Paraphrase Generation for Turkish Language

Metin Binbir, Sercan Aksoy
Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi, 34220 Istanbul, Türkiye
{l1119613, l1119620}@std.yildiz.edu.tr

***Özetçe* —Doğal dil işleme alanında aynı anlamla cümle üretimi önemli bir görevdir. Türkçe dili için veri setleri ve çalışmaları az sayıdadır. Farklı yöntemlerle elde edilmiş 3 veri setine sahiptik ve bu proje kapsamında veri setlerinden birine katkı sağlayarak diğer veri setleriyle başarılarını karşılaştırdık. Türkçe üzerinde uygulamak için İngilizce olarak BoW(Bag of Words), klasik Seq2Seq(Sequence to Sequence) ve T5(Text-to-Text Transfer Transformer) mimarilerini denedik. Sonuçları neticesinde T5 mimarisinin Türkçe muadili olan MT5(Multilingual T5) kullanmaya karar verdik. El ile hazırlanmış ve bizim katkı verdiğimiz veri seti ile eğitilmiş modelin diğerlerinden daha başarılı çıktılar ürettiği proje sonucunda insanlar tarafından değerlendirilmiştir. BLEU skoruna göre ekleme yapmadığımız ve el ile hazırlanmış veri seti ile eğitilmiş model en yüksek notu almıştır.**

Anahtar Kelimeler*—Doğal Dil İşleme, Derin Öğrenme, Aynı Anlamlı Cümle Üretimi, Makine Çevirisi, BoW, Seq2Seq, T5, MT5, Transformer, BLEU, C4.

***Abstract*—Paraphrase generation is a significant task in NLP for generation new sentences which are semantically same with given sentences. For the Turkish, datasets and studies in this task are very rare. We had three datasets which are generated with different methods. In this project scope, we expanded one of the data set and compared the success of it with the others. We have tried BoW, basic Seq2Seq and T5 models in English language to decide which architecture we are going to use for implementation in Turkish. As a result of the evaluations, we have decided to study with T5 model. For Turkish, MT5 model is equivalent of T5 model. Consequently, the trained model with the manually created dataset with our insertion was the most successful model according to human evaluation. According to BLEU score, manually generated dataset without our augmentation had the highest value.**

Keywords*—Natural Language Processing, Deep Learning, Paraphrase Generation, Machine Translation, BoW, Seq2Seq, T5, MT5, Transformer, BLEU, C4.

I. Introduction

Generating paraphrased sentences is a task which is generally used for preparing phase for the other tasks. To prepare other tasks, this project may create or expand larger data sets, summarize texts, give different options to choose different translations of a given sentence.

Datasets play a crucial role in generating paraphrased phrases. Unfortunately, there are few datasets which are generated by native speakers. To get humanly successful outputs, neural network should get inputs which are generated by native speakers. Otherwise, it is normal to see that generated paraphrases are syntactically similar to input sentence or completely different meaning than input sentence.

In this study, we used MT5 model which is multilingual version of T5 and based on transformer architecture [1]. MT5 model is created by Google for flexible and many tasks in NLP. We trained the model with five different datasets. The datasets are generated from different studies. Result of this, size of datasets are different and their corpus have various words.

Our contributions:

- We expanded the dataset which are generated by native speaker.
- We compared different datasets to understand requirements of datasets for this task.
- We produced a model which generates Turkish paraphrased sentences for the user inputs.

II. Related Work

For this task, neural network should be trained with pairs as one of the sentences is the similar meaning but different syntax of the other sentence. We have studied on several models to figure out which architecture has the best results for the Turkish in this task. For the early steps in this specific area, Seq2Seq was leading architecture. Because neural network has to know complete meaning of the input sentence so that it could generate a new one begin to end.

We have tried basic Seq2Seq, BoW [2] implementation and T5 models using same dataset and compared their results in English. The T5 model was producing new phrases that we liked a lot better. As a result, we chose T5 model to study with. T5 language model is developed by Google in 2020. They published the model with a corpus which they used for training known as C4(Colossal Clean Crawled Corpus) [3]. T5 language model needs sentence pairs for training. One of the sentences is the sample of given input and the other one is representative of the output for the given purpose. After a while, we discovered that some of the datasets were created by machine translation and we searched for it also.

