SiameseBERT: A Bert-Based Siamese Network Enhanced with a Soft Attention Mechanism for Arabic Semantic Textual Similarity

Rakia Saidi[1] [a], Fethi Jarray[1,2] [b] and Mohammed Alsuhaibani[3] [c]

[1] *LIMTIC Laboratory, UTM University, Tunis, Tunisia*

[2] *Higher Institute of Computer Science of Medenine, Gabes University, Medenine, Tunisia*

[3] *Department of Computer Science, College of Computer, Qassim University, Buraydah, Saudi Arabia*

Keywords: Semantic Textual Similarity, Siamese Networks, BERT, Soft Attention, Arabic BERT.

Abstract: The assessment of semantic textual similarity (STS) is a challenging task in natural language processing. It is crucial for many applications, including question answering, plagiarism detection, machine translation, information retrieval, and word sense disambiguation. The STS task evaluates the similarity of data pairs of text. For high high-resource languages (e.g. English), several approaches for STS have been proposed. In this paper, we are interested in measuring the semantic similarity of texts for Arabic, a low-resource language. A standard approach for STS is based on vector embedding of the input text and application of similarity metric on space embedding. In this contribution, we propose a BERT-based Siamese Network (SiameseBERT) and investigate the most available Arabic BERT models to embed the input sentences. We validate our approach via Arabic STS datasets. The araBERT-based Siamese Network model achieves a Pearson correlation of 0.925. The results obtained demonstrate the superiority of integrating the BERT embedding, the attention mechanism, and the Siamese neural network for the semantic textual similarity task.

1 INTRODUCTION

Semantic textual similarity (STS) seeks to quantify the degree of similarity between two input texts or sentences. STS is essentially modelled as a multi-classification problem, where each class corresponds to a similarity score or label. A variety of traditional techniques have been proposed to train classifiers on manually annotated corpora, but they have not shown remarkable results because they tend to discard the internal structure of sentences. Deep neural networks (DNN) outperform traditional methods for many natural language processing(NLP) tasks, including questions answering (QA) (M. Hammad and Al-Zboon, 2021; H. Al-Bataineh and Al-Natsheh, 2019), word sense disambiguation (WSD)(Saidi et al., 2022) and semantic similarity of sentences. In this work, we investigate the application of DNN to Arabic STS and, more precisely, deep contextualized models such as BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018).

Siamese networks are one of the successful DNN architectures deployed when resemblance between objects is the goal (Neculoiu et al., 2016; T. Ranasinghe and Mitkov, 2019; Feifei et al., 2020). In this study, we will explore the effectiveness of integrating Siamese architecture and BERT embedding for Arabic STS.

We investigated the BERT embedding model because it shows brave results for several Arabic NLP tasks such as part of speech tagging (POS tagging) (Saidi et al., 2021), WSD (Saidi and Jarray, 2022) and sentiment analysis(Chouikhi. et al., 2021; Chouikhi et al., 2021).

We summarize the core contributions of the paper as follows:

- We propose a hybrid approach which integrates Siamese networks BERT embedding to solve Arabic STS.
- We discretize the existing Arabic STS datasets.
- We achieve state-of-the-art performance on three discretized datasets.

The rest of the paper is organized as follows: Section 2 presents the state of the art. Our approach is

[a] https://orcid.org/0000-0003-0798-4834

[b] https://orcid.org/0000-0003-2007-2645

[c] https://orcid.org/0000-0001-6567-6413

Saidi, R., Jarray, F. and Alsuhaibani, M.
SiameseBERT: A Bert-Based Siamese Network Enhanced with a Soft Attention Mechanism for Arabic Semantic Textual Similarity.
DOI: 10.5220/0011624800003393
In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART 2023) - Volume 3*, pages 146-151
ISBN: 978-989-758-623-1; ISSN: 2184-433X

explained in Section 3. The experimental setup is presented in Section 4. Results and discussion are presented in Section 5. We conclude this paper with a summary of our contributions and discuss future extensions.

2 RELATED WORK

Measuring the degree of similarity between two documents plays an important role in NLP. Therefore, similarity-predicting methods have been studied intensely and can be divided into two categories: namely, classical machine learning and DNN methods.

