UBIQUITOUS KNOWLEDGE PROCESSING

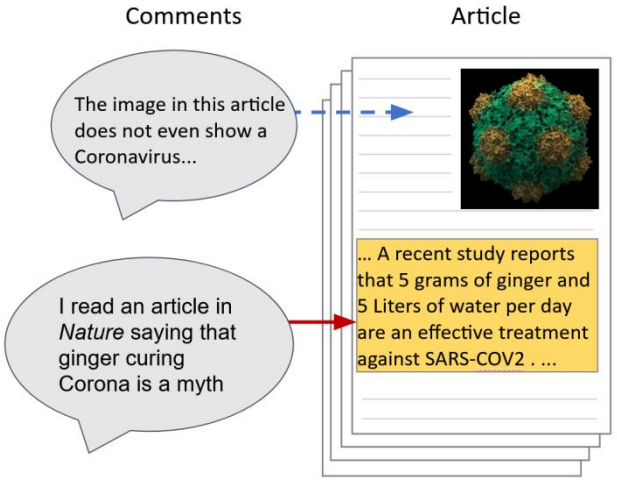
Bachelor/Master Thesis

TECHNISCHE UNIVERSITÄT DARMSTADT

Structuring Online Comments

Motivation

We have unprecedented access to information stored in digital documents, but not all of it is trustworthy [1]. Online discussions on platforms such as twitter, reddit or hypothes.is can help to navigate the uncertainty by providing background information and critical comments. However, it is often hard to get an overview of the comments, because they are spread over various sources and they can be related to any aspect of a document.
The goal of this thesis is to develop NLP techniques that link comments from various sources to the exact part of a document they comment on. This would allow aggregating comments by the aspect they discuss, greatly improving their accessibility.



First, we will collect data that allows self-supervised learning [2], and then we will employ recent transformer models that can handle large documents [3] to learn the task of linking.

Task Description

- Compile a dataset of documents and the comments on these from various sources
- Analyze the dataset to find self-supervision signals that can be exploited in training
- Train and evaluate state-of-the-art machine learning models on the task of linking

References

[1] https://www.theguardian.com/media/2016/dec/18/what-is-fake-news-pizzagate
[2] https://www.fast.ai/2020/01/13/self_supervised/
[3] Caciularu, Avi, et al. "Cross-Document Language Modeling." arXiv preprint arXiv:2101.00406 (2021)

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

Contact

Prof. Dr. Iryna Gurevych

Jan Buchmann

thesis@ukp.informatik.tu-darmstadt.de