University *of Ljubljana*
Faculty *of Computer and Information Science*

Paraphrasing sentences

Gregor Gabrovšek, Miha Marinko, Andraž Andolšek Peklaj

Abstract

During this phase of our project, we focused on constructing a dataset using abstracts from diploma theses, which we utilized to train a text-to-text model for generating paraphrases. However, the results of our initial attempts to generate high-quality paraphrases were unsatisfactory. Further research and experimentation are necessary to enhance the model's performance and produce more accurate and meaningful paraphrases.

Keywords
paraphrase, language, generation

Advisors: Slavko Žitnik

Introduction

Paraphrasing is the process of restating a sentence or text in different words, while preserving its meaning and essence. Paraphrasing is an essential skill in many contexts, such as writing, journalism, and academic research, as it allows writers to express the same ideas in different ways, avoid plagiarism, and adapt to different audiences.

This paper[1] distinguishes traditional and neural methods for paraphrase generation. Traditional methods rely on hand-crafted rules or templates to generate paraphrases based on linguistic knowledge or external resources. Neural methods use deep learning models to learn paraphrase generation from large-scale parallel data or pre-trained language models. The paper also compares different types of data sources for paraphrase generation, such as parallel corpora, monolingual corpora, question answering datasets and web queries. It also introduces different evaluation metrics for paraphrase generation, such as automatic metrics (e.g., BLEU), human evaluation (e.g., fluency) and extrinsic metrics (e.g., downstream tasks). [2] [3]

This project aims to develop a natural language processing model that can automatically paraphrase Slovenian language sentences. We will construct a dataset of paraphrases and use it to train a sequence-to-sequence model.

1. Methods

1.1 Dataset construction

For our project, we developed a dataset consisting of abstracts from University of Ljubljana diploma theses along with their paraphrases.

To obtain the textual data, we have implemented a specialized web scraper to collect abstracts in both Slovenian and English from the Repozitorij Univerze v Ljubljani website[4]. The program required a range of pages as a parameter and then proceeded to iterate over them. The scraper extracted links to each thesis page and saved the Slovenian and English abstracts in separate files, with the thesis ID serving as the filename.

We then used Neural Machine Translation model for Slovene - English language pair RSDO-DS4-NMT 1.2.6 to translate the English abstracts into Slovenian and obtain two similar paragraphs that we considered to be paraphrases. We have made our dataset available to other teams to facilitate better results[5].

1.2 Model Training

We made an attempt to fine-tune multiple different text-to-text models, including the t5-base, google/mt5-base and cjvt/t5-sl-small which are avaliable on Hugging Face. Unfortunately, many of our training efforts did not yield satisfactory outcomes. For instance, after one epoch of training Google's mT5-base model on our dataset, the resulting paraphrases were indistinguishable from those generated by the original model, and not considered good.

Our best outcomes were obtained by fine-tuning the T5-base model on our dataset. The model produces coherent text, but the output bears too much resemblance to the input text, with only a few altered words. Going forward, we plan to explore methods of penalizing the model for generating paraphrases that are too similar to the original.

1.3 Performance Evaluation

Until now, we have solely relied on human inspection to evaluate the performance of our model. The generated results had apparent flaws and thus, did not require any artificial tests on larger samples for confirmation. Still, the model managed to produce a number of different outputs, each being at least of satisfactory quality. Further work is needed to improve the model's performance and generate more accurate and meaningful paraphrases, as the paraphrases are currently still a bit too similar to the original phrase.

References

[1] Jianing Zhou and Suma Bhat. Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5086, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[2] Lingfeng Shen, Lemao Liu, Haiyun Jiang, and Shuming Shi. On the evaluation metrics for paraphrase generation, 2022.

[3] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011.

[4] Digital repository of slovenia. Accessed: 23.04.2023.

[5] Group 7. RUL: Paragraph-level paraphrasing dataset of academic abstracts from University of Ljubljana, 2023.