Controllable Paraphrase Generation for Semantic and Lexical Similarities

Yuya Ogasa[1], **Tomoyuki Kajiwara**[2], **Yuki Arase**[1]
[1]Graduate School of Information Science and Technology, Osaka University
[2]Graduate School of Science and Engineering, Ehime University
[1]{ogasa.yuya, arase}@ist.osaka-u.ac.jp, [2]kajiwara@cs.ehime-u.ac.jp

Abstract

We developed a controllable paraphrase generation model for semantic and lexical similarities using a simple and intuitive mechanism: attaching tags to specify these values at the head of the input sentence. Lexically diverse paraphrases have been long coveted for data augmentation. However, their generation is not straightforward because diversifying surfaces easily degrades semantic similarity. Furthermore, our experiments revealed two critical features in data augmentation by paraphrasing: appropriate similarities of paraphrases are highly downstream task-dependent, and mixing paraphrases of various similarities negatively affects the downstream tasks. These features indicated that the controllability in paraphrase generation is crucial for successful data augmentation. We tackled these challenges by fine-tuning a pre-trained sequence-to-sequence model employing tags that indicate the semantic and lexical similarities of synthetic paraphrases selected carefully based on the similarities. The resultant model could paraphrase an input sentence according to the tags specified. Extensive experiments on data augmentation for contrastive learning and pre-fine-tuning of pretrained masked language models confirmed the effectiveness of the proposed model. We release our paraphrase generation model and a corpus of 87 million diverse paraphrases. (`https://github.com/Ogamon958/ConPGS`)

Keywords: controllable paraphrase generation, paraphrase corpus

1. Introduction

Paraphrases render the meaning of text using different words, phrases, and syntactic structures. Paraphrase generation (Zhou and Bhat, 2021) contributes to various downstream tasks, among which data augmentation is one of the primary applications. Examples include data augmentation for machine reading comprehension (Yu et al., 2018), task-oriented dialog systems (Jolly et al., 2020; Gao et al., 2020), machine translation (Effendi et al., 2018), and spoken dialogue systems (Okur et al., 2022). Lexically diverse paraphrases are crucial in data augmentation because they enhance the linguistic diversity of the original corpus (Qian et al., 2019). However, generating lexically diverse paraphrases is challenging because dynamic surface changes easily make sentences semantically less similar (Bandel et al., 2022).

Figure 1 visualises the distributions of semantic and lexical similarities of existing paraphrases as heatmaps; the former is measured by a fine-tuned pre-trained model with the Semantic Textual Similarity Benchmark (STS-B) (Cer et al., 2017) while the latter by a sentence BLEU score (Papineni et al., 2002). These paraphrases were generated by round-trip translation, one of the common methods of automatic paraphrase generation (Mallinson et al., 2017; Kajiwara et al., 2020), and sampled from those in the existing large-scale corpora: ParaNMT-50M (Wieting and Gimpel, 2018) and Paracotta (Aji et al., 2021). Figure 1 reveals that paraphrases generated by round-trip translation and Paracotta are semantically similar, yet their lexical similarities are also high. Sentence pairs in ParaNMT-50M are lexically diverse; however, many of them are semantically too divergent as paraphrases. These indicate that lexically diverse yet semantically similar paraphrases are scarce.

Furthermore, our experiments (Section 5 and 6) empirically reveal that appropriate levels of semantic and lexical similarities for data augmentation are dependent on downstream tasks and mixing paraphrases of diverse similarities negatively affect the data augmentation. These findings indicate that the controllability in paraphrase generation is crucial. Unfortunately, no previous studies have allowed intuitive and easy control of these similarities in paraphrase generation.

To tackle these challenges, we fine-tune a pre-trained sequence-to-sequence model employing tags indicating semantic and lexical similarities of synthetic paraphrases. These tags allow control of the similarities in generation in a simple and intuitive manner (Johnson et al., 2017). Specifically, we first generate numerous paraphrase candidates of various similarities using round-trip translation with sampling-based decoding. We select a subset of desirable paraphrases based on the semantic and lexical similarities as the fine-tuning corpus. At inference, we can specify the desired similarities of paraphrases using tags. Figure 1(d) shows the distribution of paraphrases generated by our model, where lexically diverse yet semantically sim-

LREC-COLING 2024, pages 3927–3942
20-25 May, 2024.

