

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

Hide: short — long

He taistelevat kunnes olemme kaikki kuolleet,
tai he itse ovat. Piste.
Pahoittelen, että joudutte odottamaan.
Missä tohtori Jackson on?
Hän on nyt varattu.
Meidän on tultava toimeen ilman häntä.
Samapa tuo, pyysin häntä tänne
puhtaasti kohteliaisuudesta.

ja he taistelevat kunnes
me tai he kuolemme.
Anteeksi viivästys.
Missä tri Jackson on?
Hänellä on muuta tekemistä.
Jatkamme ilman häntä.
Ei väliä, pyysin häntä tänne
vain kohteliaisuutena.

Pahoittelen, että joudutte odottamaan.
Anteeksi viivästys.

ADD

He taistelevat kunnes olemme kaikki kuolleet, tai he itse ovat.
he taistelevat kunnes me tai he kuolemme.

Silloin et ole enää entisesi.
Et ole enää sama mies.

Orig
Tein sen eteen oikeasti töitä.

Orig
Näin valvaa sen kanssa.

Label 4

Paraphrase
Upper is more general
Lower is more general
Paraphrase here but not in general
Related but not paraphrase
Unrelated
Skip

Style (tone or register)
Diff in number, person, etc

Copy to rewrite Wipe

Save