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

SUBJECT:

Adapting Learned MT Metrics to Evaluate Generative Retrieval Systems

SUPERVISOR:

- Harry Scells, PhD (main supervisor);
  University of Tübingen
  harrisen.scells@uni-tuebingen.de

- Wojciech Kusa, PhD (auxiliary supervisor).
  Naukowa i Akademicka Sieć Komputerowa – Państwowy Instytut Badawczy (NASK-PIB), ul. Kolska 12, Warszawa
  wojciech.kusa@nask.pl

DESCRIPTION:

Traditional information-retrieval metrics such as MAP and nDCG fall short when applied to generative retrieval systems, which output free-form answers or directly generated document identifiers. These outputs must be assessed not only for topical relevance but also for faithfulness, grounding, and hallucination. In parallel, the machine-translation community has pioneered learned evaluation metrics (e.g., COMET) that achieve high correlation with human judgments by leveraging multilingual embeddings and explicit reference-source comparisons.

This dissertation will:

- Survey and formalize the landscape of generative-IR evaluation, identifying key gaps in measuring factuality, grounding, and user-centric utility.
- Extend and adapt a COMET-style neural framework to the generative-IR setting by designing input representations that combine the user query, the model's generated response, and retrieved context passages or document identifiers.
- Collect and annotate a benchmark dataset of generative-IR outputs with human judgments on faithfulness, relevance, and utility, enabling supervised training and evaluation of learned metrics.
- Train and fine-tune the adapted metric model on this dataset, exploring architectures that integrate entailment modules, contextual embeddings from the retrieval corpus, and explicit grounding signals.
- Evaluate the resulting learned metric against traditional IR measures and embedding-based proxies on standard benchmarks (e.g., TREC Gen, BEIR), analyzing correlations with human annotations and conducting ablation studies to pinpoint the most informative features.

This work will offer the comprehensive framework for ML-driven evaluation of generative retrieval, providing researchers and practitioners with reliable, high-fidelity metrics that reflect both the correctness and practical usefulness of generative IR outputs.

REQUIREMENTS:

- MSc degree in Computer Science, Information Retrieval, Computational Linguistics, or a related field
- Proficiency in Python and deep-learning frameworks (e.g., PyTorch, TensorFlow)
- Experience with IR evaluation toolkits and annotation platforms
- Familiarity with LLMs, sequence-to-sequence architectures, and learned evaluation metrics (e.g., COMET, BLEURT)
- Strong background in experimental design, statistics, and correlation analysis
- Advanced English skills (written and spoken)

BIBLIOGRAPHY:

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. *arXiv preprint arXiv:2009.09025*.

Leiter, C., Lertvittayakumjorn, P., Fomicheva, M., Zhao, W., Gao, Y., & Eger, S. (2024). Towards explainable evaluation metrics for machine translation. *Journal of Machine Learning Research*, *25*(75), 1-49.

Sellam, T., Das, D., & Parikh, A. P. (2020). BLEURT: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Alaofi, Marwah, et al. "Generative information retrieval evaluation." *Information Access in the Era of Generative AI*. Cham: Springer Nature Switzerland, 2024. 135-159.

Thakur, Naman, et al. “BEIR: A Heterogeneous Benchmark for Zero-Shot Evaluation of Information Retrieval Models.” In Proceedings of the 44th International ACM SIGIR Conference (SIGIR’21), pp. 325–334. 2021.