Semantic Similarity Based Filtering for Turkish Paraphrase Dataset Creation

Besher Alkurdi and **Hasan Yunus Sarioglu** and **Mehmet Fatih Amasyali**
Department of Computer Engineering
Yildiz Technical Univeristy
Istanbul, Turkey
{besher.alkurdi, yunus.sarioglu}@std.yildiz.edu.tr
amasyali@yildiz.edu.tr

Abstract

In this study, we introduce a new method for creating paraphrase datasets from parallel bilingual corpora. We also introduce large paraphrase datasets created using this method. We utilize machine translation to create paraphrase datasets by translating the English phrases in Turkish-English parallel datasets to Turkish. Detailed pre-processing steps are applied to the text pairs. A sample from our translated datasets was annotated by native speakers for semantic similarity, and a model with the same task was chosen based on the correlation with human annotations. We then filtered the pre-processed and translated text pairs by semantic similarity calculated by the chosen model. Two pre-trained encoder-decoder architectures were fine-tuned on the datasets that we created. We present results asserting our data collection and filtering method's effectiveness.

1 Introduction

Paraphrase generation can be applied in several fields including data augmentation (Kumar et al., 2019), machine translation evaluation (Thompson and Post, 2020), chatbots (Garg et al., 2021), question answering (Zhu et al., 2017), and semantic parsing (Cao et al., 2020). A major challenge in paraphrase generation research is the lack of large paraphrase datasets, especially in languages other than English. This served as a motivation for us to create high-quality and large paraphrase datasets in Turkish. We use English-Turkish datasets and translate the sentences from English to Turkish. Semantic similarity is then calculated using a Transformer-based (Vaswani et al., 2017) model for each pair in the resulting Turkish-Turkish datasets. Pairs that have a score greater than a threshold are accepted as paraphrases. The threshold is chosen in accordance to human annotations collected by us.

Our main contributions are as follows:

- We present the largest Turkish paraphrase datasets yet consisting of approximately 800k pairs in total.
- We introduce a new method for creating a paraphrase dataset from a parallel corpus combining machine translation and semantic similarity based filtering.
- We share paraphrase generation models trained on the datasets we introduce as part of our work and evaluate them using several benchmark metrics.
- We share a manually annotated semantic textual similarity dataset consisting of 500 pairs.

The datasets and the fine-tuned models are shared publicly.[1] We hope that our work encourages more research in this area, and provides a dataset that can be used for benchmarking paraphrase generation architectures and datasets in the future.

2 Related Work

The task of finding texts with similar or identical meaning, often called paraphrase identification is a challenging task. Several approaches have been tried to create paraphrase datasets in previous work.

Manual paraphrase collection is very expensive, unscalable and implausible with limited resources. Studies in this area have usually made use of crowd sourcing to construct a paraphrase dataset (Burrows et al., 2013). The main advantage of this method is its effectiveness in constructing a high-quality dataset where diversity of the sentences can be increased without the fear of producing pairs with low semantic similarity.

Semantic similarity based mining can be employed to detect paraphrases in a corpus of texts. Each sentence is compared with every other sentence in the corpus and given a similarity score, the

[1] https://github.com/mrbesher/semantic-filtering-for-paraphrasing

sentence with the highest score is considered a paraphrase. This method suffers from quadratic runtime and thus fails to scale to large paraphrase datasets. A similar approach was employed in (Martin et al., 2020).

Machine translation can be used where a text is translated to a pivot language then to the source language again (Wieting and Gimpel, 2018), (Wieting and Gimpel, 2018), (Suzuki et al., 2017). Multiple pivot languages can be used in a similar manner. While this method is successful, it may suffer from noise caused by automatic translation from the source to the pivot language and back from it.

Other automatic approaches were used like using parallel movie subtitles (Aulamo et al., 2020), image captions of the same image (Lin et al., 2014), and texts that can be marked as paraphrases based on different conditions such as duplicate questions,[2] duplicate posts (Lan et al., 2017), and text rewritings (Max and Wisniewski, 2010).

