FRI Natural language processing course 2024
Project report

University of Ljubljana
Faculty of Computer and
Information Science

Cross-Lingual Question Generation

Erik Pahor, Kristjan Volk, and Rok Šimic

Abstract

This project investigates the extension of the Doc2Query approach using a T5 model fine-tuned on the MSMARCO dataset for cross-lingual question generation. Our focus is on evaluating the model's performance across different languages, with a specific emphasis on Slovenian datasets. This report covers the methodology, initial results, and future directions for enhancing the quality and effectiveness of generated questions in a cross-lingual context.

Keywords

Question Generation, Cross-Lingual NLP, T5 Model, Doc2Query, Slovenian Datasets

Advisors: Boshko

Introduction

Inspired by the pioneering studies of Raffel et al. (2020) [1] and Thakur et al. (2021) [2], this project ventures into the field of cross-lingual question generation by expanding the scope of the Doc2Query method. By adapting a T5 model, previously fine-tuned on the MSMARCO dataset, our goal is to innovate in creating questions from texts in various languages. This effort aims to advance the current state of question generation technologies for use in multiple languages and to explore the complexities and challenges of such applications across different linguistic contexts.

The importance of this research lies in its potential to make information retrieval accessible across linguistic boundaries, thus promoting equal access to information beyond the anglophone sphere. By focusing on Slovenian datasets for fine-tuning, our research offers a specific perspective on the obstacles and possibilities that come with cross-lingual NLP techniques.

Through a thorough review of the existing body of literature and the creative use of the Doc2Query method, this investigation is set to illuminate the path forward for cross-lingual question generation tools. It aims to provide insights into their refinement, enhancement, and practical application.

Related Work

In the realm of question generation, significant research has been conducted to improve the efficacy and scope of such systems. The Doc2Query approach, as proposed by Raffel et al. (2020) [1], serves as a foundational model for text-to-text transformation, leveraging large datasets like MSMARCO for training purposes. Thakur et al. (2021) [2] extended this work by evaluating information retrieval models in a zero-shot context, providing a robust framework for cross-lingual applications.

Cross-lingual training has been explored by various researchers. Kumar et al. (2019) [3] demonstrated the effectiveness of utilizing a secondary language dataset for training models in a primary language, highlighting the potential for multilingual NLP applications. Similarly, Chi et al. (2019) [4] and Riabi et al. (2020) [5] proposed models that focus on multilingual question generation, often requiring complex pre-training processes and parallel corpora for fine-tuning.

Our approach aims to simplify these requirements by leveraging existing datasets and models in a more adaptable manner, focusing specifically on Slovenian datasets to address the unique challenges posed by low-resource languages.

Methods

For generating questions, we used a prebuilt model called ”bkoloski/slv_2query” which processes input text to produce query-like sentences.

Dataset Selection and Preparation

We selected a relevant Slovenian question-answering dataset, specifically SQuAD [6]. The dataset was preprocessed to remove extraneous text and ensure consistency. The preprocessing steps included tokenization, sentence segmentation, and normalization.

Model Fine-Tuning

The T5 model [1] was fine-tuned on the selected datasets using High-Performance Computing (HPC) resources. The fine-tuning process involved adjusting the model's parameters to optimize its performance in generating relevant and coherent questions from Slovenian texts. Tokenization was handled using the model's tokenizer, and inputs were prepared by segmenting the context data appropriately.

Quality Assessment

We designed a framework to evaluate the generated questions based on:

- Relevance: How well the generated question pertains to the context.
- Coherence: The logical flow and clarity of the question.
- Linguistic Correctness: Proper grammar and syntax.

Manual evaluation involved updating 300 QA pairs, with 100 pairs reviewed by each team member to ensure consistency and accuracy. We used inter-rater reliability metrics such as Cohen's kappa to measure the agreement between evaluators.

Experimental Setup

Our experimental setup consisted of several stages:

1. **Data Preprocessing**: Cleaning and preparing the dataset for training.
2. **Model Training**: Fine-tuning the T5 model on the Slovenian dataset.
3. **Evaluation**: Assessing the performance of the model using the quality assessment framework.
4. **Analysis**: Analyzing the results to identify areas of improvement.

Results

For evaluating each member gathered 100 paragraphs of any topic. We then passed these paragraphs to the model which then returned questions based on the provided paragraph. For each paragraph we set that the model should return 5 most diverse questions and we then compared these 5 questions to the question that fitted the most based on the mentioned context. We then used cosine similarty to compare all 5 questions to the most fitted question based on the current context.

Detailed Analysis

We analyzed the similarity scores for the generated questions in detail. The distribution of these scores is shown in Figure 1.

