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

Paraphrasing sentences

Nik Pirnat, Anže Glušič, Martin Bavčar

Abstract

In this report, we show our work on the project for the course *Natural Language Processing*. We address the problem of paraphrasing Slovenian sentences, where we try to rephrase them while keeping meaning intact. We back-translate a corpus using a NMT model, and use the resulting pairs of text to fine-tune existing language models to perform the task of paraphrasing. We compare the results of paraphrasing with three of our models.

Keywords

Sentence paraphrasing, ParaScore, Natural language processing, Back-translation, Neural machine translation, Neural networks

Advisors: Slavko Žitnik

Introduction

Paraphrasing sentences is a natural language processing task, where the goal is generating sentences with different phrasing while still retaining the original meaning. The task can be used in different applications of natural language processing, such as text summarization, question answering and machine translation. With the growing need for such NLP applications, the need for high-quality sentence paraphrasing has also grown. However, this task is difficult to solve as natural languages are complex as for each sentence there exist many different paraphrases that are difficult to automatically evaluate.

In this paper, we firstly focus on creating a Slovene data set of sentence paraphrasing, using the method of back-translation [1]. For this method we use data sets consisting of an extensive collection of Slovene text of various genres [2], which we use to translate to English then back to Slovene, this creates a paraphrased sentence that can be added to a new data set. The newly created data set will then be used to train and evaluate our generative models that are capable of producing rephrased sentences.

With this, we also tackle the problem of creating a data set for a language that has less resources than some of the most popular ones. With translation comes a price of losing quality of paraphrases and of the overall data set, which is why we also develop a metric to evaluate the data set [3].

Overall, our goal is to contribute a quality Slovene data set for sentence paraphrasing that can serve as a tool to develop and evaluate generative models on it.

Related work

In this section, we review some of the related work that has been done in the field.

In the paper Paraphrase Generation: A Survey of the State of the Art [4] Zhou and Bhat do a general overview of the field working on the task of sentence paraphrasing. They split the approaches into 4 groups: rule-based, thesaurus-based, smt-based and neural approaches. They conclude that no method for now achieves satisfactory results and additional research needs to be done. For future work, they propose combining the use of pretrained language models with other mechanisms, such as reinforcement learning or GAN.

In their paper, Federman et al. [1] compare human, neural machine translation (NMT) and hybrid back-translation paraphrasing techniques to direct paraphrasing by humans. Hybrid techniques employed NMT with post-edits from humans. The main metrics used were fluency (*how natural the paraphrase sounds*), diversity (*how different the paraphrase is*) and adequacy (*how much meaning it retains*). They found that in general paraphrasing without translation provides better adequacy and fluency, with non-experts generally being less diverse, but more adequate. Some NMT methods outperformed human experts in diversity at a substantially lower cost. Using hybrid techniques yields the best overall results, with post-editing also helping overcome the issues of lower quality data. On large data sets, using NMT can provide results that are adequate and almost as fluent, while being slightly less diverse compared to human-based approaches, but at much lower cost.

Recent paper, about evaluation metrics for paraphrase generation [3], shows how reference-free metrics get better performance than reference-based and how commonly used metrics don't align well with human annotations. While analyzing flaws of current methods of evaluating performance, they also introduce new metrics

$$ParaScore = \max(Sim(X,C), Sim(R,C)) + \omega * DS(X,C),$$

where $Sim(X,C)$ and $Sim(R,C)$ denotes *semantic similarity* between input sentence (X), output candidate (C) and reference (R) paraphrase, which represents an optimal paraphrase. $DC(X,C)$ calculates the lexical or syntactic differences, while ω represents a hyperparameter. Its performance is better than all current metrics, even without using reference paraphrases ($ParaScore.Free$).

Methods

Data acquisition

To train and evaluate our paraphrasing model, we created a dataset by doing back-translation using NMT model for Slovene-English language pair [5] on Slovene data set ccKres [2].

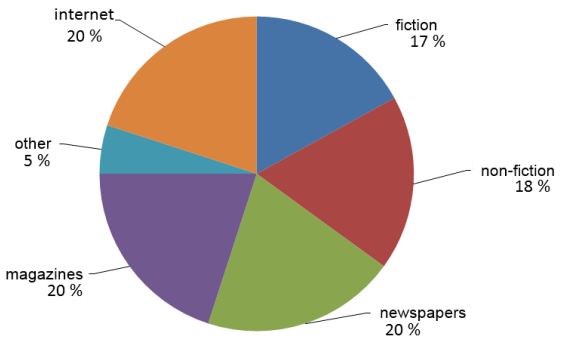

Figure 1. Representation of types of text in Kres dataset. Balanced (text types and genres) 9% sampling of Kres 1.0 (balanced Slovene corpus), which itself is sampled from Gigafida 1.0 (unbalanced Slovene corpus).

We preprocessed the dataset by cleaning the text, removing extremely long sentences that would not fit our hardware limitations and converting all text to lowercase, acquiring 109560 usable text pairs. We then split the dataset into training, validation, and test sets, with 60% of the data used for training, 20% for validation, and 20% for testing.

Paraphrasing Model and training

The model consists of an encoder that processes the input sentence and generates a fixed-length representation, and a decoder that generates the paraphrased sentence based on the encoder output and an attention mechanism that highlights the relevant parts of the input sentence.

We utilized a paraphrasing model based on the T5 architecture developed by Google. Specifically, we employed the pre-trained models "t5-sl-large" and "t5-sl-small" from the HuggingFace model hub. These models were fine-tuned on our built dataset using two Nvidia V100 32GB GPUs. Training took 21 minutes for t5-sl-small and 8 hours for t5-sl-large. We noticed that using a larger batch size (above 8) negatively affected the quality of the resulting models.