After choosing architecture, there was also one more fork in the road which is about how to compose a dataset. There are several ways to generate sentence pairs with the same meaning. As we mentioned in the previous paragraph, datasets can be created with machine translation, created manually as we did, or created with filtering the raw data using paraphrase detection models.

T5 language model is based on transformer architecture. Therefore, it process input words as tokenized vectors.

For the Turkish, there are several tokenizers. For instance, BERTurk [4], ITU NLP Pipeline and MT5 pre-trained tokenizer as we have used.

In this article, mostly we are going to mention about which dataset we used and how it effected success of the model. There were three different datasets we used. Two of them created by machine translation but from different datasets(QQP and TaPaCo) and the other one is created manually.

III. Paraphrase Dataset

Paraphrase generation task needs pair of sentences respecting to the rules we set in Evaluation chapter. We had three different datasets at the beginning. One of them was man-provided. The other two of them were created with machine translation. Datasets explanations and sizes given in Table 1.

Table 1 Datasets Information

Dataset ID	Explanation	Number of sentence pairs
DH-S	Manually prepared paraphrased questions from forum topics (no augmentation by us)	30994
DH	Manually prepared paraphrases from forum topics (augmented by us)	50994
TQP	Multilingual Sentential Paraphrase Corpus	51620
MT	Quora Question Pairs	40200
BIG	Combination of DH, MT, and TQP	142814

A. Man-provided Dataset

The dataset was filled by native speakers. It is much more valuable than other datasets which are generally used for this task. Main reason of it is that the native speaker can manipulate pairs easily respected to applied language.

1) DH-S: Questions in this dataset are from a popular forum website in Turkey (forum.donanimhaber.com). Topics which include questions are acquired. Forum generally is used for daily basis activities. Thus, it is not applicable to think that acquired sentences may agglomerated on a specific subject. For the pairs, native speakers of Turkish read questions that acquired from web site and created their pairs manually. There are 30994 question pairs in this dataset.

2) DH: To contribute and develop language datasets in Turkish, which are already had insufficient data, we have augmented 20000 more pairs into DH-S dataset manually and reached it to 50994 question pairs.

B. Machine Translation

Machine translation [5] is a very common technique to obtain paraphrased pairs to get a paraphrased sentence based on a input. Firstly, input sentence translated two times to different languages. After that, obtained translated sentence had one more translation process to its belonged language. This technique also known as backtranslation.

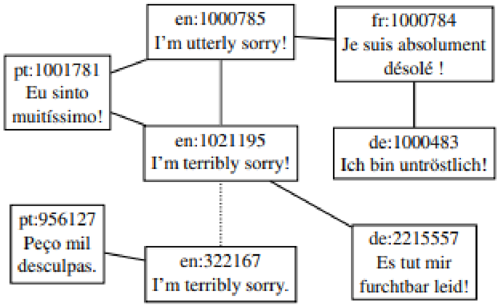

Figure 1 Paraphrased Pairs using Machine Translation

1) MT: MT dataset provided by translation of Quora Question Pairs to the Turkish language. There are 40200 question pairs in this dataset.

2) TQP: TQP dataset provided by TaPaCo which is freely available corpus for 73 languages. It is generated with machine translation also and it includes 51620 question pairs [6] [7]. For this dataset, there is a disadvantage we must highlight, in the dataset proper nouns are used in English words. It adversely affects outputs.

C. BIG

BIG dataset has provided by us. We merged and mixed DH, MT and TQP datasets explained above. There are 142814 question pairs in this dataset.

IV. Model Architecture

MT5 used as model architecture in this study. MT5 is multilingual version of T5 which is announced by Google as transformer based encoder-decoder model.

T5 model is formed by two stacks which are encoder and decoder [3]. Encoder part of the model is representative of the input sequence and decoder part is used for generation of output sequence [8].

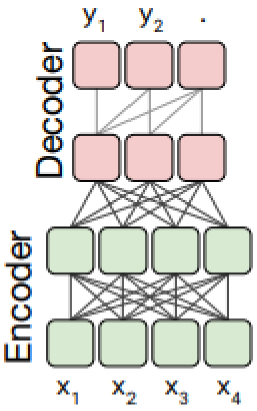

Figure 2 Encoder-Decoder Model of Casual Fully-Visible Masking

MT5's pre-trained model includes pre-defined jobs. With the strength of paired masked attention. Input sentences given with a job label appended at the beginning of it. Therefore encoder-decoder mechanism sense the job with the label and settles itself to decode in order to obey logic of label.

