FRI Natural language processing course 2023
Project report

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

Given instructions, the proposed pipeline is as follows: Get an existing dataset (ccGigafida, ccKres, . . .). We might be able to use Slovenian datasets for contextual synonym and antonym detection [5] to obtain synonyms which could be used to generate paraphrased sentences from other corpora. Use Slovene NMTs to generate the dataset of paraphrases with first reference in mind.

Initial intuition of the dataset quality will be formed manually - human evaluation on smaller subgroups of data. Given the size of the data set, the end goal would be to find a consistent way to evaluate the quality of the entire data set, without introducing bias from any chosen method. Manual evaluation will include a few suggested metrics, e.g. similarity, clarity, fluency. We could evaluate the quality of the data set generated by NMTs using contextual similarity between original sentence and the paraphrased sentence. To do this we would embed the sentences using BERT and than compute the cosine similarity between the embeddings. We can use a readability score such as the Flesch Reading Ease score or the Gunning Fog Index to evaluate the clarity of the paraphrased sentences. To estimate the fluency we could use large language models such as BERT and GPT. BERT and GPT can be used to generate fluency scores based on the model's prediction of the likelihood of the sentence being a natural language sentence. Because fluency is closely related to grammatical correctness and syntax we check for grammar and syntax errors. We also though of using ParaScore (caution - this would require tuning to make useful, which is not suitable for this stage of the pipeline).

We'll train a baseline sequence to sequence T5 model on top of the generated dataset. We could use random pattern embeddings or add a diversity term to the loss function to increase the diversity of generated paraphrases. We also consider using a multilingual base model, as it would allow us to also use existing datasets from other languages and transfer that knowledge to Slovenian language.

For training model evaluation, we turn to ParaScore. Given the width of the study [4], we are lead to believe that for the task at hand - paraphrasing sentences, the existing metrics are insufficient in handling neither lexical diverging, nor semantic similarity simultaneously, which are both present. We should also be aware of potential drawback since we introduce a hyper-parameter in need of tuning. The idea behind using ParaScore is mainly to evaluate the quality of our paraphrasing model but are keen to use a similar approach to evaluate the NMTs output, which is the input - data for model training.

Results

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] 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.