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 and evaluate the quality of the generated paraphrases.

1. Dataset generation

1.1 Translating existing datasets

We used the *Europarl Parallel Corpus*[1] for obtaining pairs of sentences for the paraphrasing dataset. We took the Slovene-English parallel corpus and translated the English sentences using machine translation. To translate from English to Slovene, we used the *Nvidia NeMO* NMT models[2] (the code is available on our Github[3]). The dataset contains approximately 600.000 pairs of sentences.

1.2 Filtering sentences in dataset

We filtered the resulting dataset by considering the amount of digits in a pair of sentences, the amount of special characters, and the sentence lengths. Furthermore, we employed a metric called parascore to remove similar sentences. The parascore metric acknowledges that paraphrase evaluation is inherently different from the evaluation of other tasks, such as machine translation and summarization. A good paraphrase must satisfy two criteria: semantic similarity and lexical divergence. Semantic similarity requires that the paraphrase remain similar meaning to the original sentence. Lexical divergence requires that the paraphrase has syntactic or lexical differences from the original sentence. Most metrics used in previous paraphrase generation research are designed for tasks other than paraphrase evaluation, which can lead to suboptimal results. To determine the appropriate threshold for filtering out sentences, we examined the distribution of scores. By visualizing the distribution of scores in a figure, we were able to make the decision where to set the threshold. This enabled us to avoid cutting out too many sentences, which could result in a loss of training data. We filtered the dataset by looking at the ParaScore [1] of pairs of sentences from the dataset, and removing pairs if the score was too low.

We also removed sentences that did not start with an alphanumeric symbol, or end with an alphanumeric symbol. When reviewing the list of removed sentences, we noticed that many of them had started with arbitrary sequences of characters, such as (SI), (AU), (HU), and others. After further investigation, we realized that these were actually identifiers, and that the sentences were otherwise appropriate for training. To identify these sentences and retain them in our dataset, we applied a regular expression filter to the start of each sentence.

We used 80% of the data for training and the remaining 20% for validation. As we are generating paraphrases and comparing the generated sentences to the prompts, we do not need a labeled dataset for testing, we will use a monolingual corpus instead.

2. Evaluation techniques

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

[1] Europarl https://www.statmt.org/europarl/

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

[3] 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

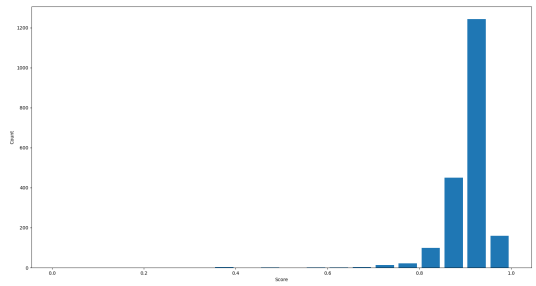

Figure 1. Distribution of Scores for Filtering Sentences i

niques.

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 Other metrics

2.2.1 ParaScore, 2022

The paper [1] 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[4].

3. Methods

3.1 T5

We tried an approach with a sequence to sequence model, finetuned for paraphrasing sentences. We used a Slovene T5 model, t5-sl-small [2] and further trained it using our paraphrasing dataset. We have not yet achieved good results with this model, it mostly changes the word order or removes parts of original sentences, but keeps the same words. We suspect that the hyperparameter settings we used for the training or inference need to be improved.

We used the resulting model for inference of paraphrases. We present a few examples of generated paraphrases in table

3.2 Vicuna

Another approach we tested is the Vicuna model [3]. This is a chatbot based on the LLaMa model, finetuned on 70.000 conversations with ChatGPT. We used the base model to generate paraphrases of English sentences. We prompted the model to paraphrase a given sentence. After that we translated both the original sentence and the resulting paraphrase with machine translation to obtain a pair of paraphrased sentences in Slovene. A few examples are presented in table 2.

Table 1. Examples of paraphrase generation with T5 finetuned for paraphrase generation.

Original	Paraphrase
Škisovo tržnico, največjo prireditev na prostem za mlade v Sloveniji, bodo otvorili tradicionalni mimohod študentskih klubov, nastop mažoretk in pozdravni govori.	Škisovo tržnico, največjo zabavo za mlade v Sloveniji, bodo otvorili tradicionalni mimohod študentskih klubov, nastop mažoretk in pozdravni govori.
Program umetne inteligence ChatGPT podjetja OpenAI je po mesecu dni blokade v Italiji spet dovoljen, je danes sporočilo podjetje.	Program umetne inteligence ChatGPT podjetja OpenAI je po mesecu dni blokade v Italiji ponovno dovoljen, je danes sporočilo podjetje.
Evropski policijski urad Europol je konec marca posvaril pred zlorabami tovrstnih besedilnih generatorjev.	Evropski policijski urad Europol je konec marca opozoril na zlorabo takšnih besedilnih generatorjev.

3.3 Vicuna para dataset and T5 finetuning

We have also generated a dataset of LLaMA paraphrases translated to Slovene. Size is 600k and original sentences are from Euparl corpus. For dataset examples look at table 3.

