Impact of Multilingual Pretraining on Cross-Language Transferability for Source Languages

Julio Perez
Matriculation number: 473816
perezduranjulio@gmail.com

Advisor:	Fabio Barth
Examiners:	Prof. Dr.-Ing. Sebastian Möller, Dr. Georg Rehm
Planned period:	08-2024 - 03-2025

1 Scientific Background

The emergence of Large Language Models (LLMs) has revolutionized natural language processing, effectively addressing diverse problems (Thirunavukarasu et al., 2023). Most LLMs are English-centered, primarily built with English data. Given that approximately 50% of all web content is in non-English languages (Petrosyan, 2024), a significant user base may face obstacles when relying solely on English-based models for tasks requiring multilingual capabilities. In recent years, cross-lingual transferability across different tasks has shown great success (Singh et al., 2024). The goal is to transfer the skills learned in specific tasks from one language to another, enabling the same abilities to be applied across different languages. However, current research remains limited and predominantly focused on target languages (Singh et al., 2024; Rathore et al., 2023; Yamaguchi et al., 2024).

Another approach to bridging the challenge of primarily English-based training corpora is further pretraining with multilingual data. Among the various multilingual language models, Occiglot exemplifies this technique. It is a further pre-trained version of Mistral, using data from five major European languages: English, German, French, Spanish, and Italian. This initiative also addresses the dominance of large technology companies in the LLM field, which often do not provide public access to pretraining data or full access to source code (Occiglot Team, 2024).

In this thesis, we aim to analyze the language transferability of the Occiglot model collection by evaluating it on the task of translating multilingual language inputs into SPARQL queries, the primary query language of the Semantic Web (The World Wide Web Consortium (W3C), 2008)). The Semantic Web aims to organize and retrieve information using standardized protocols, enhancing cross-platform integration. Ontologies enable consistent data representation across systems (van Harmelen, 2004), and various ontologies and knowledge graphs represent complex relationships across the web, facilitating structured access to vast amounts of data (Hogan et al., 2021). However, accessing this data requires proficiency in query languages like SPARQL, posing challenges for non-expert users unfamiliar with these technical languages. While efforts have been made to bridge this gap (González-Mora et al., 2020), existing approaches often target specific user groups, require significant intermediate preprocessing (Dannélls et al., 2013), or restrict user input (Kaufmann and Bernstein, 2010).

In contrast to previous research, this work focuses on the source language rather than the target language in the context of translation tasks. Specifically, German constitutes the source language, and SPARQL the target language. This research proposes fine-tuning Occiglot using publicly available datasets for non-English SPARQL query translation. The results will be analyzed and compared with those of Mistral

7B, extending the study by Rangel et al. (Rangel et al., 2024).

2 Scientific Goals

This research builds on Yin et al. (2021), who explored transformer architectures for text-to-SPARQL translation. We extend their work by evaluating the effectiveness of LLMs in generating SPARQL queries from German, providing valuable insights into multilingual LLM capabilities and cross-lingual transferability (Yin et al., 2021).

The goals for this work are:

1. **Data Curation and Preparation:** Generate two training data sets, one in German and the other in English:
   (a) A combined dataset of German questions and corresponding SPARQL queries on DBpedia from the following QALD datasets: 3 (Cimiano et al., 2013), 5 (Unger et al., 2015), 6 (Unger et al., 2016), 7 (Usbeck et al., 2017), 9 (Usbeck et al., 2018), multilingual datasets from the ESWC shared task challenge, curated by native speakers.
   (b) A second dataset containing the same questions and corresponding SPARQL queries, but each question in English.
   (c) Format these datasets in a question-and-answer format to train the LLMs using supervised learning.

2. **Multilingual SPARQL Generation:** Fine-tune Occiglot and Mistral 7-B models to generate SPARQL queries from natural language inputs in both German and English. This will be done separately for each language using the efficient Low-Rank Adaptation (LoRA) fine-tuning method (Hu et al., 2021). Additionally, hyperparameters will be optimized as described in (Feurer and Hutter, 2019). This process will involve four distinct fine-tuning tasks: one for each model and language combination.

