Abstract

Semantic textual similarity (STS) is a natural language processing (NLP) task to quantitatively assess the semantic similarity between two text snippets. STS is a fundamental NLP task for many text-related applications, including text deduplication, paraphrase detection, semantic searching, and question answering. Measuring STS is a machine learning (ML) problem, where a ML model predicts a value that represents the similarity of the two input texts. These machine learning models can be categorised in to two main areas; supervised and unsupervised. Supervised STS ML models have been trained on an annotated STS dataset while the unsupervised STS ML models predict STS without being trained on annotated STS data. In the first part of the thesis we explore supervised and unsupervised ML models in STS. We explore embedding aggregation based state-of-the-art STS methods, sentence encoders, Siamese neural networks and transformers in STS. Furthermore, for each STS method we analyse the ability of the model to perform in a multilingual and multi-domain setting. On the process, we develop new state-of-the-art unsupervised STS method based on contextual word embeddings and new state-of-the-art supervised STS method based on Siamese neural networks.

The second and third parts of the thesis, focus on applying the developed STS method in the applications of translation technology; translation memories

(TM) and translation quality estimation (QE). We identify that the edit distance based matching and retrieval algorithms in TMs are less efficient and we propose a TM matching algorithm based on the STS methods we developed in the first part of the thesis. We empirically show that this algorithm outperforms edit distance based matching algorithms. As the next application, we utilise the STS architectures we developed, in translation quality estimation. We show that STS architectures can be successfully applied in QE by changing the input embeddings in to cross-lingual embeddings. Based on that , we develop TransQuest - a new state-of-the-art QE framework that won the WMT 2020 QE shared task. We release TransQuest as an open-source library and by the time of writing this, TransQuest has more than 8,000 downloads from the community.