A handful of research on Turkish paraphrase dataset creation have been shared. (Karaoğlan et al., 2016) conduct a study resulting in 2,472 text pairs annotated by humans. (Demir et al., 2013) present a paraphrase dataset consisting of 1,270 paraphrase pairs from different sources. The mentioned datasets are not shared publicly. (Bağcı and Amasyali, 2021) present a combination of translated and manually generated datasets focusing on question pairs, and train a BERT2BERT architecture on it. None of the existing studies provide a comprehensive dataset in Turkish to the best of our knowledge.

3 Dataset Creation

The dataset creation process pipeline consists of several steps to ensure that the dataset is of high quality. Firstly, English-Turkish parallel texts with only one source and one target were downloaded using Opus Tools (Aulamo et al., 2020).

We considered using the datasets shared on OPUS (Tiedemann, 2012).[3] The following datasets where downloaded, examined, filtered and machine translated:

- **OpenSubtitles2018:** A large database of movie and TV subtitles across 60 languages[4] compiled, pre-processed and aligned by (Lison and Tiedemann, 2016).

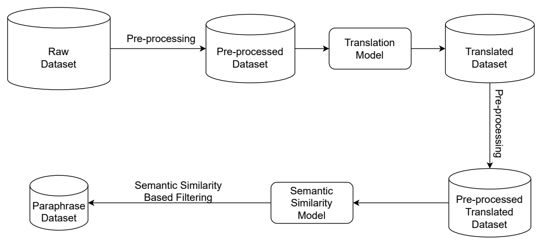

Figure 1: Dataset Creation Pipeline

- **TED2013:** A parallel dataset of TED talk subtitles originally provided by Web Inventory of Transcribed and Translated Talks.[5] The talks were translated automatically, leading to significant noise.

- **Tatoeba v2022-03-03:** A collaborative collection of sentences and translations, compiled using crowdsourcing.[6]

Text pairs were pre-processed according to the characteristics of each dataset to remove unwanted text pairs. An example is removing the explanations done in the TED dataset indicated by two hyphens before and after the explanation while keeping text in square brackets as they made the statements more understandable.

Machine translation is applied on the whole dataset from English to Turkish. At this stage the dataset contains valid text pairs that are candidates to be paraphrases. Source and translated sentences were removed if one includes the other and pre-processing steps were applied again to remove noisy texts generated by the translation model. After that, semantic similarity between text pairs is measured and pairs with a high semantic similarity score are chosen as paraphrases. The steps, illustrated in Figure 1, ensure a robust process to create a high-quality paraphrase dataset.

4 Translation and Semantic Similarity Based Filtering

4.1 Translation

Due to the huge volume of data that we aimed to translate, usage of online machine translation services was unfeasible due to restrictions set by the providers. We chose a machine translation model provided by OPUS-MT project (Tiedemann and

[2]https://www.kaggle.com/c/quora-question-pairs
[3]https://opus.nlpl.eu/
[4]http://www.opensubtitles.org/
[5]https://wit3.fbk.eu/
[6]https://tatoeba.org

Thottingal, 2020) and shared publicly on Hugging Face.[7]

4.2 Human Annotations for Ground Truth Semantic Similarity

To filter the pairs further, we considered using a semantic similarity metric to remove pairs with low semantic similarity. We had several models to achieve the task of semantic similarity scoring to choose from. For model selection, we sampled 250, 150, and 100 pairs from OpenSubtitles2018, Tatoeba, and TED2013 respectively. The samples were then annotated by 6 native Turkish speakers, with each pair assigned to two different annotators. Following (Creutz, 2018), each pair could be assigned one of the labels described in Table 1. If the annotators disagreed and the score difference was less than two, the label indicating less semantic similarity was chosen. Otherwise, the label was discarded.

A bot was created on Telegram[8] to ease the process of annotation collection and the scores collected from annotators were used later to determine a threshold to filter out low quality paraphrases.

Annotators disagreed by two points on 16 samples from OpenSubtitles2018 (OST), 5 samples from TED2013 (TED), and 3 samples Tatoeba (TAT). Therefore, a total of 24 samples were removed. The distribution of the labels in each dataset is shown in Table 2.

The desired phrase pairs are the ones labeled as near-synonyms or synonyms. 66.32%, 70.94% and 88.44% of the pairs in TED, OST and TAT respectively can be considered paraphrases accordingly. The results show a need for further filtering as phrases with different meanings are expected to affect the model's performance.

