Topic #2: Massively Multilingual Transformers meet Massively Multilingual Lexical Supervision

Level: Demanding, appropriate for ambitious master students, ideally with prior experience with deep-learning-based natural language processing and deep learning libraries (e.g., PyTorch)

Short description: Pretrained Transformer-based language models are still only distributional in nature and could benefit from clean linguistic knowledge, e.g., relations between words and concepts, that are encoded in external knowledge resources [1, 2]. Injecting linguistic constraints into pretrained LMs like BERT [3] has been mostly limited to monolingual English setup (i.e., monolingual English lexico-semantic knowledge, e.g., from WordNet, injected into a pretrained monolingual English Transformer, e.g., BERT).

Rich massively multilingual lexico-semantic resources such as BabelNet [4], however, do exist, and offer a plethora of *multilingual* lexico-semantic knowledge that could be used (1) to enrich the knowledge stored in pretrained multilingual Transformers such as multilingual BERT, XLM-R [5], or mT5 [6] and (2) to better align the representation spaces of individual languages in the massively multilingual representation space that these transformers span. For the latter, there is evidence that the representation spaces spanned by multilingual transformers are not purely driven by semantics (i.e., meaning) but also by language(s) [7]: accordingly, injecting large-scale cross-lingual knowledge at the lexical level from BabelNet, across a large number of languages, should lead to better semantic alignment of language-specific subspaces of massively multilingual transformers.

This thesis will investigate a number of learning objectives (e.g., feeding monolingual and cross-lingual word pairs into Transformer and predicting lexico-semantic relations from BabelNet; or obtaining node representations from the multilingual Transformer and then predicting nodes in the knowledge graph from the neighbourhood) and regimes (e.g., full vs. adapter-based fine-tuning of pretrained multilingual transformers). The final evaluation of the lexically-enhanced multilingual transformers will be in (zero-shot) cross-lingual transfer for downstream NLP task, with the emphasis on language understanding tasks such as question answering and natural language inference. To this end, we will use the standard benchmarks such as XTREME [8] or XGLUE [9].

References:

[1] Lauscher, A., Vulić, I., Ponti, E. M., Korhonen, A., & Glavaš, G. (2020, December). Specializing Unsupervised Pretraining Models for Word-Level Semantic Similarity. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 1371-1383).

[2] Wang, R., Tang, D., Duan, N., Wei, Z., Huang, X., Cao, G., ... & Zhou, M. (2020). K-adapter: Infusing knowledge into pre-trained models with adapters. arXiv preprint arXiv:2002.01808.

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). 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)* (pp. 4171-4186).

[4] Navigli, R., Bevilacqua, M., Conia, S., Montagnini, D., & Cecconi, F. (2021). Ten years of BabelNet: A survey. In Proc. of International Joint Conference on Artificial Intelligence (IJCAI).

[5] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020, July). Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 8440-8451).

[6] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., ... & Raffel, C. (2021, June). 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 (pp. 483-498).

[7] Cao, S., Kitaev, N., & Klein, D. (2019, September). Multilingual Alignment of Contextual Word Representations. In International Conference on Learning Representations.

[8] Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020, November). Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In International Conference on Machine Learning (pp. 4411-4421). PMLR.

[9] Liang, Y., Duan, N., Gong, Y., Wu, N., Guo, F., Qi, W., ... & Zhou, M. (2020, November). XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 6008-6018).