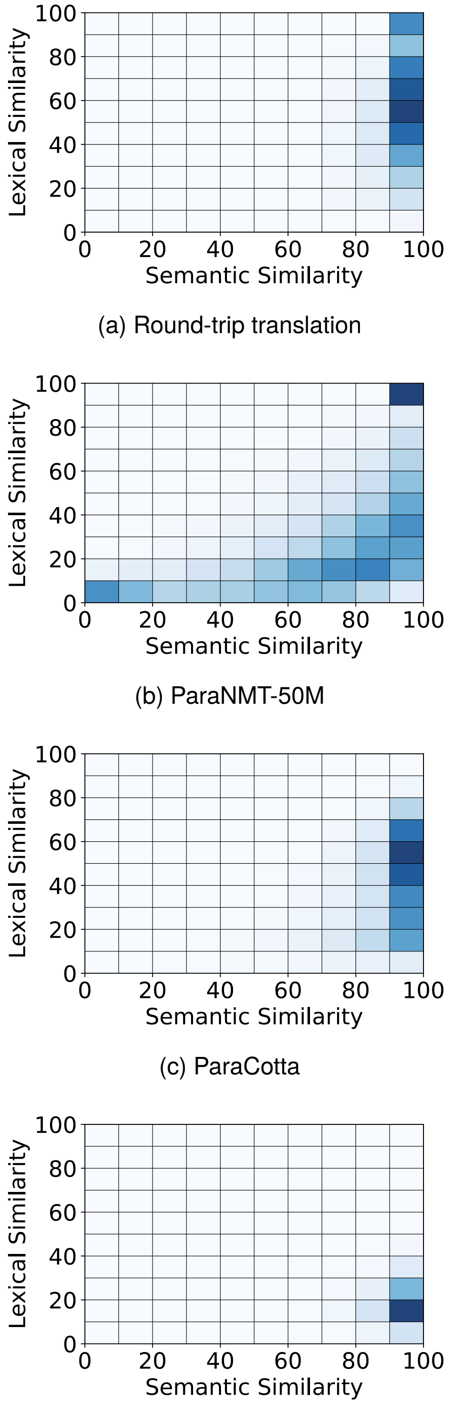
ilar paraphrases are generated successfully.

We conduct extensive experiments to investigate the effects of our model on data augmentation for contrastive learning (Gao et al., 2021; Liu et al., 2021) and pre-fine-tuning of pretrained language models (Phang et al., 2018; Arase and Tsujii, 2019). The experimental results confirm that data augmentation with our controllable paraphrase generation model boosts the performance of the downstream tasks. In addition to our codes, we will release our model so that users can generate paraphrases for their own problems.[1] Furthermore, we also publish an 87 million paraphrase corpus generated by our model for off-the-shelf usage of lexically diverse paraphrases.

2. Related Work

Diverse paraphrase generation has been actively studied; however, the controllability has been out of the scope. In addition to round-trip translation, back-translation is a common approach to generate paraphrases, as represented by ParaNMT (Wieting and Gimpel, 2018). ParaBank (Hu et al., 2019) extended ParaNMT by adding lexical constraints to the decoder (Hokamp and Liu, 2017; Post and Vilar, 2018) derived from the Paraphrase Database (Pavlick et al., 2015). Similarly, the methods of Niu et al. (2021) and Zeng et al. (2019) increased the lexical diversity in paraphrase generation by forcing the decoder to avoid outputting tokens in source sentences. There have been approaches adding linguistic knowledge to input, i.e., parse trees (Iyyer et al., 2018), sentence structures and keywords that should be used in generation (Yang et al., 2022), and exemplar sentences (Hosking and Lapata, 2021; Chen et al., 2019; Bao et al., 2019). Conversely, Maddela et al. (2021) have increased the diversity in a training corpus by preprocessing of word deletion and sentence splits, while Goyal and Durrett (2020) explored pre-ordering of source sentences for syntactic diversity. The other approaches work in a latent space using reinforcement learning with multiple paraphrase generation models (Qian et al., 2019), conditional generative adversarial networks (Cao and Wan, 2020), perturbation of latent representations (Gupta et al., 2018), and applying dropouts while specifying keywords and styles (Chen et al., 2022).

