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

Paraphrasing sentences

Gregor Gabrovšek, Miha Marinko, Andraž Andolšek Peklaj

Abstract

This project aimed to develop a natural language processing model for automatic paraphrasing of Slovenian paragraphs. The project includes dataset construction, model training using different AI models, and evaluation using both manual and automatic techniques. The results highlight the challenges faced in generating high-quality paraphrases and present examples of generated paraphrases.

Keywords
paraphrase, language, 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.

Zhou and Suma [1] distinguish 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.

In the next section, we describe our methods used for construction of a paraphrase dataset, how we trained the AI models and our techniques for evaluation of generated paraphrases. In Results sections we present examples of generated text and evaluation results of different models. Discussion section we reflect on the results of our models and describe possible improvements.

1. Methods

1.1 Dataset construction

For our project, we developed a dataset consisting of abstracts from University of Ljubljana diploma theses along with their paraphrases.

To obtain the textual data, we have implemented a specialized web scraper to collect abstracts in both Slovenian and English from the Repozitorij Univerze v Ljubljani website[4]. The program required a range of pages as a parameter and then proceeded to iterate over them. The scraper extracted links to each thesis page and saved the Slovenian and English abstracts in separate files, with the thesis ID serving as the filename.

Then, we used the Neural Machine Translation model for the Slovene-English language pair RSDO-DS4-NMT 1.2.6 to translate the English abstracts into Slovenian and obtain two similar paragraphs that we considered to be paraphrases.

The dataset consisted of about twenty thousand paraphrased abstracts. We have made our dataset available to other teams to facilitate better results [5].

While training text generation models on our dataset, we noticed that they were generating paraphrases very similar to the input text. To tackle this problem, we utilized the BLEU score to calculate the similarity scores between the original and translated texts (the distribution of the scores can be seen in Figure 1.) We used metric to prune our dataset of entries that were too similar to each other. In the final dataset, we got rid of the quarter the most similar paraphrases according to the BLEU score.

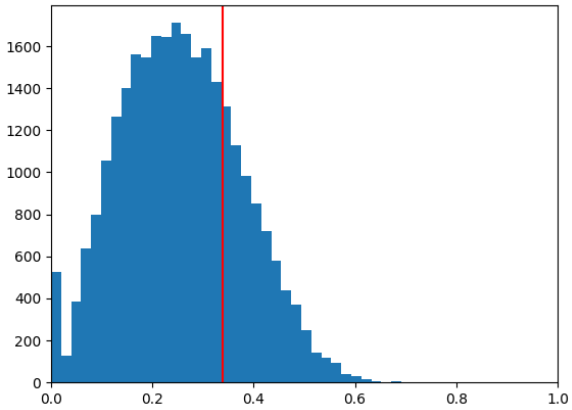

Figure 1. The distribution of BLEU similarity scores of the text pairs in our original dataset. For the pruned dataset, we removed 25% of the text pairs that were too similar. The cutoff is displayed as a red vertical line in this figure.

1.2 Model Training

We fine tuned multiple different text-to-text AI models, including *google/mt5-base*, *cjvt/t5-sl-small* and *cjvt/t5-sl-large*, which are available on Hugging Face.

Training was done using *pytorch* (and the *lightning* library) and integrated with Hugging Face, where our trained models and prepared datasets are available publicly.

When training and evaluating the models, we used two different generation methods: Greedy and TopK. Greedy sampling always generates the most probable next token, while TopK means that several possible sequences are considered at any time, and the k best candidates are returned, which helps get a more wholistic idea of model performance.

1.2.1 google/mt5-base

Initial training was done on the multilingual *google/mt5-base* model on our dataset, which was described in section 1.1. The model had problems fitting to our task, and was mostly just repeating the input, with minor changes at best. Another problem was that occasionally, text generated towards the end could end up getting repeated multiple times.

Our next step was experimenting with language models, which were specifically trained on Slovene datasets, beginning with *t5-sl-small*, originally trained by cjvt.

1.2.2 cjvt/t5-sl-small

The small T5 was much smaller, compared to the multilingual model, and was thus much faster to train, but produced only marginally better results, due to its better understanding of the underlying language.

Our main problem seemed to be in the dataset itself, since a large portion of training pairs was very similar, and not really paraphrases (figure 1). Once we obtained the pruned dataset, models were retrained and significant improvements can be observed for some sentences (table 1), while others remained very similar to the original.

In an attempt to improve performance across a wider range of input sentences, we experimented with the bigger, *cjvt/t5-sl-large*, model.

1.2.3 cjvt/t5-sl-large

Since our dataset was relatively small compared to the model, we used a smaller learning rate to prevent overly drastic changes, our experiments with the larger model were not fruitful and generated completely nonsensical text. In addition, attempts to tune various parameters, such as batch size, learning rate, and gradient accumulation steps were very time consuming when each training run takes much longer (especially when considering that sometimes, one might have difficulties even getting a job allocation on the arnes cluster).

