FRI Natural language processing course 2021
Project report

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

Paraphrasing sentences

Franc Benjamin Demšar, Lenart Rupnik, Gregor Zadnik

Abstract

In this report we describe the basic idea of a pipeline for generating a model which paraphrases slovenian sentences. We incorporate ideas from related works and present our initial datasets.

Keywords

Paraphrase, Translation, Slovenian, English

Advisors: Slavko Žitnik

1. Introduction

The task of paraphrasing sentences can be tackled in many ways. Traditional approaches include rule-based and thesaurus-based, while state-of-the-art approaches rely on neural networks [1].

Our approach uses back-translation to generate paraphrases from our chosen Gigafida [2] Slovenian dataset. We will then use this combined data to learn generative models which generates a paraphrase for the input sentence.

2. Related works

For a general overview of the task of paraphrasing we turned to Zhou and Bhat, 2021 [1]. We decided to use the translation method to generate training data.

Federmann et al., 2019 [3] explore the possible approaches to translating text. They found that Neural machine translation (NMT) performs better than human translation in many cases. It is also mostly free which makes it a good option for us students. Another important finding was that paraphrasing works best when translating into and from related languages.

For our evaluation metrics we drew inspiration from Shen et al., 2022 [4] who compare existing paraphrasing metrics and propose their own: ParaScore. We will evaluate the model automatically with their proposed metrics (METEOR, BLEU-4, Rouge-1) and manually with subjective metrics such as adequacy, diversity and fluency.

3. Methodology

4. Links

- https://viri.cjvt.si/gigafida/
- https://aclanthology.org/D19-5503.pdf
- https://www.kaggle.com/datasets/aemreusta/paraphrased-articles-using-gpt3
- https://github.com/afader/oqa
- https://slovenscina.eu/prevajalnik

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] Gigafida.

[3] Christian Federmann, Oussama Elachqar, and Chris Quirk. Multilingual whispers: Generating paraphrases with translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 17–26, Hong Kong, China, November 2019. Association for Computational Linguistics.

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

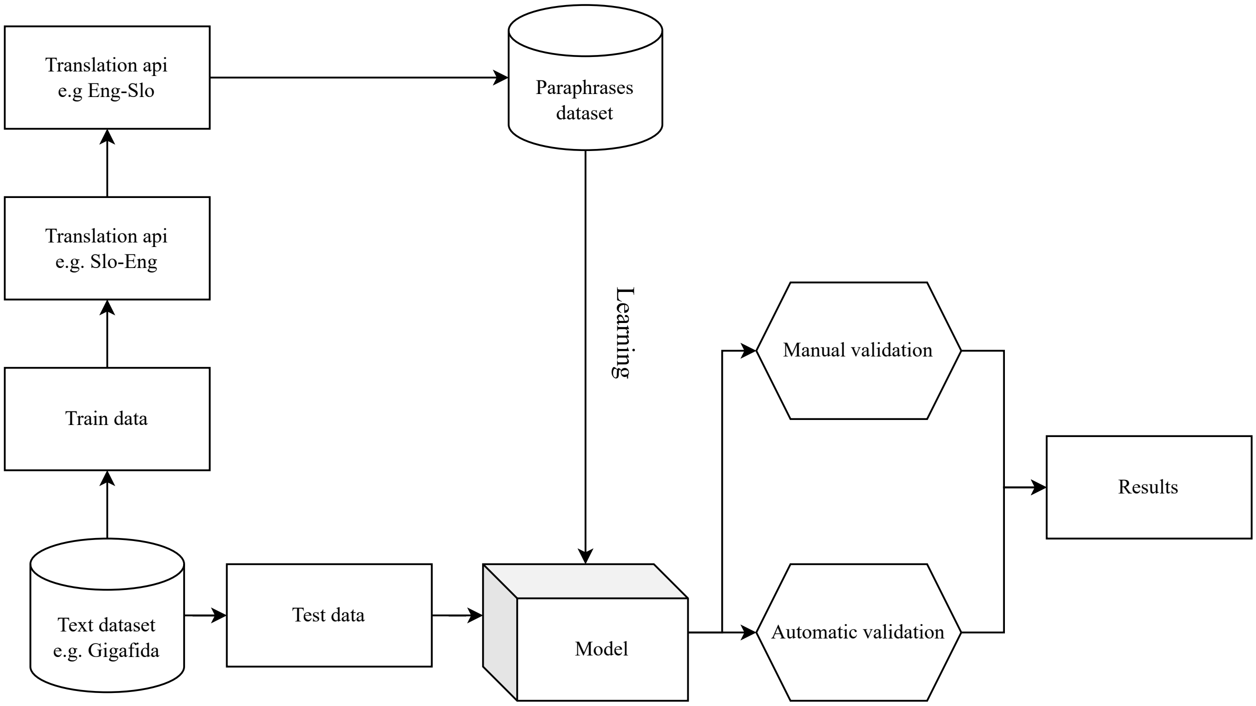

Figure 1. Visualization of a basic pipeline. Detailed visualization of a pipeline we will use to make a model which generates paraphrases.