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
November 2021

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 are available for everyone with CC-BY-SA license in https://turkunlp.org/paraphrase.html

Paraphrase dataset

- **104,645 Finnish paraphrases** collected from news articles, movie subtitles, discussion forum messages, and university student translations and exercises

- 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

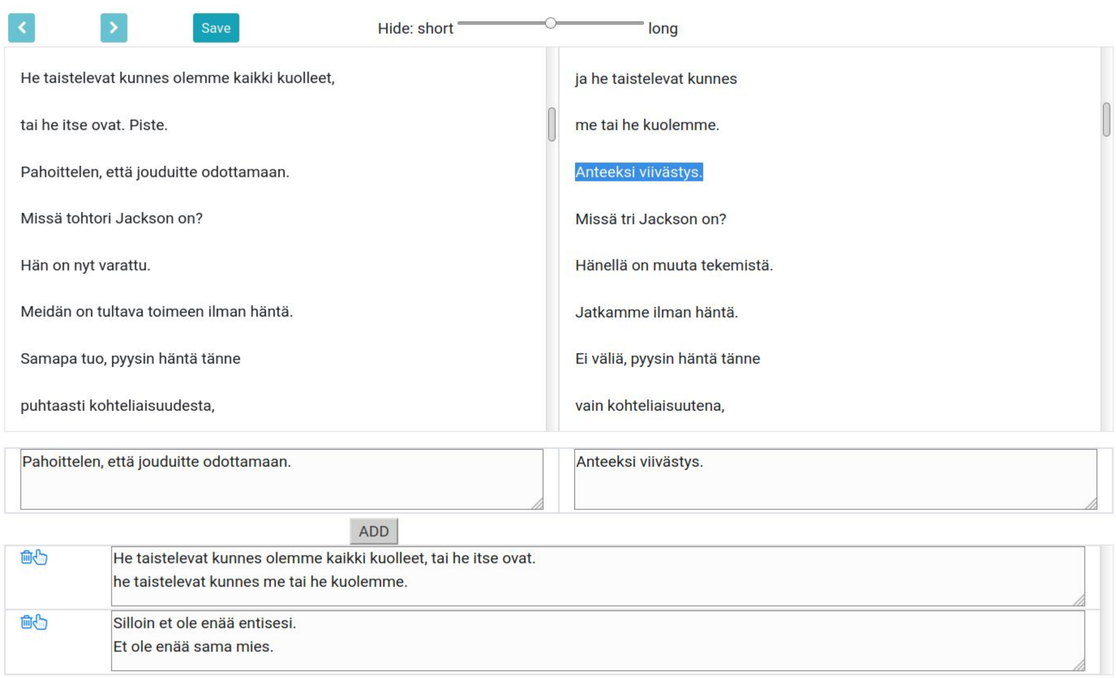
Deep learning models

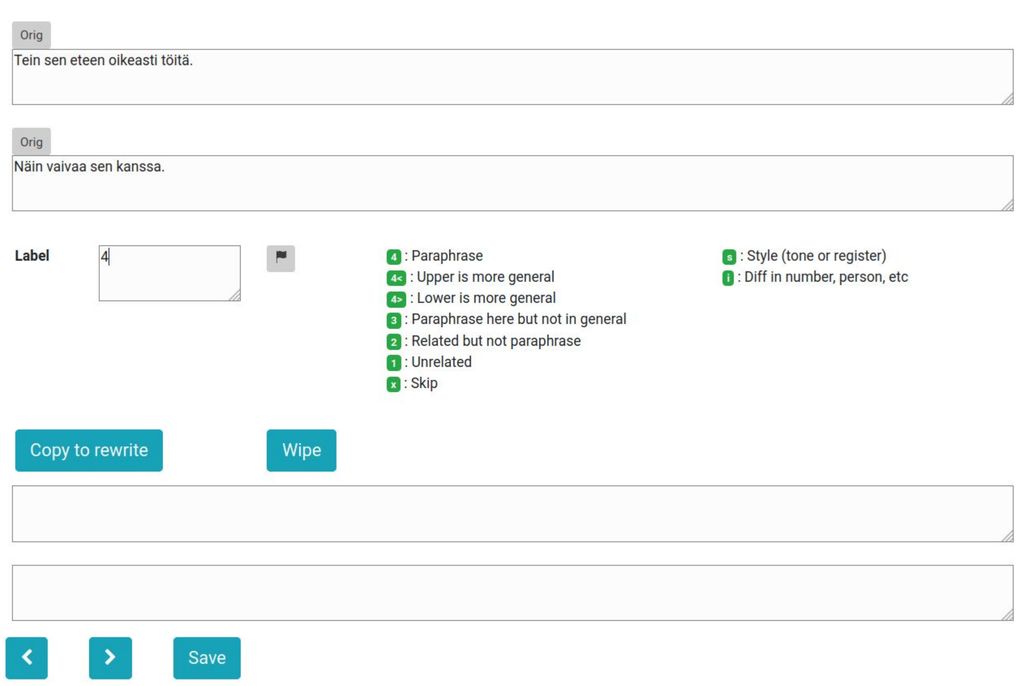
- Finetuning deep language models (e.g. BERT)
  - **Paraphrase classifier** - Given a candidate pair, decide whether it is a paraphrase or not
  - **Paraphrase retrieval** - Find paraphrase candidates from a massive collection of text
  - **Semantic search** - Given a query phrase, find a corresponding paraphrased segment from a document

Other possible directions:
- Paraphrase generation
- Machine translation evaluation
- Text rephrasing

Project outcomes

- Turku Paraphrase Corpus (Kanerva et al. 2021, final manuscript in preparation)
- Annotation guidelines (Kanerva et al. 2021)
- Finetuned deep learning models for paraphrasing
  - Paraphrase classifier
  - Sentence embeddings (SBERT)
  - Extractive paraphrase detection for semantic search
- A collection of automatically gathered paraphrase candidates (500K positive and 5M negative)
- Quantitative evaluation of the paraphrase pairs (Chang et al. 2020)





UNIVERSITY OF TURKU

TURKUNLP .ORG