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

Project 3: Paraphrasing sentences

Matej Vatovec

Abstract

Paraphrasing is the act of restating or rephrasing a piece of text. In this article we tested two models for paraphrase generation. We fine tunned and tested the cjvt/t5-sl-small model and compared it to chatGpt3. The biggest challenge was the lack of data, which we solved by using Back-translation. At the end we compared if the automatic evaluation metrics (BLEU, ROUGE-N and ROUGE-L) match the manual evaluation.

Keywords
Paraphrasing, NLP

Advisors: Slavko Žitnik

Introduction

Paraphrasing is a widely studied natural language generation task. It aims to generate alternative sentence phrasings that expresses the same meaning. It can be used in tasks such as question answering, machine translation, semantic parsing and data augmentation [1]. In recent years, the emphasis has gradually switched from traditional approaches to more advanced neural approaches, resulting in new directions and architectures.

Related work

One of the more classical approaches is Rule-based approach. D. Lin and P. Pantel [2] used it for question-answering. They proposed a unsupervised algorithm for automatic discovery of rules. However, the limitations of this and similar algorithms have led to decreased performance.

More recent approaches are based on neural networks. With the recent advances (especially the sequence-to-sequence framework) the use of NN became more popular for paraphrase generation. Most of current existing approaches are based on sequence-to-sequence models with combination of Encoder-Decoder architecture [1].

A. Gupta [3] proposed a method based on a combination of deep generative models (VAE) with sequence-to-sequence models LSTM). It is simple, modular and can generate multiple paraphrases, for a given sentence.

For better quality of generated paraphrases a framework was proposed by E. Egonmwan and Y. Chali [4]. It combines the effectiveness of two models: transformer and sequence-to-sequence. The transformer layer learns to capture long term dependencies, while the GRU-RNN encoder generates the state vector for decoding.

Another challenge is the evaluation of such methods. In general two types are commonly used: automatic evaluation and human evaluation. A more recent study [5] compared those two types. They found out that most commonly used metrics do not align well with human annotation and proposed a new evaluation metric ParaScore. Based on their experimental results the method significantly outperforms existing metrics.

Data

A large amount of data is required to construct a solid model for paraphrasing. We used the Gigafida v1.0 for our first dataset [6]. It is a reference corpus of written Slovene. It contains newspapers, magazines, selected online texts, literary texts, factual texts, textbooks and other material. To obtain pairs of sentences we used a method called Back-translation. We translated each sentence from Slovene to English and back to Slovene. For this we used the google translation API. Because there are a lot of sentences, this takes some time, and we won't be translating the whole dataset. Overall we acquired around 5 million sentence pairs. Therefore our model will be fine tuned on approximately 800000 pairs of sentences.

After the data was acquired we proceeded with the preprocessing. Since we started running out of time we decided to use only a subset of data for fine tuning (500000 sentence pairs). The first thing we did was to remove pairs of sentences that had None values. Next we removed pairs where sentences were the same after back-translation. This will

hopefully result in more diverse paraphrases.

Before we can fine tune the model, we must first encode the input data in a format that the model will understand. This job is done by the tokenizer. It splits the input strings into sub-word tokens and converts them to ids.

Methodology

T5 model

For the first model we decided to use the T5 model. More precisely we used the cjvt/t5-sl-small. Which is the checkpoint with 60 million parameters. T5 is an encoder-decoder model pre-trained on a multi-task mixture of unsupervised and supervised tasks. It has 8 encoder and 8 decoder layers. It was trained on the following corpora: Gigafida 2.0, Kas 1.0, Janes 1.0, Slovenian parliamentary corpus siParl 2.0 and slWaC.

ChatGpt3

As our baseline model we will be using one of the best models for paraphrasing Slovene sentences. ChatGpt3 is based on the OpenAI's GPT-3.5. It's a task-specific GPT that was fine-tuned to target conversational usage.

Evaluation metrics

Automatic evaluation

We used two methods for automatic evaluation. The first one is BLEU. It was originally developed for measuring translation performance, but it can also be used for paraphrases. It's quick and language independent. It requires two components: a closeness metric (word error rate) and a corpus of good quality human reference paraphrases. In its simplest form BLEU is the quotient of the matching words under the total count of words in hypothesis sentence [7].

