CSE 256 Natural Language Processing : Analysis Study

Spring 2024
University of California, San Diego

updated: April 12, 2024

Analysis Study: Understanding the Limitations of NLP Systems (25% of final grade)

For this assignment, you will choose an NLP task and select one paper that includes code related to this task. Run the code and analyze and understand the limitations of the existing system. For example, you might try to understand the errors made by a machine translation model and what kinds of sentences it fails on. If you don't speak a second language, you can approximate this by performing a back-translation - first translate an English sentence into a second language, then translate it back to English, and analyze the errors made in this back-translation. Another example involves understanding the limitations of current language models on specific tasks. For inspiration on tasks that remain challenging for even the largest language models, see this paper: Bubeck et al. [2023].

To get started finding models and papers with code, see: `https://paperswithcode.com/`. and `https://huggingface.co/models`

The analysis project accounts for 25% of the final grade: 2% for a brief literature survey and proposal (due **May 10th**), and 23% for the final analysis study and report (due **June 7th**).

Part 1: Literature Survey and Proposal (2%) 1-2 pages

Due: May 10, 2024.

Perform a brief literature survey on the topic of your chosen task, and find papers on this task that have code available. Pick one of the papers and its associated codebase. Submit a short report telling us: 1) the topic and task you picked; 2) the 3-5 papers you read, including proper citations and a 3-5 sentence summary of each; and 3) a brief description of the codebase you plan to use and the task it is designed to solve, as well as the dataset(s) and evaluation metrics you plan to use to analyze the system.

Part 2: Analysis Study (23%) 5-8 pages

Due: June 07, 2024.

Part 2.1: Analysis (3-5 pages)

Perform a detailed analysis to understand the limitations of the existing system. Identify and provide concrete examples of failure cases. Determine if there are any semantic or syntactic commonalities among these challenging examples. We would like to see a manual error analysis. For instance, annotate some failed examples to highlight various properties, discuss the results, and hypothesize about the reasons for the system's failures.

Part 2.2: Looking Forward (2-3 pages)

How can the difficult cases you identified in Part 2.1 be addressed? Outline what changes in terms of data, models, algorithms, or evaluation metrics are necessary to resolve these issues. Develop a comprehensive plan for improving the system's performance.

Part 2.3 (bonus): Implementation (up to 2% extra credit)

If you have time and are interested, you can try to implement some of the improvements you proposed in Part 2.2. This is optional and will be graded as extra credit.

What Grade to Expect:

A detailed final report template will be provided. In the meantime, here is a guideline on what grade you might expect based on the quality of your submission:

- **A+**: Exceptional or surprising. Significantly surpasses the quality and depth of most other submissions.
- **A**: Includes a thorough analysis and detailed, forward-looking plans for further development.
- **A-**: Slightly less comprehensive in analysis and development planning than assignments awarded an A.
- **B+**: Analysis or development plans are somewhat lacking or underdeveloped.
- **B or B-**: Missing one or more key elements such as detailed analysis or forward-looking plans.
- **C+ or below**: Lack of effort or incomplete submission.

References

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712, 2023. doi: 10.48550/ARXIV.2303.12712. URL `https://doi.org/10.48550/arXiv.2303.12712`.