Doctoral School of Information and Biomedical Technologies
Polish Academy of Sciences (TIB PAN)

SUBJECT:

Multilingual Discourse Relations Parsing

SUPERVISOR:

dr hab. Maciej Ogrodniczuk (Institute of Computer Science, Polish Academy of Sciences)

DESCRIPTION:

One of the primary challenges in Discourse Relation Parsing is the limited availability of annotated data. The annotation process is intricate, beginning with the segmentation of documents into text spans known as Elementary Discourse Units (EDUs), which must then be connected through semantic-pragmatic relations. This procedure is not only time-consuming but also demands a high level of expertise and linguistic comprehension from the annotators.

To address this issue, works such as [3] explore unsupervised learning methods, utilizing attention masks from Transformer based language models to extract EUDs, while others like [2] focus on semi-supervised approaches utilizing pseudo-labeling and self-training to expand existing annotated datasets. Recent publications, such as [1] have investigated the use of Large Language Models (LLMs) due to their impressive performance in other tasks. However, they found that LLMs struggle with discourse relation parsing, highlighting the necessity for developing dedicated training pipelines.

A common approach to address data scarcity issue found in other NLP fields is to combine data from various languages. This strategy has been particularly successful with Large Language Models. By integrating widely available English data with data from less-resourced languages, researchers have achieved cross-lingual transfer of capabilities. This has been demonstrated in several areas, including pretraining [6] and instruction fine-tuning [5]

The purpose of this study is to investigate cross-lingual transfer specifically in the context of Discourse Relation Parsing. Initially, the study seeks to unify various discourse ontologies to re-annotate existing datasets using the methodology proposed in [4]. This process will result in the creation of a novel, harmonized multilingual discourse dataset.

The candidate is tasked with developing a discourse parser using the harmonized multilingual dataset to explore cross-lingual transfer in discourse analysis. This involves integrating existing state-of-the-art (SOTA) approaches from the literature, including fine-tuning existing pre-trained language models using both supervised and unsupervised methods, as well as proposing novel training pipelines utilizing the latest advances in Deep Learning.

The candidate is expected to have Master's degree in computer science or similar technical field, knowledge of recent natural language processing methods and proficiency in deep learning frameworks such as PyTorch or TensorFlow. Previous research contributions such as publications, grants or patents in the field of LLM construction and fine-tuning are required.

BIBLIOGRAPHY:

[1] Chunkit Chan, Cheng Jiayang, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. Exploring the potential of ChatGPT on sentence level relations: A focus on temporal, causal, and discourse relations. In Yvette Graham and Matthew Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 684–721. Association for Computational Linguistics, March 2024.

[2] Chuyuan Li, Chloé Braud, Maxime Amblard, and Giuseppe Carenini. Discourse relation prediction and discourse parsing in dialogues with minimal supervision. In Michael Strube, Chloe Braud, Christian Hardmeier, Junyi Jessy Li, Sharid Loaiciga, Amir Zeldes, and Chuyuan Li, editors, *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 161–176. Association for Computational Linguistics, March 2024.

[3] Chuyuan Li, Patrick Huber, Wen Xiao, Maxime Amblard, Chloe Braud, and Giuseppe Carenini. Discourse structure extraction from pre-trained and fine-tuned language models in dialogues. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2562–2579. Association for Computational Linguistics, May 2023.

[4] Maciej Ogrodniczuk, Glowińska Katarzyna, Kopeć Mateusz, Agata Savary, and Zawislawska Magdalena. *Coreference. Annotation, Resolution and Evaluation in Polish*. Walter de Gruyter, 2014.

[5] Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. Aya model: An instruction finetuned open-access multilingual language model. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939. Association for Computational Linguistics, August 2024.

[6] 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.