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

Paraphrasing sentences

Matej Kranjec, Domen Vilar, Timotej Petrič

Abstract

Keywords
Keyword1, Keyword2, Keyword3 ...

Advisors: Slavko Žitnik

Introduction

The task of paraphrasing a sentence involves restating the meaning of the sentence in a different form, while preserving its original intent. Paraphrasing is an important task in natural language processing (NLP), with various applications in text simplification, summarization, and machine translation. In this paper, we propose a methodology for generating paraphrases of Slovene sentences using backtranslation and evaluate the quality of the generated paraphrases.

1. Dataset generation

1.1 Backtranslation

To generate our dataset, we plan to use backtranslation, a technique for generating paraphrases by translating a sentence from one language to another and then back to the original language. We will use the *Slovene NMT*[1] and *RSDO-DS4-NMT* model for machine translation of Slovene-English pairs[2].

By examining some test sentences using the proposed *Slovene NMT* backtranslation methodology (Slovene → English → Slovene), we determined, that the generated "paraphrased" sentences are of equal structure - only some words or word forms change in them. This is why we propose intermediate machine translation steps to also translate from English to some other languages and then back. For example: Slovene → English → German → English → Slovene. We plan to evaluate and decide which of the intermediate languages works best in combination with the Slovene language to produce the best results.

To translate from English to foreign languages other than Slovene, we used the *Nvidia NeMO* NMT models[3] (the code is available on our Github[4]).

The paraphrasing could also probably be improved by using some other language pairs, for example: Slovene → English → French → German → English → Slovene. But this would also almost certainly degrade the performance, because the languages are further apart from each other in terms of linguistic similarities. Therefore, we plan to focus on the intermediate languages that are more closely related to Slovene. Further tests are necessary.

Original:
V kvalifikacijah za drugo tekmo Lillehammerja je bil s 144 metri najdaljši Timi Zajc, ki pa je bil predvsem vesel pristanka brez padca.

Backtranslated (Slovene → English → Slovene):
V kvalifikacijah za drugo tekmo Lillehammerja je bil s 144 metri najdaljši Timi Zajc, še posebej vesel pa je bil, da je pristal brez padca.

1.1.1 Source dataset

For the dataset to use for backtranslation we plan to use ccGigafida corpus of Slovene texts[5]. This corpus is available in txt and xml format supplying Part of Speech (POS) tags. We plan to use xml format which can be helpful to select appropriate sentences that for example contains at least one verb. One example of inappropriate sentence in the corpus is: "10. SAMBA DE JANIERO - BELLINI (1)". We will also explore if using only certain genres of texts can be helpful while

[1] Github Slovene NMT. https://github.com/clarinsi/Slovene_NMT

[2] RSDO-DS4-NMT model weights. https://www.clarin.si/repository/xmlui/handle/11356/1736

[3] Nvidia NeMO models. https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/main/nlp/machine_translation/machine_translation.html

[4] Our github backtranslation sourcecode. https://github.com/UL-FRI-NLP-Course-2022-23/nlp-course-skupina-8/blob/slo_nmt/src/translation/run_translation.py

[5] ccGigafida https://www.clarin.si/repository/xmlui/handle/11356/1035, also on huggingface https://huggingface.co/datasets/cjvt/cc_gigafida

creating the appropriate dataset.

1.2 Translating existing datasets

The *MS-COCO dataset*[6] for Paraphrases dataset is a subset of the Microsoft Common Objects in Context (MS-COCO) dataset that contains pairs of paraphrased sentences in English language. We could use this to translate both sentences and create a paraphrasing dataset.

1.3 MRPC dataset

The MRPC dataset [7] is a commonly used paraphrase identification dataset. It stands for *Microsoft Research Paraphrase Corpus* and contains sentence pairs that are labeled as either paraphrases or non-paraphrases.