1.2.4 Other

We also intended to train and evaluate the different models on other teams' datasets, but ran into issues during execution. For example, when retraining one of our models on the *europarl* dataset, our trainer ran out of memory at the very end of the training loop (even though as much as possible was allocated), which was a highly precarious situation (and ultimately too tricky to resolve).

1.3 Performance Evaluation

To be able to generate good paraphrases, we needed to decide, how we will evaluate them. As automatic techniques cannot completely evaluate this metric, we also employed manual evaluation.

For evaluation we created short, but more generalized dataset where each entry consisted of few sentences. We needed texts that were more representative of real-world texts, as our training dataset consisted of scientific jargon. We gathered paragraphs from news and articles on rtvslo.si, 24ur.com, slo-tech.com and sl.wikipedia.org.

1.4 Manual Evaluation

To manually evaluate each paraphrase we needed to define simple metrics, that could be understandable by everyone in similar way and still capture as much information about the quality of the paraphrase.

We have defined three scores, where all generated paraphrases can be graded with a integer value between one and five. First of the scores was adequacy, which represents, how similar the meaning of paraphrase is to the original text. If the essence of paraphrase's meaning was the same, the score would be higher and if the meaning was lost or wrongly interpreted, the score would be lower.

We also defined a readability score. This value would be low when the paraphrase included many grammatical and syntactical errors or the structure of the sentences was not natural. The grade would be high if the text was easily readable and couldn't be differentiated from human-written text.

As our models had the tendency for generation of text that was very similar to the input, but the goal of paraphrasing is to generate a different text, we needed to evaluate this. Similarity score would be high when the paraphrase had mostly different

text, but still kept the similar length to the original. The score would be low, when the generated text mostly just copied inputted text.

All three members of our group graded twenty example paraphrases for all different models and generation strategies that we tested. We combined the results and averaged them to get more stable results.

1.5 Automatic Evaluation

For automatic evaluation we used multiple metrics that are commonly used in natural language processing.

1.5.1 Cosine Distance

Cosine distance is a metric used in natural language processing to measure the similarity between two texts. In paraphrase evaluation, it helps assess the similarity between a machine-generated paraphrase and a reference paraphrase. By representing the texts as vectors, cosine distance calculates the angle between them, indicating their degree of similarity.

1.5.2 BLEU

BLEU (Bilingual Evaluation Understudy) is a metric used for evaluation of machine-generated translations or paraphrases. It measures the similarity between a machine-generated sentence and reference sentences. The score ranges from 0 to 1, with a higher score indicating higher similarity to the reference sentences. It is computed based on n-gram precision, which measures the percentage of n-gram matches between the generated and reference sentences.

1.5.3 METEOR

METEOR (Metric for Evaluation of Translation with Explicit Ordering) is a metric used to evaluate the quality of machine-generated translations or paraphrases. It measures the similarity between a generated sentence and reference sentences by considering various linguistic factors such as exact word matches, stemmed matches, and semantic similarity. The score ranges from 0 to 1, with a higher score indicating better quality.

1.5.4 BERT Score

BERT Score is used to assess the quality of machine-generated translations or paraphrases. It uses the contextual embeddings produced by BERT (Bidirectional Encoder Representations from Transformers) models to measure the similarity between a generated sentence and reference sentences. By considering the contextual information encoded in the embeddings, BERT Score captures the semantic similarity and takes into account both word-level and structural similarities. The score ranges from 0 to 1, with a higher score indicating higher quality. BERT Score has shown effectiveness in evaluating the performance of language generation systems, providing a fine-grained analysis of their output quality.

2. Results

After performing the training of all of the described base models with our dataset, we examined and evaluated the results.

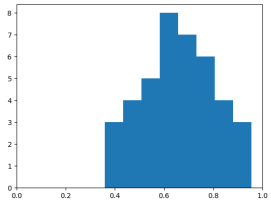

Figure 2. Histogram of BLEU scores, calculated for outputs of models trained on the MT5 Auto model with the Greedy generation method.

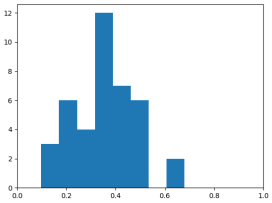

Figure 3. Histogram of BLEU scores, calculated for outputs of models trained on the MT5 Auto model with the TopK generation method.

In the table 1 there are presented some examples of the paraphrases. On the left side is the original text and on the right side is the artificially generated text. In the examples we can notice some of the problems, that will be further mentioned in the following presentation of results.

2.1 MT5 fine tuned on pruned RUL dataset

The MT5 model successfully generated paraphrases of the given text. Most of the phrases included all the meaningful data from the original and slightly changed the structure of the text.

