example, chapter 2 verse 2 is paired with chapter 2 verse 3 and the pair is labelled '2' and vice versa, shown in Table 1. For this case, we opted to remove the duplicates since they were redundant considering that the relationship is naturally bidirectional.

ID	SS	SV	TS	TV	Label
91	2	2	2	3	1
98	2	3	2	2	1

Table 1: Duplicated pairs of verses with same label, where SS stands for source Soura (chapter) , SV is the source verse, TS is target Soura and TV is target verse

The other 172 records of duplicated pairs of verses were labelled differently. For example, chapter 1 verse 5 paired with chapter 73 verse 9 were labelled '2'; however, they were labelled '1' when ordered differently, shown in Table 2. Since the label assignments were mainly influenced by Ibn Kathir's Tafsir, it is difficult for the authors to interpret Ibn Kathirs's comments in order to assign which appropriate label to pair of verses. So, it was important for this research to remove all pairs of verses that were labelled differently to ensure the dataset.

ID	SS	SV	TS	TV	Label
66	1	5	73	9	2
7339	73	9	1	5	1

Table 2: Duplicated pairs of verses with different label, where SS stands for source Soura (chapter) , SV is the source verse, TS is target Soura and TV is target verse

4.2.2 Label '1' Pairs Removal

The purpose of the paper is to binary classify pairs of Quranic verses for semantic relatedness using AraBERT. Label '1' pairs are related, however, the degree of similarity/relatedness are weaker than label '2'. Pairs of verses that share the label '1' have fewer words and concepts in common, which could affect the results based on AraBERT limitation on classical Arabic.

> والوزن يومئذ الحق فمن ثقلت موازينه فأولئك هم المفلحون
>
> And the weighing [of deeds] that Day will be the truth. So those whose scales are heavy - it is they who will be successful. [7:8]
>
> فإذا نفخ في الصور فلا أنساب بينهم يومئذ ولا يتساءلون
>
> So when the Horn is blown, no relationship will there be among them that Day, nor will they ask about one another [23:101]

In this example, the two verses are labelled '1' and they are describing events that happen on the judgment day. Both verses are related in which they share the topic of judgement day; however, they are mentioning different events that happen on the judgement day. Both verses also do not share any Arabic lexical item except for يومئذ 'That Day' which references to the judgement day.

Also, this is a preliminary research to use the AraBERT model for semantic similarity and the authors chose the extreme similarity label, which is label '2', for training and testing. Therefore, pairs of verses that have been assigned label '1' were removed from the dataset. In future studies, the authors may expand to include and train label '1' for multi-classification research.

4.2.3 Dataset Balancing

After deducing duplicated pairs of verses and label '1', we found that the dataset was imbalanced between label '2' and '0'. Label '2' has 2548 records while label '0' has 857 records. This was an issue because the dataset is imbalanced and could produce poorly results as we will mention it in the results section. We also needed pairs of verses that are not related to train and test against label '2'. Therefore, we randomly generated 2548 pairs of verses from the Holy Quran that are not in the QurSim dataset and labelled them as '-1'. Since the QurSim dataset is according to Ibn Kathir's Tafsir, we assume that these randomly generated pairs of verses are not related according to Ibn Kathir. Therefore, it is fair to say that we are building a model to test the relatedness of pairs of verses that is only based on Ibn Kathir's opinions and interpretations.

5 Results and Discussions

5.1 Results

Dataset	Testing		Training	
	v02	v2	v02	v2
Label '2' Label '0'	90.2%	88.7%	88%	87%
Label '2' Label '-1'	**92.1**%	88.6%	88%	87%
Label '2' Label '0' & '-1'	87.9%	88.6%	87%	85%

Table 3: Results of both versions of AraBERT with three datasets.

We have experimented with the latest two different versions of AraBERT, which are AraBERTv0.2 and AraBERTv2 (Antoun et al., 2020). AraBERTv2 version uses Farasa to segment the words into stems, prefix and suffix. AraBERTv0.2 does not require any Farasa segmentation. We could not use the LARGE version of the models as we only used the BASE version due to hardware limitations. The BASE model of both AraBERT versions has 136M parameters and 200 million sentences with a size of 77G of text. Furthermore, there are three groups of experiments based on dataset parameters which we will entail in this section. Summary of the experiments are shown in Table 3.

