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

Paraphrasing sentences

Uroš Škrjanc, Jaka Mesarec, and Grega Rovščček

Abstract

Paraphrasing sentences is a fundamental task in language processing with numerous applications, ranging from machine translation to plagiarism detection and text simplification. Deep learning models have gained significant popularity in this field, and this study explores the potential of t5-sl-small and gpt-sl-base models for paraphrasing in the Slovenian language. By adopting a translation-based approach, the models generate paraphrases by translating sentences into an intermediate language and back into Slovene. We utilized three online translation services such as Clarin, Deepl, and Google Translate to generate our paraphrases. We evaluated the performance of the models using various metrics, including BLEU and METEOR, as well as human evaluation. The results indicate that the models exhibit promising performance in generating meaningful and syntactically correct paraphrases. Our research revealed that the T5 model, trained using Clarin-translated sentences, demonstrates superior performance in generating high-quality paraphrases. However, the models do exhibit certain limitations, offering directions for future research in this domain.

Keywords

Paraphrase generation, Natural language processing, T5 model, GPT model, Quality evaluation

Advisors: Slavko Žitnik, Aleš Žagar

Introduction

Paraphrasing sentences plays a crucial role in various aspects of language processing, ranging from machine translation and text summarization to information extraction, question answering, syntactic and semantic analysis, and sentence parsing. It serves as a valuable tool for plagiarism detection by altering sentences that exhibit similarity. Furthermore, sentence paraphrasing can aid in text simplification for complex or technical content and optimize website content for improved search engine discoverability.

In recent years, the field of paraphrasing has witnessed a significant surge in the use of artificial neural networks, mirroring the broader trend of deep learning's impact across diverse domains of machine learning and artificial intelligence. While machine paraphrasing may exhibit lower accuracy in terms of capturing the exact similarity of meaning compared to human paraphrasing, it offers distinct advantages in terms of speed, cost-effectiveness, and the ability to generate paraphrases with a more varied vocabulary. This is particularly true when the intermediate language utilized in the paraphrasing process belongs to a different language group, and there exists a substantial corpus of texts for training.

In this paper, our motivation lies in exploring the potential of deep learning models, specifically the t5-sl-small model and the gpt-sl-base model, for paraphrasing in the Slovenian language. Slovene, as a unique language with its own syntactic and semantic intricacies, presents an interesting challenge for paraphrase generation. To tackle this, we adopt a translation-based approach, leveraging the process of translating a sentence into an intermediate language and back into the original language. By ensuring accurate translation, the resulting paraphrased sentences should retain the same meaning as the original ones. Additionally, to further assess the models, we employ three different online translators for our intermediate translation: Google Translate, Clarin, and Deepl.

Through rigorous experiments and detailed analysis, we aim to gain deep insights into the strengths and limitations of the t5-sl-small and gpt-sl-base models in the context of Slovene paraphrase generation, and shed light on potential directions for future research to enhance paraphrasing performance.

The rest of this paper is organized as follows. Section 2 provides an overview of related work in the field of paraphrase generation, highlighting the advancements and limitations of existing approaches. In Section 3, we present the methodology used in this study. The results and analysis of our experiments are presented in Section 4. Section 5 delves into the discusion

based on the presented results and suggests some potential areas future research. Finally, Section 6 concludes the paper, summarizing the key findings and contributions of this work.

1. Related works

Several studies have explored the task of sentence paraphrasing using deep learning models and have investigated evaluation metrics to assess the quality of generated paraphrases.

Federmann et al. [1] introduced a multilingual approach to paraphrase generation by leveraging neural machine translation models. They reported impressive results, demonstrating the effectiveness of strong neural machine translation models in generating paraphrases. Their approach not only outperformed human-generated paraphrases in terms of efficiency, cost, and diversity, but also provided a valuable alternative for paraphrasing tasks.

