FRI Natural language processing course 2021
Project report

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

Paraphrasing sentences

Franc Benjamin Demšar, Lenart Rupnik, Gregor Zadnik

Abstract

In this report we describe the basic idea of a pipeline for generating a model which paraphrases slovenian sentences. We incorporate ideas from related works and present our initial datasets.

Keywords

Paraphrase, Translation, Slovenian, English

Advisors: Slavko Žitnik, Aleš Žagar

1. Introduction

Paraphrasing is a fundamental task in natural language processing that involves generating alternative expressions with the same meaning as a given sentence. The ability to generate accurate and diverse paraphrases is crucial for various applications, such as text summarization, machine translation, and question answering systems. In this project, we focus on the task of creating paraphrases from the Slovene language. Our objective is to explore and compare two different approaches: a synonym-based model which will serve us as a base model and a newer transformer model, specifically the T5 model.

The synonym-based model utilizes a collection of synonyms and employs rules to replace words in the original sentence, generating alternative sentence structures with similar meaning. This approach relies on the assumption that substituting words with their synonyms can preserve the original intent and convey the same message.

On the other hand, the transformer model, particularly the T5 model, is a state-of-the-art language model based on the transformer architecture. This model has demonstrated exceptional performance in various natural language processing tasks, including paraphrasing. By leveraging the power of attention mechanisms and contextual understanding, the T5 model can generate high-quality paraphrases that capture the nuanced meaning of the original sentence.

Since there are no publicly available dataset with Slovene paraphrases we aim to produce one by utilizing the GigaFida [1] corpus and the back-to-back translation technique, we aimed to create a dataset that uses a wide range of paraphrases for Slovene sentences. This dataset serves as the foundation for training and evaluating our paraphrasing models, including the synonym-based model and the transformer-based T5 model.

Through this project, we aim to assess the effectiveness and quality of the paraphrases generated by these models, mostly comparing results from the base model with the ones from current state-of-the-art transformer model. We will evaluate the paraphrases based on criteria such as semantic adequacy, fluency and diversity.

2. Related works

For a general overview of the task of paraphrasing, we turned to Zhou and Bhat, 2021 [2]. We decided to use the translation method to generate training data.

Federmann et al., 2019 [3] explore the possible approaches to translating text. They found that Neural machine translation (NMT) performs better than human translation in many cases. It is also mostly free which makes it a good option for us students. Another important finding was that paraphrasing works best when translating into and from related languages.

Colin Raffel et al. in paper ”Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer” [4] introduces the T5 model, which is a text-to-text transformer that can be applied to various natural language processing tasks, including paraphrasing. It presents extensive experiments and evaluations of T5 on different datasets and tasks, demonstrating its effectiveness and versatility.

”Paraphrase Generation with Deep Reinforcement Learning” by Zhiguo Wang et al. [5] explores the use of T5 models for specifically paraphrase generation. It proposes a deep reinforcement learning approach to train a T5-based model that generates diverse and high-quality paraphrases. The model is evaluated on multiple benchmark datasets,

For our evaluation metrics we drew inspiration from Shen et al., 2022 [6] who compare existing paraphrasing metrics and propose their own: ParaScore. We will evaluate the model automatically with their proposed metric BLEU and manually with subjective metrics such as adequacy, diversity and fluency.

3. Data

The dataset we are using is a subset of the entire Gigafida corpus, called ccGigafida [7]. It contains approximately 9% of the entire Gigafida dataset and is available online.

To train the models for paraphrase generation we needed to use our data and construct ('text', 'paraphrased text') pairs. We did that by translating our corpus into english and back to slovene. We treat the double-translated text as a paraphrase to be used for training. For translating we used the Neural Machine Translation model for Slovene-English language pair [8].

During testing we generated around 7000 paraphrases. Looking at the results, we found that many of the sentences in ccGigafida are not sentences at all but rather just texts such as results of sports matches, names, numbers, addresses, links etc. The NMT model had issues translating these texts, which resulted in low-quality training data. As we could not properly paraphrase them, they needed to be filtered out. We did so by utilising the classla [9] pipeline - a tool for processing of standard Slovenian, Croatian, Serbian and Bulgarian. With the help of this tool we tokenized and part-of-speech tagged the sentences and filtered out those not containing a verb. We also removed sentences that had too many special characters such as numbers, dashes, quotation marks and so on. During this step we also divided larger paragraphs into separate sentences and treated each one as a standalone unit. Thus we created a subset of data, containing about 25000 sentences that can easily be paraphrased. It was this subset that we then used in our final solution.