Conversely, previous studies have aimed to collect or generate paraphrases of specific similarity values. ParaCotta (Aji et al., 2021) collected paraphrases by selecting round-trip translation pairs with low sentence BLEU scores. However, Figure 1(a) confirmed that lexically diverse paraphrases are hard to obtain by simple round-trip translation. Chowdhury et al. (2022) train a model with a corpus with a specific translation edit rate (Snover et al., 2006) value, while Meng et al. (2021)



(a) Round-trip translation

(b) ParaNMT-50M

(c) ParaCotta

(d) Ours: ⟨SIM95⟩ & ⟨BLEU0_5⟩

Figure 1: Heatmaps of semantic and lexical similarities derived from 50k paraphrase samples generated by previous studies and our model. The darker the colour of a cell becomes, the higher the ratio of paraphrases of the corresponding semantic and lexical similarities becomes.

[1] https://github.com/Ogamon958/ConPGS

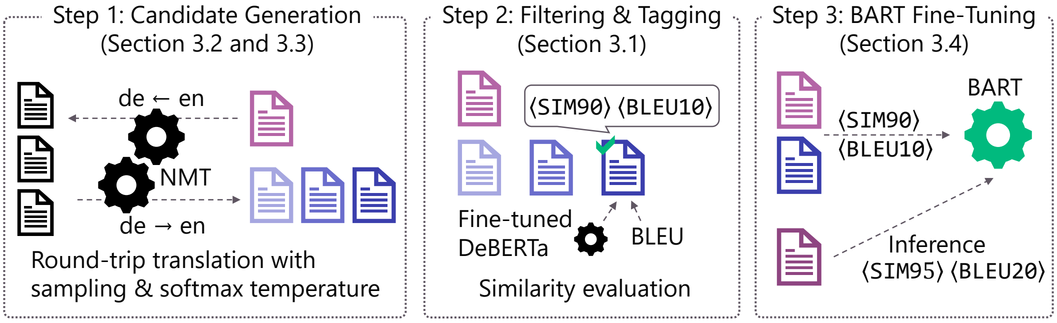


Figure 2: Overview of construction of our controllable paraphrase generation model

do so with a corpus of specific similarities.

None of these previous methods controls the level of similarity in paraphrase generation. An exception is Bandel et al. (2022); their model first learns to estimate expected semantic, syntactic, and lexical similarities between paraphrases using existing paraphrase corpora, i.e., ParaBank. Based on these values, their model allows to manually specify 'offsets' relative to the estimated similarities, determining how far these similarities can differ in a generated paraphrase. There is no simple way to set their reasonable values because these offsets depend on each input sentence. In contrast, our semantic and lexical similarities are absolute values, which can be set intuitively based on how similar or diverse we want the generated paraphrase to be. Also, unlike our study, which conducts an extensive extrinsic evaluation to investigate the effects of generated paraphrases on data augmentation, Bandel et al. (2022) showed only intrinsic evaluation results.

3. Controllable Generation

Figure 2 shows the overview of the construction of our model. We first construct a training corpus, where semantic and lexical similarities are attached as tags. We then fine-tune a pretrained sequence-to-sequence model, on which we can control the similarities in paraphrase generation using the tags.

3.1. Similarity Estimation

Sentence pairs are preprocessed before similarity estimation to remove symbols other than a space, comma, and period so that superficial differences do not affect the estimation. **Semantic similarity** is estimated by the DeBERTaV3 (He et al., 2023)[2] in the range of $[0, 1]$, which was fine-tuned with the STS-B.[3] Details of the model training are in the Appendix A. Note that this model is used only for labelling the training corpus, i.e., it is independent of our paraphrase generation model itself. **Lexical similarity** is measured by sentence BLEU after lower-casing, which has been commonly used to assess lexical similarity (diversity) (Chen and Dolan, 2011; Tian et al., 2017; Jiang et al., 2020).

Both semantic and lexical similarity scores are multiplied by 100 to have a range of $[0, 100]$. We define lexically diverse paraphrases as sentence pairs having semantic similarity higher than 70 while lexical similarity is smaller or equal to 45. We bin semantic and lexical similarities by the interval of 5, whose tags are ⟨SIM70⟩, ⟨SIM75⟩, ..., ⟨SIM95⟩ for the former (6 tags in total) and ⟨BLEU0_5⟩[4], ⟨BLEU10⟩, ⟨BLEU15⟩, ..., ⟨BLEU40⟩ for the latter (8 tags in total), respectively.

