Project Proposal - CSCI 7000

Investigating Effectiveness of Cross-Lingual Models in Discourse Knowledge Transfer across Languages

Karan Praharaj

March 11th, 2022

Discourse-level analysis could be interesting to study within the context of large multilingual language models, considering that many of these models are trained with short context spans, such as pairs of adjacent sentences. The goal of this project would be to evaluate how effectively various transformer-based pre-trained cross-lingual models transfer discourse related information across languages in a zero-shot setting. Zero-shot setting refers to the scenario where the model is evaluated on examples in languages that it has not been trained on. In our view, this makes for an interesting and challenging problem, seeing that it brings together two difficult facets of representation learning: cross-linguality and discourse-level analysis.

The models that we intend to use for this analysis are multilingual BERT (Devlin et al., 2019), RemBERT (Chung et al., 2020), XLM (Lample and Conneau, 2019), XLM-RoBERTa (Conneau et al., 2019), mT5 (Xue et al., 2021) and distil-mBERT (Sanh et al., 2019). This list is currently tentative and may change depending on the availability of computational resources. The datasets that will be used for training and evaluation are the Penn Discourse Tree Bank (PDTB) (Prasad et al., 2008) for Implicit Discourse Relation Classification, and Rhetorical Structure Theory Discourse Treebank(RST-DT) (Carlson et al., 2001) for Rhetorical Structure Classification, XNLI (Conneau et al., 2018) for Natural Language Inference, and XQuAD (Artetxe et al., 2019) for Question Answering. The XQuAD dataset is a cross-lingual extension of the SQuAD dataset, and contains translations of the SQuAD development set in 10 languages. The idea is to compare the zero-shot language-wise performance of these models on each dataset, and also an overall performance comparison for each of the tasks, on the entire test sets.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856, 2019. URL http://arxiv.org/abs/1910.11856.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the SIGDIAL 2001 Workshop, The 2nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, Saturday, September 1, 2001 to Sunday, September 2, 2001, Aalborg, Denmark*. The Association for Computer Linguistics, 2001. URL https://aclanthology.org/W01-1605/.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. Rethinking embedding coupling in pre-trained language models. *CoRR*, abs/2010.12821, 2020. URL https://arxiv.org/abs/2010.12821.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019. URL http://arxiv.org/abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291, 2019. URL http://arxiv.org/abs/1901.07291.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie L. Webber. The penn discourse treebank 2.0. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*. European Language Resources Association, 2008. URL http://www.lrec-conf.org/proceedings/lrec2008/summaries/754.html.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.