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

In this report we present 3 approaches to generating data for training paraphrasing models. The data is generated using backtranslation, by matching texts which have been translated by different human translators (i.e. movie subtitles), and by making simple, meaning-preserving augmentations on a corpus of sentences. We evaluate the obtained data based on accuracy, fluency, and diversity. We also explore parascore metric for automatic evaluation of the paraphrases. Finally, we train two T5 and BART models on the prepared dataset and evaluate their results usigng parascore.

Keywords

Sentence paraphrasing, Back-Translation, Paraphrase-Mining, Data Augmentation, Transformers, ParaScore.

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 feature different phrasing or structure, yet still accurately convey the original meaning. Creating high-quality paraphrases is a challenging problem in 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 deep learning to automatically generate paraphrases.

In [1], the authors compare translation-based paraphrase gathering methods: human (experts and non-experts), automatic, and hybrid techniques. The automatic technique involves using a neural machine translation (NMT) model for back translation. NMT-based paraphrases are more diverse than those by non-experts but fall short in adequacy and fluency compared to expert paraphrases. NMT models struggle with inputs containing slang, typing errors, and negation, often leading to poor paraphrases. Paraphrases generated by closely related pivot languages show higher adequacy and fluency, while those from less related pivot languages offer more diversity. The authors conclude that NMT-based paraphrase generation is cost-effective and diverse, but post-editing can enhance quality with minimal additional effort. Existing paraphrase language models build on large encoder-decoder models. These models operate similarly to machine translation models, where the input is encoded and the decoder generates a new sequence. However, paraphrase models aim to produce output that differs from the input while preserving information, in contrast to machine translation models that generate similar representations of the input sentence. 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** and **copy** mechanisms improve paraphrase generation by emphasising important information and determining whether to generate new tokens or copy from the input. To enhance diversity, **variational autoencoders** and **reinforcement learning** (RL) are used. RL, particularly with pretrained language models, enables unsupervised paraphrasing and addresses the training-evaluation mismatch (i.e we can directly optimise the metric of interest) RL-trained models also exhibit cross-lingual transfer without additional fine-tuning [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.

Methods

Back Translation

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.

Paraphrase Mining

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. Bellow you can take a look at two example paraphrases (from the dataset) that we find to be of high quality.

```
--------------------------------------------------------
"Ko te enkrat razglasijo za neprištevnega, potem vsako
tvoje nadaljnje dejanje pripišejo tej norosti."
"Ko si enkrat razglašen za norega, so vsa tvoja dejanja
zaznamovana s tem."

"Ali sta seznanjena z današnjim stanjem na področju
mentalnega zdravljenja, gospoda?"
"Vesta, kakšno je današnje stanje na področju
mentalnega zdravja?."
--------------------------------------------------------
```

We should point out that the approach also has the following downsides: a subtitle author may include additional context (derived from the film plot) in their translations 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.

Data Augmentation

Due to the limited size of our training set, we undertake an exploration of different data augmentation techniques. The following transformations were utilised:

1. **Paraphrase swapping**: Each paraphrase pair $\{p_1, p_2\}$ is used twice, enabling the model to learn mappings: $p_1 \rightarrow p_2$ and $p_2 \rightarrow p_1$. This effectively doubles the size of our dataset and introduces no additional noise.

2. **Paraphrase duplication**: A dataset is constructed in the form of $\{p_i, p_i, 0\}$, where p_i represents some sentence and 0 indicates that the pair is not a paraphrase. This can be useful to penalise models for generating sentences which are too similar or have different meaning.

3. **Synonym replacement**: Random synonyms in sentences are replaced to introduce lexical variations. Initially, we explored automatic techniques using WordNet, but found them to be too noisy. Instead, we manually replaced synonyms to ensure high-quality data augmentation. Hand replacement allows us to preserve the semantic similarity and grammatical compatibility.

T5 Model

T5 (Text-To-Text Transfer Transformer) is a transformer-based model typically used for translation and text summarization. We fine-tune the t5-sl-large model ([7]) on the task of paraphrase generation. For 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. To compute the Parascore on Slovene sentences we used the crosloengual-bert model ([8]).

BART Model

BART is a sequence-to-sequence model, which combines a bidirectional encoder over noisy or masked data with an auto-regressive decoder. When introduced, BART achieved state-of-the-art results for text summarisation [9], as well as several other text-to-text tasks. Therefore we're interested to see how well it performs compared to the T5 model.

