CS605 Natural Language Processing for Smart Assistants

Assignment 2

Language Model and Applications

Due in 11:59pm, June 6th, 2023. Submission includes code files and a pdf file. Please zip them together, rename it with your student id, and submit it in elearn.

Commonsense Reasoning Challenge (80 marks)

In this section, we are to solve a real task using what we have learned: Commonsense Reasoning task. In this task, you need to choose either Alternative1 or Alternative2 as the answer to the given question. Here is an example:

Question: Tom has glaucoma. What happened as a result?

Alternative1: He is afflicted with obesity.

Alternative2: Doctor advised him to use coleus.

Answer: 2

1) Dataset

You are given two files: train.jsonl, and eval.jsonl. The file train.jsonl includes 14912 samples with ground truth answers. You can split it into any ratio you want into training set and validation set (for hyper-parameter tuning). Another file eval.jsonl includes 2131 samples as public leaderboard (you can submit you answer as follows to check the performance) and 2130 samples as the private leaderboard (you need to submit your answer but won't see the performance before due date, which we will use for grading). Each data sample in a line is formatted as:

{"Id": "train-14900", "Question": "Jack murdered his boss. What happened as a result?", "Alternative1": "He was caught by the police.", "Alternative2": "He was taken away by the police.", "Answer": 2}

Where "ID" is the unique id of the sample, " Question" is the given question to answer. "Alternative1" or "Alternative2" is the target options to classify as the answer. "Answer" denotes the ground truth option; the label is either 1 or 2.

Note that there is **no** answers in eval.jsonl! You are to predict the answers as the result file.

2) Method

There is no restriction of your technical models. Namely, you could use a typical machine learning approach (e.g., naive bayes classifier) or deep learning approach (e.g., RNN).
Note that hard coding is not allowed.

3) Timeline

- From today: train.jsonl released, you may train the model first.
- 1 Jun 2024, 00:00: eval.jsonl and the sample submission will be released in elearn assignment folder.
    - Submit your result to the below public leaderboard (max 10 submissions per day).
    - https://www.kaggle.com/t/de01eab22a4d4290a5ac2365e277f85c
- Before 6 Jun 2024, 23:59: Select one submission under My Submissions tab in Kaggle to be used in the private leaderboard. Otherwise, the submission that performed best in the public leaderboard will be used in the private leaderboard.
- Before 6 Jun 2024, 23:59: Submit code files and project report (**max 3 pages**) in eLearn.
    - Report: Describe the algorithm you use, the way you prepare the dataset, as well as your experiments.

4) Submission

- Code files to reproduce your submission in Kaggle. You need to provide a .ipynb file including all your codes. Optionally, you can submit your well-trained models or a google drive link if it takes too much time to run your codes.
- A 3-page pdf, including the methodologies you used, the results and model comparisons.