In paper [1] the authors propose two new metrics, word position deviation (WPD) and lexical deviation (LD), to better characterize paraphrase pairs without expert human annotation. The authors apply these metrics to better understand the MRPC dataset and propose improvements to the dataset.

Improved dataset is available online[8].

1.4 Word replacements

In addition to backtranslation, we could also consider using other data augmentation techniques to increase the diversity of the dataset. One such technique is word replacement, where we could use the *Paraphrase.org*[9] project to replace some words or sets of words in the sentence with their synonyms, this way generating new sentences.

2. Evaluation techniques

To evaluate the quality of the generated paraphrases, we will use a combination of automatic and manual evaluation techniques.

2.1 Automatic Evaluation

Further analysis of similarity score is necessary to evaluate how good we constructed the proposed dataset. This article proposes technique for automatic evaluation of machine translation https://aclanthology.org/P02-1040.pdf. We could try to use this evaluation metric for evaluating how good the paraphrases are.

2.2 Manual Evaluation

We are going to manually evaluate the quality of paraphrased sentences for a smaller subset of the dataset. We will manually evaluate the semantic similarity between sentences, and automatically compute and penalize overlap between the source and paraphrase using ROUGE-N https://aclanthology.org/W04-1013/ [2] and some combination of other metrics, that are listed below.

2.3 Other metrics

2.3.1 ParaScore, 2022

The paper [3] explores the effectiveness of automatic metrics for paraphrase evaluation and proposes a new metric called ParaScore. The authors found that reference-free metrics perform better than reference-based metrics and that commonly used metrics do not align well with human annotation.
Code is opensource[10].

2.3.2 ROUGE-P, 2022

The authors propose a novel metric called ROUGE-P, which evaluates paraphrases based on adequacy, novelty, and fluency. Adequacy measures how well the paraphrase conveys the same meaning as the original sentence. Novelty measures how different the paraphrase is from the original sentence. Fluency measures how well the paraphrase reads and sounds like natural language. They also provide empirical evidence that current natural language generation metrics are insufficient in measuring these desired properties of a good paraphrase.
We found no code available.

2.3.3 Neural embeddings

We will also consider an automatic technique for evaluation of paraphrases using neural embeddings. One approach would be to use the *Slovene T5* model, the *SloBERTa* model[11], *CroSloEngualBERT* [12], or the *LLaMA* model from Facebook (or the *Alpaca-LORA* [13] model, which can, if need be, even be trained on more "*resource constrained*" environments), or some other large language model (LLM) like *BLOOM* https://arxiv.org/abs/2211.05100, to transform the input sentences into neural embeddings and then compare the paraphrased sentences in that space with distance measures such as cosine distance. We will also consider penalizing overlaps here.

2.4 Training a generative model for paraphrases

We are going to fine-tune a pre-trained t5 model for paraphrasing sentences using the described dataset. We will use https://huggingface.co/cjvt/t5-sl-large as a checkpoint model.

As mentioned above in subsection 2.3.3, Alpaca-LORA is also a viable option for training.

Methods

Results

Discussion

Acknowledgments

[6] MS-COCO https://cocodataset.org/

[7] MRPC dataset https://www.microsoft.com/en-us/download/details.aspx?id=52398

[8] Improved MRPC dataset https://huggingface.co/spaces/tlkh/paraphrase-metrics-mrpc, github: https://github.com/tlkh/paraphrase-metrics

[9] http://paraphrase.org/

[10] ParaScore sourcecode https://github.com/shadowkiller33/ParaScore#citation

[11] https://www.clarin.si/repository/xmlui/handle/11356/1397

[12] https://www.clarin.si/repository/xmlui/handle/11356/1330

[13] https://github.com/tloen/alpaca-lora

References

[1] Timothy Liu et al. Towards better characterization of paraphrases. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8592–8601, 2022.

[2] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[3] Lingfeng Shen, Lemao Liu, Haiyun Jiang, and Shuming Shi. On the evaluation metrics for paraphrase generation. In *The 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.