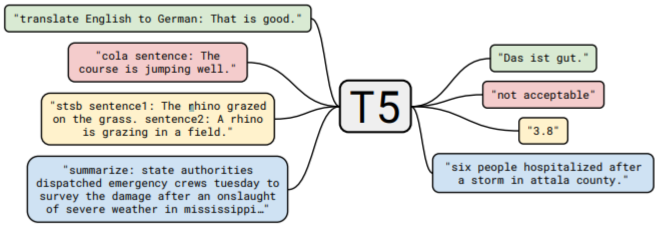
"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

Figure 3 Pre-defined Tasks of T5 Model

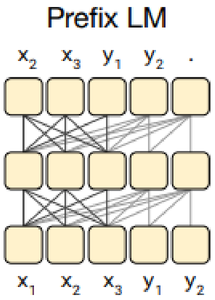

Figure 4 Encoder-Decoder Model with a Prefix Job Defined

We fine-tuned the MT5 base pre-trained language model as adding prefixes into input sentences. We used "cumleniz:" prefix which means "your sentence:" in Turkish. In Figure 4, we showed the mechanism of our encoder decoder network. **Dataset format shown below:**

<s>input sequence === target sequence </s>

Input format shown below:

cumleniz: input sequence </s>

We used MT5-base pretrained tokenizer for Turkish. Actually, this tokenizer was trained with a variety of languages, therefore it is not necessary to indicate the language it was trained with when using it.

For DH dataset, whose size is 50994 question pairs, one epoch took approximately 1 hour to train. In our project, we trained the model in 3 epochs using Google Colab.

V. Experiment

Firstly, we researched models to use for this task. Paraphrase generation is very compatible with the Seq2Seq models because of its goal to generate new sentence respected to meaning of input sentence. We run Seq2Seq basic model and its BoW implementation at the beginning. We tried them for English language. The results were not enough to start study with them. There were not many projects outside the English language. It was making to test it in Turkish difficult. Also, some of them were already trained models. We tested the models using Huggingface's Transformers library but it was not suitable for our purpose.

After a while, we explored T5 model. The results were better than other models we had tried. We could not find any Turkish Paraphrase Generation study with MT5 model, but there were several studies in different languages.

First thing we did after choosing an architecture is to try to train a language model. We would like to see all processes in this study. Therefore, instead of using pre-trained T5 model, we trained an English language model.

T5 English model was created by us. We needed to put it through its steps with tasks to see if it worked. On a sentiment analysis task, we fine-tuned our language model. F1 score was enough to understand that language model was working properly.

While we were studying on MT5 models in different languages, one model added to Huggingface's Transformers library [9] about Paraphrase Generation in Turkish.

We tested that model also, it was working properly. We scanned the dataset for this model and tried to test it with various sentences. Model tried to fill its missing knowledge with the high probability to occur words in Turkish but it was quite obvious.

There was a study in Russian language we were following [10]. Their dataset was huge, so we manipulated the parameters more suitable for our dataset and for the Turkish. Apart from them, we tried to decrease the epoch number to 2. However, it was not quite good. Therefore, we increased the epoch number to 3. It satisfied our expectations. We tried to increase batch size and maximum sequence length, but due to GPU limits in the Google Colab, we could not succeed. The parameters of training shown in Table 2.

Table 2 Parameters

Parameters	Values
Number of Epoch	3
Batch Size	4
Learning Rate	3e-4
Maximum Sequence Length	150

To see evaluation of the model, we added datasets library [11]. We used its BLEU score function which calculating the success of output sentence with n-grams between input sentence. After noticing that it is not suitable for our purpose, we managed to not consider BLEU scores as evaluation of this task.

VI. Evaluation

In this project, we have scored the outputs by two methods and suggested a method for the evaluation. General

NLP tasks needs large amount of data to train. This is because human languages may contain thousands of words. Especially, agglutination languages like Turkish are able to generate limitless new words using suffixes theoretically. Therefore, to generate successful outputs, it is necessary to add into dataset various of sentences which may occur in reality.

For completion of the project, it is essential to be tested with large amount of data. It provides much more healthy results for the reality. Same reason with the training part, the model should evaluate itself for the various inputs.

In paraphrase generation study while evaluating results, this suggestions must be observed:

- Output sentence must conform to the grammatical structure.
- Output sentence must has a meaning in a way.(It can not be random words in a line.)
- Output sentence must has the similar meaning with the input sentence.
- Output sentence should be syntactically as different as possible from the input sentence.