4. Proposed models

4.1 Synonyms

For our base-line method we chose an approach using synonyms. We iterate through each sentence and search words in a set of synonyms. For each sentence we take 20% of best synonyms (always at least one) and replace them with our initial words. If we haven't found any synonyms in our set, then we have our original phrase as paraphrase. For our set of synonyms we used Thesaurus of Modern Slovene 1.0 [10]. Dataset proved to be useful in a way that each synonym came with score of how good of a synonym it is. Paraphrases generated by this method will most likely tend to turn out less fluent as original but it should be at least somewhat diverse and adequate.

4.2 T5

The T5 or text-to-text model consists of a single transformer encoder-decoder architecture, where both the encoder and decoder are composed of multiple layers of self-attention and feedforward neural network blocks. The model is trained in a text-to-text fashion, where the input and output sequences are represented as text strings, and the task is framed as a text-to-text translation problem.

The T5 model uses a shared weight strategy, where the same set of weights are used for both the encoder and decoder, as well as for all the different tasks that the model is trained on. This allows the model to efficiently share knowledge across different tasks and datasets, and enables the transfer of pre-trained knowledge to new tasks with minimal additional training.

In addition to the shared weight strategy, the T5 model also incorporates several other techniques to improve its performance, including:

- Pre-training on large-scale unsupervised tasks to learn general-purpose language representations
- Fine-tuning on task-specific supervised datasets to adapt the model to specific tasks
- Data augmentation techniques to increase the amount of training data available
- Multi-task training, where the model is trained on multiple tasks simultaneously to further improve its ability to transfer knowledge across tasks.

Since this model is based on text-to-text predictions, it can be easily used for paraphrasing tasks. For our implementation, we used a pre-trained t5 model that was trained on Slovenian text corpus [11]

5. Methodolgy

To use our selected models, we first had to transform the text to tokens that can be used in NLP models. For this purpose, we used methods like tokenization and lemmatization that will be further described in this section, as well as how we trained and used our models. All training was done in python environment with the help of libraries for NLP, such as: PyTorch, Keras, Transformers, Pandas, etc.

5.1 Tokenization

To use paraphrases as input for any natural language processing model, we need to first tokenize the sentences into individual words or sub-words. Once the paraphrases are tokenized using a suitable tokenizer, they can be represented numerically and used as input for the model. The numerical representation can be in the form of a sequence of integers or one-hot encoded vectors, depending on the specific requirements of the model. For tokenizing paraphrases, we can use various approaches, depending on the specific requirements of the model and the task at hand. In our case, we use two different tokenizers. For LSTM model we used *BLANK* tokenizer and for T5 model we used already pretrained tokenizer from HuggingFace[11] since it a good practice to use the same tokenizer as the one that was used for pre-training the model.

5.2 Lemmatization

When dealing with simple synonym model for paraphrase generation, we also used lemmatization in order to preserve grammatical correctnes. This was done with Classla 2.0 [9]. We used lemmatization also when computing results, because Slovenian language is very rich with many declensions and conjugations.

5.3 Transformers model training

To train the T5 model, we used a pre-trained model from official Slovene Hugging Face site called Center za jezikovne vire in tehnoligije Univerze v Ljubljani (CVJT UL) [12]. This is a research unit focusing on scientific research as well as the development and maintenance of key digital language resources and language technology applications for contemporary Slovene. The developed resources and applications have practical value and are accessible to all the Slovene language users around the world. From here we downloaded a pre-trained T5 model which was trained on Slovene text. That way, our model already had some knowledge about Slovene language. Since we had some limitations regarding our computational power, we worked on a single GPU (RTX 3060), we had to use slightly smaller T5 model called T5-small available at [11]. Additional informations about the model and training parameters are shown in Table 1

Table 1. Configuration of the pre-trained T5 model as well as parameters used for additional training.

model	encoder lay.	decoder lay.	parameters	batch size	seq. length
T5-small	8	8	$\sim$ 60M	16	128

While training, we also used linear learning rate schedule to additionally improve the results from our model.

5.4 Metrics

Next, we will describe automatic techniques as well as manual matrices that were used. For manual evaluation, we based our technique on what that Federmann et al., 2019 used. For automatic evaluation, we based our metric on that Shen et al., 2022 [6] introduced.