3. **Comparative Analysis:** Conduct a comparative analysis of the performance of both models fine-tuned on English and German data in generating SPARQL queries from German. The models' performance will be evaluated within each language using the F1 score. This assessment will measure the impact of pretraining on language gaps, replicating and extending the findings of Rangel et al. (2024) (Rangel et al., 2024).

3 Research Question

Comparative Performance of Fine-tuned LLMs for Multilingual SPARQL Generation

This research addresses the following question:

- RQ: How does further multilingual pretraining affect Mistral 7-B cross-language transferability and performance on the task of SPARQL generation from text?

This comparison will evaluate:

- The cross-lingual transferability of Occiglot and Mistral: By comparing fine-tuned models on mixed-language and German-only datasets, we can measure the performance gap for text-to-SPARQL tasks.

- The impact of multilingual LLM pretraining on cross-lingual transferability: This will be assessed by comparing the differences in results between Occiglot and Mistral.

By exploring these aspects, we aim to advance multilingual text-to-SPARQL generation and the effectiveness of open-source LLMs.

4 Realization Plan

4.1 Data Acquisition and Preparation

Using supervised learning, we require a dataset of natural language queries paired with SPARQL expressions. We will use the QALD datasets (QALD 3, 5, 6, 7, 9)

(Cimiano et al., 2013; Unger et al., 2015, 2016; Usbeck et al., 2017, 2018), focusing on DBpedia, a multilingual dataset from the ESWC shared task challenges 3, 5, 6, 7 and 9. These datasets, containing 1648 multilingual questions and SPARQL queries from DBPedia, are curated by native speakers. First, we build two datasets from this with the same number of data points each. We will extract all German text and SPARQL query pairs, and also extract their corresponding English versions separately.

4.2 Model Fine-tuning

Our approach involves fine-tuning Occiglot and Mistral on the two datasets separately using a supervised approach. This process involves adapting model parameters for the text-to-SPARQL task, utilizing question-answer pairs where the questions are in natural language and the answers are the corresponding SPARQL queries. Specifically, we will employ the LORA fine-tuning strategy, as implemented in the PEFT library by Huggingface (Yu et al., 2023). This strategy, which offers results comparable to full parameter fine-tuning while requiring less computational power, is a key part of our methodology (Dettmers et al., 2024). Additionally, we will be using the Tree-structured Parzen estimator algorithm to identify the optimal hyperparameters for our fine-tuning tasks, leveraging a tool such as Optuna (Bergstra et al., 2011; Optuna Team, 2024).

4.3 Evaluation Metrics and Benchmarking

Effectiveness will be measured using the f1 score, used for the QALD challenges (Usbeck et al., 2019).

By implementing these strategies, we aim to comprehensively understand the effectiveness of multilingual LLMs for text-to-SPARQL generation and the role of cross-lingual transfer in the task.

4.4 Timeline

- Literature Review: Expand the research context by reviewing additional literature and writing the introduction chapter. (6 weeks)
- Data extraction, preparation, and writing the data collection chapter. (6 weeks)
- Fine-tuning of models, collecting results, and writing the analysis chapter. (6 weeks)
- Write the conclusion chapter and abstract. (3 weeks)
- Preparation of presentation and defense. (5 weeks)

References

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24.

Philipp Cimiano, Vanessa Lopez, Christina Unger, Elena Cabrio, Axel-Cyrille Ngonga Ngomo, and Sebastian Walter. 2013. Multilingual question answering over linked data (qald-3): Lab overview. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 321–332, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dana Dannélls, Aarne Ranta, Ramona Enache, Mariana Damova, and Maria Mateva. 2013. Multilingual access to cultural heritage content on the semantic web. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 107–115.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Matthias Feurer and Frank Hutter. 2019. Hyperparameter optimization. *Automated machine learning: Methods, systems, challenges*, pages 3–33.