4.3 Semantic Similarity Based Filtering

Several semantic similarity models were considered to filter the text pairs. The goal is to capture the closeness in meaning between two input texts. The models we considered utilize BERT as a baseline (Devlin et al., 2019). Among those are Bi-encoders, two identical encoders that compute embeddings of sentences separately. Cosine similarity is then calculated between the embedding pair. We considered three pretrained models of this kind:

- **distiluse-base-multilingual-cased:** Presented in (Reimers and Gurevych, 2020), this model creates multilingual sentence embeddings. The training objective is to map translated sentences' embeddings to the embeddings of the original sentences.

- **multilingual-l12:** This model maps texts to a 384 dimensional dense vector space, the model is shared on Huggingface.[9]

- **emrecan:** This model was fine-tuned on a machine translated version of STS-b[10] and NLI (Budur et al., 2020) to map texts to a 768 dimensional dense vector space. Contrary to the models mentioned before, this model is trained on Turkish datasets.

Cross-encoder networks (Reimers and Gurevych, 2019) accept sentence pairs as inputs, and output the semantic similarity between sentences. Following (Beken Fikri et al., 2021), and using their STS-b (Cer et al., 2017) dataset, which was translated by the authors using Google Cloud Translation API.[11] We fine-tuned BERTurk[12] starting from 5 random seeds for 4 epochs and used the model with the highest correlation score with the similarity labels on the development set split provided by the authors.

We chose the semantic similarity model that filtered out the least amount of pairs labeled as synonyms or near-synonyms. The goal is to remove pairs labeled as having distant meanings or no relevance. We chose thresholds for each model such that after filtering out pairs below the thresholds in the sample annotated by humans, 95% of the kept pairs are labeled as synonyms or near-synonyms. The percentage of the valid pairs kept can be seen in Table 3 for every model. Emrecan was chosen for filtering due to its superiority to the other models.

Table 4, shows the number of text pairs before any filtering was applied in the raw column and the number of kept pairs after pre-processing, prior to translation. The number of pairs kept after semantic similarity based filtering is shown in the last column.

[7]https://huggingface.co/Helsinki-NLP/opus-tatoeba-en-tr
[8]https://telegram.org/
[9]https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2
[10]https://huggingface.co/datasets/emrecan/stsb-mt-turkish
[11]https://github.com/verimsu/stsb-tr
[12]https://huggingface.co/dbmdz/bert-base-turkish-cased

Category	Description	Example
Eş Anlamlı *Synonyms*	İki cümle birbirlerinin yerine kullanılabilir ve temelde aynı anlama gelmektedir. *The two sentences can be used interchangeably and essentially mean the same thing*	Ona yaklaşmayın, hasta olabilir. Ondan uzak durun! Hasta olma ihtimali var. *Do not get close to him. He might be sick.* *Stay away from him! There is a chance that he is sick.*
Yakın Anlamlı *Near-synonyms*	Cümlelerin tarzları farklı olsa da iki cümlenin aynı anlama geldiğini düşünmek mümkün. *Even though the style of the sentences is different they can be thought to have the same meaning.*	O, saçını yapma tarzını değiştirdi. Saçının şeklini değiştirmiş. *She changed the way she does her hair.* *She changed the shape of her hair.*
Uzak Anlamlı *Distant Meanings*	İki cümlenin neden yan yana geldiği anlaşılabilir ancak aynı anlama geldikleri söylenemez. *It can be explained why the sentences are coupled together but one cannot consider them to have the same meaning*	Farklı roller için de seçmelere katılmıştım Birkaç rol için bekledim. *I attended the auditions for different roles.* *I waited for some roles.*
Alakaları Yok *No Relevance*	Cümleler arasında bir bağlantı yok. Farklı anlamlara sahipler. *The sentences have no connection. They have different meanings.*	Afedersin bana benim iki elim yeter. Üzgünüm, sadece ikisini alabilirim. *Execuse me, my two hands are enough for me.* *Sorry, I can only take two of them.*

