FRI Natural language processing course 2021
Project report

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

Sentence Paraphrasing

Blaž Rolih, Grega Šuštar, and Žiga Rot

Abstract

(Abstract will be written when the whole paper is finished.

Keywords
Paraphrasing, Transformers, (to be expanded)

Advisors: Slavko Žitnik

Introduction

Paraphrasing is the process of expressing a given sentence or text in different words while retaining its original meaning. It is a longstanding problem for natural language learning that can be applied in machine translation, question answering, semantic parsing... Traditional methods for paraphrasing rely on rule-based techniques, which have limitations in handling complex sentence structures and capturing the nuances of meaning. To address these challenges, recent research has explored the use of neural networks for paraphrasing. In particular, the current state-of-the-art models use transformers, that are capable of capturing long-range dependencies in sentences, to generate paraphrases. In our work, we will implement and evaluate approaches to this problem using neural networks.

Related Work

To design a sentence paraphrasing system, one must consider 3 basic aspects. The training data, model selection, and evaluation metrics for the evaluation are all key components of a well functioning system.

Zhou et al.[1] overview the current state-of-the-art approaches, and thus cover most of the available datasets for this task. While there are specialised datasets for word-level paraphrase generation, sentence-level datasets are a scarce resource, and even the ones listed come with drawbacks. Quora [2] and WikiAnswer [3] databases consist of questions only, TwitterURL [4] has noisy labels due to automatic labeling and MSCOCO [5] is a computer vision database and is thus not primarily focused on paraphrases.

To tackle this problem, automatically generated databases are often used. By employing machine translation, one can gather only the desired sentences, and translate them into a different language and back. Federmann et al. [6] show that such artificial datasets are better than ones produced by human experts from a word diversity aspect.

For the model aspect, it comes as no surprise that deep neural models are the most commonly used. Like in other fields of natural language processing, deep learning techniques have significantly raised the performance ceiling [1]. While there seem to be no task-specific architectures, many of the commonly used models in NLP make an appearance here as well. Due to the problem's nature, most of them employ an encoder-decoder architecture. As in other NLP areas, there has been a shift from the recurrent- to the transformer-based models [7]. Recently, fine-tuned large language models have also been proven to quickly adapt to new domains, while being highly data-efficient.

Once an appropriate architecture is selected and trained, model evaluation remains. In the context of paraphrase generation, this is not a trivial topic. Popular NLP performance metrics, like BLEU [8], ROUGE [9], and BERTScore [10], do not correctly evaluate generated paraphrases [11]. But these metrics can be used to evaluate some aspects, like lexical divergence. Since a good balance between semantic similarity and lexical divergence makes a good paraphrase, frequently used metrics, which only focus on one aspect (and can also punish the other), can not properly score generated paraphrases. Shen et al. [11] propose a new, problem-specific, metric named ParaScore. It combines scoring for both semantic similarity and lexical divergence into a single score, making it better aligned with the experts' scores. Similar idea comes from Patil et al. [12] with metric named $ROUGE_p$. Since these automatic evaluation methods are not always reliable, we can use human judgment to evaluate the results [6]. To get more information about the results, we can also use Pearson correlation of human evaluation with machine output to determine its reliability [13].

Initial Ideas

To obtain the data we will find an appropriate dataset on clarin.si (ccGigafida, ccKres,...) and use back-translation to obtain a dataset of paraphrases. Alternatively, we can try to find a corpus containing Slovene translation to a different language and use machine translation to translate the non-Slovene part of the dictionary.

After we are done gathering the data, we plan to apply neural-based models to it. We will first fine-tune an existing model. Here, pretrained BART or T5 will be probably used. If left with enough time, a plan is to also train a custom model from scratch. In that case, the previously mentioned BART or T5 become useful as a comparison also.

For the evaluation, we plan to use standard metrics such as BLEU, ROUGE, BERTScore in some aspects, as well as more appropriate ParaScore [11] and $ROUGE_P$ [12]. To determine the validity of automatic metrics, manual human scores in terms of appropriateness, fluency, and syntactic diversity will be used. This way, we can not only determine the quality of our results, but also the quality of automatic metrics used to evaluate our method and therefore calibrate final scores.

References

[1] 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.

[2] Zhiguo Wang, Wael Hamza, and Radu Florian. Bilateral multi-perspective matching for natural language sentences. *CoRR*, abs/1702.03814, 2017.

[3] Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. Paraphrase-driven learning for open question answering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1608–1618, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[4] Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. A continuously growing dataset of sentential paraphrases. In *Proceedings of The 2017 Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pages 1235–1245. Association for Computational Linguistics, 2017.

[5] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'a r, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[6] 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.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[8] 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, 2002.

[9] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[10] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[11] Lingfeng Shen, Lemao Liu, Haiyun Jiang, and Shuming Shi. On the evaluation metrics for paraphrase generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3178–3190, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[12] Omkar Patil, Rahul Singh, and Tarun Joshi. Understanding metrics for paraphrasing, 2022.

[13] Daniel M. Cer, Mona T. Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity - multilingual and cross-lingual focused evaluation. *CoRR*, abs/1708.00055, 2017.