Initial results from finetuned T5 models look promising, by offering much more diverse and interesting text outputs, but often lacking original meaning. For examples look at table 4.

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

Table 2. Examples of paraphrase generation with Vicuna and machine translation.

Original	Paraphrase
ker je to moj drugi mandat Komisije, vam lahko povem, da vidim veliko večji poudarek - in to je bilo omenjeno v razpravi - na izvajanju.	Po mojih izkušnjah je večji poudarek na izvajanju v tem drugem mandatu Komisije To je bila tema razprave v razpravi.
menim, da je konferenca na Baliju odlična priložnost za vse nas, da se zberemo in razmislimo o potrebnih ukrepih glede podnebne politike po letu 2012.	Konferenca na Baliju je primerna priložnost za skupno preučitev potrebnih ukrepov za podnebno politiko po letu 2012.
To je najbolj prazno upanje, kar jih je.	Ta izjava izraža najbolj žalostna pričakovanja.
Ker torej obe stranki v sporu neposredno ali posredno zagotavljata finančno podporo svoji industriji, je možna le ena razumna rešitev.	Glede na to, da obe stranki v sporu zagotavljata finančno podporo za svoji panogi, je lahko edini logični izid.
v imenu skupine Verts / ALE. - (ES) Gospa predsednica, v imenu svoje skupine bi rada najprej čestitala gospodu Schmittu za to poročilo in prav tako poudarila, da je pomembno povečati število ur poučevanja telesne vzgoje v šolah.	zastopam skupino Verts / ALE in bi rada čestitala gospodu Schmittu za njegovo poročilo Poleg tega bi rada poudarila, kako pomembno je povečati število ur, ki so namenjene pouku telesne vzgoje v šolah.

LLaMa para. tran.	Odbor je lahko zaradi svojega raznolikega strokovnega znanja Agenciji ponudil specializirano pomoč pri nalogah, za katere je bil zadolžen v skladu s Pogodbo.
orig. tran.	Odbor lahko zaradi svojega raznolikega strokovnega znanja zagotovi agenciji edinstveno podporo pri nalogah, ki so mu dodeljene s Pogodbo.
orig. sl	Zaradi različnih strokovnih znanj lahko odbor agenciji zagotovi edinstveno podporo pri nalogah, določenih s pogodbo.
LLaMA para. tran.	Znižanje uvoznih tarif za banane iz Latinske Amerike bi jih naredilo bolj konkurenčne na trgu.
orig. tran.	če bomo danes zmanjšali uvozne tarife za banane iz Latinske Amerike, bo ta pridelek postal bolj konkurenčen.
orig. sl	Če bomo danes zmanjšali tarifne stopnje za banane iz Latinske Amerike, bo ta proizvod postal bolj konkurenčen.
LLaMa para. tran.	Tunizija ne služi le kot model za tiste, ki si prizadevajo za padec arabskih avtokratov, ampak zagotavlja tudi precedens, ki lahko pomaga razložiti nadaljnje dogodke na Bližnjem vzhodu.
orig. tran.	Tunizija dejansko ni le zgled za tiste, ki si želijo padca arabskih samodržcev, ampak še naprej predstavlja precedens, ki nam lahko pomaga razumeti poznejši razvoj v zapleteni vzhodni enačbi.
orig. sl	Tunizija je dejansko ne le vzor vsem tistim, ki si želijo padca vseh arabskih samodržcev, ampak še naprej predstavlja precedens, ki nam lahko pomaga razumeti poznejši razvoj v zapleteni vzhodni enačbi.

Table 3. Comparison of LLaMA paraphrase translated to eng, original text translated to eng and original sl text. This are sample sentence pairs from *euparl-llama-slx3_enx2_600k* dataset. We can see that LLaMA paraphrase translation offer much mode diverse paraphrases. When comparing original translated and original Slovene sentences many words

4. Further work

Further training of T5, possibly try GPT2 training. Evaluation needs further definitions. Sentence filtering further work to gain more refined datasets for model training.

5. Results

6. Discussion

7. Acknowledgments

References

[1] 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.

[2] Matej Ulčar and Marko Robnik-Šikonja. Sequence to sequence pretraining for a less-resourced slovenian language, 2023.

Input	Evropski policijski urad Europol je konec marca pos[3]varil pred zlorabami tovrstnih besedilnih generatorjev.
Out1	Europol je 11. marca opozoril na zlorabe teh besedilnih generatorjev.
Out2	Europol je konec marca ugotovil, da so bile storjene zlorabe jezika.
Out3	Evropski policijski urad Europol je konec marca pozval Evropsko komisijo, naj preneha z uporabo besedilnih generatorjev.
Out4	Evropska policijska agencija (Eurapol) je v začetku marca objavila poročilo o zlorabah besedilnih generatorjev.
Out5	Evropski policijski urad Europol je v začetku aprila obvestil Europol in opozoril, da je ta program predmet zlorabe jezikovnih generatorjev.

Table 4. Example generations from T5 model finetuned with *euparl-llama-slx3_enx2_600k* dataset.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.