César González-Mora, Irene Garrigós, and Jose Zubcoff. 2020. An apification approach to facilitate the access and reuse of open data. In *Web Engineering*, pages 512–518, Cham. Springer International Publishing.

F. van Harmelen. 2004. The semantic web: What, why, how, and when. *IEEE Distributed Systems Online*, 5(03):4.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. Knowledge graphs. *ACM Comput. Surv.*, 54(4).

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Esther Kaufmann and Abraham Bernstein. 2010. Evaluating the usability of natural language query languages and interfaces to semantic web knowledge bases. *Journal of Web Semantics*, 8(4):377–393. Semantic Web Challenge 2009 User Interaction in Semantic Web research.

Occiglot Team. 2024. Announcing Occiglot: Polyglot Language Models for the Occident. https://occiglot.eu/posts/occiglot-announcement/. Online; accessed 03 June 2024.

Optuna Team. 2024. Optuna: A hyperparameter optimization framework. Accessed: 2024-07-04.

Ani Petrosyan. 2024. Common languages used for web content 2024, by share of websites. Accessed: 2024-06-20.

Julio C. Rangel, Tarcisio Mendes de Farias, Ana Claudia Sima, and Norio Kobayashi. 2024. Sparql generation: an analysis on fine-tuning openllama for question answering over a life science knowledge graph.

Vipul Rathore, Rajdeep Dhingra, Parag Singla, and Mausam. 2023. ZGUL: Zero-shot generalization to unseen languages using multi-source ensembling of language adapters. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6969–6987, Singapore. Association for Computational Linguistics.

Vaibhav Singh, Amrith Krishna, Karthika NJ, and Ganesh Ramakrishnan. 2024. A three-pronged approach to cross-lingual adaptation with multilingual llms. *arXiv preprint arXiv:2406.17377*.

The World Wide Web Consortium (W3C). 2008. SPARQL Query Language for RDF. https://www.w3.org/TR/rdf-sparql-query/. Online; accessed 02 June 2024.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature Medicine*, 29(8):1930–1940.

Christina Unger, Corina Forascu, Vanessa López, Axel-Cyrille Ngonga Ngomo, Elena Cabrio, Philipp Cimiano, and Sebastian Walter. 2015. Question answering over linked data (QALD-5). In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015*, volume 1391 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Christina Unger, Axel-Cyrille Ngonga Ngomo, and Elena Cabrio. 2016. 6th open challenge on question answering over linked data (qald-6). In *Semantic Web Challenges*, pages 171–177, Cham. Springer International Publishing.

Ricardo Usbeck, Ria Gusmita, Muhammad Saleem, and Axel-Cyrille Ngonga Ngomo. 2018. 9th challenge on question answering over linked data (qald-9).

Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Bastian Haarmann, Anastasia Krithara, Michael Röder, and Giulio Napolitano. 2017. 7th open challenge on question answering over linked data (qald-7). In *Semantic Web Challenges: 4th SemWebEval Challenge at ESWC 2017, Portoroz, Slovenia, May 28-June 1, 2017, Revised Selected Papers*, pages 59–69. Springer.

Ricardo Usbeck, Michael Röder, Michael Hoffmann, Felix Conrads, Jonathan Huthmann, Axel-Cyrille Ngonga-Ngomo, Christian Demmler, and Christina Unger. 2019. Benchmarking question answering systems. *Semantic Web*, 10(2):293–304.

Atsuki Yamaguchi, Aline Villavicencio, and Nikolaos Aletras. 2024. Vocabulary expansion for low-resource cross-lingual transfer.

Xiaoyu Yin, Dagmar Gromann, and Sebastian Rudolph. 2021. Neural machine translating from natural language to sparql. *Future Generation Computer Systems*, 117:510–519.

Yu Yu, Chao-Han Huck Yang, Jari Kolehmainen, Prashanth G Shivakumar, Yile Gu, Sungho Ryu Roger Ren, Qi Luo, Aditya Gourav, I-Fan Chen, Yi-Chieh Liu, et al. 2023. Low-rank adaptation of large language model rescoring for parameter-efficient speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.