Final project 2024

Natural Language Processing

Reichman University

Finally, it's time for the final project! This year, you have the option to work on either a project we have designed for you or one based on your own idea. Regardless of your choice, the primary deliverable is a report that summarizes your project idea, background, motivation, methodology, results, and conclusions. Let's dive into some additional details about these two options.

Option 1: Working on your own idea

If you choose to work on your own idea, it must include an innovative component. We encourage you to think about new methods and implementations, and we are happy to support you throughout the ideation process. Please reach out to us (Kai, David, Amir, Kfir) to discuss potential ideas or initial thoughts about topics you would like to explore further. We are excited to hear about your ideas!

Once you have an idea for the project, please send a short project proposal by email to kfir.bar@runi.ac.il. Your proposal should include a description of your idea and what you plan to do, along with the specific datasets you intend to use. We will provide some initial feedback and approval status. Once approved, you can start working on your project and submit the report by the deadline. The structure of the report is discussed in the following section.

Grading

Your grade will be based on the following components:

Percentage	Component
15%	**Problem definition** - is the problem well defined and is the motivation set correctly?
30%	**Experiments** - Do the experiments you chose to run prove your claim? Do they cover everything? Did you discuss and analyze your results?
30%	**Novelty** - how novel is your idea?
10%	**Report** - is the report clear and easy to read? Does it cover all the required components?
15%	**General impression**

Important dates

30/6/2024 - Deadline for sending us your project proposals

10/8/2024 - Deadline for submitting your final project reports
24/8/2024 - Grades are ready

Option 2: Default Project Description

Training a Sentence BERT Model in Hebrew

The objective of this project is to train a Sentence BERT (SBERT) model in **Hebrew** using the newly released Natural Language Inference (NLI) dataset in Hebrew. The main characteristic of SBERT is its ability to calculate a single vector for the entire given document, capturing all or most of the information of the text, for text similarity purposes. SBERT is a version of BERT that has been fine-tuned using pairs of semantically similar sentences as well as pairs of sentences that are not considered to convey similar information. In a nutshell, during training, each pair is processed by the model to produce a sentence-level vector representation for each of the two sentences, individually (typically the [CLS] BERT embeddings of the model). Then, using a specific loss function the model is modified to draw vectors of similar sentence pairs closer together. This technique is known as contrastive learning. For more detailed information, refer to the original SBERT paper: https://arxiv.org/abs/1908.10084.
Once trained, you are required to evaluate the trained SBERT model on the Semantic Text Similarity (STS) benchmark in Hebrew, which we prepared for you, and compare the performance across multiple Hebrew language models.

Dataset Preparation:
Obtain the new NLI dataset for Hebrew from here: https://huggingface.co/datasets/HebArabNlpProject/HebNLI and preprocess the dataset as needed to ensure it is suitable for training.

Model Training
Train a Sentence BERT model using the NLI dataset in Hebrew, using three base models: AlephBERT, mBERT, and DictaBERT. Ensure the training process is well-documented, including the configuration of hyperparameters and the training environment.

Evaluation
Test the trained SBERT models on the Hebrew STS benchmark (provided). Record the model's performance using standard metrics for semantic text similarity.

Reporting Results
Prepare a detailed report summarizing the training process, evaluation methodology, and performance results, according to the provided template report.

Grading:
Your grade will be structured from the following components:

Percentage	Component
30%	**Experiments** - Do the experiments are reported appropriately? Did you run some additional and interesting experiments, not listed in the project description?
35%	**Implementation** - did you implement the model the right way?
15%	**Report** - is the report clear and easy to read? Does it cover all the required components?
20%	**General impression**

Important dates
10/8/2024 - Deadline for submitting your final project report
24/8/2024 - Grades are ready

Final Report

For both options, you are required to submit a report. Please use the following template to write your report: https://www.overleaf.com/read/pjngxfknwfkj. The maximum number of pages is 8, but shorter reports are welcome.

If you have any questions or need assistance, don't hesitate to contact us. Good luck, and we look forward to seeing your innovative work!

Good luck