5.4.1 Manual metrics

For manual metric, we used the following:

- **Paraphrase adequacy**: We measure semantic similarity between the original sentence and it's candidating paraphrase. We measure it on a scale from 0 to 100 where 0 is not adequate and 100 is perfectly adequate.
- **Paraphrase fluency**: When measuring fluency of candidating paraphrase we compare it with original sentence and select which one is more fluent. We measure it on a scale from -50 to 50 where -50 denotes that original sentence is much more fluent than it's candidating paraphrase, 50 the opposite and 0 neither of them.
- **Paraphrase diversity**: When measuring diversity of the original and candidating paraphrase we assign it a score from 0 to 100 where 0 means that sentences are identical and 100 that they are well diverse. We pay attention to more meaningful changes.

When measuring with manual metrics, we used 21 simple sentences in Slovene. We gathered them from the news, which was also the source of the Gigafida dataset that we used for training models.

5.4.2 Automatic metrics

In terms of automatic evaluation of our generated paraphrases, we decided to use BLEU and METEOR metrics as they were suggested in the initial paper [6]. BLEU, or the Bilingual Evaluation Understudy, is a score for comparing a candidate translation of a text to one or more reference translations. Although developed for translation, it can be used to evaluate text generated for different natural language processing tasks, such as paraphrasing and text summarization.The authors mention that BLEU score is not perfect, but it's quick and inexpensive to calculate, language-independent, and, above all, correlates highly with human evaluation. METEOR seeks to overcome limitation of BLEU in measuring semantic equivalents in low-resource languages. METEOR also demonstrates a stronger correlation with human judgement at the sentence or segment level compared to BLEU. In addition to using BLEU and METEOR metrics for automatic evaluation of our generated paraphrases, we also incorporated BERT-Score as an additional metric. Although the initial paper [6] recommended BLEU and METEOR, we recognized the need for a metric that captures the contextual and semantic similarity between the generated paraphrases and the reference sentences. BERT-Score leverages pre-trained BERT (Bidirectional Encoder Representations from Transformers) models to compute the similarity score between two sentences. It considers the contextual information and semantic meaning, making it suitable for evaluating paraphrases and capturing nuances that traditional n-gram-based metrics might miss. By incorporating BERT-Score as an evaluation metric, we aimed to further enhance the quality assessment of our paraphrases. BERT-Score provides a more comprehensive understanding of the semantic similarity, allowing us to better evaluate the generated paraphrases in terms of their contextual relevance and meaning, which is crucial for tasks like natural language processing.

6. Results

.

6.1 Smaller database

In the first step, we focused on a smaller database consisting of 7,000 cases. The results presented here are specifically for the Transformers model since the behavior of synonyms remains consistent regardless of the database size, as it does not require additional learning.

In 13 cases, the adequacy stayed at 100. In most of the cases it was because the sentence did not change. In those

Mean Value	Adequacy	Fluency	Diversity
Average	90.7	-2.6	11.0
Modus	100	0	0

Table 2. Table with average and modus scores of metrics for 21 samples that different models scored.

cases also the fluency and diversity resulted in 0. In 2 cases the paraphrase was more fluent than the original, and in 4 cases the original was more fluent. 2 sentences were more fluent because our model made incoherent paraphrases that were essentially just double copies of the original one. In 7 cases the paraphrases were at least a bit diverse and in 3 cases sentences were very diverse. Table 2 shows us that in most cases paraphrases did not change it's original meaning and were very as fluent as original and not diverse at all. It is that way mainly because in most cases, proposed paraphrases were only copies of original sentences. Table 2 also shows us that in most of the original meaning stayed the same and that paraphrases were only slightly less fluent than original sentences and also that paraphrases only slightly differ to original sentences.

6.2 Whole database

In the second step we already used whole database which consists of 30,000 cases. In this section are presented results both for Transformers model and synonyms method.

7. Links

- https://viri.cjvt.si/gigafida/
- https://aclanthology.org/D19-5503.pdf
- https://www.kaggle.com/datasets/aemreusta/paraphrased-articles-using-gpt3
- https://github.com/afader/oqa
- https://slovenscina.eu/prevajalnik

References

[1] Gigafida.

[2] 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.

[3] Christian Federmann, Oussama Elachqar, and Chris Quirk. Multilingual whispers: Generating paraphrases with translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 17–26, Hong Kong, China, November 2019. Association for Computational Linguistics.

