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) and Thakur et al. (2021), 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.

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. The dataset was preprocessed to remove extraneous text and ensure consistency.

Model Fine-Tuning

The T5 model 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.

Results

Our initial experiments with the fine-tuned T5 model show promising results in generating coherent and contextually relevant questions in Slovenian. The evaluation metrics used and the corresponding results are summarized in Table **??**.

Table 1. Evaluation Metrics for Question Generation

Metric	Pre-trained Model	Fine-tuned Model
Relevance	75%	85%
Coherence	70%	80%
Linguistic Correctness	65%	78%

While the fine-tuned model shows improvement across all metrics, challenges remain in handling nuanced linguistic differences and ensuring the accuracy of translations.

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.

Future Directions

- **Dataset Expansion**: Collect and incorporate more diverse Slovenian texts to improve model robustness.
- **Model Optimization**: Experiment with different hyperparameters and architectures to optimize performance.
- **Advanced Evaluation**: Develop more sophisticated evaluation frameworks, including automated and human-in-the-loop assessments.
- **Cross-Lingual Adaptations**: Explore adaptations for other low-resource languages to generalize the approach.

Acknowledgments

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

References