Because of the large amount of test samples, it is difficult to evaluate project manually. Therefore, we have used BLEU score. BLEU score method compares the output sentence with input sentence by finding n-grams [12]. For obeying rules as we mentioned in the previous paragraph, BLEU score should not be high. On the other hand, there is a limit which is able to be changed the input while generating same meaning of the sentence. We believe that with both BLEU scores for each model and human evaluation, we can determine a threshold value for the optimum BLEU score. BLEU scores shown in Table 3.

Table 3 Metric Scores of MT5 model on Different Datasets

Dataset	BLEU-1	BLEU-2	BLEU-3
DH-S	0.926	0.787	0.586
DH	0.768	0.606	0.420
TQP	0.415	0.255	0.147
MT	0.471	0.312	0.197
BIG	0.752	0.585	0.420

For the human evaluation of the project, five people reviewed 20 different input sentences with 5 different outputs for each model on 5 models whose datasets are different. Selected people are the native speakers of the language we are studying with. Outputs were sent them in an Excel format with the rules. We asked from them to select the best and worst generated paraphrased sentences per line which includes 5 sentences belong different models.

Table 4 Human Evaluation of MT5 MODEL ON DIFFERENT DATASETS

Dataset ID	Like	Dislike
DH-S	%19.8	%14.2
DH	%26.6	%14.6
TQP	%16	%32.2
MT	%13.4	%27.2
BIG	%24.2	%11.8

Logical statement of the success, if a model gets more upvotes, it is usual to expect downvotes are rare. However, after generating hundreds of results, we figured out that models may generate bad sentences after generating a good one. Thus, we chose upvotes for the main score system in our study, but we showed the results of both.

Many paraphrase generation tasks are implementation of machine translation. Therefore, it is really common to see high BLEU scored pairs. After studying on paraphrased projects, we saw that there are lots of studies have good scores on paraphrase detection. Even though paraphrase detection is another NLP task, evaluating the success of paraphrase generation by using high success rated paraphrase detection might be a good way to score model.

VII. Analysis

In this chapter, we measured the success of models. Models trained by different datasets and we compared them manually.

We have trained five models with different datasets. Their explanations and their sizes given in Table 1.

Table 4 shows that native speakers of Turkish have found the model that is trained with DH dataset the most successful and marked the model that is trained with MT dataset the most unsuccesful.

In Table 5, we showed that the generated paraphrases of the model trained with DH dataset are generally discrete from input sentence, because its train dataset prepared for it. DH dataset is subset of BIG dataset, thus there were similarities in some generated sentences.

Since both MT and TQP datasets are based on machine translation method, the traces can be seen in Table 5. With high probability, either input and output sentences have lots of common words, or meaning of the input and output sentences have quite different meanings from each other.

Table 5 Paraphrased Sentences Using Five Different MT5 Models Trained by Different Datasets

