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

Paraphrasing sentences

Nik Pirnat, Anže Glušič, Martin Bavčar

Abstract

In this report, we show our initial work on the project for the course *Natural Language Processing*. It addresses the problem of paraphrasing Slovenian sentences, where we try to rephrase them while keeping meaning intact. We create guidelines for our future work, which will include back-translation using neural machine translation and training generative models for direct paraphrasing.

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

0.1 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].

We preprocessed the dataset by cleaning the text, removing any special characters and punctuation, and converting all text to lowercase. 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.

0.2 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 model "t5-sl-large" from the Hugging Face model hub, which was fine-tuned on a large-scale Slovenian language dataset.

We trained the model using the Adam optimizer with a learning rate of 0.001 and a batch size of 4. The model was trained for 1 epoch.

0.3 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.

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.