MSBD 5018 Natural Language Processing

Individual Project: Evaluating Language Models

Spring 2023

1 Introduction

In this project, you will evaluate language models (LMs). This project contains two parts: you will test examples and explore different properties of LMs.

- In Section 2, you will evaluate the performance of LMs on natural language inference. You need to improve your performance on this task through prompt engineering.
- In Section 3, you need to measure the bias of LMs on a demographic that interests you. We provide an approach to evaluate the bias of masked language models.

Language Models: Language model list: (i) Masked Language Models: BERT, RoBERTa, DeBERTa; (ii) Causal Language Models: GPT2, XLNet, OPT; (iii) Seq-to-Seq Language Models: BART, T5. In Section 2, please evaluate at least **two** language models on natural language inference from the above list. In Section 3, you will evaluate biases, stereotypes, and associations in at least **two** masked language models from the above list.

There are multiple sizes for each language model, such as `roberta-base` and `roberta-large` (125 million and 355 million parameters, respectively). You can choose any size to conduct the following experiments. Also, some models may have `cased` and `uncased` versions, such as `bert-base-cased` and `bert-base-uncased`. In this project, you can use either version.

Learning Goals: By finishing the project, you will understand: (1) how to evaluate LMs with standard practices, (2) how the evaluation of LMs is sensitive to your design decisions.

2 Capabilities

In this part of the project, you will evaluate the capability of language models on natural language inference. We work with the three-way classification formulation of the task. Each input in the dataset is a pair of contexts (the premise and the hypothesis): the task is to predict whether the hypothesis is entailed (i.e. always true), contradicted (i.e. always false), or neutral (neither entailed nor contradicted) given the premise. Further discussion of this task can be viewed in [1]. We use the MultiNLI dataset [4][1]. MultiNLI is one of the largest corpora available for natural language inference, improving upon available resources in both its coverage and difficulty. MultiNLI accomplishes this by offering data from ten distinct genres of written and spoken English, making it possible to evaluate systems on nearly the full complexity of the language, while supplying an explicit setting for evaluating cross-genre domain adaptation. Performance on this dataset is measured using accuracy, and a random guess would achieve a performance of around 33%. Here we require you to conduct prompting for solving this natural language inference task. For reducing the computational cost for you, we sampled 5000 evaluation samples. There are 2500 samples in the matched setting, and the other 2500 samples in the mismatched setting. For details of these setttings, please refer to [4].

[1] Please use the evaluation set that we provide to you.

Met my first girlfriend that way.	FACE-TO-FACE **contradiction** C C N C	I didn't meet my first girlfriend until later.
8 million in relief in the form of emergency housing.	GOVERNMENT **neutral** N N N N	The 8 million dollars for emergency housing was still not enough to solve the problem.
Now, as children tend their gardens, they have a new appreciation of their relationship to the land, their cultural heritage, and their community.	LETTERS **neutral** N N N N	All of the children love working in their gardens.
At 8:34, the Boston Center controller received a third transmission from American 11	9/11 **entailment** E E E E	The Boston Center controller got a third transmission from American 11.
I am a lacto-vegetarian.	SLATE **neutral** N N E N	I enjoy eating cheese too much to abstain from dairy.
someone else noticed it and i said well i guess that's true and it was somewhat melodious in other words it wasn't just you know it was really funny	TELEPHONE **contradiction** C C C C	No one noticed and it wasn't funny at all.

Figure 1: Randomly chosen examples from the dev set of MultiNLI, shown with their genre labels, their selected gold labels, and the individual annotator labels (abbreviated E, N, C).

2.1 Prompting Approach

For the causal LMs, like GPT2, and seq-to-seq LMs, like T5, the NLI task can be solved by using prompt engineering. We provide a simple three-step procedure for prompting that involves (i) constructing a query given the input x that will be submitted to the language model, (ii) specifying decoding hyperparameters, and (iii) specifying a verbalizer to map from the language model's response to a class label $\hat{y}$. For (i), you can refer to the reference on prompt designs https://stanford-cs324.github.io/winter2022/lectures/capabilities/. For (ii), you should read the paper and the documentation on the models that you choose to use to select the decoding hyper-parameters. For (iii), we briefly explain the definition of verbalizer here. Once you specify a query and decoding hyperparameters, you will give input to the language model and the language will return a response. Depending on what you specify, this may be a probability distribution over the next words, likelihood, or a specific sampled completion. You will then need to transform this response into a predicted label $\hat{y}$. To formalize this, if the label space for the task is L and the LLM's vocabulary is V, we can define the verbalizer $v : L \to V$ that associates each label to a vocabulary entry. A natural choice for the verbalizer is simply mapping each label to its associated string (e.g. mapping the entailment category to "entailment"), though other verbalizers may work better (e.g. using words like true or correct). For example, if the LM returns a probability distribution p over V, the predicted label $\hat{y}$ can be given by:

$$\hat{y} = \arg\max_{y \in L} p(v(y)). \tag{1}$$

2.2 Outline of Deliverables

Measurement Setup. Your report should include:

1. **Method:** The method you select to conduct the experiment on the Multi-NLI task.
2. **Implementation Details:** Please also include some important implementation details, such as model size (base or large) and decoding hyper-parameters. In the prompting approach, please also explain your ideas for how to construct and choose effective prompts.

Experiment Results. Your report should include:

1. **Quantitative Results.** You should include the quantitative results for evaluating your selected language models on both the matched and mismatched development sets of MultiNLI. You do not need to report the performance on the MultiNLI test set as they are not publicly available,

but meanwhile, you should not tune your hyper-parameters or design the prompts directly on these two development sets.