Source	**yaşadığın şehirde en güzel manzaraya sahip yer neresidir? (Where is the place that has the most beatiful view in the city you have lived in?)**
DH-S	Şehirdeki en güzel manzaraya sahip mekan neresi? (Where is the place that has the most beatiful view?)
DH	hayatını sürdürdüğüm şehirde en güzel manzara alan bölge nedir? (What is the area with the most beautiful view in the city where I live?)
TQP	Senin yaşadığın şehirde en iyi manzaraya sahip yer nedir? (What is the best view in the city you live in?)
MT	Yaşadığınız şehirde en güzel manzara alabileceğin yer neresi? (Where is the best place to get the most beautiful view in your city?)
BIG	yaşadığın şehirde en güzel manzaraların neler olduğunu biliyor musun? (Do you know what the most beatiful views in the city you live in are?)
Source	**hangi saatlerde kitap okumak daha verimlidir? (At which times is it more productive to read a book?)**
DH-S	hangi saatlerde kitap okunabilir? (At which times can book be read?)
DH	hangi saatlerde kitap okumak daha yararlı? (At which times is it more beneficial for reading a book?)
TQP	Ne zaman kitap okumak daha güvenli? (When is the reading book safer?)
MT	Özel kitap okumak için hangi saatler uygun olacak? (Which times is going to be more appropriate for reading a private book?)
BIG	kitap okumak için en uygun saat nedir? (What is the most suitable time to read a book?)
Source	**son model arabalar çevreye daha mı az zarar veriyor? (Are the latest model cars harming less to the environment?)**
DH-S	son model arabaların çevreye zararları daha mı az oluyor? (Are the damages of the latest model cars to enviroment less?)
DH	son arabaların çevreye etkisi az mi? (Do recent cars have little impact on the environment?)
TQP	Son model araba çevreye daha az zarar verir mi? (Is the latest model car less harmful to the environment?)
MT	Son model arabalar çevreye zarar veriyor mu? (Do latest model cars harm the environment?)
BIG	yeni arabalar çevreye daha az zarar veriyor mu? (Do new cars harm less to the enviroment?)
Source	**insanlar boş zamanlarını neler yaparak daha verimli hale getirebilir? (What can people do to make their spare time more effective?)**
DH-S	insanlar boş zamanlarını iyileştirebilecek şeyler? (Things that can make people’s free time better?)
DH	insanlar boş zamanlarını nasıl iyileştirir? (How can people make their spare time better?)
TQP	İnsanlar boş zamanını nasıl geçirir? (How do people spend their free time?)
MT	boş zamanımızı en verimli hale getirmenin en iyi yolları nelerdir? (What are the best ways to make our spare time the most productive?)
BIG	insanlar boş zamanlarını nasıl daha verimli yapabilir? (How can people make their free time more efficient?)
Source	**ülkemizdeki konut fiyatlarında ileriki zamanlarda nasıl bir artış bekleniyor? (What kind of increase is expected in housing prices in our country in the future?)**
DH-S	ülkemizdeki konut fiyatlarında ilerleyen haftalarda ne kadar artış olur? (How much increase will be in housing prices next weeks?)
DH	ülkemizdeki konut fiyatları gelecek tarihte nasıl artacak? (How will housing prices in our country increase in the future?)
TQP	ülkemizdeki konut fiyatlarına kısa bir zaman için biraz artış bekliyor musun? (Do you expect a little increase in housing prices in our country for a short time?)
MT	Fransa’da konut fiyatlarının ileriki zamanlarda nasıl artacağını düşünüyorsunuz? (How do you think housing prices in France will increase in the future?)
BIG	ülkemizdeki konut fiyatları gelecekte nasıl artacaktır? (How will housing prices in our country increase in the future?)
Source	**okul bahçesinde top oynayan çocukları gördünüz mü? (Have you seen the children playing ball in the schoolyard?)**
DH-S	Okul bahçesinde top oynayan çocuklar var mı? (Are there any children who playing ball in the schoolyard?)
DH	okul bahçesinde top oynayan çocuklardan haberdar mısınız? (Are you aware of children playing ball in the schoolyard?)
TQP	okula gidene katlanan çocukları gördün mü? (Have you seen the children who endure the person going to school?)
MT	Okul bahçesinde top oynananlar neler? (What are the games played in the schoolyard?)
BIG	okul bahçesinde top oynayan çocuklardan haberiniz var mı? (Do you know about the children playing ball in the schoolyard?)
Source	**salgın hastalıkların etkisi ne kadar sürecek? (How long will the effects of epidemics take?)**
DH-S	salgın hastalıklarının etkisi kaçtır? (How much is the impact of epidemics?)
DH	salgın hastalıkları kaç yılda etki eder? (In how many years do epidemic diseases affect?)
TQP	sağlık hastalıklarının sebebi ne kadar sürecek? (How long will the cause of health diseases last?)
MT	salgın hastalıkları nasıl sürecek? (How will the epidemics maintain?)
BIG	salgın hastalıklar ne kadar zaman alır? (How long do epidemics take time?)
Source	**sıcak havalarda parkta yürüyen insanlar görüyor musunuz? (Do you see people walking in the park in hot weather?)**
DH-S	parkta sıcak havalarda yürüyen var mı? (Does anyone walk in the park in hot weather?)
DH	sıcak havalarda parkta duran insanlara rastlanır mı? (Do you come across people standing in the park in hot weather?)
TQP	Sıcak havalarda parkta yürüyüşe çıkan insanlar görüyor musun? (Do you see people taking a walk in the parks in hot weather?)
MT	Yüksekliklere yakın parkta yürüyen insanlar görüyor musunuz? (Do you see people walking in the park near the heights?)
BIG	sıcak havalarda parkta yürüyebilenler bulunuyor mu? (Are there people who can walk in the park in hot weather?)
Source	**deniz kenarında uçan kuşlar çok mutlu gözükmüyorlar mı? (Do not the birds flying by the sea look very happy?)**
DH-S	deniz kenarında uçan kuşlar çok mutlu görünmezler mi? (Do not the birds flying by the sea look very happy?)
DH	deniz kenarında uçan kuşlar fazla mutlu değil mi? (Are not the birds flying by the sea so happy?)
TQP	Mavi deniz kenarında uçan kuşlar çok mutlu görünmüyor mu? (Do not the birds flying by the blue sea look so happy?)
MT	Denizlerde uçan kuşlar mutlu mu? (Are the birds flying on the seas happy?)
BIG	deniz kenarında uçan kuşlar gerçekten mutlu değil mi? (Are not the birds flying by the sea really happy?)
Source	**ülke genelinde bisiklet kullanımının artması konusunda ne düşünüyorsun? (What do you think about the increase in the use of bicycles across the country?)**
DH-S	ülke genelinde bisiklet kullanımı artış yönünde düşünceniz nedir? (Cycling is on the rise across the country, what is your thought?)
DH	bisiklet kullanımının yükselmesi hakkında fikirleriniz nedir? (What are your thoughts on the rise of cycling?)
TQP	Dünya genelinde bisiklet kullanımının artırılması konusunda fikrin nedir? (What is your opinion on increasing the use of bicycles around the world?)
MT	Bisiklet kullanımının artmasının altındaki görüşleriniz nelerdir? (What are your views on the increase in bicycle use?)
BIG	ülke genelinde bisiklet kullanımının artması hakkındaki görüşleriniz nedir? (What are your views about increasing use of bicycles countrywide?)