Results

Dataset Evaluation

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.
2. **Fluency**: A measure of how well the paraphrase reads or sounds in the context of the original sentence or text.
3. **Diversity**: A measure of how different or unique the paraphrase is from the original text.

Table 1. Overall metric scoring results

	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

The average **accuracy** score for the data-set is 4.29 (in case of back-translation), with a standard deviation of 0.93. The scores provided by the annotators indicate that the dataset generally contains accurate information, with an overall high level of correctness. The relatively small standard deviation suggests that there is a moderate level of agreement among the annotators regarding the accuracy of the dataset. On average, paraphrase-mining did provide a lower accuracy score and a larger deviation among the annotators.

The average **fluency** score for the dataset is 4.28 (in case of back-translation), with a standard deviation of 0.85. The scores indicate that the dataset generally exhibits a high level of fluency, meaning that the language used is smooth and natural-sounding. Similar to accuracy, the small standard deviation suggests a moderate level of agreement among the annotators in terms of fluency. When comparing both data acquisition methods, we can see the trend where back-translation tends to just about outperform paraphrase-mining. Better overall score and better coherency in human evaluation.

The average **diversity** score for the dataset is 3.31 (in case of back-translation), with a standard deviation of 1.22. The scores provided by the annotators indicate that the dataset has a moderate level of diversity, suggesting that it contains a reasonable variety of different examples or elements. The relatively larger standard deviation in this case implies that there might be some divergence in opinions among the annotators regarding the diversity of the dataset. It is important to note that many sentences contain the same content, which results in higher accuracy but lower diversity. Paraphrase-mining yields the biggest score difference in the ability to generate a diverse enough sentence.

High validity and reliability of measurements are crucial for providing meaningful and trustable results. To further explore and interpret the discrepancies in annotators scoring, we turn to Krippendorff's alpha coefficient. It is a reliability coefficient developed to measure the agreement among observers, coders, judges, raters, or measuring instruments drawing distinctions among typically unstructured phenomena or assign computable values to them. Krippendorff's alpha in contrast is based on the observed disagreement corrected for disagreement expected by chance. This leads to a range of -1 to 1 for both measures, where 1 indicates perfect agreement, 0 indicates no agreement beyond chance and negative values indicate inverse agreement.

Table 2. Krippendorff's alpha coefficient results

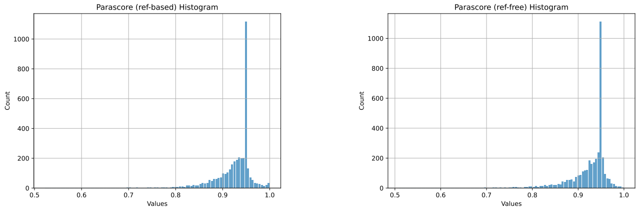
	Back-translation	**Paraphrase-mining**
Accuracy	0.44	0.34
Fluency	0.36	0.42
Diversity	0.07	0.12

Coefficient calculation only rewards the inputs where all annotators agree on the exact same value - category. The equation itself does not imply that data should be ordinal, and similar scores do not contribute positively to the outcome. This generality makes the scoring useful for any data input, but does provide worse results for big or small deviations. It is important to report both standard deviation and annotators agreements scores to interpret the human evaluation accordingly. All scores are skewed towards agreement, but are as close to randomness in case of paraphrase diversity.

Model Evaluation

T5-Base

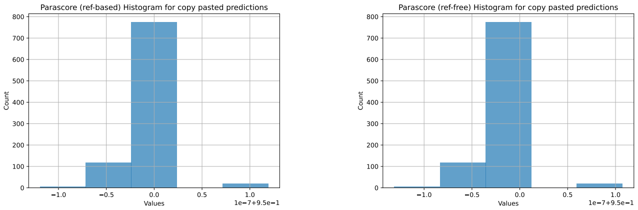
We trained the model on the original dataset (consisting of around 8000 sentences) with the following hyper-parameters: batch size $B = 8$, learning rate $\alpha = 1e-4$ and weight decay $\gamma = 1.5e-4$. We used the AdamW optimiser, and trained the model for 2 epochs (after that the validation loss plateaus). The first row in Table 3 shows the mean ref-based and ref-free parascore and the fraction of paraphrases which are identical to the input sentence. We can see that both the ref-based and ref-free parascore are high, however the copy fraction of 0.25 is also relatively high. Figures 3 (a) and (b) display the histograms of ref-based parascore and ref-free parascore, we can see that majority of values are concentrated around the 0.95 mark.



