IMPROVING THE SEMANTIC QUALITY OF BUSINESS PROCESS MODELS BY ANALYZING TEXT SIMILARITY USING SENTENCE-BERT

Oleksandr Rudskyi, Andrii Kopp
*National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine*

In this study, we explore the problem of comparing business process models with their textual descriptions. Business process models have proven to be an effective tool for visualizing and improving complex organizational operations. However, creating business process models is a complex and resource-intensive task, which sometimes leads to situations where the model itself does not match the textual description of the business process. This can lead to time and financial losses [1].

The proposed solution should extract the names of business process activities from the BPMN 2.0 file, generate the text T_1 from all these names, and compare it with the textual description of the given business process T_2. To calculate the semantic correspondence of these two texts, we propose to use the Sentence-BERT (SBERT) language model. The main idea of the SBERT model is to represent each sentence in a vector space where semantically similar sentences are close to each other [2]. It is proposed to calculate the distance between these vectors using the cosine similarity:

$$cosine\ similarity = S_C(A, B) := \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}},$$

where A_i and B_i are the coordinates of vector A and B, respectively.

Pic. 1 demonstrates the architecture of the Sentence-BERT model for calculating the semantic similarity score.
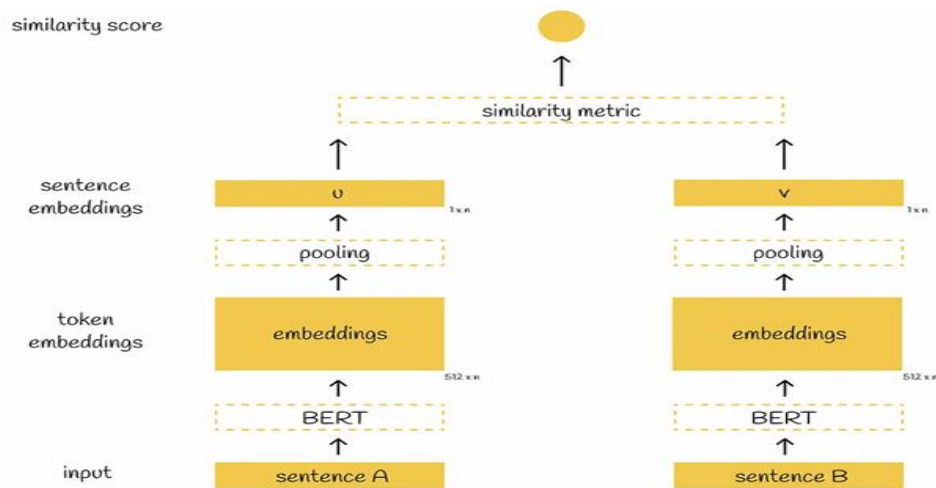
Picture 1 – The SBERT architecture for similarity score calculation

References:
1. Mroczek A., Wiśniewski P., Ligęza A. Overview of Verification Tools for Business Process Models. 2017. Vol. 13. P. 295–302.
2. Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using siamese BERT-Networks. *Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*. 2019. P. 3982–3992.