Independent project

Your work

In the project work, your task will be to

- find an interesting task that can be addressed with machine learning methods
- find (or annotate) an appropriate dataset for that task
- look around for related work
- implement software to train some machine learning model for your task
- evaluate your system
- write up a report about your results.

The projects would normally be done in groups of up to three. Please ask for permission if you want to form a larger group.

Preliminary project description

Please submit a short text that describes what you intend to work on. For suggestions of topics, see below.

You should submit this description in Canvas. The text should include a project title and one or a couple of paragraphs giving a rough outline of how you intend to work. If you have found any related work or relevant dataset, this is useful information to include. The deadline for this preliminary description is **December 10**.

The report and the presentation

When writing your report, please use one of the these style templates (available for Word and LaTeX). The text should be structured as a typical technical report, including an abstract, statement of problem, method description, experiments and results, and a conclusion. Please submit the project report via the Canvas page. The deadline for submitting the report is **January 15**.

Grading

The grading is primarily based on the quality of the written report, and to a smaller degree on the quality of the presentation. For a maximal score, the writing is expected to be concise and clear, related work to be described thoroughly, the technical approach to be evaluated systematically, and the conclusions to be insightful.

It is not necessarily the case that a more "difficult" project will give a high grade, but maybe it will be easier to come up with interesting questions for experiments and analysis if the topic isn't too similar to the course assignments or lecture demos.

Finding a topic

You can work on any project that is small and manageable enough for a short project of this kind. PhD students should ideally try to find a topic that is relevant to their research interests. It's OK if you end up with preliminary results, as long as you have some tidbit to present at the end.

If you need help deciding on a topic, here are a few suggestions. Some of these are very easy and some more challenging, and this list is not sorted by the difficulty level.

- **Adapting a lab assignment or lecture demo**. Start from one of the lab assignments or lecture demos and change it to something that is interesting for you. Maybe change assignment 1 to WSD for another language or another dataset? Maybe change assignment 2A to another entity recognition task or a part-of-speech tagging task?
- **Comparing representations for an application**. Select some NLP application, either one we saw during the lectures or something where you can find the code online. Investigate how various types of text representations affect the application's performance. For instance, you could compare standard word embeddings, ELMo, BERT, and other representations.
- **Using reinforcement learning for parsing**. Train a transition-based dependency parser (or some other similar application) using reinforcement learning, for instance by following the approach described by Fried and Klein (2018).
- **Caption generation**. Build a system that generates descriptions of images.
- **Learning semantic categories**. Use a toolkit such as `word2vec` and a database of semantic categories (e.g. FrameNet) to build a system that discovers new words belonging to a category. For instance, you could make a system that discover words that refer to different types of food.
- **Crosslingual applications**. For some application, e.g. some categorization or tagging task we've seen in the course, investigate how well *crosslingual representations* such as multilingual BERT allows us to train with one language and evaluate with other languages.
- **Shared tasks**. Every year, several NLP competitions are arranged at the SemEval conference. Participate in one of the 2021 tasks or get the data from some task from previous years. Additionally, there are shared tasks organized by CoNLL, BioNLP and other conferences, where datasets can often be accessed easily.

Alternatively, if you don't have an idea and you don't like the suggested topics, just talk to the teachers in the course.