(a) Ref-based parascore. **(b) Ref-free parascore**

Figure 1. Histograms of parascore values. Parascore is computed on the test set using the finetuned **T5-Base** model.

After further investigation we found that most of the sentences with ref-based and ref-free parascore of 0.95 actually correspond to copy paste paraphrases, this can be seen on Figure 2. This implies that using only parascore to evaluate model performance is not enough and one should also compute at least the fraction of copy-pasted paraphrases.

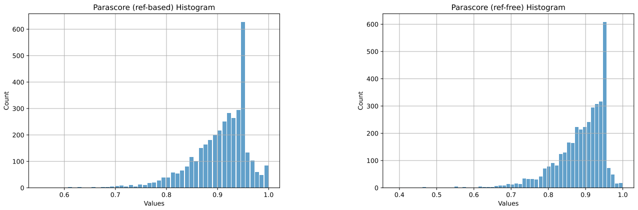


(a) Ref-based parascore. **(b) Ref-free parascore**

Figure 2. Histograms of parascore values for copy paste paraphrases. Parascore is computed on the test set using the finetuned **T5-Base** model.

T5-Aug

Keeping all hyper-parameters constant, we train the model again this time on the augmented dataset (including swapped paraphrases and paraphrases with replaced synonyms). The augmented dataset consists of around 15000 sentences. The model converged after 3 epochs. The second row in Table 3 shows the performance of the T5-Aug model, we can see that both the ref-based and ref-free parascore are lower than what we obtained with T5-Base, however the copy fraction of 0.11 is also significantly lower, implying that T5-Aug model is much more creative when it comes to generating paraphrases. Figures 3 (a) and (b) display the histograms of ref-based parascore and ref-free parascore. We can see that the fraction of paraphrases with the score equal to 0.95 is lower compared to what we observe on Figure 1. Additionally, certain cases exhibit notably lower scores, reaching as low as 0.55 for ref-based parascore and 0.39 for ref-free parascore.



(a) Ref-based parascore. **(b) Ref-free parascore**

Figure 3. Histograms of parascore values. Parascore is computed on the test set using the finetuned **T5-Aug** model.

To gain a deeper understanding of the relationship between the parascore metric and the generated paraphrases, we examine a subset of predictions. Two illustrative examples are presented below, showcasing the paraphrase with the lowest ref-based parascore and the paraphrase with the highest ref-based parascore.

```
------------------------------------------------------
"O prodaji ne vedo"
"Prodaja ne ve, kaj prodaja"
ref-based-parascore=0.55

"Libijo je novembra 2009 z močno gospodarsko delegacijo
obiskal predsednik vlade Borut Pahor."
"Premier Borut Pahor je Libijo obiskal novembra 2009
z močno poslovno delegacijo."
ref-based-parascore=1.0
------------------------------------------------------
```

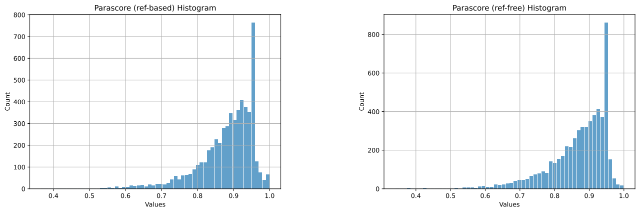
Below are two representative examples that demonstrate the paraphrase with the lowest ref-free parascore and the paraphrase with the highest ref-free parascore. The first example brings attention to another limitation of the parascore metric. Despite ”Začuti žejo po krvi.” being a valid paraphrase of ”Željan je krvi.” the ref-free-parascore equals only 0.39. Interestingly the ref-based parascore for the reference: ”Čuti žejo po krvi.” equals to 0.90. This observation suggests that relying solely on the computation of a single parascore variation, whether it is ref-based or ref-free parascore, may not be sufficient.

```
------------------------------------------------------
"Željan je krvi.",
"Začuti žejo po krvi."
ref-free-parascore=0.39

"Slovenski in Bosanski organi naj bi do 19. septembra
omogočili rešitev tega problema."
"Slovenski in bosanski organi naj bi omogočili rešitev
tega problema do 19. septembra."
ref-free-parascore=1.0
------------------------------------------------------
```