Table 1: Semantic Similarity Labeling Task Description for Human Annotators

Label	OST	TAT	TED
No Relevance	25	2	6
Distant Meanings	43	15	26
Near-synonyms	92	40	37
Synonyms	74	90	26

Table 2: The distribution of human annotations across the datasets

Model	OST	TAT	TED
BERTurk	33.73	**33.85**	42.86
Distiluse	40.36	8.46	34.92
Multilingual-l12	36.75	9.23	36.50
Emrecan	**42.68**	26.92	**46.03**

Table 3: Percentage of the Kept Valid Pairs

5 Experiments

We ran experiments to measure the quality of our constructed datasets. These are intended to be used as a baseline for future research on Turkish paraphrase generation. We train our models on the unfiltered and the filtered versions of our datasets to analyze the applied filtering method's impact on the quality of our datasets.

For our experiments, we randomly select 5% of the pairs in each dataset as development split and 5% as test split. The rest of the pairs are used for training the models. In this section we present the experimental results of the models we fine-tuned on the train splits and tested on the test splits of our datasets. We employ transfer learning using pre-trained Text-to-Text Transformer models. mT5 is a multilingual variant of T5 presented in (Xue et al., 2021). We use a pre-trained checkpoint of mT5-base provided by Google and published on Hugging Face.[13] We also utilized BART (Lewis et al., 2020) using trBART, a checkpoint of BART-base (uncased) pre-trained from scratch by (Safaya et al., 2022). The authors published the model on Hugging Face.[14]

In our initial experiments, models fine-tuned

[13]https://huggingface.co/google/mt5-base
[14]https://huggingface.co/mukayese/bart-base-turkish-sum

Name	Raw	Pre-processing	Similarity Based Filtering
OST	13,190,557	1,944,955	706,468
TAT	393,876	265,203	50,423
TED	131,874	104,238	39,763

Table 4: Number of Text Pairs in the Datasets Before and After Filtering

on the TED dataset failed to generate acceptable parahrases. We did not continue experimenting with the dataset, and thus only provide the translations and the filtered dataset without experiment results.

Our models were trained for 4 epochs with a learning rate of $1e-4$ on the OST dataset, and for 6 epochs with a learning rate of $1e-4$ on the TAT dataset. Those values yielded the highest BLEU scores of the models on the development splits after several experiments with different learning rates. Five candidate texts were generated for each source text. The candidate with the highest probability that does not consist of the same letters as the source was chosen for evaluation.

We report the following metrics: BERTScore (Zhang et al., 2019),[15] BLEU (Papineni et al., 2002),[16] ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and TER (Snover et al., 2006). The scores reported in Table 5 are the mean of 4 results from 4 training runs using the settings we described earlier.

Note that the mT5-base trained on the OST dataset outperformed the other models in both datasets. This, in our opinion, suggests generalizability and high dataset quality. To further assess the impact of our filtering method, we fine-tuned mT5-base on the unfiltered datasets and observed that despite the difference in size, the models fine-tuned on the unfiltered datasets yielded worse performance on the OST dataset and less semantically similar pairs on the TAT dataset. We believe that this is due to the fact that TAT is more carefully constructed using crowdsourcing, and thus the effect of semantic similarity based filtering is less visible. We report the score of mT5-base trained for 3 and 4 epochs on the unfiltered OpenSubtitles2018 (OST-RAW) and Tatoeba (TAT-RAW) respectively. The scores of the model on the test splits started to decrease after those epochs.

We present some generated paraphrase examples in Appendix A, to highlight the success and the failure cases of the fine-tuned models.

[15] https://github.com/Tiiiger/bert_score

[16] https://huggingface.co/spaces/evaluate-metric/bleu

6 Conclusion

We detailed an approach for creating paraphrase datasets from parallel text coprora using machine translation and semantic similarity based filtering. For filtering, we chose a semantic similarity model that kept the most paraphrases in the datasets based on similarity labels we collected from human annotators for a sample of our datasets. We present the paraphrase datasets we created with benchmark results of Text-to-Text Transformer models trained on our datasets across a variety of metrics.

