FRI Natural language processing course 2023
Project report

University of Ljubljana
Faculty of Computer and Information Science

Paraphrasing sentences

Gregor Gabrovšek, Miha Marinko, Andraž Andolšek Peklaj

Abstract

To be abstracted.

Keywords
paraphrase, sentence, 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

We will create a dataset consisting of pairs of sentences or paraphrases. We will obtain the initial sentences from the ccGigaFida dataset[4] and generate potential paraphrases using backtranslation (translating from Slovene to English and back). Any differences between the starting and ending sentences (even just a single word) will indicate that they are paraphrases. To increase the size of our dataset and add more variation, we will also attempt to translate an existing Twitter paraphrase dataset[5] into Slovenian. Another option that we will use is to translate the TaPaCo dataset on Huggingface [6], which includes more than 200 thousand paraphrased sentences in English. We will evaluate the quality of the translated paraphrases to determine if they are suitable for use in our model.

1.2 Model Training

We will train a sequence-to-sequence model using simpletransformers from Huggingface with our dataset. One sentence in each pair will serve as the input during training, while the other will serve as the target text. We will use the Google's mt5-base model available on Huggingface [7]. The model will be fine-tuned on our constructed dataset to learn how to paraphrase Slovenian sentences effectively.

1.3 Performance Evaluation

To assess the model's performance, we will compare the similarity between sentences in terms of meaning. This evaluation method will help us determine how well the model can paraphrase sentences in Slovenian. We will use various metrics such as BLEU score, PINC score, ROUGE score, and METEOR score to evaluate the model's performance. We will also perform a manual evaluation of a smaller sample to get an idea of how well the model performs. Automatic tests cannot capture all important features of paraphrase, such as similarity, clarity, and fluency.

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] Nataša Logar, Tomaž Erjavec, Simon Krek, Miha Grčar, and Peter Holozan. Written corpus ccGigafida 1.0, 2013. Slovenian language resource repository CLARIN.SI.

[5] Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. A continuously growing dataset of sentential paraphrases. In *Proceedings of The 2017 Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pages 1235–1245. Association for Computational Linguistics, 2017.

[6] Yves Scherrer. TaPaCo: A Corpus of Sentential Paraphrases for 73 Languages, March 2020.

[7] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer, 2021.