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

0.1 Backtranslation

We plan to use backtranslation to generate the dataset. https://github.com/clarinsi/Slovene_NMT and https://www.clarin.si/repository/xmlui/handle/11356/1736 model to translate Slovene → English → Slovene.

By examining some test sentences using the proposed *Slovene NMT* backtranslation methodology (Slovene → English → Slovene), we determined, that the generated "paraphrased" sentences are by a large part of equal structure - only some words or word forms change in them. This is why we want to propose additional machine translation tools and add intermediate steps of translating from English to some other languages and then back. For example: Slovene → English → German → English → Slovene. We plan to manually evaluate and decide which of the intermediate languages works best in combination with the Slovene language to produce the best results. https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/main/nlp/machine_translation/machine_translation.html

The paraphrisation 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. Further tests are necessary.

0.2 Evaluation techniques

Further analysis of similarity score is necessary to evaluate how good we constructed the 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.

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/ [1].

We will also consider an automatic technique for evaluation of paraphrases using neural embeddings. One approach would be to use the BERT *SloBERTa* model https://www.clarin.si/repository/xmlui/handle/11356/1397, *CroSloEngualBERT* https://www.clarin.si/repository/xmlui/handle/11356/1330, or the *LLaMA* model from Facebook (https://github.com/shawwn/llama-dl), 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.

0.3 Dataset

For the dataset to use for backtranslation we plan to use ccGigafida corpus of Slovene texts: https://www.clarin.si/repository/xmlui/handle/11356/1035 https://huggingface.co/datasets/cjvt/cc_gigafida 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 creating the appropriate dataset.

0.4 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.

Methods

Results

Discussion

Acknowledgments

References

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