Papineni et al. [2] introduce BLEU, which addresses some of the limitations of previous metrics (NIST and ROGUE). They describe the key components of BLEU, including the modified n-gram precision, the brevity penalty, and the geometric mean. They explain the rationale behind these components and how they contribute to a more robust evaluation. They present experiments and evaluations conducted on various machine translation systems, showing that BLEU correlates well with human judgments and provides consistent results.

Another metric was proposed by Banerjee and Lavie [3] as a alternative to other metrices. According to the authors Meteor correlates better with human evaluation. The authors proved this by correlating Meteor scores with human judgments of translation quality using Arabic-to-English and Chinese-to-English datasets.

Prakash et al. [4] conducted research on generating paraphrases via pivoting using neural machine translation. Their evaluation demonstrated superior performance of their model compared to several existing methods, including sequence-to-sequence and bi-directional LSTM models, based on metrics such as BLEU, METEOR, and TER.

Although Zhou and Bhat [5] highlighted significant advances in state-of-the-art results, they acknowledged the existing limitations and expressed the need for further improvement. In their article, they discussed potential research directions that hold promise in achieving better outcomes.

A paper on the topic of evaluation metrics in paraphrase generation by Shen et al.[6] considers different metrics of paraphrase evaluation and investigates their correlation with human judgment. It poses questions about quality and limitations of existing metrics and presents a detailed analysis. It also proposes a new metric, called ParaScore, which, by their findings, significantly outperforms existing metrics.

Methods

To gather a suitable dataset for paraphrasing, we selected a substantial number of sentences from the ccGigafida corpus. In order to streamline the process, we initially focused on sentences ranging from 5 to 25 words in length. These sentences were then preprocessed by removing any characters that could potentially hinder the translation process. Subsequently, we randomly sampled a portion of these sentences to be translated using three popular online translators: Clarin, Google Translate, and Deepl. We translated the sentences from Slovene to English, and then back to Slovene. A quick manual inspection of the resulting translations, where we removed sentences that stayed the same, enabled us to obtain approximately 27,000 sentences from Clarin and Google Translate, while Deepl contributed around 11,000 sentences. These sets of sentences were deemed suitable for training and evaluating our neural networks.

Having acquired an appropriate dataset, we proceeded to build two distinct types of paraphrasing models (T5 and GPT2). To ensure comprehensive coverage, we trained each of these models on all three compilers, resulting in a total of six unique models. The training process consisted of multiple iterations, with each model being trained in incremental steps of 50 epochs.

We utilized the existing models cjvt/t5-sl-small and cjvt/gpt-sl-base as our starting point. These models were initially subjected to 50 epochs to establish a foundation. Subsequently, the resulting models were further fine-tuned through an additional three steps of 50 epochs, culminating in a total training duration of 200 epochs for each model.

To create diverse and representative training data, we randomly selected 10,000 sentences for each training session.

Once the models were trained, we evaluated their performance using a set of previously unseen sentences that were not included in the training process. We randomly selected 100 sentences from the database of untranslated sentences that were not used in the model training. As with the training, sentences containing a minimum of 5 and a maximum of 25 words were selected for testing, and the same sentences were used to test all models.

When dealing with the output sentences generated by the T5 model during paraphrasing, we faced the challenge of selecting the most suitable sentence. The T5 model often produced multiple sentences as its output, and we needed to determine which one best captured the intended paraphrase. To tackle this issue, we employed a metric called perplexity [7].

Perplexity

Perplexity is a measure commonly used in natural language processing to assess the quality or fluency of a language model. It quantifies how well a language model predicts a given sequence of words. Lower perplexity values indicate that the model is more certain and accurate in predicting the next word in a sequence, suggesting a higher level of fluency [8].

By incorporating perplexity into our selection process, we improved the dependability and coherence of the chosen sentence. This allowed us to ensure that the paraphrase we

selected not only captures the original meaning effectively but also maintains a high degree of naturalness when expressed in Slovene.

