**CS 5740: Natural Language Processing** **(Due: February 21, 11.59 p.m.)**

# Homework #1

Instructors: Tanya Goyal

Submission by (netIDs): kl2235

**Logistics**: Read all the instructions carefully before you start working on the assignment, and before you make a submission.

- This is an additional section, meant only for students enrolled in a graduate version of the course, **CS 5740**. If you are enrolled in CS 4740, DO NOT complete this section.
- All CS 5750 students need to complete this assignment **individually**. This assignment is to be done in addition to the regular HW1 assignment.
- Please refer to the course policies on the course webpage to understand how this extra CS5740 component is graded.
- Please typeset your assignment in LaTeX. You can use this template for your written answers. Please include your netID at the top of this page. (Search for "FILL YOUR NETID" to find the appropriate place to enter your netIDs).
- You will submit this assignment on gradescope. We have created a new course on gradescope and added all CS5750 students in there. Please let the instructors know if you have not been added.
- The University Academic Code of Conduct will be strictly enforced.

## Section C: Natural Language Inferences (S/U points)

In this part of the homework assignment, you will dig deeper into the Natural Language Inferences (NLI) task from the programming component of HW1. As we explain in the ipynb, NLI is a classification task that decides whether a *hypothesis* $h$ can be inferred from a *premise* $p$.

In the assignment, we assumed that entailment decisions are objective, and assigned a binary label to whether the hypothesis is entailed or not. However, there are many confounders that influence human judgments on the relations between $h$ and $p$.

Consider the example below (Pavlick and Kwiatkowski, 2019):

$$p = \textit{Paula swatted the fly.}$$
$$h = \textit{The swatting happened in a forceful manner.}$$

In this case, whether $h$ can be inferred from $p$ depends on the interpretation of the word *forceful*. In fact, if we take ask for annotations from 10 of our friends, it is very likely that we might see disagreements in their entailment judgments.

To understand this further, please read the following paper of research done in the area: Pavlick and Kwiatkowski. "Inherent Disagreements in Human Textual Inferences." Transactions of the Association for Computational Linguistics (TACL) (2019). pdf link

**(1)** Based on the paper, what are the confounders that influence humans when they judge the inference relations in NLI? Identify at least two. How does section 4 reflect one or more of the confounders?

***(Solution)***

As stated in section 3, the first confounder is different heuristics, and the second one is different context interpretation.

In section 4, the author mentioned that if these confounders are noise, then the distribution of labels should be single gaussian, but it was GMM. Hence, we can reflects these confounders through the use of Gaussian

Mixture Models, which indicates disagreements are caused by stable alternative interpretations rather than random noise

**(2)** In section 5, why does softmax distribution misrepresent human disagreements? How can we make the model do better at capturing human disagreements?

***(Solution)***

In current models like BERT, it uses softmax to predict the probabilities of different labels. However, the probabilities indicate how confident the model is, and it can not close to the distribution of human labels.

As the author mentioned, we could improve NLI models by training them to learn the full distribution of human labels. Instead of forcing the model to predict a single "correct" label through majority voting of averaging, the model should be trained to output the observed label distribution. For example, for a (p, h) pair, the model could predict entailment=0.5, neutral=0.5, and contradiction=0.0.

**(3)** According to section 2, NLI has both a formal logic definition (that $h$ is true in every possible world where $p$ is true) and an informal definition (that typical humans will think $h$ is true given $p$). In reality the informal definition has prevailed in the NLP community. What is the tradeoff of using the informal definition over the formal definition when modeling the task and training objective? Does this insight hold for other NLP tasks too?

***(Solution)***

The pros for informal definition: the data is easy to collect (only need people to label the dataset), and this kind of definition is close to the real world application; the cons for formal definition: the label is unstable cause different backgrounds or cultures may have different judgements upon the labels, so the model cannot learn the stable labels.

I think this tradeoff also applies to many other NLP tasks, such as sentiment analysis and summarization, cause these tasks don't have single objectively correct answers. For these tasks, the models could be trained to approximate human judgements rather than the formal definitions.

**(4) Important:** Please specify you use of Generative AI for this assignment. Clearly specify the intent behind the use (understand the paper, polish your answers, write the first draft of your answer, etc.)

If you used GenAI tools in any capacity, please post a link to your chat (you can share view-only access to your chats in ChatGPT, Claude, Gemini, and other popular offerings). If you are unable to share your chat, please list the exact prompt used here.

Note that we will run your written answers through an automatic AI detection tool. In general, we will only penalize answers that were entirely written by GenAI. On the other hand, using these tools to polish your answers is permitted, but we require access to your chat/prompt to distinguish between this case and the former. Please refer to the Gen AI policy on the course webpage for more details.

***(Solution)***

I use ChatGPT to understand the paper and refine my answers.

Example prompts include:

- Help me polish this paragraph for clarity and academic tone;
- Explain more in detail the main argument of Pavlick and Kwiatkowski (2019) to help me understand it more clearly;
- What is the specific meaning of human disagreement? Can it be interpreted as noise?
- Why does modeling the distribution of human labels help show that disagreement is not random noise?