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

Project 3: Paraphrasing sentences

Matjaž Zupančič Muc, Marko Novak, and Klemen Laznik

Abstract

The abstract goes here.

Keywords

Sentence paraphrasing,

Advisers: Slavko Žitnik

Introduction

Paraphrasing plays an important role in language understanding tasks, such as question answering, machine translation and semantic parsing. Additionally, it serves as a useful method for data augmentation. The goal of paraphrase generation is to produce alternative versions of a given sentence that may feature different phrasing or structure, yet still accurately convey the original meaning. Creating high-quality paraphrases is a challenging problem in natural language processing (NLP), as it requires a deep understanding of the underlying semantics and syntax of the input sentence. In recent years, there has been a growing interest in using machine learning techniques, particularly transformer models, to automatically generate paraphrases.

Authors of [1] compare translation based paraphrase gathering using human (experts and non-experts), automatic and hybrid techniques. The automatic technique is based on back translation using a neural machine translation (NMT) model. They first translate a source sentence from one language to multiple target languages (i.e pivot languages) and than back to the original language. They found that NMT based paraphrases have a higher diversity compared to paraphrases written by human non experts, but NMT based paraphrases don't reach the adequacy or fluency level provided by expert paraphrases. They found that NMTs corrupt inputs such as slang and typing errors leading to bad paraphrases. They also found that NMTs struggle with negation, i.e NMTs tend to loose or add negation to a sentence. Finally they found that paraphrases generated by pivot languages which are closely related have a higher adequacy and fluency level, while paraphrases generated by pivot languages that are not closely related have higher diversity. They conclude that NMT based paraphrase generation is cheap and diverse, although NMTs produce less fluent outputs post editing could be used to improve the quality with little additional expenditure.

Existing paraphrase generating language models build on large sequence-to-sequence / encoder-decoder language models. The general principle of those models is similar to models used in machine translations where the first part of the model creates an encoded representation of the input, and the decoder part generates a new sequence from the encoding. However, while machine translation models are trained to generate an output which closely represents the input sentence, for paraphrasing the output needs to differ from the input while maintaining information. In Paraphrase Generation: A Survey of the State of the Art [2], the authors present several improvements on top of the encoder-decoder models, which are used by the current state-of-the-art models to achieve better results.

Attention mechanism helps models emphasise the important information within the input. This information is passed to the decoder part of the model as an extra input vector. On the other hand, the **copy** mechanism helps determine whether to generate a new token or copy an input token at each step of decoding. This helps to improve the form of generated paraphrases, but limits the diversity of generated text. The two common approaches to improve the diversity of generated paraphrases are **variational autoencoders** (VAE) and **reinforcement learning** (RL); the benefit of using RL with a pretrained language model is that it doesn't require a large labelled dataset and can perform paraphrasing in unsupervised setting. RL also addresses the problem in the missmatch between the training goal and the evaluation metric. The networks are often trained to maximize the probability of generating the references paraphrases and that evaluated using metrics for paraphrase generation. RL can be trained to maximize the reward indicated as a desired evaluation metric. The models trained this way have also demonstrated the

ability to perform cross-lingual transfer without additional finetuning [3].

Paraphrase evaluation is a challenging problem, and existing metrics have limitations. In [4], the authors propose a new metric named ParaScore, which takes into account both semantic similarity and lexical divergence between the input sentence and candidate paraphrases. ParaScore is a reference-based metric and is defined as the maximum value between the semantic similarity of the candidate paraphrase to the reference paraphrase and the input sentence to the reference paraphrase, plus a weighted measure of lexical divergence. The authors also propose a reference-free version of ParaScore by removing the reference paraphrase from the metric. Authors conducted experiments on four datasets and compared ParaScore to several baselines, including BLEU, Rouge, METEOR, BERT-iBLEU, and iBLEU. The results show that ParaScore performs significantly better (i.e it achieves a higher correlation with human judgement of paraphrase quality) than all the other metrics on all the datasets and is more robust than other metrics. The ParaScore metric addresses some of the limitations of existing metrics, such as BLEU, which is a widely used metric but does not consider lexical divergence, and iBLEU, which considers lexical divergence but is sub-optimal. ParaScore provides a better evaluation of paraphrasing and has the potential to be used in applications such as text simplification and data augmentation.

Methods

INFO: I got rid of the proposed methodology from last submission (we'll add the relevant bits back before the final / third submission).