During the evaluation, we focused on several key aspects:

1. **Distance or similarity:** We assessed the degree of similarity between the original sentences and their translated counterparts.
2. **Word differences:** We measured the number of words that differed between the translated sentences and their original counterparts.
3. **Syntactic correctness:** We evaluated the syntactic correctness of the translated sentences, ensuring that they maintained grammatical accuracy.
4. **Sound assessment:** We conducted manual assessments to gauge the overall sound and naturalness of the translated sentences.

To further help us with the evaluation, we decided to try the metrics listed bellow, as they are some of the standard metrics used in this field of research:

1. BLEU
2. METEOR

By utilizing these metrics, we sought to gain a comprehensive understanding of the performance of our paraphrase generation models. Now, let us provide a brief overview of the mentioned metrics.

BLEU

BLEU is a widely used metric that serves as a standard measure for evaluating machine translation and paraphrase generation. It calculates the n-gram overlap between the generated paraphrases and a set of reference sentences, offering a quantitative assessment of similarity [2, 9]. Range of BLEU is form 0 to 1.

METEOR

METEOR is another evaluation metric that addresses some of the limitations of BLEU. It incorporates additional linguistic features such as synonymy, stemming, and word order to provide a more comprehensive evaluation of paraphrase quality [10, 3]. Range of METEOR is from 0 to 1.

Human evaluation

In addition to the established evaluation metrics, we devised a simple yet effective metric to assess the quality of the generated paraphrases. This metric encompasses four distinct criteria, each assigned a specific point value:

- **Relevance:** This criterion evaluates the semantic consistency of the paraphrased text compared to the original. Scores ranging from 0 to 3 are assigned based on the level of semantic coherence.
- **Difference:** The paraphrased expression should demonstrate noticeable differences from the original sentence. Point values from 0 to 3 are assigned to reflect the degree of differentiation.
- **Fluency:** This criterion focuses on the fluency and naturalness of the paraphrased expression. Scores ranging from 0 to 3 capture the overall smoothness and coherence of the generated text.
- **Preservation of dates, names, and numbers:** It is important to consider the proper preservation of specific entities such as dates, names, and numbers during paraphrasing. A binary score of 0 or 1 is assigned to indicate whether these elements are appropriately maintained.

Each category within the metric is assigned a specific range of values, and the paraphrased data is evaluated accordingly. The final score is determined by summing the scores from all four categories. If the paraphrased text is identical to the original, the score is equal to 0. Conversely, if all criteria are successfully met, the score reaches its maximum value of 10. This metric provides a comprehensive assessment of the quality and effectiveness of the paraphrasing process. The inspiration for this metric stems from an article on the evaluation of paraphrase generation [11].

ParaScore

Let us also mention ParaScore. Parascore, as proposed by Shen et al. [6], is a metric that takes into account both the semantic similarity between words and the lexical variety. It aims to generate paraphrases that are not only semantically similar but also employ different words, enhancing diversity in the output.

Although this metric looked promising, we decided not to use it in our final evaluation of the models, as we did not find a way to implement it, where it would produce meaningful results for Slovene paraphrase evaluation.

Results

In this section, we will go over the results obtained from our evaluation of the paraphrasing models. We will look at the performance of both the T5 and GPT models, each trained on sentences generated by the before mentioned translators (Clarin, Deepl, and Google Translate).

To gain insight into the paraphrasing quality exhibited by the models, we begin by examining a representative example sentence that underwent paraphrasing using all of the models. This carefully selected sentence offers a glimpse into the overall performance of the models, providing a preliminary understanding of the quality of the generated paraphrases. We will only show the results of the models trained on 50 epochs and 200 epochs, to keep it tidy. It is also important to note that this example serves as a starting point for our evaluation and is not meant to represent the complete range of outputs from the models.