3.2. Candidate Generation and Selection

Figure 1(a) revealed that simple round-trip translation may end up generating sentence pairs semantically similar but with high lexical overlap despite round-trip translation being one of the most common methods for automatically generating paraphrases. Therefore, we employ Top-k sampling (Fan et al., 2018) while applying the temperature in the softmax computation in the decoder to increase lexically diversity:

$$\frac{\exp(\boldsymbol{z}_i/T)}{\sum_j \exp(\boldsymbol{z}_j/T)},$$

where $\boldsymbol{z}$ is an input vector and $T \in \mathbb{R}_+$ is the temperature that controls the softness of the output probability distribution. The larger T makes the probability split more evenly among the vocabulary. Therefore, when combined with Top-k sampling, the model tends to output diverse tokens.

On the flip side, the round-trip translation with these settings produces semantically less similar sentences that are no longer paraphrases. As a

[2] https://huggingface.co/microsoft/deberta-v3-large

[3] DeBERTa outperformed BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) on the validation set in our preliminary experiment.

[4] As we had a smaller number of candidates with BLEU score less than 10, we merged them into one bin.

straightforward remedy, we generate a huge number of paraphrase candidates and evaluate their semantic and lexical similarities as described in Section 3.1. Finally, we extract only lexically diverse paraphrases with similarity tags attached.

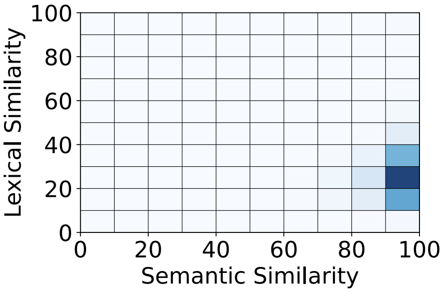
3.3. Training Corpus Construction

Following Kajiwara et al. (2020), we performed English to German then German to English round-trip translation using the de facto standard translators released by Ng et al. (2019).[5] We empirically set the size of the Top-k sampling and temperature T, which were searched in $k = \{10, 20, 30, 40\}$ and $T = \{1.0, 2.0, 3.0, 4.0\}$ respectively. We observed the distributions of semantic and lexical similarities of generated candidates using held-out sentences for development. Furthermore, we sampled small sets of candidates and manually evaluated the fluency and semantic and lexical similarities. Based on these observations, we decided to use the 2 settings of $(k, T) = (20, 3.0), (30, 2.0)$, which confirmed to produce relatively larger numbers of candidates being semantically similar while lexically diverse. We applied these settings combinatorial with directions of forward and backward translations, i.e., two settings times two directions, which gave us 4 candidates per input sentence.

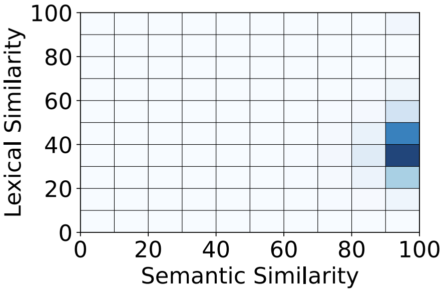
As inputs to round-trip translation, we used the English side of English-German WikiMatrix (Schwenk et al., 2021) and about 30M English sentences sampled from NewsCrawl (Akhbardeh et al., 2021). Consequently, we obtained 120M of candidate pairs from which we selected lexically diverse paraphrases. Remind that we have 6 and 8 tags of semantic and lexical similarities, respectively. We ensure that the distribution of numbers of paraphrases for the combinations of semantic and lexical similarities is balanced.[6] Finally, we split the corpus into a training set of 5M pairs and validation and test sets of $2,700$ pairs each, respectively.

3.4. Paraphrase Generation Model

Our lexically diverse paraphrase generation model was developed by fine-tuning BART (Lewis et al., 2020)[7] with the corpus constructed in Section 3.3. At inference, we can input tags of desired lexical and semantic similarities in generated paraphrases. We set the beam size to 5 and constrained the output length to be 0.75 to 1.5 times the input length.



(a) ⟨SIM95⟩ & ⟨BLEU20⟩



(b) ⟨SIM95⟩ & ⟨BLEU40⟩

Figure 3: Distribution of semantic and lexical similarities on paraphrases generated by our model

4. Profile of Our Paraphrases

Figures 1(d), 3(a), and 3(b) show the distributions of semantic and lexical similarities of paraphrases generated by our model when specified tags of ⟨SIM95⟩&⟨BLEU0_5⟩, ⟨SIM95⟩&⟨BLEU20⟩, and ⟨SIM95⟩&⟨BLEU40⟩, respectively. The darker the colour of a cell, the higher the ratio of paraphrases with the corresponding semantic and lexical similarities is. For each setting, we generated paraphrases for 50k sentences sampled from English Wikipedia[8] and evaluated these similarities as described in Section 3.1. Obviously, the darkest cells match well with the specified tags. These figures indicate that our paraphrase generation model preserves controllability.

