UL FRI Data Science - Natural Language Processing, 2022-2023
Project Report

Paraphrasing sentences in Slovene

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

Abstract

With the exponential growth of the natural language processing field using deep learning methods, especially transformers, we can see many *state-of-the-art* models that have great performance on various tasks such as sentiment analysis, machine translation, text generation, sentence paraphrasing and other. But, the most of these methods are trained and used only on data written in English, our research goal is to develop a method for sentence paraphrasing in Slovenian language. We fine-tunned an already pre-trained GPT model (the Slovenian variant) on Slovenian data, and we achieved a decent performance on our test data having a high mean grade from the human evaluators for our model. Although the model performance have many flaws, this is the first of the many steps to enrich the family of transformer-based models to other languages such as Slovenian.

Keywords
Natural Language Processing, Sentence paraphrasing, Transformers

[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 creation

Sentence similarity in machine translation dataset

We explored different methods to create a dataset suitable for fine-tuning a Language Model (LM). One of the approaches we considered was using a machine translation dataset. Specifically, we used the cjvt/rsdo4_en_sl dataset [10], which consists of one million translated sentences from English to Slovene. Our goal was to find similar sentences in English and extrapolate their Slovene counterparts, thereby creating a parallel dataset for training our LM.

To achieve this, we used cosine distance on BERT embeddings [11] to measure the similarity between sentences in English and Slovene. We found that despite using a state-of-the-art language model, the results were not satisfactory. Out of the first 10,000 sentences, only 200 pairs had a similarity score higher than 95%. This was not sufficient for fine-tuning a LM, as we required a larger dataset to improve the performance of our model. Therefore, we concluded that using a machine translation dataset alone was not effective in creating a high-quality parallel corpus.

Slovene paraphrasing with GPT prompting

Another approach we tried for creating a Slovene paraphrasing dataset was using GPT prompting. We used the cvjt/gpt-sl-base model [12] to generate paraphrases for the given input sentence. To do this, we prompted the GPT model to complete the original sentence with a paraphrased version. For instance, given the input sentence ”Today is a beautiful day,” we prompted the model with the phrase ”{input_sentence}, oziroma drugače,” which translates to ”{input_sentence}, or in other words.”

We found that this approach produced some promising results, with the GPT model generating a range of plausible paraphrased sentences for a given input. However, there were some issues with this approach. One of the main problems was the slow speed of the model. The cvjt/gpt-sl-base model [12] is a large language model with more than 125 million parameters, and generating paraphrases for a single sentence could take up to several minutes. Additionally, it was sometimes challenging to determine where the GPT model ended the paraphrased sentence. This made it challenging to identify the start and end points of the generated paraphrase, making it challenging to create a high-quality dataset.

Machine translation of a paraphrase dataset

The ParaBank2 dataset [13] is a large corpus of 18 million paraphrased English sentences that has been utilized as a data source for our research. Our objective was to obtain the Slovene translations of the English sentences, which we approached through two methods.

First, we attempted to use machine translation tools such as DeepL, Google Translate, and other alternatives to generate the translations. Although these tools produced more accurate translations than the Slovene neural machine translation (NMT) model [14], the translation speed was too slow, even when employing proxies to circumvent rate limiting. Consequently, we found this approach unsuitable for amassing a sufficiently large dataset for model fine-tuning.

As an alternative, we utilized the SloveneNMT model [14] to translate the ParaBank2 dataset locally. As there is no rate limiting associated with this model, we selected the top 100,000 rows of the ParaBank2 dataset [13], which are ordered in descending order of paraphrase quality, for translation. This process was computationally demanding and required 10 hours of processing on an RTX 3090 with 24GB of VRAM. We should note that during translation, some paraphrased sentences were translated into identical sentences, resulting in duplicates that were removed before model fine-tuning.

Building a Model

For building our model, we used the GPT-sl-base language model [12], which is a Slovene variant of the GPT family of models. This model was based on the BigScience workshop fork of Megatron and was pre-trained on large Slovene corpora, including Gigafida, KAS, slWaC, and MaCoCu. With about 110 million parameters, it has 12 transformer layers, a dimension of 768, and 16 attention heads. This model can process sequences up to 1024 tokens in length and has a tokenizer trained on a smaller subset of corpora, with a vocabulary of 60k tokens.

Before fine-tuning the GPT-sl-base model on the generated dataset, we split the translated dataset into three separate sets for training, evaluation, and testing. Specifically, we used a split of 60%, 20%, and 20% for the training, evaluation, and testing sets, respectively. This allowed us to train the model on a sufficiently large dataset while also evaluating its performance on unseen data to ensure that the model was not overfitting to the training set.

To fine-tune this pre-trained GPT-sl-base model on our generated dataset, we used transfer learning. The idea behind transfer learning is to take a pre-trained model, which has already learned features on large datasets, and use it to adapt to a new, smaller dataset. We fine-tuned the GPT-sl-base model using the dataset we created in the earlier stages of our project. During the fine-tuning process, we trained the model for 3 epochs with a batch size of 16 and a learning rate of 5e-5. We trained the model on a Kaggle platform, on a double Tesla T4 GPU with 16GB of memory, and the training took approximately 12 hours before capping out due to Kaggle limitations.

The results of the fine-tuned model were quite promising, with a training loss drop from 0.36 to 0.076. Although we were not able to train the model for the intended number of epochs, we still managed to achieve a considerable improvement in the model's performance on the generated dataset. This demonstrates that the GPT-sl-base model is an effective starting point for the development of a Slovene paraphrasing LM. Further work could explore the use of other pre-trained models or larger datasets to achieve even better results.

