COM SCI 263: Natural Language Processing
Final Project

Overview

This course project is designed to immerse you in the authentic experience of conducting research within the field of Natural Language Processing (NLP), enhance your understanding of NLP tasks and principles, sharpen your problem-solving skills, and stimulate innovation. Offering a curated selection of four project topics, the course also encourages creativity and individual initiative by allowing students the flexibility to propose their own topics. Through this hands-on project, you will navigate the research process from designing steps like hypothesis formulation and experimental setup, to implementation steps like experiment execution, culminating in a comprehensive report and a presentation of your findings. In summary, this project provides an opportunity to apply theoretical knowledge to practical challenges, preparing you for future endeavors in NLP and related fields.
Note: We will offer up to 10% bonus to groups who exhibit high creativity in the projects or tackle the optional tasks.

Submission

There are several important milestones for the final project. Please submit each of them to gradescope following the description below:

1. **Project planning report due: Wednesday, April 17th at 11:59PM. (10%)**
   - Teaming information
   - The research topic you choose to work on (can be your own topic)
   - The distribution of work among teammates
2. **Project midterm report due: Monday, April 29th at 11:59PM. (20%)**
   - Introduction to the problem
   - Literature Review and baseline model selection
   - Description of your approach
   - Project progress until this point
   - Weekly plan and distribution of work for the subsequent weeks.
3. **Course project final presentation starts: Wednesday, May 22nd at 11:59PM. (30%)**
   - The same expectation/rubric as the Assignment 1 (paper presentation).
   - Week 8 presentations will be given 10% bonus while week 9 presentations will be given 5% bonus.
4. **Project final report due: Wednesday, June 05th at 11:59PM. (40%)**
   - A 4-8 pages research paper-like writeup, preferably using this template.
   - Extend your midterm report to include more details about your approach.
   - Dataset and evaluation framework.
   - Experimental results with quantitative and qualitative analysis of your method vs. baselines.

Project Idea 1: Creating Persona Chatbots

Dialogue persona is a personal profile that users or developers can inject into the model. Creating models with personas has been an important research direction in NLP research, as assigning personas to dialogue models can help improve engagement and personalization in their conversational interactions with users. Existing methods of persona assignment suffer from two major limitations: (1) the coherence of persona throughout a conversation, and (2) the degradation of generation and other abilities due to persona adoption. This project therefore aims to explore potential methods to resolve the limitations of current methods, and work towards building better persona-assigned dialogue models.

Tasks:

1. Identify **at least 1 baseline** approach to inject a persona into a generative language model.
2. Design methods to improve over the baseline you identified to inject a persona into a generative language model. Please see below for an example of a json data containing information for “persona-elaine”.

   ```
   # Example of persona
   {"persona-elaine": {
          "name": "Elaine",
          "gender": "female",
          "profession": "student at UCLA",
          "nationality": "China"
   }}
   ```

3. Design/Identify a dataset (with **at least 50** data entries) and an evaluation framework to test the persona-assigned chatbot in multiple dimensions:
   (a) Coherency of persona: How well does the model “keep” a persona throughout a conversation.
   (b) Generation ability: How fluent/natural are model generations when assigned personas.
   (c) Other abilities (reasoning, multilingual, etc.): How do persona-assigned models perform on these tasks.
4. Evaluate the persona-assigned models (both the baseline and your model) on the framework and report the results.
5. Prepare the project final presentation.
6. Draft a report with both quantitative and qualitative results.

Project Idea 2: Evaluating LLM for downstream applications

Recent advancements in Large Language Models (LLMs) have allowed for high-quality long context generation. The powerful generation ability of LLMs facilitates the development of multiple downstream applications, such as mathematical tutoring, reference letter writing, and code generation. However, evaluating the LLMs' abilities on these tasks has been a challenging research direction. Since these applications fall into drastically different fields, field-specific evaluation datasets and methods are required to provide accurate and detailed insights into the models' performance. This project aims at targeting one such downstream application that you can think of (be creative!), and you are expected to build a domain-specific evaluation framework to measure LLM's performance on this application task.

Tasks:

1. Define a downstream application of LLMs that you want to study.
   (a) Example: Math tutoring, reference letter writing, code generation.
2. Identify **1-2 LLMs** you plan to evaluate. It can be ChatGPT, or other open-source LLMs.
3. Design/Identify a set of test data (prompt + ground truth answer) with **at least 50** data entries to probe for LLM's generations.
   (a) Example: For math tutoring, a set of math questions with their solutions.
4. Design methods and automated metrics to evaluate LLM performances.
   (a) Example: For reference letter writing, might want to consider aspects like fluency, positivity, fairness, etc..
   (b) Must provide an explanation for metric construction, i.e. why do we need these specific aspects to evaluate LLMs for this application?
5. Evaluate the LLM's generations using the metrics and report the results.
6. (Optional) Perform human annotation on LLM outputs and compare human judgment with metric outcomes.
   - Can be used to justify the design of evaluation methods.
7. Prepare the project final presentation.
8. Draft a report with both quantitative and qualitative results.

