COMP 550 Final Project Proposal

Jiaxuan Chen (ID 260886901), Junyu Li (ID 260957495), Xinyi Zhu (ID 260924159)

Research Objective

To perform a comparative analysis between two cross-lingual language models, Multilingual BERT(mBERT) and XLM-RoBERTa (XLM-R), on the multilingual amazon reviews dataset.

Data format

https://github.com/huggingface/datasets/tree/master/datasets/amazon_reviews_multi#dataset-description
The multilingual amazon reviews dataset contains reviews in 6 languages. For each language, the samples are split into training, validation, and test sets. All the data are in JSON files, in which each record contains the review text and title, the star rating, etc. The corpus is balanced across stars, so each star rating constitutes 20% of the reviews in each language. We will use 10k training samples for each language since the whole corpus is too huge.

Overview

mBERT is a transformer-based masked language model pretrained on a large corpus of multilingual data, that can be used for various multilingual tasks, such as cross-lingual natural language inference (XNLI). XLM-R has a similar architecture to mBERT, with some changes namely removing the next sentence prediction object, training with bigger batch sizes, and dynamically changing the masking pattern. [1][2] The previous works have shown that XLM-R outperforms mBERT, especially on low-resource languages[3]. Our main goal is to compare how well these two models perform on a new dataset. The task is to train these two classification models in each language, then test it in itself along with other languages. We will use baseline mean absolute error (MAE) as suggested in previous work [4] to evaluate the performance, instead of classification accuracy, since the star ratings for each review are ordinal.

Experimental procedures

We will first dive into the truncated dataset and perform data preprocessing as seen fit. Unhelpful features will be dropped. The next step of our research is to study the structure, strategies, and implementation of the two multilingual classification models. We will then finetune the models and apply supervised along with the zero-shot cross-lingual transfer for fine-grained classification. Validation sets will be used to make sure our models do not overfit the data. Note that each time we only finetune on data from one source language and test on all languages (source or non-source). In other words, there will be 6*6=36 results. We will report and compare the MAE for the two models we used. In the end, we may use sample growing subsets of the data to explore the impact of dataset size on testing error.

Discussion:

The questions that interest us are:

1. Do the models perform better when the training and testing set are from the same source language?
2. What languages that one model trained on performs better than another model?
3. How does the size of the dataset affect the accuracy of the models?
4. Is the XLM-R model better than mBERT in all ways?
5. What are some strengths or weaknesses of each model?

Reference:

1. Moberg, John. “A Deep Dive into Multilingual NLP Models - Peltarion.” Peltarion, 2020, https://peltarion.com/blog/data-science/a-deep-dive-into-multilingual-nlp-models
2. Libovický, Jindřich et al. “How Language-Neutral is Multilingual BERT?” ArXiv abs/1911.03310 (2019): n. Pag. https://arxiv.org/abs/1911.03310
3. Conneau, Alexis, et al. “Unsupervised Cross-Lingual Representation Learning at Scale.” ArXiv.org, 2019,https://arxiv.org/abs/1911.02116.
4. Keung, Phillip & Lu, Yichao & Szarvas, György & Smith, Noah. (2020). The Multilingual Amazon Reviews Corpus. https://arxiv.org/abs/2010.02573
5. Wang, Cindy, and Michele Banko. “Practical Transformer-Based Multilingual Text Classification.” Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers, 2021, https://doi.org/10.18653/v1/2021.naacl-industry.16.
6. Documentation: https://huggingface.co/transformers/model_doc/xlmroberta.html