Language as Data.

Projects (WS 2024/25)

Project 3: Using Pre-Trained Language Models

In this project, you will work with pre-trained language models. For deliverable 1 and 2, you use a monolingual English model and for deliverable 3, you will use a multilingual model.

Submit the project report as a pdf and the code in a zip archive through Stud.IP. The filenames should read *03_firstname-lastname_firstname-lastname_firstname-lastname.{pdf,zip}* with the first and last names of each group member. The reports are due on Monday evening (20:00 CET), **27tth of January**. We expect you to submit descriptions for all deliverables. If you run into any problems working on this project, describe the problem and/or give a partial analysis. If you need additional computing resources, contact us early. We can give you access to the cluster, but this won't work on short notice.

1. Deliverable 1: Task Analysis

Select a task from the **GLUE benchmark** for English and report relevant task statistics. Describe the task setup using examples from the dataset. Manually analyze the instances and select ten instances that you expect to be easy to solve and ten instances that you expect to be difficult.
Reflect on the annotation setup:

- What are the characteristics of the annotator sample?
- How detailed are the annotation guidelines? Are the scales anchored with examples?
- Is agreement between annotators reported? Are the raw annotations available?
- How were conflicting annotations merged?
- What is your opinion on the quality of the dataset (both the selection of instances and the annotations)?
- Explain your categorization of "easy" and "difficult" instances.

2. Deliverable 2: Finetune

Finetune GPT-2 on the GLUE task and compare the result to finetuning a randomly initialized GPT-2 model (no pre-training). Try at least two finetuning variants and compare the performance. Analyze the performance on the instance level for your 20 selected instances and reflect on the results.

If the above works well, feel free to try out other models or finetuning strategies.

3. Deliverable 3: Multilingual Model

Use a multilingual model and prompt translations for one of your target languages.

Option 1: Qualitative Evaluation
Select 40 prompts such that you systematically vary one characteristic (e.g., register, domain, syntactic complexity) and analyze how it affects the translation quality. For example, you could design minimal pairs of sentences that only differ in tense or gender.

Option 2: Quantitative Analysis
Extract translations from a parallel corpus (e.g., from the **Workshop of Machine translation**) and evaluate the quality using the BLEU score. Reflect on the strengths and weaknesses of the BLEU score.

4. Deliverable 4: Project Summary

Provide a brief summary of your project's key findings and reflect on the work process.