The experiments were performed using Google Colab. AraBERT along with Transformers library were installed while performing the experiments. The model was fine tuned for semantic similarity task. All of the experiments have batch size of 32, learning rate is 2e-5 and 8 epochs.

The first dataset has label '2' and label '0' of QurSim dataset after the filtration process. Label '2' has 2548 records while label '0' has 857 records. The dataset was split into training set, validation set and testing set. The training set has 2052 records, the validation set has 229 records and the testing set has 1124 records. AraBERTv0.2 scored better accuracy score metric than AraBERTv2. Although the dataset is small and imbalanced, both AraBERT versions achieved good accuracy score, which is shown in Table 3.

The second dataset has label '2' from QurSim dataset and label '-1' for pairs that were generated randomly. Both labels have the same number of records, which are 2548 pairs of verses. The dataset was split into training set, validation set and testing set. The training set has 3072 records, the validation set has 342 records and the testing set has 1682 records. We found that AraBERTv0.2 performed better than AraBERTv2 with the accuracy score of 92%. AraBERTv0.2 scored the best accuracy score compared to the other datasets. This is due to generated pairs of verses labelled '-1' being not related.

The third and final dataset comprises of label '2' and label '0' augmented with '-1'. They also have the same number of records of 2548 pairs of verses. The reason to combine label '0' and label '-1' for the experiment is to test the BERT model if it would identify the label '0' and '-1' as one class. The dataset was again split into training set (3072 records), validation set (342 records) and testing set (1682 records). The results turned out to be worse than previous two for AraBERTv0.2 while AraBERTv2 maintaining consistency with its accuracy score, which similar to the previous two experiments.

ID	SS	SV	TS	TV	Label
5648	32	27	80	24	1
5649	32	27	80	25	1
5650	32	27	80	26	1
5651	32	27	80	27	1
5652	32	27	80	28	1
5653	32	27	80	29	1
5654	32	27	80	30	1
5655	32	27	80	31	1
5656	32	27	80	32	2

Table 4: Series of verses are related to a single verse

5.2 Discussion

Although the results were promising, we have looked at the output files of these experiments and we found common problems that we will examine in this section.

5.2.1 Dataset

In regards to the dataset, we found there are a series of verses that are discussing and describing a particular topic and are related to one verse. However, in the QurSim dataset, those series of verses are paired with a single verse, an example shown in Table 4.

> متاعا لكم ولأنعامكم
>
> [As] enjoyment [i.e., provision] for you, and for your livestock. [80:32]
>
> أولم يروا أنا نسوق الماء الى الأرض الجرز فنخرج به زرعا تأكل منه أنعامهم وأنفسهم أفلا يبصرون
>
> Do they not see how We conduct the water to a dry land, and with it We produce vegetation, from which their livestock eat, and themselves? Do they not see? [32:27]

In this example, both verses [80:32] and [32:27] are related as they are mentioning the livelihood for people and the cattle. In verse [32:27], Allah mentions His kindness by providing water to dry land, and herewith bring forth crops and vegetation for people and their cattle to eat. The verse [80:32] is the last verse of a series of verses describing the same meaning of verse [32:27]; however, they were broken into several small verses. Therefore, the model failed to predict that both verses are related.

5.2.2 Lexical Synonyms

We also found that in QurSim dataset, a very few pairs of verses that are strongly related have less similar words; however, they use different words to achieve the same meaning. Therefore, AraBERT did not predict a few of those verses to be related. Also, there are phrases in classical Arabic that are pertain to Islamic teachings that the model did not predict correctly between related verses.

> ثم جعلناه نطفة في قرار مكين
>
> Then We placed him as a sperm-drop [or "as a zygote"] in a firm lodging [i.e., the womb] [23:13]
>
> ألم نخلقكم من ماء مهين
>
> Did We not create you from a liquid disdained [or "insignificant fluid"]? [77:20]

For this example, both verses [23:13] and [77:20] are related as they mentioned sperm as part of human creation. In verse [23:13], Allah draws the analogy of the sperm or zygote as an insignificant fluid or a disdained liquid to serve the purpose of its weakness in comparison to the power of the Creator. The model did not understand the analogy and the context of these verses which is why the model failed to correctly predict to be related.

