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, 2021a) 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 singlelingual paraphrasing models on four different languages and investigate the potential benefits of a multilingual paraphrasing model. We will start by creating a simple singlelingual paraphrasing model for each of the following languages: English, Slovenian, German, Czech. After these models are evaluated, we will try and improve them by creating one multilingual model, which will be trained on all four languages. This model will use the paraphrase translation approach, which we already described, in order to paraphrase sentences. After comparing the single- and the multilingual models for each language, we expect to prove that the multilin-

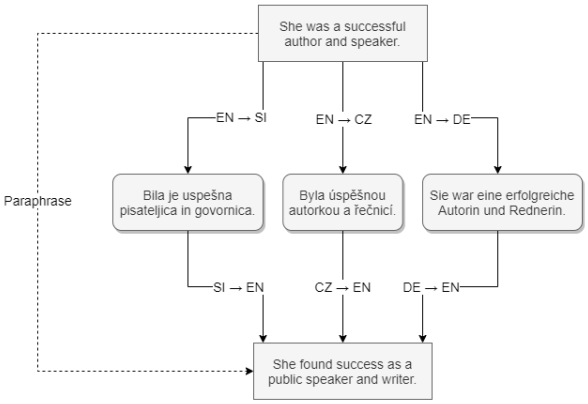

Figure 1. Multilingual paraphrasing of a sentence using the translation approach. Example of paraphrasing a sentence in English (EN) by translating it into Slovenian (SI), German (DE) or Czech (CZ) language.

gual paraphrasing is in general more effective for paraphrasing sentences.

Related work

The most popular approach nowadays for paraphrase generation is to use sequence-to-sequence systems (Zhou and Bhat, 2021a), 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

The described approach uses 4 different datasets (1 for each language). To obtain the datasets, we decided to leverage the PPDB paraphrasing dataset Pavlick et al. (2015) which is the largest English one available. It contains both a single-word to single-word paraphrases and multi-word to multi-word paraphrases in the amount that should be sufficient for our purpose.

With that being said, the PPDB dataset mostly contains phrases which may not be complex enough, as mentioned in Zhou and Bhat (2021b). If this issue appeared during our work, we plan on enhancing our raw English data by Lan et al. (2017), which is a dataset that contains paraphrases collected from Twitter by linking tweets using shared URLs.

We will use the Hugging Face Abid et al. (2023) framework to obtain a multi-lingual one. Dataset will first be separated into 4 exclusive sets and each of them will be used as a base for a particular language. The Slovene, German and Czech datasets will then be translated to their target language using Hugging Face pretrained models that are available as a part of the python library. The English one will be left as-is.

Also, to enhance our data, we plan on leveraging existing datasets that contain data in our target languages. There are already multilingual datasets that build on top of the PPDB dataset, such as Ganitkevitch and Callison-Burch (2014), with different amounts of paraphrases for individual languages.

Methods

Use the Methods section to describe what you did an how you did it – in what way did you prepare the data, what algorithms did you use, how did you test various solutions ... Provide all the required details for a reproduction of your work.

Below are LaTeXexamples of some common elements that you will probably need when writing your report (e.g. figures, equations, lists, code examples ...).

Equations

You can write equations inline, e.g. $\cos\pi = -1$, $E = m \cdot c^2$ and α, or you can include them as separate objects. The Bayes's rule is stated mathematically as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{1}$$

where A and B are some events. You can also reference it – the equation 1 describes the Bayes's rule.

Lists

We can insert numbered and bullet lists:

1. First item in the list.
2. Second item in the list.
3. Third item in the list.

- First item in the list.
- Second item in the list.
- Third item in the list.

We can use the description environment to define or describe key terms and phrases.

Word What is a word?.

Concept What is a concept?

Idea What is an idea?

Random text

This text is inserted only to make this template look more like a proper report. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam blandit dictum facilisis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Interdum et malesuada fames ac ante ipsum primis in faucibus. Etiam convallis tellus velit, quis ornare ipsum aliquam id. Maecenas tempus mauris sit amet libero elementum eleifend. Nulla nunc orci, consectetur non consequat ac, consequat non nisl. Aenean vitae dui nec ex fringilla malesuada. Proin elit libero, faucibus eget neque quis, condimentum laoreet urna. Etiam at nunc quis felis pulvinar dignissim. Phasellus turpis turpis, vestibulum eget imperdiet in, molestie eget neque. Curabitur quis ante sed nunc varius dictum non quis nisl. Donec nec lobortis velit. Ut cursus, libero efficitur dictum imperdiet, odio mi fermentum dui, id vulputate metus velit sit amet risus. Nulla vel volutpat elit. Mauris ex erat, pulvinar ac accumsan sit amet, ultrices sit amet turpis.