BART-Base and BART-Aug

We trained BART model using the same approach and data as with the T5 model. While the T5 model we used as a starting point has been trained on large Slovenian corpus, we didn’t find an equivalent BART model, so we used a multilingual version of the model which includes Slovenian data. Compared to the T5 model, BART achieved a lower and more varying parascore, which is likely due to not being pretrained specifically for Slovenian language. However, it achieved a significantly lower copy factor, both on base and augmented data (Table 3). Comparing the score distributions (Figure 3, Figure 4), we can see BART had a much higher fraction of paraphrases with score below 0.7, but at the same time also had a higher peak at 0.95.



(a) Ref-based parascore. **(b) Ref-free parascore**

Figure 4. Histograms of parascore values. Parascore is computed on the test set using the finetuned **BART-Aug** model.

Table 3. Performance of all models on the test set. Columns **Para-ref** and **Para-ref-free** denote ref-based and ref-free parascore, while column **Copy Frac.** denotes the fraction of paraphrases that are identical to the source sentence.

Model	Para-ref	Para-ref-free	Copy Frac.
T5-Base	0.91 ± 0.043	0.93 ± 0.04	0.25
T5-Aug	0.91 ± 0.056	0.88 ± 0.066	0.11
BART-Base	0.90 ± 0.066	0.88 ± 0.073	0.15
BART-Aug	0.88 ± 0.078	0.87 ± 0.084	0.09

Exploring Encoder-Decoder Attention

In this section, our aim is to explain the extent to which we can comprehend the behaviour of the model. To achieve this, we intentionally create a relatively complex paraphrase and examine the encoder-decoder attention maps of the fine-tuned T5-Aug model. Our objective is to determine whether the model possesses the capability to establish the same connections that we, as human writers, made during the paraphrase composition process. As depicted in Figure 5, the cross-attention illustrates the relationship between the paraphrase (left column) and the source sentence (right column). Notably, on the left plot, we observe robust attention connections between words such as: ”vodstvo-vlada,” ”implementiralo-sprejela,” ”strategije-ukrepe,” and ”napredka-razvoja.” We should note that these synonymic connections are not trivial, as their establishment necessitates the incorporation of the entire sentence context. This observation highlights the model’s ability to grasp the contextual nuances and capture the intricate dependencies required to effectively recognise and generate appropriate paraphrases. On the second plot, we observe an even more contextually driven connection formed by the model, i.e when encoding the word ”okolje,” the model attends to words like ”zmanjšanje ogljičnega odtisa.”

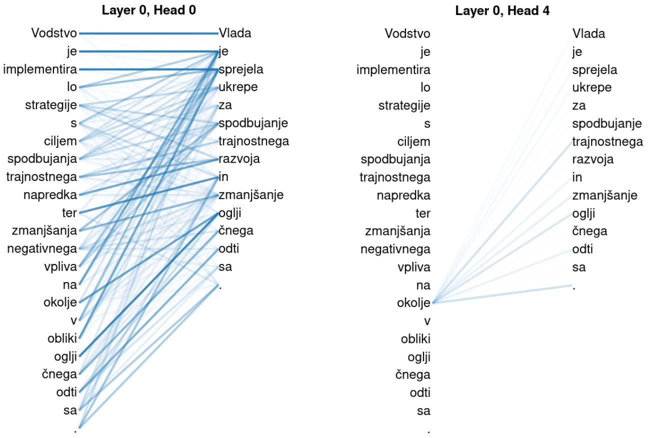


Figure 5. Encoder-Decoder Attention paraphrased sentence (left column), original sentence (right column)

Correlation Between ParaScore and Human Judgement

Given completed evaluation using automated computation model Parascore, we’ve already covered the obvious pitfalls when occurring paraphrased sentences which are just duplicates. The algorithm converges to a value of 0.95, and rarely exceeds that value. This brings up an inherent problem, where human evaluation punishes diversity by giving a low score, as opposed to parascore which provides the best scores exactly on duplicate sentences. We could be weighing individual metrics by some agreement, but the point of paraphrasing is to provide a different form of a sentence.