Evaluation metrics

To evaluate the quality of our paraphrasing model, we developed a custom score that measures the similarity between the generated paraphrase and the original sentence. Our custom score is loosely based on the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric, which is commonly used in natural language processing tasks.

We designed our metric to take into account the semantic similarity between the original sentence and the paraphrase, the ratio of lengths and word overlap. Specifically, we used pre-trained word embeddings to compute the cosine and jaccard similarity between the embeddings of the original sentence and the paraphrase, then weight them and incorporated this into our custom score.

We evaluated the performance of our model using our custom score on the test set, as well as several other commonly used metrics, including BLEU (Bilingual Evaluation Understudy) and ROUGE.

Results

We evaluated the performance of our proposed T5-based paraphrasing model (referred to as "T5 Model") and compared it against a baseline model that only substitutes some words in the paraphrases (referred to as "Baseline Model"). The evaluation was conducted using a diverse test dataset consisting of original phrases and their corresponding paraphrases.

We employed three evaluation metrics to assess the quality of the paraphrases generated by the T5 Model and the Baseline Model: BLEU score, ROUGE-1 score, and custom metric score.

The BLEU score measures the n-gram overlap between the generated paraphrases and the reference phrases, providing a measure of grammatical correctness and lexical similarity. Higher BLEU scores indicate better paraphrase quality. ROUGE-1 Score

The ROUGE-1 score measures the overlap of unigram sequences between the generated paraphrases and the reference phrases. It evaluates the similarity in word usage and content preservation. Higher ROUGE-1 scores indicate better paraphrase quality.

The custom metric score captures multiple aspects of paraphrase quality, including n-gram overlap, length difference, semantic similarity, and fluency score. The custom metric combines these factors using weighted coefficients to provide a comprehensive measure of paraphrase quality. Higher custom metric scores indicate better paraphrase quality.

We present the results of our evaluation in Table 1, comparing the performance of the T5 Model and the Baseline Model across the evaluation metrics.

Model	BLEU Score	ROUGE-1 Score	Custom Metric Score
T5-small Model	0.101	0.482	0.326
T5-large Model	0.123	0.502	0.303
Baseline Model	0.078	0.213	0.229

Table 1. Performance comparison of the T5 Model and Baseline Model

We can see that both the T5 models considerably outperform the naive baseline model, while the large T5 model achieves slightly better results than its small counterpart.

While evaluating our model we noticed it runs into issues when the original text contains names as we can see in the examples below.

Original paragraph:

Iz Bosne in Hercegovine je prišel Sekula Dugandžić, iz Hrvaške Petar Grgec ..., iz Jugoslavije Vladimir Kepić in Dobrosav Milojević, iz Slovenije pa Stane Novak, Irena Polanec, Ciril Povše in Marko Skok.

Generated paragraph:

Iz Bosne in Hercegovine je prišel Sekula Dugandžić iz Bosne in Hercegovine, iz Hrvaške Petar Grgec..., iz Jugoslavije, Vladimir Kepić in Dobrosav Milosav Milosav Milosav Milosav Milosav.

We see that our model continued repeated a random name until the end of the paragraph.

Discusion

In addition to evaluating the T5-small-sl Model and the Baseline Model, we also experimented with the T5-large-sl Model to further explore the performance of different T5 variants. The evaluation results of all three models are summarized in Table 1.

The T5-small Model achieved a BLEU score of **0.101**, indicating a moderate level of paraphrase similarity compared to the original sentences. The ROUGE-1 score was **0.482**, suggesting that the generated paraphrases had limited overlap with the reference sentences in terms of unigram matches. The custom metric score for the T5-small Model was **0.326**, indicating a moderate overall performance when considering both semantic similarity and fluency.

Similarly, the T5-large Model demonstrated slightly improved performance compared to the T5-small Model. It achieved a higher BLEU score of **0.123**, indicating a better level of paraphrase similarity. The ROUGE-1 score was **0.502**, suggesting increased overlap with the reference sentences in terms of unigram matches. However, the custom metric score for the T5-large Model was slightly lower at **0.303**, indicating a relatively weaker performance in terms of semantic similarity and fluency compared to the T5-small Model.

The Baseline Model, as expected, performed worse than both T5 models across all evaluation metrics. The BLEU score was **0.078**, indicating a lower level of paraphrase similarity compared to the T5 models. The ROUGE-1 score was **0.213**, suggesting a lower overlap in unigram matches with the reference sentences. The custom metric score for the Baseline Model was **0.229**, indicating poorer performance in terms of semantic similarity and fluency.

Overall, the T5-small Model and the T5-large Model outperformed the Baseline Model in generating high-quality paraphrases. The T5-large Model showed a slight advantage over the T5-small Model in terms of paraphrase similarity and overlap with reference sentences, although the custom metric score was slightly lower. This indicates that the T5-large Model has the potential to capture more complex linguistic patterns and generate more accurate paraphrases.

It is important to note that although the T5 models demonstrated promising performance, we ran out of time to explore other advanced models such as OpenAI's Davinci due to time constraints. Future work could consider incorporating the Davinci model and comparing its performance against the T5 models to gain further insights into the effectiveness of different language models for paraphrase generation tasks. Also, future research could focus on fine-tuning the T5 models on larger and more diverse training datasets to improve their performance.

In conclusion, our results highlight the potential of T5-based models, specifically the T5-small Model and the T5-large Model, for paraphrase generation tasks. These models outperformed a baseline approach in terms of paraphrase similarity and overlap with reference sentences. Further investigation with other advanced models and continued research efforts are necessary to advance the state-of-the-art in paraphrase generation and unlock their full potential in practical applications.

References

[1] 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.

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

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

[4] 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.

[5] 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.