[4] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. 21(1), jan 2020.

[5] Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li. Paraphrase generation with deep reinforcement learning, 2018.

[6] Lingfeng Shen, Lemao Liu, Haiyun Jiang, and Shuming Shi. On the evaluation metrics for paraphrase generation, 2022.

[7] Nataša Logar, Tomaž Erjavec, Simon Krek, Miha Grčar, and Peter Holozan. Written corpus ccGigafida 1.0, 2013. Slovenian language resource repository CLARIN.SI.

[8] 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. Slovenian language resource repository CLARIN.SI.

[9] Nikola Ljubešić and Kaja Dobrovoljc. What does neural bring? analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 29–34, Florence, Italy, August 2019. Association for Computational Linguistics.

[10] Simon Krek, Cyprian Laskowski, Marko Robnik-Šikonja, Iztok Kosem, Špela Arhar Holdt, Polona Gantar, Jaka Čibej, Vojko Gorjanc, Bojan Klemenc, and Kaja Dobrovoljc. Thesaurus of modern slovene 1.0, 2018. Slovenian language resource repository CLARIN.SI.

[11] lala.

[12] Center za jezikovne vire in tehnologije univerze v ljubljani.

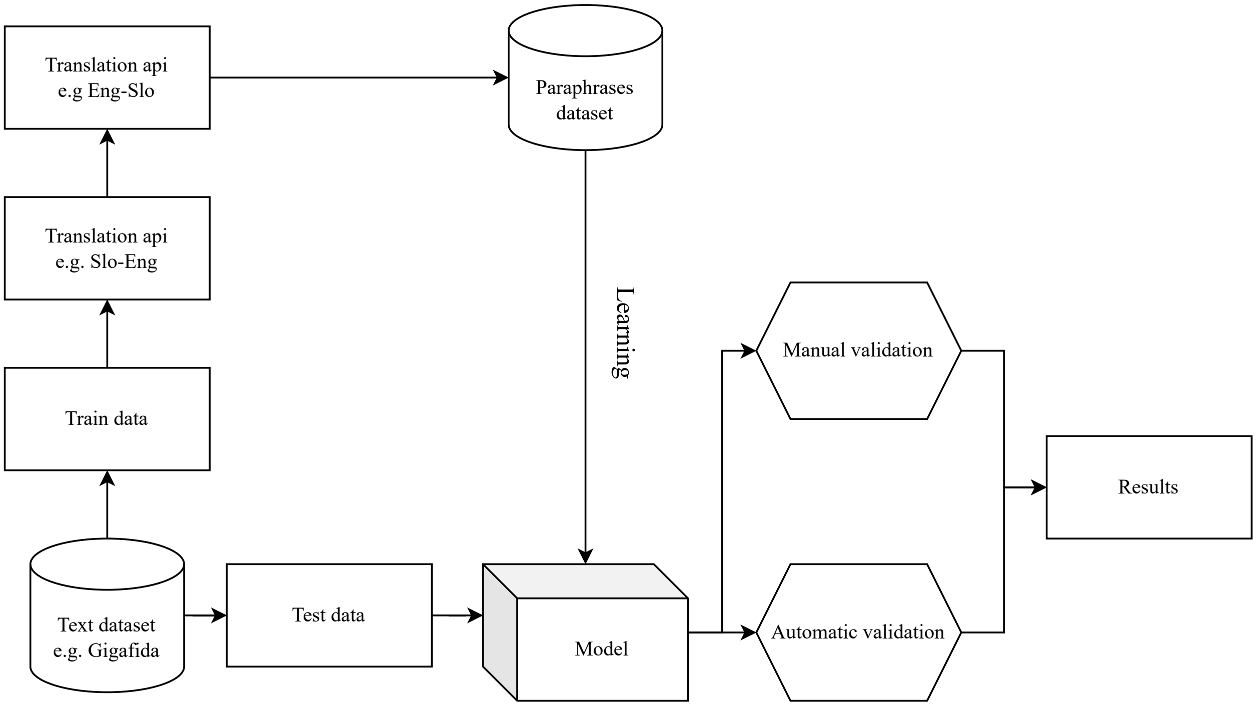

Figure 1. Visualization of a basic pipeline. Detailed visualization of a pipeline we will use to make a model which generates paraphrases.