FRI Natural language processing course 2023
Project report

University of Ljubljana
Faculty of Computer and Information Science

Paraphrasing sentences

Jannik Weiß, Nikolay Vasilev, and Jan Jenicek

Abstract

The abstract goes here. The abstract goes here.

Keywords

Keyword1, Keyword2, Keyword3 ...

Advisors: Slavko Žitnik

Introduction

Paraphrasing is the process of restating a text or sentences using different words or sentence structures while retaining the original meaning. Its main goal is to help people clarify information, adapt the information to a specific audience, avoid plagiarism and improve communication. Paraphrase generation enables people to express themselves more clearly and effectively, but it is also a fundamental task in the natural language processing (NLP) (Zhou and Bhat, 2021) as it enables computers to better understand and interpret human languages.

One approach of generating paraphrases is by using the machine translation techniques. This approach is based on the assumption that by translating a sentence into another language and then translating it back to the original language, the generated sentence can be a valid paraphrase. This kind of paraphrasing is also called multilingual paraphrasing, as mentioned in Federmann et al. (2019) and in Guo et al. (2019), and is based on a model that must be trained on at least 2 languages. It leverages the similarities and differences between languages to generate paraphrased sentence with same meaning as the original. In Figure 1 we can see example of that process. In this case the original sentence is in English (EN) and is translated into German (DE), Slovenian (SI) or Czech (CZ) language in order to generate its EN paraphrase.

The paper aims to provide a comprehensive evaluation of monolingual paraphrasing models on four different languages and investigate the potential benefits of a multilingual paraphrasing model. We will start by creating simple monolingual

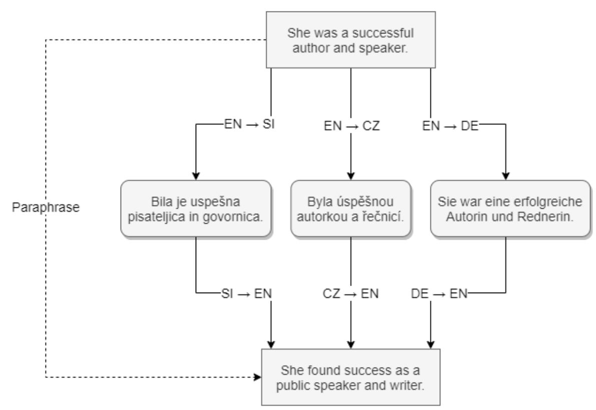

Figure 1. Multilingual paraphrasing of a sentence using the translation approach. Example of paraphrasing a sentence in English (EN) by translating it into Slovenian (SI), German (DE) or Czech (CZ) language.

paraphrasing models using the following languages: English, Slovenian, German, Czech. After these baseline models are evaluated, we will try and improve them by creating one multilingual model, which will be trained on all four languages. We then compare the single- and the multilingual models in order to find out how much of an advantage (if any) the multilingual model has over a monolingual model.

Related work

The most popular approach nowadays for paraphrase generation is to use sequence-to-sequence systems (Zhou and Bhat, 2021), as was first done by Prakash et al. (2016). As is the case with other NLP tasks modern approaches are based on the Transformer architecture (Vaswani et al., 2017). Two main training approaches can be identified to develop paraphrase generation systems:

1. Training on paraphrase data. This intuitive training approach uses sentence pairs which are paraphrases of each other as training data. A recent example of this is Bandel et al. (2022) which also introduces a mechanism to control for semantic, syntactic and lexical dimensions of the generated paraphrase.

2. Training on translation data. Guo et al. (2019) and Thompson and Post (2020) formulate paraphrase generation as translation from one language to the same language, having trained a multilingual model on parallel datasets. This approach makes the system capable of multilingual paraphrase generation by definition. The multilingual training makes this approach better than pivoting (a.k.a. backtranslation) for paraphrase generation, because it avoids the potential errors made in the two translations. However, if the parallel dataset itself is created using machine translation this approach is prone to translation bias.

Apart from specific training methods, research has shown that in general multilingual models can outperform monolingual models on various NLP tasks (Lample and Conneau, 2019).

Evaluation of paraphrase generation is not trivial, due to the generative nature of the task. Like in translation or summarization there are multiple correct answers to any particular paraphrase generation problem. While syntactic and lexical differences can be evaluated using rather simple metrics, evaluating semantic similarity is more involved. A current state of the art metric is Parascore (Shen et al., 2022).

Dataset

We create 4 paraphrase datasets for our approach, one for each language. We aim to create similarly sized datasets for each language, which is why we don't use off-the-shelf popular paraphrase datasets like the PPDB by Pavlick et al. (2015) or the Tatoeba by Tiedemann (2020). In our research these mostly contain enough data for English and often for German, but next to nothing for Slovene.