Alzahrani (Alzahrani, 2016) addressed the problem of semantic similarity by comparing the semantics of brief sentences in Arabic and English. Dictionary and machine translation methods were used to assess the similarity of cross-lingual writings from a monolingual perspective. The averaged maximum-translation similarity methodology achieved a correlation score of 0.7206. Li et al. (Li et al., 2006) used machine translation to assess the similarity of the phrase vector. The results show that the machine translation-based similarity technique achieved a correlation of 0.86.

Moreover, Alameer and Malallah (Alian and Awajan, 2017) used a hybrid similarity measure strategy that used the N-gram, cosine and Dice similarity measure. Alian and Awajan (Alian and Awajan, 2021) proposed a support vector machine to measure sentence similarity using lexical, semantic, and syntactic-semantic variables as hand-crafted features. The effectiveness of this method is tested using three datasets from SemEval 2017 (Arabic paraphrasing benchmark, MSRvid, and SMTeuroparl). This technique yields a correlation of 0.354 when applied to the Arabic paraphrasing benchmark.

Nagoudi and Schwab (Nagoudi and Schwab, 2017) used a recurrent neural network based on word alignment and several weighting techniques of word embeddings to extract sentence representations. Alsaleh et al. (Alsaleh et al., 2021) employed contextualized embedding mBERT and AraBERTv2 to classify pairs of verses from the QurSim dataset as semantically related or unrelated. They achieved an accuracy of 92% with AraBERTv0.2.

3 PROPOSED APPROACH: SIAMESEBERT

Given a sentence pair (a,b), the task is to predict the similarity score between sentences a and b. In this contribution, we propose SiameseBERT, a Siamese-based network that uses contextual embeddings from BERT as its backbone enhanced with a soft-attention mechanism.

Our approach consists in integrating contextual embeddings from BERT into a Siamese network. Fig.1 depicts the basic architecture of our approach. It broadly includes four blocks: (1) Word Embedding, (2) Interaction Layer, (3) Aggregation Layer, and (4) Prediction Layer. Let's explain each block.

1. **Word Embedding:** it consists of encoding each word using the BERT pre-training language model. Thus, each input word is represented as a 768-dimentional vector. Mathematically, let w_i^a represents word i of sentence a and w_j^b represents word j of sentence b. The embedding of each word is as follows: $x_i^a = BERT(w_i^a)$ represents the encoding of the word i of sentence a and $x_j^b = BERT(w_j^b)$ where $BERT$ refers to the output of the embedding layer.
2. **Interaction Layer:** the attention mechanism measures the alignment between every pair of words across both sentences and enables information to be shared between them. Mathematically, let e_{ij} represent the alignment model scores between w_i^a and w_j^b. In this paper, we adopt a Dot-Product attention mechanism where the alignment score function is calculated as $e_{ij} = x_i^{aT} x_j^b$. Let $\tilde{x_i^a}$ be the context vector of w_i^a. It is a weighted sum of the word embedding and normalized alignment scores. Therefore $\tilde{x_i^a} = \sum_{j=1}^{l_b} \frac{e^{e_{ij}}}{\sum_{k=1}^{l_b} e^{e_{ik}}} x_j^b$. Similarly $\tilde{x_j^b} = \sum_{i=1}^{l_a} \frac{e^{e_{ij}}}{\sum_{k=1}^{l_a} e^{e_{kj}}} x_i^a$.
3. **Aggregation Layer:** First, we apply a simple maximum-pooling of words in the same sentence to obtain the encoding of the sentence. Second, we concatenate the sentence embeddings to get the sentence pair embedding. Mathematically, let $pool^a$ be the element-wise max-pooling of sentence a over x_i^a and let $\widetilde{pool^a}$ be the element-wise max-pooling of context embedding of sentence a over $\tilde{x_i^a}$. We have the following formulas: $pool^a = \max_{1 \leq i \leq l_a} x_i^a$ and $\widetilde{pool^a} = \max_{1 \leq i \leq l_a} \tilde{x_i^a}$. Similarly, we obtain $pool^b = \max_{1 \leq j \leq l_b} x_j^b$ and $\widetilde{pool^b} = \max_{1 \leq j \leq l_b} \tilde{x_j^b}$ for sentence b. Finally, let X_{ab} be the encoding of the pair of sentences (a,b).

