Updated plans, UTU, Digicampus/Autograding
15.4.2020
contact: Kaapo Seppälä, 040 744 4182, kamise@utu.fi

What has been done

Work in progress

1. We have machine-translated several datasets into Finnish, these datasets are SNLI, MNLI, XNLI, MRPC, and RTE. These are important datasets in the field of natural language understanding, which allows the training and evaluation of models for text similarity detection.
2. We have trained Finnish Sentence BERT models, which specialize in representing a Finnish sentence in vector forms to the machine.
3. We are training bilingual BERT models, which facilitates Finnish benefitting from English resources to a certain extent.
4. We are constructing the FinParaphrase corpus, which currently has more than 7000 annotated sentence pairs. The data sources used include student generated text, news headings, and movie subtitles. The completed corpus is expected to benefit paraphrase research for the Finnish language with a wide variety of applications.

Publications

- Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, Sampo Pyysalo: Multilingual is not enough: BERT for Finnish. 2019. https://arxiv.org/abs/1912.07076 The paper is about BERT models for the Finnish language allowing for dense vector representation of text units (such as phrases)
- In preparation: Bilingual English-Finnish BERT model. This work strives to provide a bilingual language model allowing us potentially to utilize existing English language data to strengthen Finnish models.

Initial prototype plan

The prototype will focus on clustering similar answers. When given an essay or a short text segment, it will return similar essays for the examiner, highlighting the essential parts that make the essays similar. The purpose of this functionality is to allow the examiner quick orientation in large sets of answers, improving grading efficiency and consistency. A docker environment will be set up so that the prototype can be tested. The prototype will provide an API allowing a set of documents (answers to a particular question in the exam) to be indexed and subsequently answer queries for similar answers. Only a rudimentary test web

interface will be provided, the API and the indexing engine behind will be the main contribution.

Updated schedule:

Completed:

June 2019	Literature review (educational NLP, essay clustering, and automatic grading)
July 2019	Literature review Annotations (qualitative research essays)
August 2019	Tested different ways of representing and clustering sentences and how to evaluate the clustering methods.
September 2019	Literature review (sentence representation) Tested sentence representation methods Result write up and analysis
October 2019	Literature review (sentence representation, evaluation methods) Trained and tested sentence representation models
November 2019	Literature review (sentence representation) Trained and tested sentence representation models
December 2019	Annotations (accounting essays), exploration on annotation scheme
January 2020 (three weeks of vacations)	Annotations (accounting essays) Preparation to train new models
February 2020	Training of bilingual BERT
March 2020	Training of bilingual BERT Working with machine translated datasets Start of paraphrase corpus construction

Ongoing

April-May 2020	Paraphrase corpus construction Training of bilingual BERT Prototype development
June 2020	Paraphrase corpus construction Training of bilingual BERT Prototype development **Prototype testing (from 15.6.2020)**