> قم اليل إلا قليلا
>
> Arise [to pray] the night, except for a little [73:2]
>
> تتجافى جنوبهم عن المضاجع يدعون ربهم خوفا وطمعا ومما رزقناهم ينفقون
>
> Their sides shun their beds, as they pray to their Lord, out of reverence and hope; and from Our provisions to them, they spend. [32:16]

As for this example, both verses [73:2] and [32:16] are related as they mention praying during the night; however, the model did not predict they were related. The verse [73:2] mentions "Arise the night" as staying up at night, which is correct literal translation; however, the phrase in the classical Arabic religious context entails that a person staying up the night to pray. Therefore, the model did not predict both verses to be related.

6 Conclusion

The paper presented experiments on the QurSim dataset using fine-tuned AraBERT model to classify pairs of verses either to be semantically related or not. The paper applied data filtration to the QurSim dataset to avoid redundancy and generate unrelated pairs of verses. Also, the paper used both versions of AraBERT and the experiments suggested that AraBERTv0.2 has better results than AraBERTv2 across all three datasets. The best performance was achieved by AraBERTv0.2 with 92% accuracy score. However, by examining the results, AraBERT could not identify some of the classical Arabic lexical synonyms and religious context which could be a result of classical Arabic limitation in the AraBERT corpus. Finally, this study has a lot of potential for improvement on both the datasets and fine-tuning the AraBERT model.

References

Hesham Al-Bataineh, Wael Farhan, Ahmad Mustafa, Haitham Seelawi, and Hussein T Al-Natsheh. 2019. Deep contextualized pairwise semantic similarity for arabic language questions. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1586–1591. IEEE.

Hana Al-Theiabat and Aisha Al-Sadi. 2019. The inception team at NSURL-2019 task 8: Semantic question similarity in Arabic. In *Proceedings of The First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019 - Short Papers*, pages 112–117, Trento, Italy. Association for Computational Linguistics.

Basma Alharbi, Hind Alamro, Manal Alshehri, Zuhair Khayyat, Manal Kalkatawi, Inji Ibrahim Jaber, and Xiangliang Zhang. 2020. Asad: A twitter-based benchmark arabic sentiment analysis dataset.

Khalid Aloufi. 2019. Quran dataset.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

I. A. El-Khair. 2016. 1.5 billion words arabic corpus. *ArXiv*, abs/1611.04033.

Nizar Habash, Owen Rambow, and Ryan Roth. 2009. Mada+tokan: A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, Cairo, Egypt. The MEDAR Consortium.

Chadi Helwe, Ghassan Dib, Mohsen Shamas, and Shady Elbassuoni. 2020. A semi-supervised BERT approach for Arabic named entity recognition. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 49–57, Barcelona, Spain (Online). Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality.

Reham Mohamed, Maha Ragab, Heba Abdelnasser, Nagwa M. El-Makky, and Marwan Torki. 2015. Al-bayan: A knowledge-based system for Arabic answer selection. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 226–230, Denver, Colorado. Association for Computational Linguistics.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Nicole Peinelt, Dong Nguyen, and Maria Liakata. 2020. tBERT: Topic models and BERT joining forces for semantic similarity detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7047–7055, Online. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works.

Haitham Seelawi, Ahmad Mustafa, Hesham Al-Bataineh, Wael Farhan, and Hussein T. Al-Natsheh. 2019a. Nsurl-2019 shared task 8: Semantic question similarity in arabic.

Haitham Seelawi, Ahmad Mustafa, Hesham Al-Bataineh, Wael Farhan, and Hussein T. Al-Natsheh. 2019b. NSURL-2019 task 8: Semantic question similarity in Arabic. In *Proceedings of The First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019 - Short Papers*, pages 1–8, Trento, Italy. Association for Computational Linguistics.

Abdul-Baquee Sharaf and Eric Atwell. 2012. Qursim: A corpus for evaluation of relatedness in short texts.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

I. Zeroual, Dirk Goldhahn, T. Eckart, and A. Lakhouaja. 2019. Osian: Open source international arabic news corpus - preparation and integration into the clarin-infrastructure. In *WANLP@ACL 2019*.