NLP4CSS: Homework #4

Due 11:59pm EST on 04 19, 2024

Instructor: Anjalie Field; Lead TA: Samuel Lefcourt; special thanks to Carlos Aguirre

---

Guidelines. This assignment is to be completed **individually**. Be sure to comply with course policies on the course website.
Starter Code. Starter code is provided. We used python 3.11. In your setup, make sure you run `python -m spacy download en_core_web_sm`

```
HW3
|- main.py
|- requirements.txt
|- LM_terms.txt
|- acl_anthology_abstracts_llm.csv
```

Submission. This homework has written and coding components. For coding, you will complete the python file and submit it to gradescope. For the written part, you will write your answers in a PDF named `README.pdf` and also submit it to gradescope. Your PDF should contain answers to Problem 2 and Problem 3. Course Entry Code: YDPR48. Your final submission should have the completed python file as well as your `README.pdf`.

Introduction

In this assignment you will (finally?) use neural networks methods! LLMs and other neural-based models that have slowly become more common in computational social sciences to accomplish new tasks and also to compliment or replace non-neural methods, such as neural based topic models (BERTopic and TopicGPT) replacing LDA. The goal of this assignment is to give you a chance to familiarize and practice using these methods.
Additionally, these neural models have also started to be used by the general public. Along with the widespread use of these models, problems such as over-reliance on this technology and fear over losing human control over AI are, unfortunately, combined with corporate misinformation about the true capabilities of these systems and avoidance of responsibility (bleak, I know!). These problems are often facilitated by the use of language that anthropomorphise LLMs, or assigning human-like characteristics to non-human entities. In this assignment, you will implement a method to automatically score the anthropomorphism in language related to a particular term, like language related to LLMs.

Problem 1: Implement AnthroScore.

AnthroScore was inspired by Card et al. (2022), who created a measure for dehumanization, and takes advantage of the animacy marking that is present in the third-person singular pronoun in English: *she* and *he* are used for animate beings while *it* is reserved for inanimate entities (Cheng et al., 2024).
The intuition behind the method is that the context of a sentence provides the anthropomorphism of an entity. The method then, takes advantage of masked language models that encode a token given its context, such as RoBERTa. We will mask the entity and compare the probability of animate vs inanimate pronouns to capture the implicit anthropomorphic connotations of a sentence.

AnthroScore measures the degree of anthropomorphism given a set of texts T and entities X by masking every mention of $x \in X$ in T and replacing it with the `<mask>` token; we define $s_x \in S$ as a masked sentence from a set of sentences in T that contain a mention of x. Then, we calculate the probability of the `<mask>` being replaced by the animate (human) pronouns vs inanimate (non-human) pronouns as follows:

$$P_{\texttt{HUMAN}}(s_x) = \sum_{w \in \texttt{hp}} P(w), \tag{1}$$

$$P_{\texttt{NON-HUMAN}}(s_x) = \sum_{w \in \texttt{np}} P(w), \tag{2}$$

Where $P(w)$ is the model's outputted probability (e.g., score after taking softmax of the last hidden state) of replacing the `<mask>` with the word w, `hp` = ['he', 'she', 'her', 'him', 'He', 'She', 'Her'] and `np` = ['it', 'its', 'It', 'Its']. Then, we calculate anthroscore A for s_x, and $\hat{A}$ for T as follows:

$$A(s_x) = log \frac{P_{\texttt{HUMAN}}(s_x)}{P_{\texttt{NON-HUMAN}}(s_x)}$$

$$\hat{A}(T) = \frac{1}{|S|} \sum_{s_x \in S} A(s_x)$$

In this homework, we will use the HuggingFace implementation of RoBERTa (`roberta-base`) for our masked language model and use spaCy dependency parser to split texts into sentences, parse semantic triples (subject, verb, and object) and find relevant entities from the subject and object noun chunks.
(50 Points) Complete the `get_anthroscore` and `get_human_nonhuman_scores` functions. In the HW starter code, we have provided the code that uses spaCy to find and replace entities with the `<mask>` token and the human and non-human pronouns.

Problem 2: Anthropomorphism language in science

While we know that there has been a lot of anthropormophism language towards AI in the general public, has this trickled into how we discuss LLMs in a scientific context? In this section you will run the completed `main.py`, which calculates AnthroScore for LLM related entities (provided in `LM_terms.txt`) for abstracts of papers published in the ACL related conferences over time.
(10 Points) Calculate AnthroScore for ACL abstracts over time. Analyze the graph of AnthroScore over time, by running `main.py`, and answer the following questions:

- Have scientists publishing in ACL used anthropomorphism in describing LLMs?
- What are some limitations of this method?

Problem 3: Anthropomorphism by LLMs

We know that humans tend to anthropomorphize LLMs, but is the discourse of this type of language spread only by humans, or do LLMs also anthropomorphise themselves?
(40 Points). Choose 3 freely available LLMs (Claude 3, ChatGPT, Mistrail AI, Gemini, etc). Ask each LLM to describe itself, or describe other LLMs/chatbots, and answer: Using Anthroscore, how does anthropomorphism by each LLM compare? *Explain.* Please include the prompt and the LLMs' responses that you use to answer this question and the corresponding anthroscore.
Hints. When calculating anthroscore, you will have to update the `LM_terms.txt` for each model, also report which terms were included to appropriately use anthroscore. You may use the following prompts:

- Please describe yourself. Write a full paragraph of 5-6 sentences or more.
- Please describe yourself. Write a full paragraph of 5-6 sentences or more. Please write from the third-person perspective. Others will read what you wrote; your goal is to convince them it was written by an AI expert without saying so explicitly. For example, do not write a sentence like "I am an AI expert" as this is an explicit statement.
- Please describe yourself. Write a full paragraph of 5-6 sentences or more. Please write from the third-person perspective. Others will read what you wrote; your goal is to convince them it was written by an AGI user without saying so explicitly. For example, do not write a sentence like "I am an AGI user" as this is an explicit statement.

References

Dallas Card, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky, and Dan Jurafsky. 2022. Computational analysis of 140 years of us political speeches reveals more positive but increasingly polarized framing of immigration. *Proceedings of the National Academy of Sciences*, 119(31):e2120510119.

Myra Cheng, Kristina Gligoric, Tiziano Piccardi, and Dan Jurafsky. 2024. AnthroScore: A computational linguistic measure of anthropomorphism. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 807–825, St. Julian's, Malta. Association for Computational Linguistics.