PD08: Authorship verification using LLMs

Background

Cloze test proved to be a useful tool for testing text comprehension. Some universities use it during a disciplinary procedure when a student is suspect from submitting a work authored by someone or something else (plagiarism, contract cheating, unallowed use of generative AI). Authors of the text are more likely to fill in correct words.

The project aims to find a method that identifies words to be masked such that the cloze test can reliably discriminate between authors and non-authors. LLMs are trained to predict the word in given context. Previous experiments showed that nouns that the model would not guess correctly are good candidates.

Goal

- To extend the existing project by conducting more experiments with LLMs and users
- To improve existing method (better discriminate between authors and non-authors)

Tasks

- Employ more language models to identify masked word (so far only MT-5 was used)
- Experiment with probability of the word in given context (so far only rank was used)
- Investigate the influence of language (English, German, etc.; native / non-native)
- Investigate the influence of time (authors forget their text and achieve lower scores)

The project aims to find a _____ that identifies words to be masked such that the cloze test can reliably _____ between authors and non-authors. LLMs are trained to predict the word in given context. Previous _____ showed that nouns that the model would not guess correctly are good candidates.

Tomáš Foltýnek
foltynek@fi.muni.cz

Terry L. Ruas
ruas@gipplab.org