(The numbers in the model name represent the number of epochs the model was trained with)

- **Original sentence:** Po svoji naravi je svetovljanka in se znajde v katerikoli deželi, povsod je lahko doma.
- **GPT-Clarin-200:** Po svojem ustroju je svetovljanka in se znajde v katerikoli deželi, povsod je lahko doma.
- **GPT-Clarin-50:** Po svoji legi je namreč svetovljanka in se povsod znajde v kakršnihkoli krajih, povsod je lahko doma.
- **GPT-Deepl-200:** Po svoji naravi je pravo nekakšen svet, je povedala po svoji naravi.
- **GPT-Deepl-50:** Po svoji naravi je svetouvenka in se znajde v katerikoli deželi, povsod je lahko doma.
- **GPT-Google-200:** Po svoji naravo je svetovljanka in se znajde v katerikoli deželi, povsod je lahko doma.
- **GPT-Google-50:** Po svoji naravi je namreč svetovljanka in se znajde v katerikoli dežele, povsod je lahko doma.
- **T5-Clarin-200:** Po svoji naravi je svetovljan in se lahko prosto giblje v vseh deželah in povsod je lahko doma.
- **T5-Clarin-50:** Po svoji naravi je svetovljan in lahko se znajde v katerikoli deželi, kjerkoli je lahko doma.
- **T5-Deepl-200:** Po svoji naravi je svetovljan in se lahko znajde v katerikoli državi, kjerkoli je doma.
- **T5-Deepl-50:** Po naravi je popotnik in se lahko znajde v katerikoli deželi in kjerkoli na svetu.
- **T5-Google-200:** Po svoji naravi se znajde v vsaki deželi, povsod je lahko doma.
- **T5-Google-50:** Po svoji naravi se znajde v katerikoli deželi, povsod je lahko doma.

For evaluating the models, we employed a consistent set of 100 sentences across all models. The outputs generated by the models were assessed using BLEU and METEOR metrics, along with human evaluation. It is important to mention that human evaluation was conducted solely on the models trained for 50 epochs and 200 epochs, as evaluating all the sentences for each model would be very time-consuming. By evaluating the models at these specific epochs, we aimed to capture the substantial evolution of the models and ensure that we didn't overlook any significant changes in performance. This approach provides us with a comprehensive understanding of the model's progression without necessitating evaluation at every intermediate epoch. The results are presented in table 1

Table 1. Evaluation Results for T5 and GPT Models

Models	Bleu	Meteor	Human Evaluation
T5			
Clarin 50	0.432	0.681	6.31
Clarin 100	0.423	0.658	/
Clarin 150	0.417	0.664	/
Clarin 200	0.427	0.684	6.41
Deepl 50	0.419	0.655	6.44
Deepl 100	0.383	0.638	/
Deepl 150	0.413	0.650	/
Deepl 200	0.410	0.661	6.06
Google 50	0.443	0.693	5.39
Google 100	0.438	0.687	/
Google 150	0.441	0.694	/
Google 200	0.440	0.707	5.17
GPT			
Clarin 50	0.321	0.558	3.82
Clarin 100	0.271	0.506	/
Clarin 150	0.273	0.514	/
Clarin 200	0.236	0.497	3.52
Deepl 50	0.320	0.586	3.85
Deepl 100	0.254	0.523	/
Deepl 150	0.201	0.440	/
Deepl 200	0.182	0.438	3.06
Google 50	0.303	0.566	3.19
Google 100	0.310	0.584	/
Google 150	0.276	0.531	/
Google 200	0.288	0.544	3.02

Discussion

Based on the evaluation results presented in Table 1, we can observe the performance of both T5 and GPT models for paraphrasing. The Bleu and Meteor scores were used as automated metrics, while human evaluation provided additional insights.

