Final Project Instructions

Deep Natural Language Processing Class, 2022

Due dates:
Group formation - 16th of May
Project proposal - 31th of May
Paper submission - 28th of June

1 Final Project

The final group projects amounts to 50% of your grade. The main goal of this assignment is to prepare you for conducting full-blown NLP research, i.e creating a research proposal, performing literature review, running experiments and finally writing 'a baby academic paper'. It consists of three parts: the project proposal, the final project report and the paper presentation. All are required to be considered for the full mark.

The final thesis does not have to be 'grandiose'. We mainly will mainly grade how well you present the problem, execute experiments to prove your hypothesis, and discuss the results keeping up with academic standards.

2 Finding yours teammates

The final group project is all about collaboration - an essential skill for conducting large experimental projects that require good mathematical modeling, creating neural architectures, building training pipelines, collecting data and finally setting-up infrastructure.

We highly **recommend** forming groups of 3 people. The maximum number of people in the group is 4 and the minimum is 2. Only in extraordinary situations we will allow exceptions to these quotas. The groups can be formed of people from different groups (these include mixing of classmates from UW and UJ) - anything goes here.

We encourage you to find yourself partners that you want to work with and it is your duty to actively find your group. If you feel like you do not know enough about your classmates, we will help you finding partners at the laboratory session on **16th of May**. At the end of this class you must **finalize** the formalization of all groups.

3 Choosing a Project Thesis

The good project starts with asking interesting research questions. Since plenty of work is being published every day about deep learning and natural language processing field, one

might get lost targeting right topics and problems.

That is why, we provide three low-risk tracks that groups can take: 1) open-domain Polish QA system evaluated on the PolEval 2021 challenge, 2) intent classification for Slavic languages with MASSIVE dataset or 3) reproducing an NLP paper. They offer a low-risk research roadmap while providing a great deal of potential experimental ideas. We recommend to choose on of them as your final project. We will also discuss in next section how to propose your own project.

1. Open-domain QA Polish model

Each year Institute of Computer Science at Polish Academy of Sciences is running amazing challenges tackling Polish NLP problems. You can read more about the project here. Your project would consist of building an open-domain question answering system based on Task 4 from PoleEval 2021.

The goal of the task is to develop a solution capable of providing answers to general-knowledge questions typical for popular TV quiz shows, such as 1 z 10. The evaluation will be carried out on the test-B dataset. Asking Google on the fly is **not** a viable solution.

Checking if the two answers match will depend on the question type: For non-numerical questions, assess textual similarity. A Levenshtein distance should be computed between the two (lowercased) strings and if it is less than $\frac{1}{2}$ of the length of the gold standard answer, the candidate answer is accepted.

For numerical questions (e.g. In which year...), assess numerical similarity. Specifically, use a regular expression to extract a sequence of characters that could be interpreted as a number. If such sequences can be found in both answers and represent the same number, the prediction is accepted.

2. Intent classification for Slavic languages

Intent classification is a backbone of conversational AI systems. Recently introduced MASSIVE dataset offers a challenge of modeling intents across 51 different languages. Each language consists of 19,521 datapoints spanning 18 domains with 60 intents to model. It is a great testbed for researching multilingual approaches and models. Due to the size of the dataset and the scale of the testbed, we will focus on Polish language (along with other Slavic languages if the resources permit). The intent performance is measured with intent accuracy with two data splits - high and low. The authors of the testbed provided XLM and mT5 baselines that are good starting points for next experiments.

Best ideas and models could be published at the Massively Multilingual NLU 2022 workshop held alongside the EMNLP 2022.

3. Reproducing an NLP Paper

Reproducing a paper or a part of an experimental set-up sounds like a rather boring exercise but in fact it is one of the best way into getting to a particular field. This track would require to choose a recent paper (published since 2018) from any conference or arXiv that touches on

any NLP topic and reproducing particular results from the paper. This could be a particular experiment, full reproduction or a salience analysis of the main model.

The reproducibility itself is a major problem in the machine learning word since majority of the papers are almost impossible to reproduce!

If you are struggling in finding a suitable paper contact us!

3.1 Proposing Your Own Idea

Understandably, the best project is the one defined by yourself with novel research questions. However, given the rigid timeline of the course it is the most risky option with many potential unknown traps. Nevertheless, if you strongly believe in the topic, you can propose your own research project by writing 200 words proposal answering following questions: 1) What are the main research goals? 2) What is novel about this project? 3) What is the industrial impact (if any)?. Self-proposed projects require an approval from your TA and the course lecturer.

Here is a list of main topics considered at ACL 2022 - there are many areas you can consider:

- Computational Social Science and Cultural Analytics
- Dialogue and Interactive Systems
- Discourse and Pragmatics
- Ethics and NLP
- Generation
- Information Extraction
- Information Retrieval and Text Mining
- Interpretability and Analysis of Models for NLP
- Language Grounding to Vision, Robotics and Beyond
- Linguistic Theories, Cognitive Modeling, and Psycholinguistics
- Machine Learning for NLP
- Machine Translation and Multilinguality
- NLP Applications
- Phonology, Morphology, and Word Segmentation
- Question Answering
- Resources and Evaluation
- Semantics: Lexical
- Semantics: Sentence-level Semantics, Textual Inference, and Other Areas
- Sentiment Analysis, Stylistic Analysis, and Argument Mining
- Speech and Multimodality
- Summarization
- Syntax: Tagging, Chunking and Parsing

The ultimate deadline for proposing your own topic is **20th of May**.

4 General Instructions

The group project consists of two major steps. The first one - the project proposal is a build up for the final report (10% of all course points). The final report is a major bulk of work where you write about experimental efforts (40% of all course points).

4.1 Project Proposal

The project proposal consists of writing the two first sections of your final report, i.e. Introduction and Related Work. The proposal must not take less than 500 words and not more than 1000 words. The project proposal should make an argument why the chosen topic is important. It should also introduce main research questions, a plan of the experimental phase as well as recent literature review (Google Scholar could be useful here).

Overall, the project proposal should answer following questions:

- What is the main research questions you are trying to tackle?
- What NLP task(s) will you address?
- What dataset(s) will you use?
- What baseline(s) will you use?
- What experiments will you carry out?

As the proposal will be written while performing experiments, it might be challenging to write about the exact plan of the experimental phase with strong statements. Please write this part with the best intentions.

The proposal (paper) **must** consists of the following sections: Title, Authors with emails, (Optional) Collaborator, Supervisor, Introduction, Related Work. The project proposal deadline submission is **31th of May**.

4.2 Final report

The final project report should conform to the ACL submission guidelines. The template of the report will be based on the submission template of the long paper (8 pages long) in the camera-ready version. Please have a look at some examples of submissions here (all of them are amazing papers as well!):

- SQuAD: 100,000+ Questions for Machine Comprehension of Text
- Deep Contextualized Word Representations
- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

The report (paper) **must** consists of the following sections: Title, Authors with emails, (Optional) Collaborator, Supervisor, Abstract, Introduction, Related Work, Model or Problem, Experimental Set-up, Results and Discussion, and Conclusions. The literature does not add up to the 8 pages limit.

The final report deadline submission is **28th of June**.

5 Submission instructions

Project proposals and final reports should be send via Moodle/Pegaz with a standard due dates policy.