Topic #1: **Pre-training Multilingual Document Encoders**

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: Multilingual text encoders like multilingual BERT [1] and XLM-R [2] have become the base for cross-lingual transfer for downstream NLP tasks (i.e., we have labeled training data in the source language but limited amount of annotated instances in the target languages). These, as well as the specialized multilingual transformers for sentence-level tasks, such as mUSE [3], LASER [4], or multilingual Sentence-BERT [5], show good transfer performance for sentence-level tasks as well as in unsupervised cross-lingual similarity-based tasks such as cross-lingual information retrieval [6]. However, for tasks at the document level (e.g., document classification; learning to rank, i.e., supervised document retrieval), there still lacks a principled (i.e., pre-trained) model for producing document-level representations.

In this thesis, the task is to pre-train a multilingual document-level encoder, starting from multilingual short-text encoders (e.g., mBERT). Learning a document-level representation from a sequence of sentence/paragraph representations for a specific task has been done monolingually (for English) and through task-specific for tasks like document classification [7, 8] or text segmentation [9]. A second Transformer network is stacked on top of the sentence-level encoder which is trained to provide document representations from a sequence of sentence representations produced by the sentence-level encoder. The goal of this thesis is to pre-train this document-level Transformer for a number of languages, and in a task-agnostic fashion, so that it can be fine-tuned for downstream cross-lingual transfer for any document-level task (e.g., document classification). For the pre-training we would leverage Wikipedia as a source of (near-)comparable document-level data (i.e., we have the articles on the same concept/entity in a number of languages). One would pretrain the document-level encoder, on top of the multilingual sentence/paragraph encoder (e.g., mBERT), by means of contrastive objectives: an objective would force the document-level representation of some Wikipedia article in L1 to be more similar to the representation of the same Wikipedia article in L2 than to representations of other articles in L1, in L2, and other languages (L3, L4, ...).

Once pre-trained, the multilingual document-level encoder would be evaluated in zero-shot downstream transfer for document-level tasks: (1) document classification, and (2) document retrieval. We fine-tune our pre-trained multilingual document encoder on task-specific data in the source language (typically English) and then make predictions on the unseen data in the target language.

References:

[1] 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).

[2] 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).

[3] Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., ... & Kurzweil, R. (2020, July). Multilingual Universal Sentence Encoder for Semantic Retrieval. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 87-94).

[4] Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Transactions of the Association for Computational Linguistics, 7, 597-610.

[5] Reimers, N., & Gurevych, I. (2020, November). Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation. In Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 4512-4525).

[6] Litschko, R., Vulić, I., Ponzetto, S. P., & Glavaš, G. (2021). Evaluating Multilingual Text Encoders for Unsupervised Cross-Lingual Retrieval. Proceedings of the European Conference on Information Retrieval (ECIR) (pp. 342-358)

[7] Yu, J., Jiang, J., Khoo, L. M. S., Chieu, H. L., & Xia, R. (2020). Coupled Hierarchical Transformer for Stance-Aware Rumor Verification in Social Media Conversations. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1392-1401)

[8] Pappagari, R., Zelasko, P., Villalba, J., Carmiel, Y., & Dehak, N. (2019, December). Hierarchical transformers for long document classification. In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 838-844). IEEE.

[9] Glavaš, G., & Somasundaran, S. (2020). Two-Level Transformer and Auxiliary Coherence Modeling for Improved Text Segmentation. Proc. of the AAAI Conference of the Association for Advancement of Artificial Intelligence. (pp. 7797-7804)