Homework 6: Ethics and Impacts

Due May 3rd at 10pm

To submit your assignment, place your write-up and code in a Google Drive folder that is shared with me.

Part 3: Assembling on Your Probe Task

Your main task in this assignment is to make progress on your probe task. I would like you to submit a preliminary dataset for your task. You should have **at least 32 frame sentences**. For each frame sentence, you should construct variants that highlight your contrast of interest.

For instance, in my example Grammatical Diversity probe, I picked the syntactic structure *positive so don't I*. A single frame sentence is shown in 1; its two variants are shown in 1a and 1b below.

1. Went here the other night with a girlfriend. Sure it's trendy, but **so aren't** most NYC clubs.
   (a) Variant 1: Went here the other night with a girlfriend. Sure it's trendy, but **so aren't** most NYC clubs.
   (b) Variant 2: Went here the other night with a girlfriend. Sure it's trendy, but **so are** most NYC clubs.

The exact requirements for your dataset vary by your choice of task, as described below.

Grammatical Diversity probe

If you choose the Grammatical Diversity option, you will test how your chosen phenomenon affects the downstream performance of NLP models on several tasks: sentiment analysis, question-answering, entailment, and sentence probability.

You must construct 8 frame sentences for each task. It is unlikely that the same examples will work for all four tasks, given the differences in format they require. However, it is ok if you reuse some sentences between tasks.

Your examples, whenever possible, should be based on text drawn from the test corpora for each task. This minimizes the chance of worse performance due to genre mismatch.

I have provided you with access to the following datasets for each task:

- Sentiment
  - Yelp: reviews dataset (note: count 3-4 stars as POS and 1-2 stars as NEG)
  - Cardiff Twitter: sentiment dataset
  - you can also find and adapt tweets from other sources
- Question-answering
  - RACE: question-answer pairs from English exams in China

  - QuAIL: reading comprehension questions from blogs, fiction, user stories and news
- Entailment
  - SNLI: human-written sentence pairs annotated with textual entailment information
  - MNLI: crowd-sourced sentence pairs annotated with textual entailment information
- Sentence probability
  - Any data is fair for this task. I recommend using example sentences from the Yale Grammatical Diversity project or linguistics papers that you have read on the topic.

You can find these datasets in the CS 232 Final Project Resources folder. **Warning**: some of these datasets are very large! I would not recommend downloading them all at once.

If possible, you should base your frame sentences on paraphrases of items from these datasets. This will ensure that your dataset fits the format of each task.

You should keep in mind the threats to validity discussed by Blodgett et al. (2021). Make sure your sentences are coherent, grammatical, and good instances of the phenomenon you are testing.

You will store your dataset in TSV files (tab-separated values). **You will turn in four files, one for each task.** For each frame sentence, you should record its original source and the task for which it is constructed. You can see examples of my files in the CS 232 Final Project Resources folder.

Here is how you should format your dataset for each task:

Sentence probability

A TSV file, where the first five fields are as follows:

NUMBER CONDITION TASK SOURCE EXAMPLE

You may have additional fields as required.

- NUMBER: An ID number. Each frame sentence should have a unique number.
- CONDITION: A for sentence with grammatical feature of interest; B for plain variant
- TASK: sentence probability
- SOURCE: original source
- EXAMPLE: the sentence itself

Sentiment

A TSV file, where the first six fields are as follows:

NUMBER CONDITION TASK SOURCE EXAMPLE LABEL

You may have additional fields as required.

- NUMBER: An ID number. Each frame sentence should have a unique number.
- CONDITION: A for sentence with grammatical feature of interest; B for plain variant
- TASK: sentiment
- SOURCE: original source
- EXAMPLE: the sentence itself
- LABEL: the correct sentiment label (positive or negative)

Entailment

A TSV file, where the first seven fields are as follows:

NUMBER CONDITION TASK SOURCE SENTENCE1 SENTENCE2 LABEL

You may have additional fields as required.

- NUMBER: An ID number. Each frame sentence should have a unique number.
- CONDITION: A for sentence with grammatical feature of interest; B for plain variant
- TASK: entailment
- SOURCE: original source
- SENTENCE1: the first sentence
- SENTENCE2: the second sentence
- LABEL: the correct label (entailment, contradiction, or neutral)

Question-answering

A TSV file, where the first eleven fields are as follows:

NUMBER CONDITION TASK SOURCE TEXT QUESTION ANSWER1 ANSWER2 ANSWER3 ANSWER4 LABEL

You may have additional fields as required.

- NUMBER: An ID number. Each frame sentence should have a unique number.
- CONDITION: A for sentence with grammatical feature of interest; B for plain variant
- TASK: QA
- SOURCE: original source
- TEXT: the text passage basis for the question
- QUESTION: the question
- ANSWER1: answer option A
- ANSWER2: answer option B
- ANSWER3: answer option C
- ANSWER4: answer option D
- LABEL: the correct answer (A, B, C or D)

Cultural Assumptions probe

If you choose the Cultural Assumptions probe, you will test an aspect of how GPT-3 models culture.

You must construct 32 frame sentences for your task. You should decide what kind of prompting paradigm you will use: intra-sentence prediction or inter-sentence prediction.

You must also design an evaluation metric. How will you measure the cultural biases of GPT-3 based on your frame sentences? Will you compare the probabilities of sentences? Will you compare the probabilities of specific sentence completions? Will you come up with a metric for evaluating the top-k sentence completions that GPT-3 produces?

You should keep in mind the threats to validity discussed by Blodgett et al. (2021). Make sure your sentences are coherent, grammatical, and target the particular aspect of cultural bias you are interested in.

Dataset

You will submit your dataset as a single TSV file (tab-separated values). The format will depend somewhat on which prompting paradigm you choose. You can see examples of my files in the CS 232 Final Project Resources folder.

I chose to have 6 conditions for each of my sentences: five countries and a neutral version.

My dataset is formatted as a TSV file with the following four fields:

NUMBER EXAMPLE COUNTRY

- NUMBER: An ID number. Each frame sentence should have a unique number.
- EXAMPLE: the sentence itself
- COUNTRY: the name of the country that is being targeted

I first wrote a file containing 32 frame sentences. Then I used a script (stub_to_prompt.py) to substitute in the city names for each of the different country conditions. I then manually edited the output to construct the neutral condition and make sure that it sounded natural.

Evaluation metric

You are responsible for designing your own evaluation metric, since it depends on how you set up your task. I compared each country-specific sentence to the neutral condition.

I took the top 5 most probable next words according to GPT-3 for each sentence.

Since the top 5 words are not always the same for all conditions of a sentence, I added a sixth OTHER category, and calculated how much probability GPT-3 assigned to all other words in the vocabulary, by summing over the top 5 probabilities and subtracting from 1.

Then I calculated, for each country condition, the sum of probability differences between the country-specific version and neutral version of each sentence. This is a rough way of quantifying the divergence between the probability distributions for the country-specific and neutral conditions.

You can see how I did this in CA_GPT3_scoring.py script if you want to do something similar.

You can also come up with other ways of evaluating your task. For instance, you might define a set of words that you think the model might generate, and score the raw generated text using this list.

For this homework, you do not need to have fully implemented your evaluation metric. **You must submit a short description of your evaluation metric (1-2 paragraphs)**, so that I can give you feedback before your final paper is due.