The results also included some noticeable problems that were occurring in the generated paraphrases. Most of them had very high resemblance to the original input, specially the greedy method, which can be seen in the BLEU score distribution in figure 2.

Our dataset included a lot of entries with similar beginning, such as ”Magistrsko delo”. The model learned these common phrases and sometimes included them in the beginning of the generate paraphrase, even tho they did not have anything to do with the context. Our generation method also caused a similar problem. We used a keyword ”V slovenskem kontekstu” to instruct the model, what to do. Sometimes this also appeared in the results.

A relatively profound problem was repeating of parts of sentences. In some cases model concatenated multiple identical n-grams from the input text.

The output sometimes also included words that are not included in Slovene language.

Table 1. Examples of generated paraphrases

google/mt5-base	
46-letno žensko so soočili z 10 vprašanji, med njimi pa je bilo ključno vprašanje, kaj je bila zadnja stvar, ki ji jo je sin rekel, preden je moril na osnovni šoli, poroča Telegraf.rs.	Magistrsko delo so žensko soočili z 10 vprašanji. med njimi je bilo glavno, kaj se ji je pri tem zgodilo pred preporodom na osnovni šoli, poroča telegraf.rs.
Čeprav pokopališča vzbujajo najgloblja čustva, kot je žalost, si člani združenja prizadevamo ustvariti okolje, ki človeku prinaša pozitivna čustva in vzdušje umirjenosti,	V slovenskem prostoru velja, da pokopališča vzbujajo najgloblja čustva, kot je žalost, člani združenja se v evropsko pomembnih pokopaniščih srečujemo z okoljem, ki ljudem prinaša pozitivna čustvena in umirjenost.≪ lidija pliberšek, predsednica združenje evropo najpomembnejših pokopišč, prizadeva ustvariti okolje, v katerem bi človeka prinesla pozitivne čustvovanja in vzdušje umerjenosti. v raziskavi sta navedene odgovore na področju ureditve okolja, iz katerega lahko doživljamo pozitivni čustvi in duševnem spokoju≪, pravi lidia plibšek.
cjvt/t5-sl-small	
Tudi Priscilla Presley, soproga pokojnega kralja rokenrola Elvisa Presleyja, se je poklonila Tini. Zapisala je, da je bila glasbenica ena najljubših ustvarjalk pokojnega Elvisa. Ko je stopila na oder, je bila to čista čarovnija. Zapustila je izjemno dediščino in vsi jo bodo zelo pogrešali.	To je bil tudi priscilla presley, soproga pokojnega kralja rokenrola elvisa presneyja, poklonjena svoji družini. priscille pressleyeva žena pokojne kraljeve žene elvis preseya, se je poklonila tini. zapisala je, da je bila glasbeno ustvarjalka, katere glavno poslanstvo je bilo ohranjati bogastvo in vse bo dobro pogrešala.
Ruske sile so ponoči ponovno izvedle več napadov z brezpilotnimi letalniki, je sporočil Kijev. Rusija še naprej terorizira Ukrajino z nočnimi napadi, je ob tem sporočil ukrajinski predsednik Volodimir Zelenski. Wagnerjeve enote so začele umik iz Bahmuta.	Ukrajina je ponoči izvedla več napadov s protipiracijskimi letalniki, ki so bili pošiljani po ukrajinskem vzhodu. rusija je v noči ponovno izvedla nove napade z brezpilotnimi letalnik. na posnetku je ukrajinski predsednik volodimir zelenski.

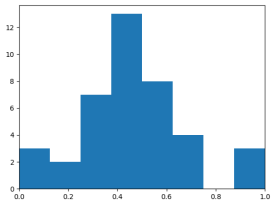

Figure 4. Histogram of BLEU scores, calculated for outputs of models trained on the T5 Small model with the Greedy generation method.

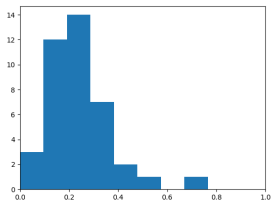

Figure 5. Histogram of BLEU scores, calculated for outputs of models trained on the T5 Small model with the TopK generation method.

2.2 T5 small fine-tuned on pruned RUL dataset

The T5 small model included similar problems as the MT5, but it was a bit more creative in the text generation. The creativeness can be see in smaller BLEU score for both greedy 4 and TopK 5 method. This means there are bigger differences between the texts, but this also brings additional downsides, such as changing the meaning of the text. On multiple occurrences the structure of the sentences stayed relatively similar to the original, but few words were changed and the opposite meaning was perceived.

The paraphrases sometimes also included made-up words and syntactical errors, that are not correct in Slovene language.

