Approaching Semantic Text Similarity with Hybrid Methods: a Case Study on French

Antonella Fadda[1], Rémi Cardon[2], Natalia Grabar[3], Thomas François[4]

[1]FIAL, UCLouvain – antonella.fadda@student.uclouvain.be

[2]Cental, UCLouvain – remi.cardon@uclouvain.be

[3]UMR STL, UdLille – natalia.grabar@univ-lille.fr

[4]Cental, UCLouvain – thomas.francois@uclouvain.be

Abstract

The difficulty in understanding texts is a daily struggle for many people. To overcome this problem, Natural Language Processing (NLP) offers various solutions, namely text simplification. The main difficulty in developing systems for text simplification is the lack of resources, such as parallel corpora or lexicons. One common approach for parallel corpora development is extraction of sentences that share the same meaning, from comparable corpora. Doing so requires evaluating the semantic similarity between sentence pairs. In this article, we propose to investigate this task in the light of the recent developments in NLP. Concretely, we will work on the French language, using two corpora : DEFT'20 and CLEAR. DEFT'20 is a French corpus containing 1,010 sentence pairs annotated with their degree of similarity on a 0-5 scale. CLEAR is a French comparable biomedical corpus made for text simplification out of three different sources, Wikipedia/Vikidia, drug leaflets, and medical literature summaries. We report on experiments with state of the art language models for French (general such as CamemBERT and FlauBERT) and with classic feature-based machine learning approaches (e.g. Random Forest with similarity measures such as Manhattan distance, Levenshtein distance, Dice coefficient, etc.). As we observe that the top-performing systems of the DEFT 2020 campaign on the task achieve similar results as the language models in isolation, we closely analyze the strengths and weaknesses of the two approaches in order to identify how complementary they are. We evaluate our experiments in two ways: (1) by their performance on the DEFT'20 corpus and (2) by their ability to identify parallel sentences from the CLEAR comparable corpus.

Keywords: semantic text similarity ; sentence alignment ; natural language processing.