7 Future Work

Our approach results in a high-quality paraphrase dataset, but has a downside of filtering out valid pairs with low lexical similarity depending on the semantic similarity metric used. We plan on combining lexical and semantic similarity into a new filtering metric to obtain a dataset that has more diverse pairs. We will compare the effectiveness of models trained on the current datasets and the diverse dataset in data augmentation for different tasks. Furthermore, we also plan to test the effect of curriculum learning (Bengio et al., 2009) on the newly created diverse datasets, and similar to (Li et al., 2018) we will evaluate the output of the models with the help of human annotators on multiple aspects like clarity, fluency, and semantic similarity.

8 Acknowledgment

This study was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) Grant No: 120E100.

References

Mikko Aulamo, Umut Sulubacak, Sami Virpioja, and Jörg Tiedemann. 2020. OpusTools and parallel corpus diagnostics. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3782–3789. European Language Resources Association.

OST Test Dataset

Model	Train Dataset	BERTScore Cased	BERTScore Uncased	BLEU	ROUGE-L	METEOR	TER
mT5-base	OST	**89 ± 0.01**	**92.05 ± 0.01**	**46.26 ± 0.09**	**74.8 ± 0.02**	**72.97 ± 0.13**	**36.4 ± 0.04**
trBART	OST	77.8 ± 0.17	87.92 ± 0.13	33.59 ± 0.32	64.65 ± 0.33	62.62 ± 0.45	50.96 ± 0.4
mT5-base	TAT	84.95 ± 0.38	89.23 ± 0.24	29.37 ± 0.83	66.64 ± 0.46	63.14 ± 0.86	49.29 ± 1.13
trBART	TAT	74.21 ± 0.32	85.25 ± 0.28	23.45 ± 0.29	59.32 ± 0.6	54.93 ± 0.5	57.22 ± 0.59

TAT Test Dataset

Model	Train Dataset	BERTScore Cased	BERTScore Uncased	BLEU	ROUGE-L	METEOR	TER
mT5-base	TAT	94.07 ± 0.36	95.75 ± 0.25	61.66 ± 1.34	84.67 ± 0.62	**82.72 ± 0.42**	22.43 ± 1.27
trBART	TAT	84.42 ± 0.33	94.09 ± 0.26	56.58 ± 0.99	81.68 ± 0.54	78.83 ± 0.52	26.69 ± 0.76
mT5-base	OST	**94.47 ± 0.06**	**95.94 ± 0.03**	**63.87 ± 0.44**	**85.18 ± 0.19**	82.46 ± 0.27	**21.41 ± 0.21**
trBART	OST	82.65 ± 0.25	92.47 ± 0.16	48.71 ± 0.69	76.45 ± 0.32	73.26 ± 0.6	34.79 ± 0.28

Table 5: The Performance Scores of Our Models on the Test Datasets. TER score measures distance. The other metrics measure similarity.

OST Test Dataset

Model	Train Dataset	BERTScore Cased	BERTScore Uncased	BLEU	ROUGE-L	METEOR	TER
mT5-base	OST	**89 ± 0.01**	**92.05 ± 0.01**	**46.26 ± 0.09**	**74.8 ± 0.02**	**72.97 ± 0.13**	**36.4 ± 0.04**
mT5-base	OST *(Unfiltered)*	88.89 ± 0.06	91.94 ± 0.04	36.4 ± 0.23	73.87 ± 0.09	72.16 ± 0.16	37.58 ± 0.15
mT5-base	TAT	84.95 ± 0.38	89.23 ± 0.24	29.37 ± 0.83	66.64 ± 0.46	63.14 ± 0.86	49.29 ± 1.13
mT5-base	TAT *(Unfiltered)*	88.95 ± 0.2	92.08 ± 0.14	38.13 ± 0.4	68.39 ± 0.23	65.87 ± 0.22	45.13 ± 0.33