In some rare cases the model outputted an empty string. This happened only on the greedy generation method, but it is possible that with further testing we could notice this for the other method too.

2.3 T5 large fine-tuned on the pruned RUL dataset

The results of the large T5 dataset were worse than expected. The generated paraphrases included a lot of text that was not connected with the context of the original sentences. The greedy method of generation did not work and the results of the TopK method were worse than the other two models with brief overview. For this reason we did not do the whole

Base Model	Generation Method	Evaluation metric			
		Cosine Similarity	BLEU	METEOR	BERT Score
MT5 Auto	Greedy	0.91	0.65	0.91	0.93
	TopK	0.67	0.35	0.70	0.87
T5 Small	Greedy	0.74	0.47	0.75	0.86
	TopK	0.51	0.23	0.53	0.83
T5 Large	Greedy	0.36	0.25	0.35	0.70
	TopK	0.41	0.13	0.39	0.78

Table 2. Evaluation scores for all three trained models. For each model, we evaluated the Greedy and TopK generation method.

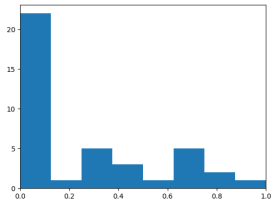

Figure 6. Histogram of BLEU scores, calculated for outputs of models trained on the T5 Large model with the Greedy generation method.

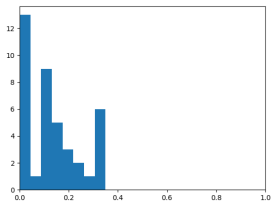

Figure 7. Histogram of BLEU scores, calculated for outputs of models trained on the T5 Large model with the TopK generation method.

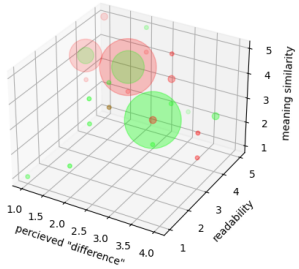

Figure 8. 3d scatter of manual evaluation scores. Greedy method. Red: google/mt5-base. Green: cjvt/t5-sl-small

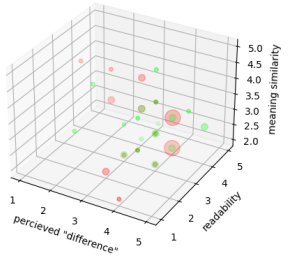

Figure 9. 3d scatter of manual evaluation scores. TopK method. Red: google/mt5-base. Green: cjvt/t5-sl-small

manual evaluation. As can be seen on Figure 7 and 6, the BLEU scores are somewhat low and equal to zero in many cases.

2.4 Comparison of the models

The metrics that were calculated with our automatic evaluation are presented in Table 2 and the manual evaluation results can be seen in the Figures 8 and 9.

Judging by the cosine similarity, we can observe that the model based on MT5 Auto in general produced rather similar phrases to the original ones. BERT Score indicates that they were also similar in meaning.

On the other hand, the model based on T5 Large produced paraphrases that were not very similar to the input paragraph. However, it turned out that the meaning of the output was also not a good match.

The model based on T5 Small, however, resulted in a moderate similarity score while still having a relatively high BERT score. Therefore, based on our evaluation techniques, we find the model *t5-sl-small* the best of the tested. It is closely followed by the *mt5-base*.

3. Discussion

The development of a paraphrase generation model is a challenging task, and our results indicate that there is room for improvement. While our model has shown some promising capabilities, we have also encountered limitations and areas where further enhancement is needed.

One notable issue we observed is that a significant number of generated paraphrases are nearly identical to the input. This suggests that the model struggles to introduce substantial variations in the output, potentially hindering its usefulness in certain applications. Improving the diversity and creativity of the generated paraphrases should be a key focus for future iterations of the model.

Another shortcoming we identified is the occasional omission of important information from the input during paraphrase generation. This can lead to incomplete or inaccurate outputs, which significantly impacts the reliability and utility of the model. Additionally, we observed instances where the model generated made-up text that was not present in the original input. This issue is concerning as it introduces potentially

incorrect information into the paraphrases.

To overcome the problems, we could employ many different changes. As we have seen, the generated text can include phrases such as ”Magistrsko delo” that appeared very commonly in our dataset. We should use a dataset, that would be more generalized and would not include similar repeating patterns.

Text-to-text models are constantly improving and changing. Our methods of fine-tuning and generation of paraphrases could also be attempted on other models, where the results might be better.

We could also customize the training of the models. Currently we cannot effectively control, how similar are the paraphrases to the original text. We believe this could be achieved with custom loss functions and with new ways of preprocessing the dataset.

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] Digital repository of slovenia. Accessed: 23.04.2023.

[5] Group 7. RUL: Paragraph-level paraphrasing dataset of academic abstracts from University of Ljubljana, 2023.