

Improving Linguistic Variations in a Synthetic Compositional Reasoning Benchmark

Semantic Computing Group
David M. Schmidt
daschmidt@techfak.uni-bielefeld.de

In light of rapidly increasing capabilities of large language models (LLMs), some even proclaiming the “age of LLMs”, the question arises where the limits of current LLMs lie. The reasoning abilities of LLMs and especially the abilities of LLMs to work and reason in a compositional way have been investigated by numerous related works in recent years, many of them indicating fundamental limitations in LLMs when it comes to truly compositional tasks [1]. Recent work [2] focusing on question answering over linked data (QALD), i.e., the generation of SPARQL queries for given natural language questions, reached a similar conclusion.

However, the benchmark proposed in [2] is synthetic and limited w.r.t. linguistic variations of the respective questions. Therefore, in this thesis, you will explore different approaches to enrich the CompoST benchmark and make the questions sound more natural without violating the specific properties of the dataset (i.e., that certain benchmark items together contain all necessary information to answer another question). This will include extending the existing Lemon-based [3] SPARQL verbalization as well as potential LLM-based techniques.

Prior knowledge regarding LLMs, semantic web, QALD, or prompt optimization techniques is helpful but not required. Existing code from [2] is written in Python, so you should be familiar with Python. The thesis can be written in English or German.

Related literature

- [1] Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., West, P., Bhagavatula, C., Bras, R. L., Hwang, J. D., Sanyal, S., Welleck, S., Ren, X., Ettinger, A., Harchaoui, Z., & Choi, Y. (2023). Faith and fate: Limits of transformers on compositionality. arXiv. <https://doi.org/10.48550/ARXIV.2305.18654>
- [2] Schmidt, D. M., Schubert, R., & Cimiano, P. (2026). Compost: A benchmark for analyzing the ability of llms to compositionally interpret questions in a qald setting. In D. Garijo, S. Kirrane, A. Salatino, C. Shimizu, M. Acosta, A. G. Nuzzolese, S. Ferrada, T. Soulard, K. Kozaki, H. Takeda, & A. L. Gentile (Eds.), *The Semantic Web – ISWC 2025* (Vol. 16140, pp. 3–22). Springer Nature Switzerland. <https://doi.org/10.48550/arXiv.2507.21257>
- [3] <https://lemon-model.net/>

The Semantic Computing Group researches and develops methods that enable machines to acquire relevant knowledge as well as linguistic capabilities. Using methods from *natural language understanding* and *machine learning*, we are aiming at machines that are capable of knowledge acquisition by reading unstructured textual data. In particular, the group focuses on methods for information extraction, semantic parsing, ontology learning, sentiment analysis, entity linking, as well as question answering.

More information is available at:

<https://www.uni-bielefeld.de/fakultaeten/technische-fakultaet/arbeitsgruppen/semantic-computing>

Interested? @mail to daschmidt@techfak.uni-bielefeld.de