We use the ParaCrawl dataset as a starting point. This large parallel dataset contains lots of sentences in lots of languages. We extract the Slovene-English, Czech-English and German-English subsets and use machine translation to create Paraphrase data in all four languages. To ensure we have the same amount of data for all languages, we only use records that exist in all 4 languages - 76,879 records in total.

The paraphrase datasets for individual languages are obtained by back-translation of the other language from the pair. This is English sentences for the Slovene, Czech and German paraphrase datasets. For the English one, we decided to back-translate the German sentences as it proved to provide the most consistent results.

We used the Hugging Face Abid et al. (2023) framework to obtain the back-translations. We implemented two methods of translation, the inference API and the local translation, both via pre-trained models. We eventually used the local translation as the use of the inference API proved not to be feasible.

From the pretrained models, we decided to use the ones from *Helsinki-NLP*, namely:

- *opus-mt-en-sla* for the Slovene dataset.
- *opus-mt-en-cs* for the Czech dataset.
- *opus-mt-en-de* for the German dataset.
- *opus-mt-de-en* for the English dataset.

Dataset evaluation

We evaluate the quality of our datasets via human evaluation of a dataset sample and in direct comparison to other popular paraphrase datasets. We evaluate semantic similarity and lexical divergence and calculate a score base on their combination. For this we create a small framework which mixes sentences from different datasets and presents them to the user, which reduces the bias a user might have when knowing which dataset they are evaluating. In Table 1 the results of our human evaluation of our English paraphrase dataset in comparison to the Tatoeba English-English dataset split can be seen.

Language	Our Dataset	Tatoeba
en-en	0.256	0.307

Table 1. Average human evaluation scores of our datasets vs. Tatoeba

Methods

Monolingual Models

In order to best be able to compare our models we use a mt5 model Xue et al. (2022), which is a T5 model trained on multilingual data. There exist separate monolingual models for each of our languages, but they will differ for example in architecture, training methods or training data. By using the same multilingual model for the monolingual baseline we don't use the model in ifs full capacity, but we have the same conditions for all models we want to compare. This puts us in a better position to interpret the results later.

We create a generic training script that we can easily adapt to the different languages to create baseline models. We use the Parascore metric to evaluate our models on the validation dataset split during training.

Multilingual Models

We finetune a single multilingual model based on a pretrained mt5 model for multilingual paraphrase generations in our four languages. The training data for this model is simply the combination of the four datasets we created earlier. We train this model once with an equal amount of data as the baseline models and once with all the data. The former means we include only one fourth of each dataset, and the latter means we include all data, which makes the dataset four times larger compared to the data available to a monolingual model.

Results

Singlelanguage

Our monolingual baseline for english performs already pretty well. After 8 epochs of training the finetuned mt5 model reaches a Parascore score of 0.87.

Multilanguage

Discussion

Conclusion

References

Abubakar Abid, Matthew Carrigan, Lysandre Debut, Sylvain Gugger, Dawood Khan, Merve Noyan, Lucile Saulnier, Lewis Tunstall, and Leandro von Werra. Hugging face library, 2023. URL https://huggingface.co/.

Elron Bandel, Ranit Aharonov, Michal Shmueli-Scheuer, Ilya Shnayderman, Noam Slonim, and Liat Ein-Dor. Quality controlled paraphrase generation, 2022. URL https://arxiv.org/abs/2203.10940.

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. doi: 10.18653/v1/D19-5503. URL https://aclanthology.org/D19-5503.

Yinpeng Guo, Yi Liao, Xin Jiang, Qing Zhang, Yibo Zhang, and Qun Liu. Zero-shot paraphrase generation with multilingual language models, 2019. URL https://arxiv.org/abs/1911.03597.

Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291, 2019. URL http://arxiv.org/abs/1901.07291.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–430, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-2070. URL https://aclanthology.org/P15-2070.

Aaditya Prakash, Sadid A. Hasan, Kathy Lee, Vivek V. Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. Neural paraphrase generation with stacked residual LSTM networks. *CoRR*, abs/1610.03098, 2016. URL http://arxiv.org/abs/1610.03098.

Shen, Liu, Jiang, and Shuming Shi. On the evaluation metrics for paraphrase generation, 2022. URL https://arxiv.org/abs/2202.08479.

Brian Thompson and Matt Post. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation*, pages 561–570, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.wmt-1.67.

Jörg Tiedemann. The tatoeba translation challenge – realistic data sets for low resource and multilingual mt, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. Google's mt5, 2022. URL https://huggingface.co/google/mt5-base.

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. doi: 10.18653/v1/2021.emnlp-main.414. URL https://aclanthology.org/2021.emnlp-main.414.