TAT Test Dataset

Model	Train Dataset	BERTScore Cased	BERTScore Uncased	BLEU	ROUGE-L	METEOR	TER
mT5-base	TAT	94.07 ± 0.36	95.75 ± 0.25	61.66 ± 1.34	84.67 ± 0.62	82.72 ± 0.42	22.43 ± 1.27
mT5-base	TAT *(Unfiltered)*	91.61 ± 0.12	93.93 ± 0.09	34.74 ± 0.62	**86.6 ± 0.2**	**84.85 ± 0.23**	**18.23 ± 0.25**
mT5-base	OST	**94.47 ± 0.06**	**95.94 ± 0.03**	**63.87 ± 0.44**	85.18 ± 0.19	82.46 ± 0.27	21.41 ± 0.21
mT5-base	OST *(Unfiltered)*	91.97 ± 0.07	94.2 ± 0.05	37.02 ± 0.16	84.05 ± 0.19	81.59 ± 0.28	22.76 ± 0.32

Table 6: A Comparison Between the Performance of mT5 Model Checkpoints Trained on Our Filtered and Unfiltered Datasets. TER score measures distance. The other metrics measure similarity.

Ahmet Bağcı and Mehmet Fatih Amasyali. 2021. Comparison of turkish paraphrase generation models. In *2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, pages 1–6. IEEE.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Figen Beken Fikri, Kemal Oflazer, and Berrin Yanikoglu. 2021. Semantic similarity based evaluation for abstractive news summarization. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 24–33, Online. Association for Computational Linguistics.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Emrah Budur, Rıza Özçelik, Tunga Gungor, and Christopher Potts. 2020. Data and Representation for Turkish Natural Language Inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8253–8267, Online. Association for Computational Linguistics.

Steven Burrows, Martin Potthast, and Benno Stein. 2013. Paraphrase acquisition via crowdsourcing and machine learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(3):1–21.

Ruisheng Cao, Su Zhu, Chenyu Yang, Chen Liu, Rao Ma, Yanbin Zhao, Lu Chen, and Kai Yu. 2020. Unsupervised dual paraphrasing for two-stage semantic parsing. *arXiv preprint arXiv:2005.13485*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Mathias Creutz. 2018. Open subtitles paraphrase corpus for six languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Seniz Demir, Ilknur Durgar El-Kahlout, and Erdem Unal. 2013. A case study towards turkish paraphrase alignment. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 188–192.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sonal Garg, Sumanth Prabhu, Hemant Misra, and G Srinivasaraghavan. 2021. Unsupervised contextual paraphrase generation using lexical control and reinforcement learning. *arXiv preprint arXiv:2103.12777*.

Bahar Karaoğlan, Tarık Kışla, Senem Kumova Metin, Ufuk Hürriyetoğlu, and Katira Soleymanzadeh. 2016. Using multiple metrics in automatically building turkish paraphrase corpus. *Research in Computing Science*, 117:75–83.

Ashutosh Kumar, Satwik Bhattamishra, Manik Bhandari, and Partha Talukdar. 2019. Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3609–3619.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1224–1234, Copenhagen, Denmark. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li. 2018. Paraphrase generation with deep reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3865–3878, Brussels, Belgium. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2020. Muss: multilingual unsupervised sentence simplification by mining paraphrases. *arXiv preprint arXiv:2005.00352*.

Aurélien Max and Guillaume Wisniewski. 2010. Mining naturally-occurring corrections and paraphrases from Wikipedia's revision history. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Ali Safaya, Emirhan Kurtuluş, Arda Goktogan, and Deniz Yuret. 2022. Mukayese: Turkish NLP strikes back. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 846–863, Dublin, Ireland. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Yui Suzuki, Tomoyuki Kajiwara, and Mamoru Komachi. 2017. Building a non-trivial paraphrase corpus using multiple machine translation systems. In *Proceedings of ACL 2017, Student Research Workshop*, pages

36–42, Vancouver, Canada. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

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

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

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

Shuguang Zhu, Xiang Cheng, Sen Su, and Shuang Lang. 2017. Knowledge-based question answering by jointly generating, copying and paraphrasing. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2439–2442.

A Paraphrase Examples

We present in Table 7, 8 examples of paraphrases generated by the models fine-tuned on our train datasets. An abbreviation of the dataset each model was fine-tuned on is provided in parenthesis. We tried to choose representative examples showing cases of both failure and success.