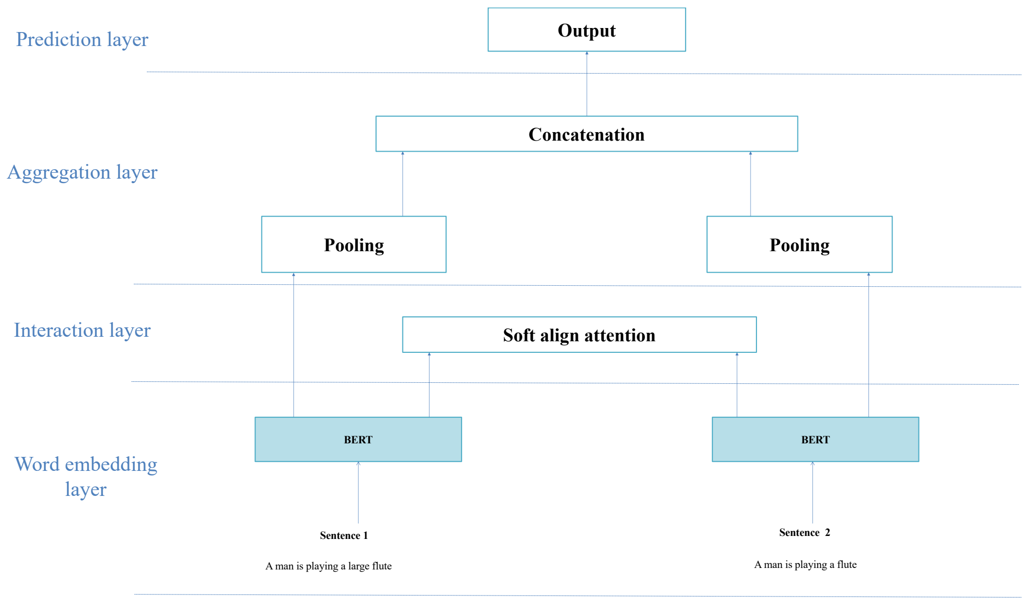

Figure 1: SiamBERT architecture for Arabic STS. The left side of the figure shows the boundaries of the four blocks.

It is defined as the concatenation of four vectors $pool^a, \widetilde{pool^a}, pool^b$ and $\widetilde{pool^b}$. Mathematically $X_{ab} = [pool^a, \widetilde{pool^a}, pool^b, \widetilde{pool^b}]$

4. **Prediction Layer:** We fed the sentence pair embedding into a fully connected network. The entire model is end-to-end trained with softmax as the activation function and sparse categorical cross entropy as the loss function. Mathematically, we have $\widehat{Y_{ab}} = F(X_{ab})$ where and F is the prediction layer.

4 EXPERIMENTS

4.1 Hyperparameters Setting

In this work, we investigate four available pre-trained Arabic BERT models: AraBERT (Antoun et al., 2020), Arabic-BERT (Safaya et al., 2020) and CAMeL-BERT[1] and the multilingual mBERT (Libovickỳ et al., 2019), which can also handle Arabic texts. Table 1 displays the characteristics of each model. The following hyperparameters are used to fine-tune all the models. 12 Transformer blocks, 768 hidden layer blocks, and 12 self-attention heads make up the entire set. We use Adam (Kingma and Ba, 2014) as the optimizer, with a sequence length of 128, a batch size of 6, a learning rate of 10^{-5}, and a dropout probability of 0.2. We kept the best model so far and adjusted for 50 epochs.

4.2 Evaluation Metrics

We evaluated the predicted similarities using two metrics: Pearson correlation and MSE.

- Between two sets of data, the Pearson correlation coefficient measures the linear correlation. It effectively measures covariance in a normalized manner, with the result always lying between 0 and 1. It is determined by dividing the product of the standard deviations of two variables by their covariance. Similar to covariance itself, the measure can only take into consideration linear correlations between variables and excludes many other forms of linkages or correlations.
- The mean squared error (MSE) measures the average of the squares of the errors between the predicted similarities and the human-annotated similarities.