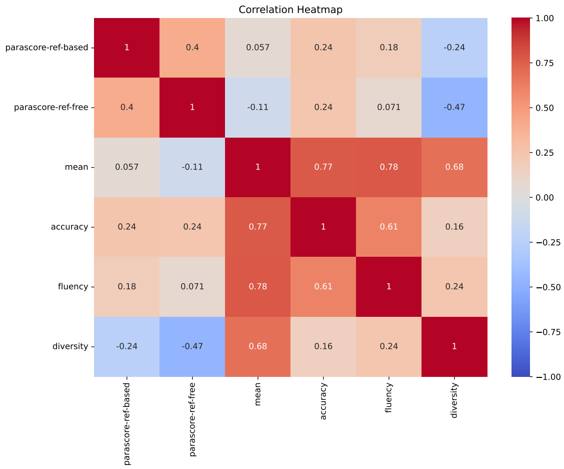


Figure 6. Correlation heat map among human evaluation using three metrics (accuracy, fluency and diversity) and both parascores - reference and non reference based

To further understand the correlation of human evaluation and parasocre on model predictions, we’ve allocated one annotator to evaluate additional 40 paraphrases in the same manner as in data set annotation. A table with all the human evaluated scores and both parascores is compared to find some correlation, to check for validity of automated evaluation of paraphrase quality.

As mentioned throughout parascore analysis, there is a greatly negative correlation between the parascores and diversity. This is caused by parascore giving relatively maximal scores on duplicated and the minimal score from a human annotator. What is also expected is that although the reference based and non reference based parascores are different, they are next best correlated amongst the human evaluation. To further break down the overall mean scores for human evaluation, it would seem that mean score is completely random when compared to any of the parascores. This is the exact reason we checked for individual metric, which show that mean correlation is close to zero because we combine the higher accuracy correlation and negative diversity correlation.

Going back to the Parascore introductory article, it is not providing one of the main conclusions, which is to punish low diversity accordingly. It does though focus on semantics, which is captured with accuracy, and lexical diversity. We’ve proven that accuracy is scored in a correlated fashion, but the diversity is scored inversely. Although the human evaluation pool is rather small, it is enough to see trends in validity of automated paraphrase scoring.

Discussion

The **Back Translation** method contributed to the majority of our dataset. We believe that utilising the same sentence twice, once to generate a paraphrase with $sl \rightarrow en \rightarrow sl$ translation and the another paraphrase using $sl \rightarrow en \rightarrow de \rightarrow en \rightarrow sl$ was a great way utilising the original set of sentences. Additionally this approach may have prevented the model to learn copy-pasting, as we provided it with multiple possible paraphrases of the original sentence. The **Paraphrase-Mining** approach using movie translations was intriguing and novel in our understanding. It yielded a small collection of human-like paraphrases, however, the manual filtering required for matching paraphrases proved to be time-consuming, limiting the size of the generated dataset. We have shown that the proposed **data-augmentation techniques** are effective at increasing the size, and diversity of the dataset, and thus yield better performing models. We found our **experiment on encoder-decoder attention** to be intriguing as it provided some insights into the connections made by the model during paraphrase generation. However, we acknowledge that our small-scale experiment cannot fully elucidate the inner workings of the model.

To be fair, we cannot provide a guarantee that our solution surpasses the performance of the back-translation method. We also believe that the primary purpose of paraphrases generated through automatic techniques should be to expand the dataset size, rather than serving as the fundamental basis of the system. To achieve a system that surpasses the performance of the backtranslation technique, we advocate for the manual correction and validation of paraphrases generated automatically. Furthermore, the utilisation of a substantial collection of high-quality human-written paraphrases would significantly enhance the system's strength and effectiveness. In hindsight, if we had the opportunity to start from scratch, we would prioritise the creation of a vast paraphrase dataset using models such as GPT-3.5. By leveraging the language comprehension capabilities of these advanced (and expansive) models, we could engineer a more robust and cost-effective solution.

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.

[7] Matej Ulčar and Marko Robnik-Šikonja. Sequence to sequence pretraining for a less-resourced slovenian language. *arXiv preprint arXiv:2207.13988*, 2022.

[8] M. Ulčar and M. Robnik-Šikonja. FinEst BERT and CroSloEngual BERT: less is more in multilingual models. In P Sojka, I Kopeček, K Pala, and A Horák, editors, *Text, Speech, and Dialogue TSD 2020*, volume 12284 of *Lecture Notes in Computer Science*. Springer, 2020.

[9] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019.