Evaluating SIAMESE Architecture Neural Models for Arabic Textual Similarity and Plagiarism Detection

Ahmed Amine Aliane
Research Center For Scientific
And Technical Information
Email: aamine@cerist.dz; aaliane781@gmail.com

Hassina Aliane
Research Center For Scientific
And Technical Information
Email: ahassina@cerist.dz; ahassina4@gmail.com

2020 4th International Symposium on Informatics and its Applications (ISIA) | 978-1-7281-9652-7/20/$31.00 ©2020 IEEE | DOI: 10.1109/ISIA51297.2020.9416550

***Abstract*—Semantic text similarity in NLP is the study of the degree of resemblance between texts using a certain metric. It has many applications in tasks such as question answering, information retrieval, document clustering, topic detection, topic tracking, questions generation, machine translation, text summarizing and others. Nowadays, neural models are outperforming existing state of the art approaches in major NLP tasks and it is not surprising to see the STS community researchers adopt these models although there are still few works for Arabic language. As Siamese neural architecture has proven recently its relevance for STS in other languages, we evaluate in this work three models within this architecture for Arabic Textual similarity and plagiarism detection: BiLSTM and CNN which we call basic models and a BERT Transformer model.**

I. Introduction

Semantic text similarity in NLP is the study of the degree of resemblance between texts using a certain metric. It has many applications in tasks such as question answering, information retrieval, document clustering, topic detection, topic tracking, questions generation, machine translation, text summarizing and many others. Nowadays, neural models are outperforming existing state of the art approaches in major NLP tasks and it is not surprising to see the STS community researchers adopt these models. Beside the development of neural architectures, the development of more and more sophisticated distributed word representations has certainly boosted the NLP research in general and STS research in particular. Indeed, since the first word2vec model of [1], a plethora of models has emerged such as Fasttext, Glove for word embeddings, Sent2ve, Phrase and others for sentence embeddings, Doc2vec for paragraph embeddings, ... Lately new contextualized word embeddings such as ELM0 and BERT are also knowing a growing success. Within neural approaches, recent SIAMESE neural networks, with their peculiar architecture show relevance for comparison and similarity learning between two or more inputs and have been successfully used for this task especially for English language. Hence, we evaluate in this work three SIAMESE based models combined with three similarity measures for Arabic text similarity and plagiarism detection. The models are CNN, BiLSTM which we call basic models and a transformer model. We used the Mawdoo3 questions similarity dataset for STS and the 2015 Arabic Pan dataset for plagiarism detection evaluation. The remainder of this paper is organized as follows: section II presents a review of related works, section III is dedicated to the proposed models while section IV describes experimentations and results. After a discussion of obtained results, a conclusion ends this paper.

II. Related works

The excitement for STS has grown increasingly during the last decade. In order to help the research in this field, STS has been added as a shared task to the SemEval series since 2012 making available an STS Benchmark and numerous public datasets for the SemEval tasks evaluation and for every STS enthousiast in general. The benchmark is annually updated and it includes datasets in many languages.

In the wake of the success story of Neural models in major NLP tasks, these models are increasingly being used in STS research. As pointed out by [2], there are two main approaches to STS using neural models: unsupervised approaches which use pre-trained word/sentence embeddings directly for the similarity task without training a neural network model on them and supervised approaches which do so. Unsupervised methods are particularly interesting for low resources languages where building datasets necessitates huge time and efforts. These approaches have shown decent results in the final rankings of shared tasks [3].

[4] proposed an unsupervised appproach named Sent2Vec for learning sentence embeddings which they evaluated for unsupervised sentence similarity using the cosine similarity between two sentences on the STS 2014 [5] and SICK 2014 [6] datasets. They compared the similarity scores to gold standard human judgements. Sent2Vec recently outperformed significantly the results of other sentence embeddings models such as Phrase [7]. [8] proposed the Infersent sentence embedding model which they evaluated for the unsupervised STS task 2014 [5] also using the cosine similarity. The results were competitive for the paraphrase detection dataset. [9] proposed a completely unsupervised sentence embedding model which is reported to improve performance on STS