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) works similarly to BLEU by comparing paraphrases to other human created paraphrases. We used the ROUGE-N and ROUGE-L. ROUGE-N calculates the n-gram recall between a candidate and a set of reference.

$$\text{BLEU} = \text{BP} \cdot \exp(\sum_{n=1}^{N} W_n log(p_n)$$

Where BP and p_n are:

$$p_n = \frac{\sum_{\text{n-gram} \in \text{hypothesis}} \text{Count}_{\text{match}}(ngram)}{l_{\text{n-gram}}^{\text{hyp}}}$$

$$\text{BP} = \begin{cases} 1, & \text{if } c < r \\ e^{(1-r/c)}, & \text{otherwise} \end{cases}$$

ROUGE-L calculates the weighted harmonic mean combining the precision score and the recall score [8].

$$\text{ROUGE} = \frac{\sum_{\text{unigram} \in \text{reference}} \text{Count}_{\text{match}}(unigram)}{l_{\text{ref}}^{\text{unigram}}}$$

$$\text{ROUGE-L} = \frac{\text{LCS(reference,hypothesis)}}{l_{\text{ref}}^{\text{unigram}}}$$

Where LCS is longest common subsequence

Manual evaluation

We will evaluate a couple of paraphrases manually, based on three different criteria:

- Paraphrase fluency: The paraphrase should be grammatically correct, well-structured, and fluent in the target language. It should adhere to the conventions of syntax, grammar, and vocabulary usage.
- Paraphrase adequacy: The most crucial criterion is whether the paraphrase accurately conveys the same meaning as the original sentence. The paraphrase should maintain the core message, intent, and information of the original sentence.
- Diversity: A good paraphrase should use alternative words or phrases to express the same idea as the original sentence. It should demonstrate a diverse and appropriate vocabulary while maintaining clarity.

Results

We manually evaluated the models on the 5 sentences listed below. Since the sample size is very small we tried to select sentences that are diverse in structure and length.

Sentences that were manually evaluated:

1. **Original**: To je test
   chatGPT3: To je preizkus.
   t5-sl-small: To je preizkus.
2. **Original**: Kot razlog je navedel, da je bil lačen, ker je izpustil zajtrk.
   chatGPT3 Kot razlog je navedel, da je bil lačen zaradi tega, ker je preskočil zajtrk.
   t5-sl-small: Za razlog je navedel, da je bil lačen, ker je izpustil zajtrk.
3. **Original**: Po Sloveniji so zagoreli številni kresovi.
   chatGPT3 V različnih delih Slovenije so se vžgali številni kresovi.
   t5-sl-small: V Sloveniji so zagoreli številni kresovi.
4. **Original**: SpaceX je v vesolje uspešno izstrelil svojo komercialno raketo.
   chatGPT3 Podjetju SpaceX je uspelo uspešno izstreliti svojo komercialno raketo v vesolje.
   t5-sl-small: SpaceX je uspešno izstrelil svojo komercialno raketo v vesolje.

5. **Original**: Sheldon je rekel: 'Za kosilo želim jesti kitajsko hrano.'
   chatGPT3 Sheldon je izrazil svojo željo po kitajski hrani, ko je dejal: "Za kosilo bi rad jedel kitajsko hrano."
   t5-sl-small: Sheldon je rekel: "Za kosilo želim jesti kitajsko hrano."

Model	Adequacy		Fluency		Diversity	
	t5	cGPT3	t5	cGPT3	t5	cGPT3
1	10	10	10	10	3	3
2	10	10	7	8	2	4
3	8	10	10	8	2	5
4	10	10	7	7	4	6
5	10	9	10	8	1	7

Table 1. Results of manual evaluation and chtGPT3 evaluation.

The results of manual evaluation are shown in Table 1. As we can see our model compares quite well with chatGpt3. It's main drawback is the lack of diversity. The generated sentences are very similar to original ones (the 5. sentence is exactly the same only different quotes were used). Because of this, fluency and adequacy are retained quite effectively.

The results of automatic evaluation is shown in Table 2. The ROUGE-N and ROUGE-L scores match with the results of the manual evaluation. The BLEU score, on the other hand, is quite the contrary.

Model	BLEU	ROUGE-N	ROUGE-L
t5-sl-small	$1.331 \cdot 10^{-231}$	0.76	0.726

Table 2. Automatic evaluation of t5-sl-small.

Conclusion

In this article we fine tuned the model T5-sl-small for Slovene paraphrase generation and compared it to chatGpt 3. The generated paraphrases are of good quality but not much different from the original sentences. A possible solution to this problem could be to fine tune the model on longer blocks of text. Future work can focus on fine tuning models with different architectures and collecting higher-quality data.

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, 2021.

[2] Dekang Lin and Patrick Pantel. Discovery of inference rules for question-answering. *Natural Language Engineering*, 7(4):343–360, 2001.

[3] Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. A deep generative framework for paraphrase generation. In *Proceedings of the aaai conference on artificial intelligence*, volume 32, 2018.

[4] Elozino Egonmwan and Yllias Chali. Transformer and seq2seq model for paraphrase generation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 249–255, 2019.

[5] Lingfeng Shen, Haiyun Jiang, Lemao Liu, and Shuming Shi. Revisiting the evaluation metrics of paraphrase generation. *arXiv preprint arXiv:2202.08479*, 2022.

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

[7] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[8] Lin CY ROUGE. A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization of ACL, Spain*, 2004.