Sentence encoders as a method for helping users identify and improve semantic similarity in bio-medical text

Brayn Díaz[1], Juan Pavez[1], Sebastián Rodríguez[1], Wenceslao Palma[2], Hector Allende-Cid[2], Rene Venegas[2] Eduardo N. Fuentes[1,3*]

I. INTRODUCTION

Academic writing is one of the most valuable skills a scientist can develop. A primary challenge for scientists is to coherently and concisely organize and present ideas within a manuscript. Unfortunatelly, novel computational approaches that help scientists write coherent scientific texts have been scarce and not very effective. Traditional methods that rely on bag-of-words, as well as word embeddings such as those produced by word2vec or GloVe may not be optimal due to language complexity.

Recently, positive results have been reported for sentence embeddings, which provide semantic relatedness between sentences without relying on word overlapping metrics. State of the art models of sentence embeddings are Universal Sentence Encoder and sent2vec, and the latter has an implementation trained on bio-medical data as presented in BioSentVec. Despite the rapid advancement of sentence embeddings and their applications on bio-medical text analysis, the performance of these models for helping users identify and improve semantic similarity between sentences has not been shown.

The aim of this work was to prove the effectiveness of different sentence encoders in bio-medical text to detect semantic similarity between sentences in different sections of academic texts.

II. METHODS

A. Corpus

We used three corpora: 1) A "Gold Standard" which consists of 2116 papers; 2) a first quartile (Q1) bio-medical journal comprised of 2027 papers from the Molecular Cell journal; 3) and a fourth quartile (Q4) bio-medical journal comprised of 1230 papers. The "Gold Standard" consists of papers in bio-medicine that contain top scientometrics, have in-housed copy editors with rigorous editorial processes, and that present the discursive structure prototypical of well-written papers.

B. Model

We use two models: the transformer-based Universal Sentence Encoder available on TensorFlow Hub, and BioSentVec, a sent2vec model trained on bio-medical text which is available on Github. Due to the lack of labeled data, no additional training is done. A section is modeled as a sequence of sentences, and the similarity between sentences is defined as 1 minus the arc cosine of the cosine similarity of their embeddings divided by π . We also define the *mean sequential similarity* (MSS) in a section as the mean of the similarities between successive sentences. Then, MSS is is used as a measure of semantic similarity in a text.

III. RESULTS

First, we begin calculating the average MSS in different sections (Introduction, Results, and Discussion) in academic manuscripts in the Gold Standard. The Results section (0.72 ± 0.02 with the USE model and 0.65 ± 0.01 with BioSentVec) showed important differences in comparison with Introduction (0.76 ± 0.02 with USE, 0.68 ± 0.01 with BioSentVec) and Discussion (0.78 ± 0.02 with USE, 0.67 ± 0.02 with BioSentVec), which were very similar. Next, and to compare the MSS in the Gold Standard and other journals, we selected a top Q1 journal and several Q4 journals belonging in the Molecular and Cell Biology Discipline. As a control we used a shuffled version of sentences for the Gold Standard, Q1, and Q4 journals. We found that in both models, the average MSS decreased after shuffling the sentences randomly. The standard deviation increased, though the increase is more significant in the USE model. This confirms that shuffled texts are less cohesive than the original versions.

Finally, we test the effectiveness of the MSS to indicate semantic similarity between sentences. For this we used examples of different sections from new Q1 and Q4 scientific articles and plotted their sequential similarity, comparing it with the Gold Standard MSS values as a benchmark to indicate which sentences are outliers (i.e. present higher or lower semantic similarity in comparison with the Gold Standard). We found that this qualitative approach is successful with both USE and BioSentVec models, allowing identification of adjacent sentences that are outliers in both Q1 and Q4 papers. Finally, linguists assessed the performance of both models to test which one provides the best feedback to improve cohesion.

IV. CONCLUSION

We demonstrated the effectiveness of both the USE and BioSentVec as methods for helping users identify and improve semantic similarity between sentences in bio-medical

[1]WriteWise, WriteWise Research Group, Artificial Intelligence Unit, Av. Beaucheff 935, Santiago, Chile

[2]Pontificia Universidad Católica de Chile

[3]BioPub, Scientific Writing Unit, Av. Beaucheff 935, Santiago, Chile

*__Corresponding author__: Dr. Eduardo N. Fuentes; email: ef@writewise.cl

texts. The shared tendencies between the models support sequential similarity as a metric to evaluate a text's cohesion. With both methods outliers can be easily spotted, and then specific modifications in the sentences can be carried out depending on the type of outlier.