Evaluation

To evaluate the performance of the fine-tuned GPT-sl-base model, we used the BLEU score [15], which is a commonly used metric for evaluating the quality of machine translation output. BLEU stands for ”bilingual evaluation understudy” and compares a machine-generated translation to one or more human-generated translations. The score is calculated by comparing the number of n-grams (contiguous sequences of words) in the machine-generated translation that also appear in the human-generated translation(s). The BLEU score ranges from 0 to 1, with 1 indicating a perfect match between the machine-generated and human-generated translations.

In our case, we used the BLEU score to evaluate the performance of our model in generating paraphrases in Slovene. We evaluated the model on a test set of 1000 sentence pairs, where each sentence was paraphrased by our fine-tuned GPT-sl-base model. We then calculated the BLEU score for each pair of sentences and reported the mean and standard deviation of the scores. The mean BLEU score we obtained was 0.515 with a standard deviation of 0.255.

While the BLEU score is a widely used metric for evaluating machine translation, it has some limitations. One limitation is that it only considers the number of matching n-grams in the machine-generated and human-generated translations and does **not take into account the semantic similarity** between the two. Furthermore, the BLEU score can be misleading when evaluating translations that are significantly different from the reference translation(s). In our case, as we were evaluating paraphrasing, the BLEU score could only capture some aspects of the quality of the generated sentences, and it is important to consider additional metrics, such as human evaluation, to fully assess the quality of the generated paraphrases.

Human evaluation

In addition to the previous metrics for evaluation, we added the human evaluation where we considered three people that are familiar with the problem and have different biases towards the data. Since the dataset is too large and consequently would take an enormous amount of time to manually evaluate each pair of sentences, we randomly selected 100 instances for evaluation. Our grade system was composed of five numerical grades (1, 2, 3, 4, 5), where 5 means that the meaning of the paraphrased sentence is totally matched with the meaning of the original sentence and 1 means that the meaning of the paraphrased and the original sentence is completely unmatched.

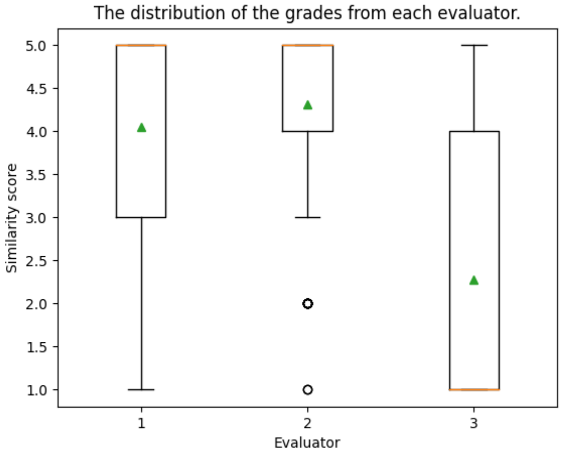

Figure 1. Box plot of evaluators on similarity of 100 sentence pairs.

As we can see on Figure 1, the results from the first two evaluators show decent performance of the fine-tunned model.

Discussion

In general the performance of the model is very accurate on short sentences and paragraphs (ex. original sentence: ”V slabšem položaju smo.”, and paraphrased: ”mi smo v slabšem položaju.”). The model can perfectly handle numbers (ex. original sentence: ”štiri dolarje in 27 centov.”, and paraphrased: ”štiri dolarje in sedemintrideset centov.”), and also there were a lot short sentences that were duplicated, mainly because of the fact that they were too short to be paraphrased. The models also can handle acronyms form paragraphs related to laws (ex. original sentence: ”Uradni list Evropske unije L 210 z dne 31. julija 2006”, and paraphrased: ”(UL L 210, 31. julij 2006”). On the other hand, the model failed to paraphrase the longer sentences, a common occurrence is to first successfully paraphrase the sentence and then add some meaningless random text, since the model failed to generate the ”end of sentence” token. Also the model uses the change of gender related pronouns to paraphrase the original sentence which is not correct.

Conclusion

In conclusion, we collected the data in Slovenian language for our research, trained a large language models such as GPT, but the Slovenian variant, and achieved decent performance which can be comparable with the paraphrasing models that are on other languages except English, since a lot of work is done only for the English language. Although the model has lower performance with the longer sentences and paragraphs, it works very well on short sentences and paragraphs which shows promising results.

Future work

Our next goal would be to improve our fine-tunned GPT model in term of performance by selecting better fine-tuning parameters, by enlargement of the dataset or consider some other some other pre-trained large language models for text generation.

Also, we would like to enrich our set of performance metrics by adding the ROUGE metric, more precisely ROUGE-L metric which will compare the Longest Common Subsequence between the generated paraphrased sentence and the true paraphrased sentence.

In addition to the computational evaluation, we would like to improve the human evaluation by considering more instances for comparison which will give us a better of the performance of our fine-tunned model.

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 computa-*

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

[11] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[12] CJVT. CJVT/gpt-sl-base. https://huggingface.co/cjvt/gpt-sl-base, 2021.

[13] J Edward Hu, Rachel Rudinger, Matt Post, and Benjamin Van Durme. Parabank: Monolingual bitext generation and sentential paraphrasing via lexically-constrained neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6521–6528, 2019.

[14] Iztok Lebar Bajec, Andraž Repar, Jure Demšar, Žan Bajec, Mitja Rizvič, Borut Kumperščak, and Marko Bajec. Neural machine translation model for slovene-english language pair rsdo-ds4-nmt 1.2. 6. 2022.

[15] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*, 2002.