4.3 Datasets Annotation

In this work, we investigated the available Arabic Semantic Textual similarity corpora[2]. This training dataset is released for the SEMEVAL 2017 Multilingual Semantic Textual Similarity: Arabic subtask (Track 1). It contains three subcorpora: Microsoft

[1] https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-ca. Last accessed 02 June 2022

[2] https://alt.qcri.org/semeval2017/task1/index.php?id=data-and-tools. Last accessed August 27, 2022

Table 1: Characteristics of utilised Arabic BERT models.

Model	Ara-BERT	Arabic-BERT	CAMeL-BERT	mBERT
Parameters	135M	110M	108M	110M
Normalization	yes	no	yes	yes
Textual Data	27GB	95GB	167GB	61GB
Tokenization	wordpiece	wordpiece	wordpiece	character

Research Paraphrase Corpus(MSR-Paraphrase)[3], Microsoft Research Video Description Corpus (MSR-Video)[4] and WMT2008 development dataset (SM-Teuroparl)[5]. The evaluation corpus for Arabic STS is used For test and validation. It contains 250 pairs(Arabic-Arabic).

Table 2: Statistical information for datasets of Arabic STS.

Data	#Pairs	#Sentences
MSR-Paraphrase	510	1020
MSR-Video	368	736
SMTeuroparl	203M	406

Two sentences are sent to the system, which is asked to return a discrete value similarity score on a scale of [0, 5], with 0 denoting complete semantic independence and 5 denoting semantic equivalence. Performance is assessed using the Pearson correlation between semantic similarity scores computed automatically.

As preprocessing steps, we convert the fine-grained and the continuous labels into discrete labels 0, 1,2,3, 4 or 5. We used the decision tree (DecisionTreeDiscretiser) [6] method to discretize the similarity scores.

5 RESULTS AND DISCUSSION

Our main experimental results using the Pearson correlation metric are shown in Table 3. We compared our results with the most famous baselines based on the STS dataset.

First, we emphasize the introduction of an attention mechanism improves the performance of SiameseBERT over the “no-attention” baseline, which justifies the insertion of such a mechanism. Second, we note that even without the attention mechanism, the proposed SiameseBERT approaches outperform the existing approaches, which proves the power of integration BERT and Siamese networks. We achieved the best performance with the highest average Pearson’s correlation coefficient on the major of experiments. The higher Pearson’s correlation obtained is **0.925** on the MSR-Paraphrase dataset. Third, it is worth mentioning that the best BERT model varies from one dataset to another. This may be due to different pre-training corpora used for building such an Arabic BERT model. Moreover, the Multilingual BERT (mBERT) didn’t achieve the best result in any dataset. Table 4 displays the results of the four Arabic BERT models on MSE metric.

The MSE metric consolidates the evaluation done by Person correlation. In fact, both metrics have the same optimal model.

6 CONCLUSION

This paper examined the incorporation of Arabic BERT models into Siamese neural networks to determine the semantic similarity between two sentences. For each data set, we determined the most performant available BERT version. SiameseBERT shows the strength of combining BERT, the attention mechanism, and Siamese for the STS task. Finally, we think that the results can be improved by improving the attention mechanism phase.

REFERENCES

Alian, M. and Awajan, A. (2017). A. q. abd alameer and s. malallah kadhem. *Iraqi Journal of Science*, pages 152–162.

Alian, M. and Awajan, A. (2021). Arabic sentence similarity based on similarity features and machine learning. In *Soft Computing*, volume 25, pages 10089–10101. Springer.

Alsaleh, A. N., Atwell, E., and Altahhan, A. (2021).

[3] http://research.microsoft.com/en-us/downloads/607d14d9-20cd-47e3-85bc-a2f65cd28042/. Last accessed August 27, 2022

[4] http://research.microsoft.com/en-us/downloads/38cf15fd-b8df-477e-a4e4-a4680caa75af/. Last accessed August 27, 2022

[5] http://www.statmt.org/wmt08/shared-evaluation-task.html. Last accessed August 30, 2022

[6] Tree discretization method, https://towardsdatascience.com/an-introduction-to-discretization-in-data-science-55ef8c9775a2. Last accessed 05 July 2022

Table 3: Pearson correlation coefficient of Arabic STS systems on the STS dataset (the higher, the better). Para stands for MSR-Paraphrase, Vid stands for MSR-Video, euro stands for SMT europarl datasets, n/a stands not applied.