Source	Ve onu sizden kimse alamaz, beyler. *And no one can take it away from you, gentlemen.*	Woodhouse tatlım biraz daha buza ihtiyacım var. *Woodhouse, honey, I need some more ice.*	Bir sandviçe yetecek kadar malzemem var. *I've got stuff that will be enough for a sandwich.*
mT5-base (OST)	Ve kimse onu sizden alamaz, beyler. *And no one can take it away from you, gentlemen.*	Woodhouse, tatlım, biraz daha buz lazım. *Woodhouse, honey, there is a need for more ice.*	Bir sandviç için yeterli malzemem var. *I've got enough stuff for a sandwich.*
trBART (OST)	beyler ve onu sizden kimse alamaz. *gentlemen and no one can take it away from you.*	woodhouse biraz daha buza ihtiyacım var. *woodhouse, I need some more ice.*	bir sandviçe yetecek kadar malzeme var. *there is enough stuff for a sandwich.*
mT5-base (TAT)	Kimse bunu sizden alamaz, beyler. *No one can take that away from you, gentlemen.*	Woodhouse tatlım biraz daha buza ihtiyacın var. *Woodhouse, honey, you need more ice.*	Sandviçe yetecek kadar malzemem var. *I've got stuff that will be enough for sandwich.*
trBART (TAT)	beyler ve onu sizden kimse alamaz. *gentlemen and no one can take it away from you.*	woodhouse biraz daha buza ihtiyacım var. *woodhouse, honey, there is a need for more ice.*	bir sandviçe yetecek kadar malzeme var. *there is enough stuff for a sandwich.*
mT5-base (OST-RAW)	Kimse onu senden alamaz, çocuklar. *No one can take it away from you, kids.*	Woodhouse, tatlım, biraz daha buz lazım. *Woodhouse, honey, there is a need for more ice.*	Bir sandviç için yeterli malzemem var. *I've got enough stuff for a sandwich.*
mT5-base (TAT-RAW)	Kimse bunu sizden alamaz, beyler. *No one can take that away from you, gentlemen.*	Woodhouse tatlım biraz daha buza ihtiyacın var. *Woodhouse honey you need more ice.*	Sandviçe yetecek kadar malzemem var. *I've got stuff that will be enough for sandwich.*

Table 7: Generated Paraphrases of Examples from the OST Dataset

Source	Tom daha sonra ne yapacağını bilmiyordu. *Tom didn't know what to do next.*	Tom asla tek başına oraya gitmezdi. *Tom would never go there by himself*	İlk olarak ne yapacaklarını merak ettiler. *They wondered what they would do first.*
mT5-base (OST)	Tom ne yapacağını bilmiyordu. *Tom didn't know what to do.*	Tom oraya tek başına gitmezdi. *Tom wouldn't go there by himself.*	Önce ne yapacaklarını merak ettiler. *They wondered what they would do before.*
trBART (OST)	tom bundan sonra ne yapacağını bilmiyordu. *tom didn't know what to do next.*	tom oraya hiç gitmezdi. *tom never went there.*	ilk olarak ne yapacaklarını merak ediyorlar. *they are wondering what they're going to do first.*
mT5-base (TAT)	Tom sonra ne yapacağını bilmiyordu. *Tom didn't know what to do next.*	Tom oraya asla tek başına gitmez. *Tom never goes there by himself.*	İlk başta ne yapacaklarını merak ettiler. *They wondered what they were going to do at first.*
trBART (TAT)	tom bundan sonra ne yapacağını bilmiyordu. *tom didn't know what to do next.*	tom oraya hiç gitmezdi. *tom never went there.*	ilk olarak ne yapacaklarını merak ediyorlar. *they are wondering what they're going to do first.*
mT5-base (OST-RAW)	Tom bundan sonra ne yapacağını bilmiyordu. *tom didn't know what to do next.*	Tom oraya hiç tek başına gitmedi. *Tom didn't go there by himself.*	Önce ne yapacaklarını merak ediyorlar. *They are wondering what they would do before.*
mT5-base (TAT-RAW)	Tom bundan sonra ne yapacağını bilmiyordu. *Tom didn't know what to do next.*	Tom oraya asla tek başına gitmez. *Tom never goes there by himself.*	Önce ne yapacaklarını merak ettiler. *They wondered what they would do before.*

Table 8: Generated Paraphrases of Examples from the TAT Dataset