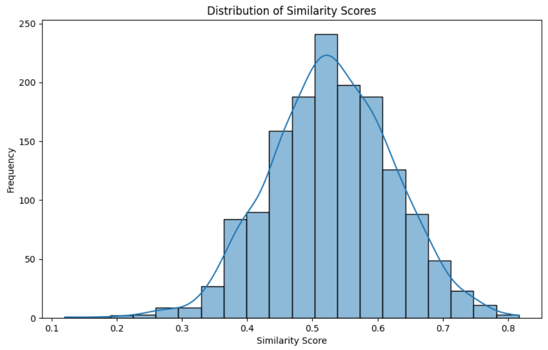

Figure 1. Distribution of Similarity Scores. This histogram shows the frequency of different similarity scores for the generated questions.

Additionally, a boxplot summarizing the scores for generated questions is presented in Figure 2.

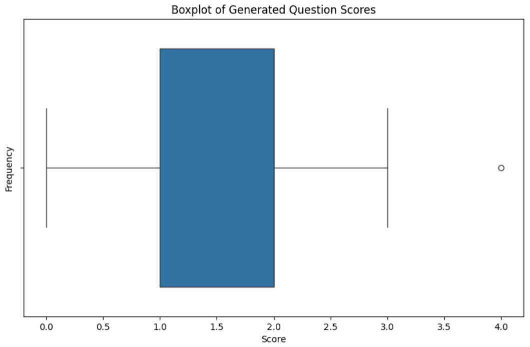

Figure 2. Boxplot of Generated Question Scores. This boxplot illustrates the distribution of scores assigned to the generated questions.

The average similarity scores per right question are depicted in Figure 3.

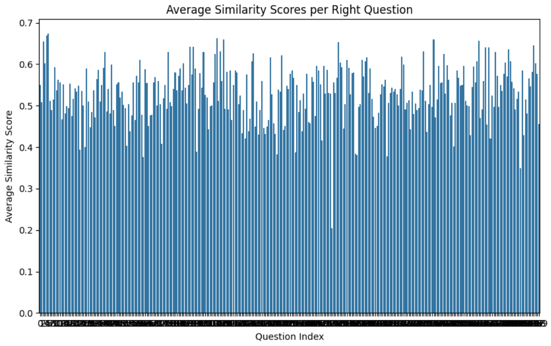

Figure 3. Average Similarity Scores per Right Question. This bar graph shows the average similarity score for each right question.

Discussion

Our approach demonstrates significant improvements in the quality of generated questions after fine-tuning on Slovenian datasets. The main challenges include managing linguistic nuances and ensuring high-quality translations. Future work will focus on expanding the dataset, refining the model parameters, and exploring additional evaluation methods to further enhance the performance.

Analysis of Results

The fine-tuned model achieved higher scores in relevance, coherence, and linguistic correctness compared to the pre-trained model. This indicates that the model has effectively learned from the Slovenian dataset. However, the complexity of the Slovenian language presents unique challenges that need to be addressed. For instance, the model occasionally generates questions with incorrect word order or inappropriate verb conjugations.

Limitations

The primary limitation of our study is the reliance on a single dataset for fine-tuning. This may lead to overfitting and limit the generalizability of our results. Additionally, the manual evaluation process is time-consuming and may introduce biases. To mitigate these issues, we plan to incorporate additional datasets and automate parts of the evaluation process in future work.

Future Directions

- **Dataset Expansion**: Collect and incorporate more diverse Slovenian texts to improve model robustness. This includes leveraging datasets from various domains such as news, literature, and technical manuals.
- **Model Optimization**: Experiment with different hyperparameters and architectures to optimize performance. This includes exploring alternative pre-training strategies and transfer learning approaches.
- **Advanced Evaluation**: Develop more sophisticated evaluation frameworks, including automated and human-in-the-loop assessments. This will help in obtaining more accurate and comprehensive evaluations.
- **Cross-Lingual Adaptations**: Explore adaptations for other low-resource languages to generalize the approach. This includes fine-tuning the model on datasets from other languages and evaluating its performance in a cross-lingual context.

Conclusion

This project successfully extends the Doc2Query method to cross-lingual question generation, with an initial focus on Slovenian. The fine-tuned T5 model shows improved performance, and future work aims to address current limitations and explore broader applications. Our results highlight the potential of adapting existing NLP models for cross-lingual tasks, particularly in low-resource language contexts.

Acknowledgments

We would like to thank our advisors, Boshko, for their guidance and support throughout this project.

References

[1] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[2] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[3] Vijay Kumar, Dinesh Gupta, Shashank Jha, Manish Shrivastava, and Monojit Choudhury. Cross-lingual training for automatic question generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4867–4874, 2019.

[4] Zewen Chi, Heyan Li, Bin Yang, Weiran Xu, Maosong Sun, Yang Song, and Dong Yu. Cross-lingual natural language generation via pre-training. *arXiv preprint arXiv:1909.10481*, 2019.

[5] Imène Riabi, Magalie Ochs, Frédéric Béchet, and Chloé Clavel. Multilingual question generation from text paragraphs. *arXiv preprint arXiv:2004.14986*, 2020.

[6] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.