Project Idea 3: Enable Auto-Regressive Language Models to Fill in The Blanks

Recent advancements in auto-regressive language models have demonstrated the models' efficacy in text generation tasks, enabling widespread application across various domains. These models proficiently predict subsequent words, thereby aiding human writers. However, limitations arise as auto-regressive language models lack foresight into future words, hindering their ability to perform text insertion and revision. Specifically, pre-trained decoder-only models typically continue the provided prefix without considering the suffix, resulting in inserted text that may be irrelevant or contradictory to the context. In addition to text insertion, text rewriting also poses a significant challenge, as modifying specific segments within a sentence requires ensuring coherence with the surrounding text. This project aims to enhance decoder-only or encoder-decoder language models to facilitate text insertion and rewriting using specialized training or decoding techniques on our designated evaluation set.

Tasks:

1. Select **1** decoder-only or encoder-decoder model you want to work on.
   (a) Decoder-only: GPT2, Llama, etc..
   (b) Encoder-Decoder: T5, T0, etc..
2. Design methods that enable models to do insertion and rewriting.
   (a) Insertion: Given a prefix and a suffix, to fill in the middle.
   (b) Rewriting: Given a prefix, a text segment to rewrite, and a suffix; rewrite the text segment that well connects to the prefix and suffix.
   (c) The methods may include, but are not limited to: dataset curation, new training objectives, attention mechanisms, or the application of specific decoding methods.
3. Design/Identify **1 baseline method** to compare against, that shares the same base model (e.g., GPT2 or T5) as your method.
4. Evaluate your proposed methods on the provided evaluation set, and compare their results with predictions from other baseline models.
   (a) Run the models on our provided evaluation set, which contains around 100 samples.
   (b) Conduct human evaluation to compare the predictions from different models. The comparison can be a simple pairwise preference comparison, to identify whether one prediction is better than another.
   (c) Provide further error analysis of the predictions from both the proposed and baseline methods.
5. (Optional) Design/Identify automatic evaluation metrics to evaluate models' ability to do insertion or rewriting.
   (a) Example: For insertion and rewriting, you might want to consider aspects like fluency, grammar correctness, and the coherence between the inserted/rewritten text and prefixes and suffixes, etc..
   (b) Check whether results of the proposed automatic evaluation metric align with that of human preference evaluation on the predictions.
6. Prepare the project final presentation.
7. Draft a report with both quantitative and qualitative results.

Project Idea 4: Cross-Cultural Social Acceptability Classification: Enhancing Model Performance Across Diverse Contexts

Social acceptability classification plays a vital role in identifying instances of harm or discomfort for individuals within a given context. However, existing approaches to social acceptability classification often reflect Western-centric perspectives, overlooking the cultural variability in social norms and acceptable behaviors. This project aims to address this gap by developing a culturally diverse dataset of socially acceptable interactions and behaviors, encompassing various cultural contexts. By assessing model performance on this dataset, we aim to enhance the ability of machine learning models to accurately predict social acceptability across diverse cultural settings. As a reference, a dataset to test **general model accuracy without culture context** can be found at https://nlpositionality.cs.washington.edu/acceptability .

Tasks:

1. Develop a comprehensive dataset with **at least 50** data entries. Each sample in the dataset should comprise of:
   (a) An interaction (a chat with an LLM).
   (b) Information on the cultural context of the interaction.
   (c) Corresponding ground truth labels indicating the social acceptability of the interaction.
2. Design an evaluation framework to test a model's ability for cross-cultural social acceptability classification (SAC) on the new dataset. Here are some aspects you should consider:
   (a) Fine-grained analysis of factors affecting SAC accuracy.
   (b) Impact on the inclusion of cultural context on SAC and model performance.
   (c) Other Dimensions (implications for real-world applications in diverse settings, etc)
3. Establish **1-2 baselines** of existing social acceptability classification models and evaluate on the new dataset.
4. (Optional) Develop novel methodologies for enhancing model adaptability to diverse cultural contexts, such as domain adaptation, cross-lingual transfer learning, and culturally aware feature engineering.
5. Prepare the project final presentation.
6. Draft a report with quantitative and qualitative analysis.