Textual paraphrase dataset for deep language modelling

Filip Ginter (PI), Jenna Kanerva, Li-Hsin Chang, Maija Sevón, Jenna Saarni, Otto Tarkka
Hanna-Mari Kupari, Jemina Kilpeläinen, Valtteri Skantsi
TurkuNLP Group
University of Turku, Finland
October 2020

Project goals

- Building a large dataset of **100,000 lexically diverse paraphrases for Finnish**
- Building a small test dataset for Swedish
- Developing deep learning models for paraphrase identification and generation
- Data and models will be made available for everyone with CC-BY-SA license

Paraphrase dataset – current status

- 25,000 Finnish paraphrases collected from news titles and movie subtitles (25% of projected total)

- Annotation process:
  - Dedicated tool for **picking paraphrase candidates** from various text samples
  - Dedicated tool for **labeling candidates** according to the **detailed annotation scheme**
  - Option to rewrite candidates to be full paraphrases

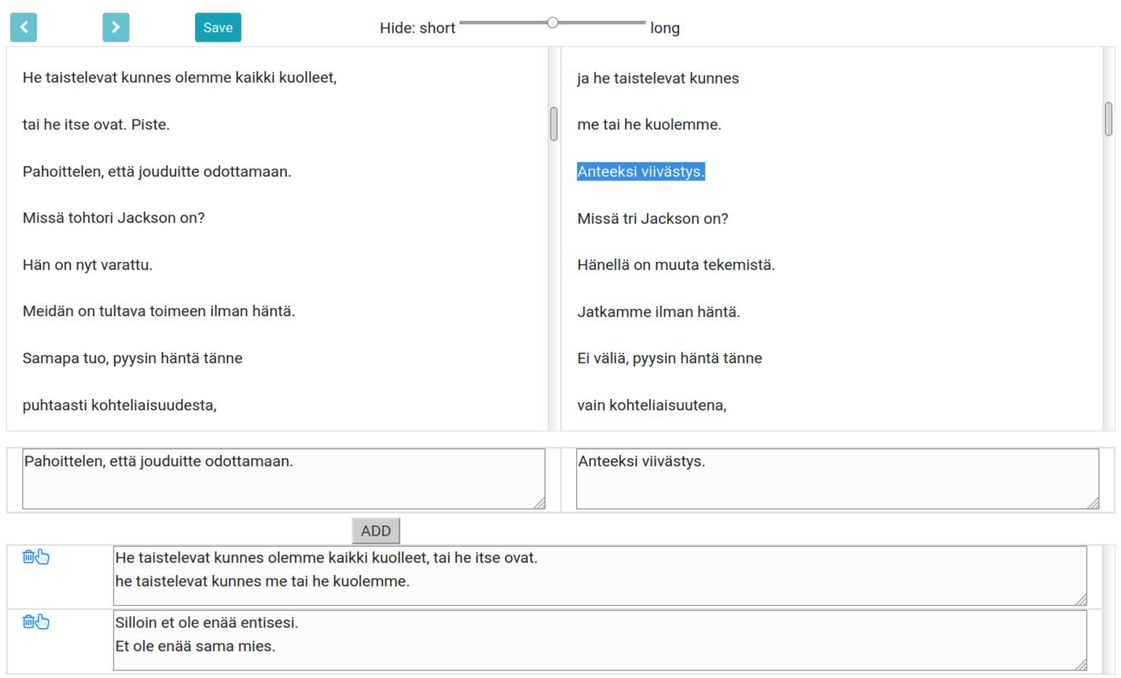
Deep learning models

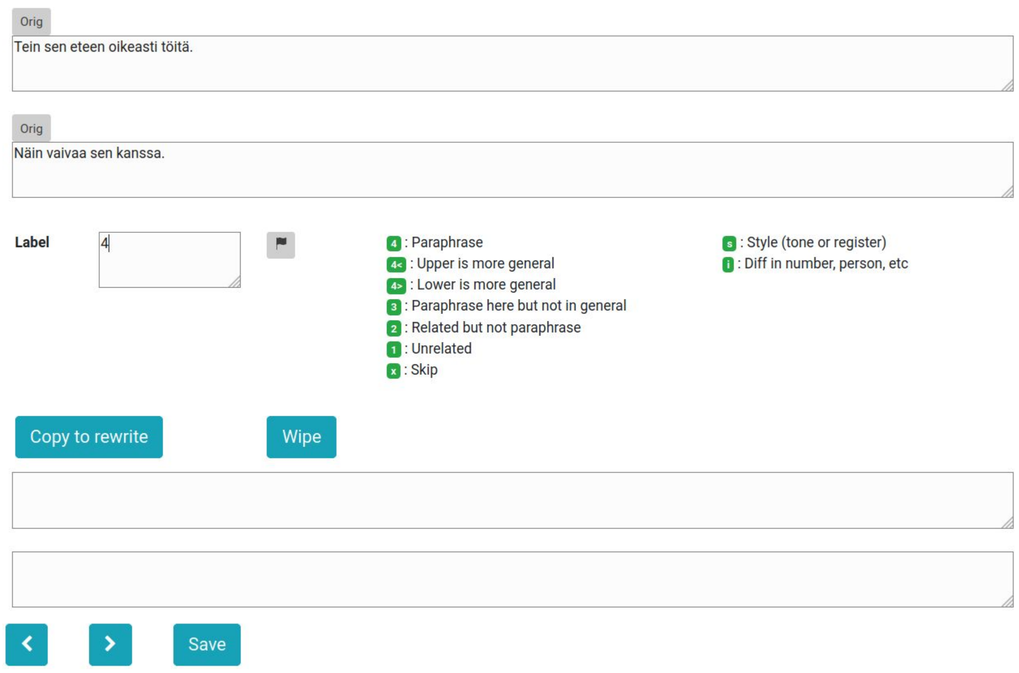
- Finetuning deep language models (e.g. BERT) for paraphrase identification

Other possible directions:
- Models for paraphrase generation
- Models for machine translation evaluation

Applications

- Semantic search – Search engine returning hits based on semantic meaning rather than surface words
- Identify repeated arguments from literature
- Plagiarism detection – Detect shallow surface modifications
- Text rephrasing – Create variability by rephrasing the text
- Training machine translation system – Evaluate the meaning of a translated sentence, not just its words





UNIVERSITY OF TURKU

TURKUNLP .ORG