For the first dataset we took the sentences from Training corpus SUK 1.0 [5], which contains 48 594 Slovenian sentences. We choose this corpus because it contains general, moderately-long sentences. For each sentence we then generated two paraphrased versions, first one by translating the original sentence to English and back to Slovene, and the second one by a two-stage backtranslation from Slovene to English, from English to German, and then back. This yielded good results both in terms of accuracy and diversity. However, as we used an online API to translate the sentences, we only generated a limited dataset due to the free API limitations.

This method is fast and generates good data without any additional effort from our side, so we can easily increase the dataset as we move forward and better determine how much data we need for model training.

The second dataset is constructed using a paraphrase-mining approach. We extract full sentences from two subtitle files of the same movie, each written by a different subtitle provider. For each sentence in the first subtitle file we find a corresponding sentence in the other subtitle file. The final dataset is a set of corresponding sentences. We found that if the two subtitle files are sampled at the same frame rate, we could somewhat align the subtitles and find the corresponding parts, however finding multiple sources of high quality subtitles both sampled at the same frame rate is not easy. As a result, we used a pre-trained BERT model trained on Croatian, Slovenian, and English corpora ([6]) to embed the sentences and discover corresponding sentences by calculating the pairwise cosine distance between the sentence embeddings. To decrease the number of false positive matches, we limited the comparison region of sentences.

This approach can have several positive aspects. Multiple subtitle providers offer variations in translations, idiomatic expressions, and other nuances that can enrich the dataset. Generated paraphrases are potentially more human-like compared to fully automatic methods like back-translation. This is because the sentences are either written or at least verified by humans, ensuring that the paraphrases capture the natural language and nuances of the source text. In contrast, back-translation relies solely on machine translation, which may not always capture the intended meaning and may produce stilted or unnatural language. However the approach also has the following downsides: a subtitle author may include additional context in their translations, based on previous movie scenes, while the other may not. This can result in poor matches between sentences, since the context of the sentences may not align. As a result, the resulting paraphrase dataset may contain inaccuracies. Therefore the approach requires additional human effort to check the matched sentences, which can be time-consuming and costly. Despite the potential issues this approach can still be faster than writing human paraphrases from scratch.

We're looking at two sequence-to-sequence models for our paraphrasing model: BART and T5. Both architectures offer large pretrained models as well as multilingual variants, which could allow us to include examples from other languages in order to improve model performance. We'd also like to include some negative training examples, i.e. pairs of similar sentences that aren't paraphrases.

Results

We evaluated the quality of 200 random paraphrase pairs generated using the paraphrase-mining technique and 100 paraphrase pairs generated using the back-translation technique. We assigned a score from 1 (poor) to 5 (excellent) for each of three metrics to each paraphrase pair.

1. Accuracy: A measure of how well the paraphrase conveys the same meaning as the original text. A paraphrase that is accurate should capture the key ideas and information of the original text while maintaining the same overall meaning and context.

2. Fluency: A measure of how well the paraphrase reads or sounds in the context of the original sentence or text. A paraphrase that is fluent should be grammatically correct, well-structured, and convey the same meaning as the original text, while also being easy to read or understand.

3. Diversity: A measure of how different or unique the paraphrase is from the original text. A paraphrase that is diverse or original should convey the same meaning as the original text, but using different words or structures, and avoiding excessive copying or close similarity to the original text

Scoring

	Back-translation	**Paraphrase-mining**
Accuracy	4.29 ± 0.93	4.12 ± 1.10
Fluency	4.28 ± 0.85	4.17 ± 1.05
Diversity	3.31 ± 1.22	3.03 ± 1.25

Krippendorff's alpha coefficient (Annotators agreement)

	Back-translation	**Paraphrase-mining**
Accuracy	0.44	0.34
Fluency	0.36	0.42
Diversity	0.07	0.12

Discussion

Acknowledgments

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] 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] Tong Niu, Semih Yavuz, Yingbo Zhou, Huan Wang, Nitish Shirish Keskar, and Caiming Xiong. Unsupervised paraphrase generation via dynamic blocking. *CoRR*, abs/2010.12885, 2020.

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

[5] Špela Arhar Holdt, Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Polona Gantar, Jaka Čibej, Eva Pori, Luka Terčon, Tina Munda, Slavko Žitnik, Nejc Robida, Neli Blagus, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Taja Kuzman, Teja Kavčič, Iza Škrjanec, Dafne Marko, Lucija Jezeršek, and Anja Zajc. Training corpus SUK 1.0, 2022. Slovenian language resource repository CLARIN.SI.

[6] Jasmina Pegan, Marko Robnik-Šikonja, Iztok Kosem, Polona Gantar, Primož Ponikvar, and Cyprian Laskowski. Slovenian datasets for contextual synonym and antonym detection, 2022. Slovenian language resource repository CLARIN.SI.