Phasellus in ligula nunc. Vivamus sem lorem, malesuada sed pretium quis, varius convallis lectus. Quisque in risus nec lectus lobortis gravida non a sem. Quisque et vestibulum sem, vel mollis dolor. Nullam ante ex, scelerisque ac efficitur vel, rhoncus quis lectus. Pellentesque scelerisque efficitur purus in faucibus. Maecenas vestibulum vulputate nisl sed vestibulum. Nullam varius turpis in hendrerit posuere.

Figures

You can insert figures that span over the whole page, or over just a single column. The first one, Figure 2, is an example of a figure that spans only across one of the two columns in the report.

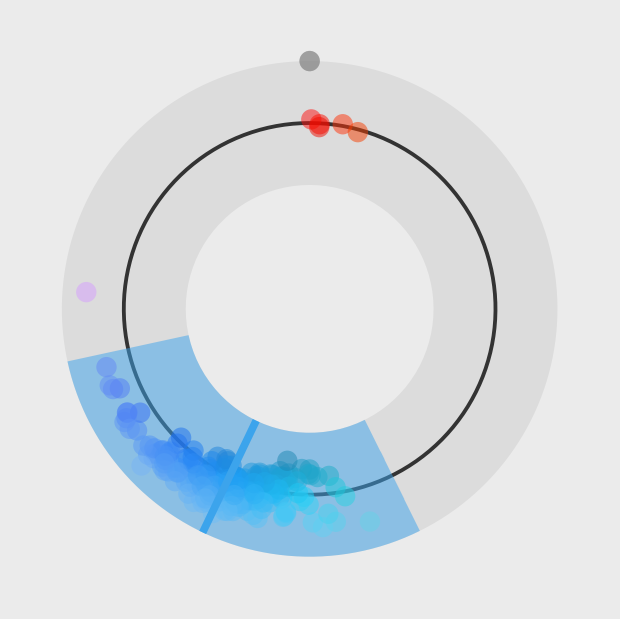

Figure 2. A random visualization. This is an example of a figure that spans only across one of the two columns.

On the other hand, Figure 3 is an example of a figure that spans across the whole page (across both columns) of the report.

Tables

Use the table environment to insert tables.

Table 1. Table of grades.

Name		
First name	Last Name	Grade
John	Doe	7.5
Jane	Doe	10
Mike	Smith	8

Code examples

You can also insert short code examples. You can specify them manually, or insert a whole file with code. Please avoid inserting long code snippets, advisors will have access to your repositories and can take a look at your code there. If necessary, you can use this technique to insert code (or pseudo code) of short algorithms that are crucial for the understanding of the manuscript.

Listing 1. Insert code directly from a file.

```
import os
import time
import random

fruits = ["apple", "banana", "cherry"]
for x in fruits:
  print(x)
```

Listing 2. Write the code you want to insert.

```
import(dplyr)
import(ggplot)

ggplot(diamonds,
           aes(x=carat, y=price, color=cut)) +
  geom_point() +
  geom_smooth()
```

Results

Use the results section to present the final results of your work. Present the results in a objective and scientific fashion. Use visualisations to convey your results in a clear and efficient manner. When comparing results between various techniques use appropriate statistical methodology.

More random text

This text is inserted only to make this template look more like a proper report. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam blandit dictum facilisis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Interdum et malesuada fames ac ante ipsum primis in faucibus. Etiam convallis tellus velit, quis ornare ipsum aliquam id. Maecenas tempus mauris