2. **Case Study.** An example of sentence pair (or prompt) should be depicted in a *figure*, demonstrating the capability of conducting natural language inference in both models' generations. You can compare the results from different models in this example to further support your quantitative results.

3 Risks

In this part of the project, you will evaluate the biases, stereotypes, and associations in Masked LMs in relation to social groups. To get started, we briefly describe the measurement method "Minimal Pairs," which focuses on multiple social domains (e.g., race, gender, and political ideology). For language models, you should experiment with two masked language models listed in Section 1.

To keep things interesting and simple, you will choose a group that is important to you (e.g., Hispanic, female, Chinese, LGBT+) and study the social domain containing that group throughout this part. You will measure biased stereotypes related to that domain, especially when mentioned in the text. For example, if the *female* group is important to you, you can select *gender/gender identity* as the domain you study in this part [3].

Minimal Pairs Previous works[2, 3] introduce the evaluation methods called minimal pairs, where a model is presented with a pair of sentences with only a few different words, contrasting a stereotypical association with an astereotypical association. In this way, the logic behind this method is that the discrepancies in the probabilities assigned by a language model to these sentences should identify whether the model has a bias/preference for a particular stereotypical association. Ideally, an unbiased language model should assign probabilities equally (without preference) to the stereotypical sentence and another astereotypical sentence in each pair.

Dataset and Social Domain Please use the dataset "CrowS-Pairs" curated by [3], which recognizes nine social domains for detecting bias and stereotypes in language models: race/color, gender/gender identity, sexual orientation, religion, age, nationality, disability, physical appearance, and socioeconomic status/occupation.

You need to choose a domain that contains the group important to you. You should identify sentence pairs that encode stereotypes related to this domain to evaluate LMs (and abandon sentence pairs related to other domains).

Some domains contain much more examples than others, i.e., race/color, gender/gender identity, socioeconomic status/occupation, nationality, and religion. If you choose one of those, you are allowed sample 80 examples to conduct the experiments.

Metric Please use the pseudo-log-likelihood proposed in Section 3 in the CrowS-Pairs paper [3]. This metric use Masked LMs to compute pseudo-log-likelihood scores for each sentence. Then, this metric measures the percentage of pairs for which a model assigns a higher pseudo-log-likelihood to the stereotyping sentence. A masked LM without any stereotypes shuold achieve the ideal score of 50%.

3.1 Outline of deliverables

Measurement Setup. Your report should include:

1. **Social Group.** The social group and domain you choose. Please also include the total number of examples under the domain you look at.

2. **Metric.** A clear explanation of the metric.

3. **Implementation Details:** Please also include some important implementation details, such as model size (base or large) and decoding hyper-parameters.

Experiment Results. Your report should include:

1. **Quantitative Results.** Quantitative results for evaluating masked language models on your bias evaluation data. You should compare the biases of the models you choose.
2. **Case Study.** Three examples of sentence pairs should be depicted in a *figure*, demonstrating specific biases in both models.

4 Logistics

4.1 Grading Rubrics

- Content clearness.
- Coverage. Whether essential elements (e.g., case study) are covered in the report.
- Insights. Whether there are unique understandings of the field when analyzing cases.
- Plagiarism. We will use Turnitin on canvas to check similarity scores. Penalties will be added to those whose similarity scores are greater than 30%.

4.2 Data Access

All of the datasets you will use in the project can be accessed through the Python package *Hugging Face Datasets*. Documentation and tutorials for how to use this package to download and interact with these datasets can be found at https://huggingface.co/docs/datasets. You can find the list of downloadable datasets at https://huggingface.co/datasets.

4.3 Model Access

All the pre-trained LMs you will use in the project can be acquired using the Python package *Hugging Face Transformers*. Documentation for this package can be found at https://huggingface.co/docs/transformers/index. You also can find documents for each model with example code in the list of supported models at https://huggingface.co/docs/transformers/index#supported-models. Meanwhile, model cards are also a great reference for you to learn how to use each language model, which can be found at https://huggingface.co/models.

4.4 Deliverables

Report. The main deliverable for this project is a report, which should ideally be written in LaTeX or otherwise in Microsoft Word, Pages for Mac, or an equivalent program, and submitted as a PDF. The name of this report should be `MSBD5018_individual_<your_itsc_id>.pdf`. Replace `<your_itsc_id>` with your itsc id without angle brackets.

Your report should properly cite and attribute all papers and resources you used in your project. Furthermore, your report should acknowledge any individuals other than yourself who helped you.

Code. We require that you submit a *.zip* file of your codebase. We do not expect to run the code you use in the project though, if we find plagiarism in the report, we reserve the right to do so. We hope not to have to do this and operate on a trust-based system for this course. The name of this file should be `MSBD5018_individual_<your_itsc_id>.zip`

4.5 Submission

All deliverables should be uploaded to Canvas by 23:59 on 20^{th} May.

References

[1] Samuel R. Bowman et al. "A large annotated corpus for learning natural language inference". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 632–642. DOI: 10.18653/v1/D15-1075. URL: https://aclanthology.org/D15-1075.

[2] Moin Nadeem, Anna Bethke, and Siva Reddy. "StereoSet: Measuring stereotypical bias in pretrained language models". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021, pp. 5356–5371.

[3] Nikita Nangia et al. "CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, pp. 1953–1967.

[4] Adina Williams, Nikita Nangia, and Samuel Bowman. "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 1112–1122. URL: http://aclweb.org/anthology/N18-1101.