For the T5 models, all three translators (Clarin, Deepl, and Google) yielded reasonably high Bleu and Meteor scores, indicating a good level of quality of the paraphrased sentences. The models trained for 50 epochs consistently achieved promising results across all translators, with Bleu scores ranging from 0.419 to 0.443 and Meteor scores ranging from 0.655 to 0.693. The models trained for 200 epochs generally demonstrated slightly higher Meteor scores, suggesting potential improvements in paraphrasing quality over extended training, however looking at the Bleu scores there was a very slight decline with extended training for all of the translators. In terms of human evaluation, the T5 models achieved scores between 5.17 and 6.44, indicating a moderate to good performance in terms of paraphrasing quality.

Turning our attention to the GPT models, we observed lower Bleu and Meteor scores compared to the T5 models, indicating a relatively lower level of quality. The GPT models trained for 50 epochs achieved Bleu scores ranging from 0.182 to 0.321 and Meteor scores ranging from 0.438 to 0.558, which is quite substantially lower compared to all T5 models. However, the models trained for 200 epochs exhibited even lower scores, suggesting that further training did not significantly improve the paraphrasing performance and actually only caused overfitting. Human evaluation scores for the GPT models were also consistently lower compared to T5, ranging from 3.02 to 3.85, indicating a need for improvement in the syntactic and semantic consistency of the paraphrased sentences.

Among the various models evaluated, the T5 model trained on sentences translated by Clarin and trained for 200 epochs

demonstrated the highest level of paraphrasing quality, based on the mean of all three metrics. This finding coincides with our initial expectations as the Clarin translator we employed is specifically designed for translating between Slovenian and English texts. Consequently, it is likely to generate more favorable training examples, thus contributing to the improved performance observed in the T5 model trained on Clarin-translated sentences.

Overall, the results suggest that the T5 models generally outperformed the GPT models in terms of paraphrasing quality. The T5 models trained on data from all three translators consistently achieved higher automated metric scores and human evaluation scores.

Conclusion

In this research paper, we conducted an in-depth investigation into the effectiveness of deep learning models for paraphrasing in the Slovenian language. Our main focus was on two specific models: T5-sl-small and GPT-sl-base. Leveraging a translation-based approach, we utilized online translation services such as Clarin, Google Translate, and Deepl to train and evaluate six distinct models.

During the evaluation process, we employed various criteria to assess the quality of the generated paraphrases. These criteria included measuring similarity to the original text in terms of meaning, identifying word differences between the original and paraphrased versions, evaluating syntactic correctness, and considering the overall naturalness and fluency of the paraphrases. To ensure comprehensive evaluation, we also employed established metrics like BLEU and METEOR, in addition to conducting human evaluation.

The results obtained observed show that these models exhibited promising capabilities, demonstrating their potential as cost-effective and efficient alternatives to human paraphrasing. The paraphrases generated by the models generally retained the original meaning, while incorporating subtle variations in word choice and sentence structure.

However, it is worth noting that the models also exhibited certain limitations. Some paraphrases displayed deviations in meaning, indicating the need for further refinement and fine-tuning of the models. Additionally, occasional syntactic errors and unnatural phrasing were observed in the generated paraphrases. These findings highlight areas for improvement and potential future research.

Acknowledgments

We would like to thank assistants Slavko Žitnik and Aleš Žagar for all the help and advice they provided.

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] 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, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[3] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[4] Aaditya Prakash, Sadid A Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. Neural paraphrase generation with stacked residual lstm networks. *arXiv preprint arXiv:1610.03098*, 2016.

[5] 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.

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

[7] Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. Masked language model scoring. *arXiv preprint arXiv:1910.14659*, 2019.

[8] Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, Jennifer C Lai, and Robert L Mercer. An estimate of an upper bound for the entropy of english. *Computational Linguistics*, 18(1):31–40, 1992.

[9] Wikimedia Commons. Bleu.

[10] Wikimedia Commons. Meteor.

[11] Anonymous ACL submission. Are we evaluating paraphrase generation accurately?

[12] Chin-Yew Lin and Franz Josef Och. ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 501–507, Geneva, Switzerland, aug 23–aug 27 2004. COLING.