UBIQUITOUS KNOWLEDGE PROCESSING

Bachelor/Master Thesis

TECHNISCHE UNIVERSITÄT DARMSTADT

Document Context-Aware Interpretable Sentence Similarity

Motivation

Measuring sentence similarity [1] is a classic topic in natural language processing (NLP). Semantic Textual Similarity (STS) [2] is a well-studied task that measures the equivalence of sentence pairs in terms of meaning by predicting similarity scores, while the idea of interpretable STS (iSTS) [3] is to explain why and how two sentences may be similar/different by supplementing STS with an explanatory text. Previous works on STS and iSTS analyze sentence pairs in an atomic fashion, without knowing the document-level context. The proposed thesis topic is based on the core idea that the meaning of a sentence should be defined by its contexts, and that the sentence similarity could be better determined and explained by taking contexts into consideration. This thesis will construct a document revision dataset containing alignments between sentences pairs with an alignment type and a similarity score. An iSTS system based on advanced sentence Transformer models such as [4] will be trained on this dataset which, given a pair of sentences and their corresponding contexts, explains what is similar and different, in the form of graded and typed sentence alignments. By systematic comparison of various systems with or without knowledge of document context, this thesis will answer the question of whether it is beneficial to measure sentence similarity in contexts.

Task Description

- Construct a document revision dataset containing sentence alignments with labels of types.
- Train an iSTS system on the dataset, given sentence pairs and the corresponding contexts.
- Compare various iSTS systems with or without knowledge of document context.

References

- [1] https://huggingface.co/tasks/sentence-similarity
- [2] Cer et. al. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In Proceedings of SemEval-2017, ACL
- [3] Agirre et. al. 2016. SemEval-2016 Task 2: Interpretable Semantic Textual Similarity. In Proceedings of SemEval-2016, ACL
- [4] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 EMNLP-IJCNLP, ACL

Analysis	■■■■□
Programming	■■■■□
Literature	■■■□□

Contact

Prof. Dr. Iryna Gurevych

Qian Ruan

thesis@ukp.informatik.tu-darmstadt.de