	With attention			Without attention		
Model	**Para**	**Vid**	**euro**	**Para**	**Vid**	**euro**
ArabicBERT	0.851	**0.773**	0.760	0.823	0.755	0.738
CAMeL-BERT	0.690	0.651	**0.804**	0.671	0.584	0.666
AraBERT	**0.925**	0.457	0.782	0.906	0.438	0.757
mBERT	0.771	0.502	0.714	0.755	0.452	0.690
(Alian and Awajan, 2021)	n/a	n/a	n/a	0.354	0.743	0.467
(Nagoudi and Schwab, 2017)	n/a	n/a	n/a	0.182	0.691	0.206

Table 4: MSE of Arabic STS systems on the STS dataset (the lower, the better). Para stands for MSR-Paraphrase, Vid stands for MSR-Video, euro stands for SMT europarl datasets.

	With attention			Without attention		
Model	**Para**	**Vid**	**euro**	**Para**	**Vid**	**euro**
ArabicBERT	0.188	**0.238**	0.371	0.251	0.289	0.302
CAMeL-BERT	0.325	0.372	**0.201**	0.332	0.0.386	0.339
AraBERT	**0.154**	0.550	0.220	0.186	0.559	0.251
mBERT	0.285	0.489	0.382	0.302	0.498	0.395

Quranic verses semantic relatedness using arabert. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 185–190. Leeds.

Alzahrani, S. (2016). Cross-language semantic similarity of arabic-english short phrases and sentences. *J. Comput. Sci.*, 12(1):1–18.

Antoun, W., Baly, F., and Hajj, H. (2020). Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Chouikhi, H., Chniter, H., and Jarray, F. (2021). Arabic sentiment analysis using bert model. In *International Conference on Computational Collective Intelligence*, pages 621–632. Springer.

Chouikhi., H., Chniter., H., and Jarray., F. (2021). Stacking bert based models for arabic sentiment analysis. In *Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KEOD,*, pages 144–150. INSTICC, SciTePress.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Feifei, X., Shuting, Z., and Yu, T. (2020). Bert-based siamese network for semantic similarity. In *Journal of Physics: Conference Series*, volume 1684, page 012074. IOP Publishing.

H. Al-Bataineh, W. Farhan, A. M. H. S. and Al-Natsheh, H. T. (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.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, Y., McLean, D., Bandar, Z. A., O'shea, J. D., and Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. *IEEE transactions on knowledge and data engineering*, 18(8):1138–1150.

Libovickỳ, J., Rosa, R., and Fraser, A. (2019). How language-neutral is multilingual bert? *arXiv preprint arXiv:1911.03310*.

M. Hammad, M. Al-Smadi, Q. B. B. and Al-Zboon, S. A. A. (2021). Using deep learning models for learning semantic text similarity of arabic questions. 11(4):3519.

Nagoudi, E. and Schwab, D. (2017). Semantic similarity of arabic sentences with word embeddings. In *Third arabic natural language processing workshop*, pages 18–24.

Neculoiu, P., Versteegh, M., and Rotaru, M. (2016). Learning text similarity with siamese recurrent networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 148–157.

Safaya, A., Abdullatif, M., and Yuret, D. (2020). Bert-cnn for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation", Barcelona (online)", International Committee for Computational Linguistics.*

Saidi, R. and Jarray, F. (2022). Combining bert representation and pos tagger for arabic word sense disambiguation. In *International Conference on Intelligent Systems Design and Applications*, pages 676–685. Springer.

Saidi, R., Jarray, F., and Alsuhaibani, M. (2022). Comparative analysis of recurrent neural network architectures for arabic word sense disambiguation. In *Proceedings of the 18th International Conference on Web In-*

formation Systems and Technologies, WEBIST 2022, Valletta, Malta, October 25-27, 2022, pages 272–277. SCITEPRESS.

Saidi, R., Jarray, F., and Mansour, M. (2021). A bert based approach for arabic pos tagging. In *International Work-Conference on Artificial Neural Networks*, pages 311–321. Springer.

T. Ranasinghe, C. O. and Mitkov, R. (2019). Semantic textual similarity with siamese neural networks. In *Proceedings of the International Conference on Recent Advances*, pages 1004–1011. Natural Language Processing (RANLP 2019).