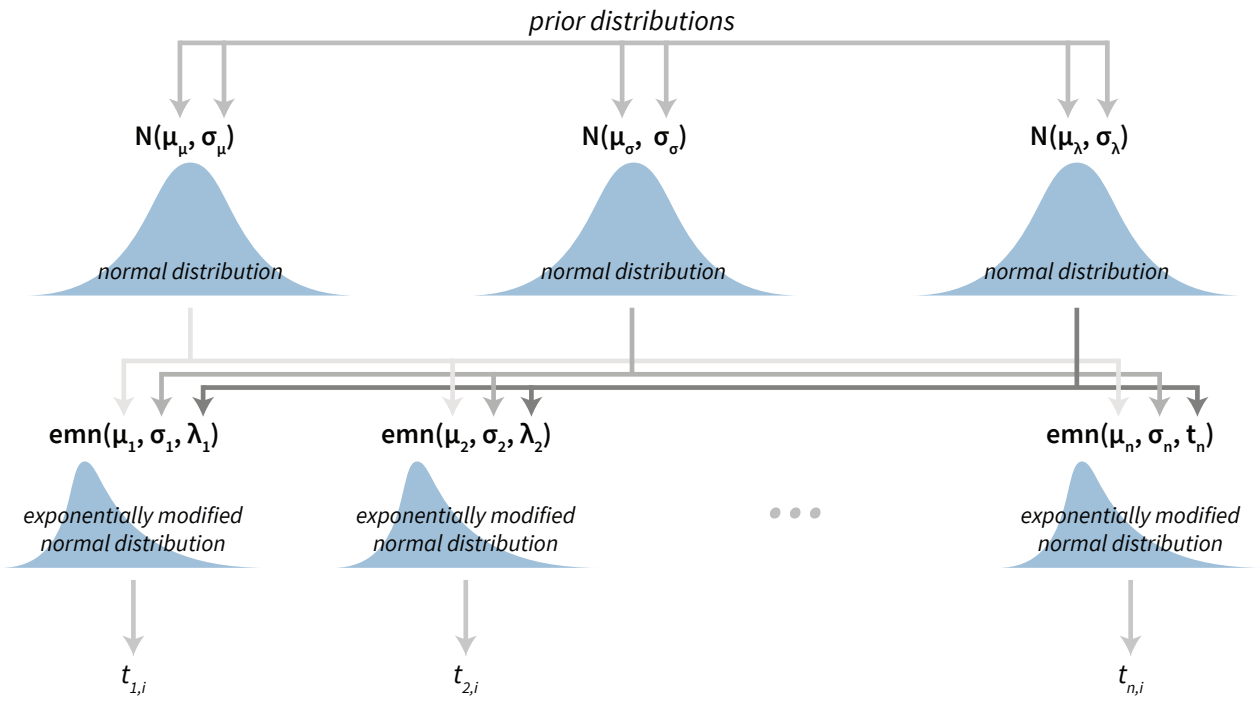

Figure 3. Visualization of a Bayesian hierarchical model. This is an example of a figure that spans the whole width of the report.

sit amet libero elementum eleifend. Nulla nunc orci, consectetur non consequat ac, consequat non nisl. Aenean vitae dui nec ex fringilla malesuada. Proin elit libero, faucibus eget neque quis, condimentum laoreet urna. Etiam at nunc quis felis pulvinar dignissim. Phasellus turpis turpis, vestibulum eget imperdiet in, molestie eget neque. Curabitur quis ante sed nunc varius dictum non quis nisl. Donec nec lobortis velit. Ut cursus, libero efficitur dictum imperdiet, odio mi fermentum dui, id vulputate metus velit sit amet risus. Nulla vel volutpat elit. Mauris ex erat, pulvinar ac accumsan sit amet, ultrices sit amet turpis.

Phasellus in ligula nunc. Vivamus sem lorem, malesuada sed pretium quis, varius convallis lectus. Quisque in risus nec lectus lobortis gravida non a sem. Quisque et vestibulum sem, vel mollis dolor. Nullam ante ex, scelerisque ac efficitur vel, rhoncus quis lectus. Pellentesque scelerisque efficitur purus in faucibus. Maecenas vestibulum vulputate nisl sed vestibulum. Nullam varius turpis in hendrerit posuere.

Nulla rhoncus tortor eget ipsum commodo lacinia sit amet eu urna. Cras maximus leo mauris, ac congue eros sollicitudin ac. Integer vel erat varius, scelerisque orci eu, tristique purus. Proin id leo quis ante pharetra suscipit et non magna. Morbi in volutpat erat. Vivamus sit amet libero eu lacus pulvinar pharetra sed at felis. Vivamus non nibh a orci viverra rhoncus sit amet ullamcorper sem. Ut nec tempor dui. Aliquam convallis vitae nisi ac volutpat. Nam accumsan, erat eget faucibus commodo, ligula dui cursus nisi, at laoreet odio augue id eros. Curabitur quis tellus eget nunc ornare auctor.

Discussion

Use the Discussion section to objectively evaluate your work, do not just put praise on everything you did, be critical and exposes flaws and weaknesses of your solution. You can also explain what you would do differently if you would be able to start again and what upgrades could be done on the project in the future.

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.

Juri Ganitkevitch and Chris Callison-Burch. The multilingual paraphrase database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4276–4283, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/659_Paper.pdf.

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.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. A continuously growing dataset of sentential paraphrases, 2017. URL https://arxiv.org/abs/1708.00391.

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.

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.

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 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.414. URL https://aclanthology.org/2021.emnlp-main.414.

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 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.414. URL https://aclanthology.org/2021.emnlp-main.414.