Paraphrasing sentences in Slovene

Drejc Pesjak[1], Ilija Tavchioski[2], Klemen Vovk[3]

Abstract

TODO

Keywords
Keyword1, Keyword2, Keyword3 ...

[1] *dp8949@student.uni-lj.si, 63180224* [2] *it8816@student.uni-lj.si, 63180383* [3] *kv4582@student.uni-lj.si, 63190317*

Introduction

The task of paraphrasing sentences is to change the original sentence while preserving the meaning it conveys. For example, *Elephants consume small plants, bushes, fruit, and they can consume up to 170 kilograms of food per day.* can be paraphrased to *An elephant can eat up to 170 kilograms of small plants, bushes and fruit per day.*. Paraphrasing can be used to aid machine translation, improve question answering [1], detect plagiarism and origin [2], and evaluate machine-generated text (i.e. summaries) [3]. This task has been greatly aided by the boom of deep learning (mainly transformers [4]). In this work, we tackle paraphrasing Slovene sentences by fine-tuning a large language model on a dataset of Slovene original and paraphrased sentences that have been obtained by translation from English, as there is vastly more (labeled) data available in English.

Related work

In related work, we can find several approaches to paraphrasing sentences in English but, also in other languages as well. Here are included rule-based approaches [5] but also translation-based approaches which are the most common today, including paraphrasing with monolingual data, where there is a corpora where the first one is in one language and the second one is in the same language but after translation from another language [6].

In addition, similar methods are developed using bilingual corpora [7] where the paraphrase is determined by shared translation between the languages. On the other hand, with the rise of transformers [4] we can find several approaches using generative models such as GPT-2 [8] where the model is fine-tuned on a predefined dataset with paraphrasing examples and performed with decent results. In addition, there are some approaches using other techniques such as deep reinforcement learning as well [9].

Methodology

In this section, we describe the methodology we used for the task of paraphrasing Slovene sentences.

Dataset Collection

We used three datasets to train and evaluate our models:

1. **TaPaCo [10]**: This dataset consists of parallel sentences in multiple languages, including Slovene, and was collected from various sources. We used the Slovene portion of the dataset, which contains around 200,000 sentences, to fine-tune our language model. The dataset is available on Hugging Face[1].

2. **English Paraphrasing Data**: We used the English paraphrasing dataset from the Paraphrase.org website[2] to generate Slovene paraphrases. We translated the original English sentences to Slovene using Google Translate and kept only the translations that were judged to be of high quality by native speakers of Slovene.

3. **Paraphrasing MT Dataset [11]**: We used the English-Slovene Machine Translation dataset from the CJVT repository[3] to find additional paraphrases for our evaluation. We used the original English sentences in the dataset and calculated the similarity between them. The Slovene translations of similar sentences were then used for training.

Building a Model

We used a fine-tuned T5 language model to generate paraphrases for Slovene sentences. We adapted the **multilingual T5 model** [12] for Slovene by fine-tuning it on the TAPACO dataset. We followed the methodology described in [13] to

[1] https://huggingface.co/datasets/tapaco
[2] http://paraphrase.org/#/download
[3] https://huggingface.co/datasets/cjvt/rsdo4_en_sl

adapt the model for Slovene. We trained the model using a batch size of 32, a learning rate of 1e-4, and for 5 epochs.

In addition to the T5 model, we also experimented with a Hierarchical Refinement Quantized Variational Autoencoder **(HRQ-VAE)** model [14] for paraphrasing Slovene sentences. The HRQ-VAE model is a generative model that uses a hierarchical refinement process to generate diverse and high-quality paraphrases. We trained the HRQ-VAE model on the TAPACO dataset using the hyperparameters recommended in the original paper.

Evaluation

To evaluate the performance of our paraphrasing model, we will use several metrics commonly used in the paraphrasing literature, including BLEU, ROUGE, and ParaScore [15]. Additionally, we will use two recently proposed metrics, P-BLEU [16] and iBLEU [17], which have been shown to better correlate with human judgment than traditional metrics.

BLEU and ROUGE are n-gram based metrics that measure the similarity between the generated paraphrase and the reference sentence(s). ParaScore, on the other hand, is a paraphrase-specific metric that considers the degree of semantic equivalence between the two sentences.

P-BLEU and iBLEU, on the other hand, are extensions of the traditional BLEU metric that consider paraphrase diversity and quality. P-BLEU measures the diversity of the generated paraphrases, while iBLEU measures the informativeness of the generated paraphrases.

We will evaluate our model on a held-out test set and report the results in terms of all five metrics. Additionally, we will perform a human evaluation to assess the quality of the generated paraphrases compared to the reference sentences.

References

[1] Wee Chung Gan and Hwee Tou Ng. Improving the robustness of question answering systems to question paraphrasing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6065–6075, 2019.

[2] Saša Petrović, Miles Osborne, and Victor Lavrenko. Using paraphrases for improving first story detection in news and twitter. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 338–346, 2012.

[3] Liang Zhou, Chin-Yew Lin, Dragos Stefan Munteanu, and Eduard Hovy. Paraeval: Using paraphrases to evaluate summaries automatically. In *Proceedings of the human language technology conference of the NAACL, main conference*, pages 447–454, 2006.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[5] Kathleen R. McKeown. Paraphrasing using given and new information in a question-answer system. In *17th Annual Meeting of the Association for Computational Linguistics*, pages 67–72, La Jolla, California, USA, June 1979. Association for Computational Linguistics.

[6] Regina Barzilay and Kathleen R. McKeown. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, page 50–57, USA, 2001. Association for Computational Linguistics.

[7] Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893, Valencia, Spain, April 2017. Association for Computational Linguistics.

[8] Sam Witteveen and Martin Andrews. Paraphrasing with large language models. *CoRR*, abs/1911.09661, 2019.

[9] Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li. Paraphrase generation with deep reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3865–3878, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[10] Yves Scherrer. TaPaCo: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6868–6873, Marseille, France, May 2020. European Language Resources Association.

[11] Andraž Repar and Iztok Lebar Bajec. Parallel corpus EN-SL RSDO4 1.0, 2021.

[12] 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, 2020.

[13] David Dale. How to adapt a multilingual t5 model for a single language. *Towards Data Science*, 2021.

[14] Tom Hosking, Hao Tang, and Mirella Lapata. Hierarchical sketch induction for paraphrase generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2489–2501, Dublin, Ireland, May 2022. Association for Computational Linguistics.

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

[16] Yue Cao and Xiaojun Wan. DivGAN: Towards diverse paraphrase generation via diversified generative adversarial network. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2411–2421, Online, November 2020. Association for Computational Linguistics.

[17] Nitin Madnani. ibleu: Interactively debugging and scoring statistical machine translation systems. In *2011 IEEE Fifth International Conference on Semantic Computing*, pages 213–214, 2011.