VIII. Conclusion

In this project, we aimed to understand effects of dataset on a model. To see difference between models, we had to use same architecture on various datasets.

T5 model is a new method announced relative to its competed ones. Basic structure of training with two sentences at one time was the main reason why we selected this architecture.

There are many ways to generate paraphrased pair datasets. While examining datasets, their cleanity was obvious for us as native speakers. Success of generated sentences and human evaluation phase were similar with what we expected.

NLP tasks need lots of data to run properly and it is difficult to acquire it by computer. Because, the main reason of the NLP is about teaching to machines the human language.

As we said in Evaluation section, we think that it is possible to use a task for evaluating another project. It is also possible to use separated NLP tasks to preprocess another projects. Therefore, we believe that human labor is inevitable for the starters.

As we explained in Evaluation and Analysis sections, manually generated dataset is indispensable successful choice. Also, it can be seen that our augmentation is contributed to success of model.

References

[1] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, “mT5: A massively multilingual pre-trained text-to-text transformer,” 2020.

[2] Y. Fu, Y. Feng, and J. P. Cunningham, “Paraphrase generation with latent bag of words,” 2020.

[3] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html

[4] Z. Jiang, W. Yu, D. Zhou, Y. Chen, J. Feng, and S. Yan, “Convbert: Improving bert with span-based dynamic convolution,” 2021.

[5] A. Sokolov and D. Filimonov, “Neural machine translation for paraphrase generation,” 2020.

[6] Y. Scherrer, “TaPaCo: A corpus of sentential paraphrases for 73 languages,” in *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 6868–6873. [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.848

[7] M. Y. Sarıgöz, “monatis/tqp: V0.1,” Apr. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.4719801

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017.

[9] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[10] A. Fenogenova, “Russian paraphrasers: Paraphrase with transformers,” in *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*. Kiyv, Ukraine: Association for Computational Linguistics, Apr. 2021, pp. 11–19. [Online]. Available: https://www.aclweb.org/anthology/2021.bsnlp-1.2

[11] Q. Lhoest, P. von Platen, T. Wolf, A. V. del Moral, Y. Jernite, S. Patil, M. Drame, J. Chaumond, J. Plu, J. Davison, S. Brandeis, T. L. Scao, V. Sanh, K. C. Xu, L. Tunstall, N. Patry, A. McMillan-Major, P. Schmid, S. Gugger, L. Debut, C. Delangue, T. Matussière, S. Bekman, and F. Lagunas, “huggingface/datasets: 1.7.0,” May 2021. [Online]. Available: https://doi.org/10.5281/zenodo.4817769

[12] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” 10 2002.