Click here to be redirected to the virtual room of the Project Expo.

Project abbreviation: TurkuParaC

Project name: Turku Paraphrase Corpus

Project coordinator: Filip Ginter

Funding: ELG, Academy of Finland

Project duration: ELG: 08/2021-07/2022; Academy of Finland: 09/2021-08/2024

Main key words: paraphrase, Finnish, Swedish, deep learning, corpus creation, natural language understanding

Background of the research topic: Paraphrase is an important target for natural language understanding, as it requires the models to identify shared meaning despite vastly differing wording.

Goal of the project: The goal is to create a large-scale manually annotated paraphrase corpus suitable for paraphrase model training and evaluation, as well as to develop machine learning models for paraphrase detection

Project abstract: The project builds a large dataset of Finnish paraphrase pairs accompanied by a small test set of Swedish paraphrases. The paraphrases are selected and classified manually, so as to minimize lexical overlap, and provide examples that are maximally structurally and lexically different. The objective is to create a dataset which is challenging and better tests the capabilities of natural language understanding. An important feature of the data is that most paraphrase pairs are distributed in their document context. The primary application for the dataset is the development and evaluation of deep language models, and representation learning in general.

Publications:

J. Kanerva & F. Ginter & LH. Chang & I. Rastas & V. Skantsi & HM. Kupari & J. Kilpeläinen & J. Saarni & M. Sevón & O. Tarkka. 2021. Finnish Paraphrase Corpus. Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa 2021) pp. 288-298.

LH. Chang & S. Pyysalo & J. Kanerva & F. Ginter. 2021. Quantitative Evaluation of Alternative Translations in a Corpus of Highly Dissimilar Finnish Paraphrases. Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age

J. Kanerva, F. Ginter, LH. Chang, I. Rastas, V. Skantsi, J. Kilpeläinen, HM. Kupari, A- Piirto, J. Saarni, M. Sevón, O. Tarkka. 2021. Annotation Guidelines for the Turku Paraphrase Corpus. arXiv 2108.07499