Session 2
6th May
15:15 h

# BioTMPy: a Deep Learning-based tool to classify biomedical literature

Nuno Alves[1], Ruben Rodrigues[1], Miguel Rocha[1]

[1] BIOSYSTEMS, Centre of Biological Engineering, University of Minho, Campus de Gualtar 4710-057 Braga Portugal

Over the last few decades, the publication rate has been massively increasing, resulting in a huge number of available scientific documents, which consequently makes the search of relevant information for a certain topic a heavy and time-consuming task. Biomedical Text Mining has been addressing this problem for a while, but there is still space for improvements. For instance, PubMed, which contains now more than 32 million citations, has only recently implemented a machine learning model to improve document ranking, and is still trying to improve their system by implementing a Deep Learning model, needing for now further studies.

Following this line of thought, a deep learning-based tool named BioTMPy was developed to facilitate the search of relevant documents. BioTMPy is divided into separate modules that ease distinct processes of a document relevance pipeline. More precisely, modules to load datasets in different formats, convert them into distinct data structures, perform data analysis, and implement several deep learning models with their associated methods to perform hyperparameter optimization, cross validation, etc. Additionally, the package provides some examples on how to integrate all the modules together to perform a complete pipeline for document relevance.

To validate the developed pipelines, BioTMPy was later applied on a BioCreative's challenge from 2019. This challenge addressed the search of relevant documents for the topic of "mining protein interactions and mutations for precision medicine". With a comparison between different pre-trained embeddings, BioWordVec seemed to show on this data a slightly better performance over GloVe, "pubmed_pmc" and "pubmed_ncbi". Additionally, a model with a pre-trained BERT model (BioBERT) and a Bi-LSTM managed to surpass the best challenge's submission with a difference of 7.25% for average precision and 3.15% for the f1-score.

In addition, a web service was implemented to provide an effortless use of the developed model, allowing a user to order documents by relevance regarding the topic mentioned above. This means that, after gathering a corpus, one can use BioTMPy to develop a pipeline to retrieve the most relevant documents for a certain topic and consequently make it available to the public.