Table 1 shows paraphrases generated by our model with different semantic and lexical similarities (see Appendix B for more examples). Specifying ⟨SIM95⟩ commands the model to generate paraphrases with almost equivalent meanings to the source. The examples show that generated paraphrases satisfy this condition while achieving lexical diversity according to the specified tags. When ⟨SIM70⟩ was specified, which commands the model to have moderate semantic diversity in paraphrases, the expression “leg” is converted

[5] the wmt19-en-de and wmt19-de-en models under https://huggingface.co/facebook/

[6] We sampled twice the number of paraphrases for ⟨BLEU0_5⟩ as this bin covers a two times larger range.

[7] https://huggingface.co/facebook/bart-base

[8] https://huggingface.co/datasets/princeton-nlp/datasets-for-simcse/resolve/main/wiki1m_for_simcse.txt

Source: Maria Sharapova has been forced to withdraw with a leg injury .	
Tags	Generated paraphrases
⟨SIM95⟩ ⟨BLEU0_5⟩	Maria Sharapova **withdrew** with an injury to **her** leg.
⟨SIM95⟩ ⟨BLEU40⟩	Maria Sharapova **had to pull out** with a leg injury.
⟨SIM70⟩ ⟨BLEU0_5⟩	Maria Sharapova **withdrew from the tournament** with an **ankle** injury.
⟨SIM70⟩ ⟨BLEU40⟩	Maria Sharapova has been forced to **pull out of the French Open** with a leg injury.

Source: The group included four children, Turkish official says .	
Tags	Generated paraphrases
⟨SIM95⟩ ⟨BLEU0_5⟩	Four children **were among** the group, a Turkish **government** official **said**.
⟨SIM95⟩ ⟨BLEU40⟩	Four children **were among** the group, Turkish official says.
⟨SIM70⟩ ⟨BLEU0_5⟩	Four children **were among** the **victims**, a Turkish **government** official **said**.
⟨SIM70⟩ ⟨BLEU40⟩	Four children **were among** the **victims**, Turkish official says.

Table 1: Example paraphrases generated by our model with various semantic and lexical similarities (**bold words** are not included in source sentences)

to “ankle” and generated sentences have the additional phrases “the tournament” or “the French open” which drifts the meaning of the source in the first example. In the second example, the word “group” is replaced by “victim”, which alters the original meaning.

Lexically Diverse Paraphrase Corpus We further constructed a large-scale English corpus consisting of 87M of lexically diverse paraphrases for off-the-shelf usages using our paraphrase generation model. Source sentences were sampled from Wiki-40B (Guo et al., 2020). We first conducted language identification[9] and selected English sentences between 10 to 100 tokens. We input these sentences with tags; combining four semantic similarity tags of ⟨SIM80⟩, ⟨SIM85⟩, ⟨SIM90⟩, and ⟨SIM95⟩, and four lexical similarity tags of ⟨BLEU0_5⟩, ⟨BLEU10⟩, ⟨BLEU15⟩, and ⟨BLEU20⟩. In total, we have 16 different combinations of tags and corresponding paraphrases.

5. Effects on Contrastive Learning

We evaluate the effects of our controllable paraphrase generation on data augmentation for improving sentence representations using contrastive learning. As a representative method, we apply our model to SimCSE (Gao et al., 2021).

5.1. Preliminary: SimCSE

SimCSE fine-tunes a pre-trained masked language model using contrastive learning that pulls semantically close embeddings together while pushing apart semantically distant embeddings. SimCSE can be conducted using either a raw corpus or a natural language inference (NLI) corpus. When using the raw corpus, the same sentence is input to the pretrained model twice and applied dropouts, which serve as a positive pair. Conversely, negatives are sampled from the mini-batch. When using the NLI corpus, entailment pairs serve as positives and contradictive pairs serve as negatives.

Following the original experimental settings of SimCSE, in this evaluation, BERT-base[10] was fine-tuned using 1M sentences sampled from English Wikipedia as the raw corpus and 280k pairs from MNLI (Williams et al., 2018) and SNLI (Bowman et al., 2015) as the NLI corpus. We used the official implementation by Gao et al. (2021).[11]

The effect of SimCSE was evaluated on unsupervised STS using STS12-16 (Agirre et al., 2012, 2013, 2014, 2015, 2016), STS-B, and SICK-R (Marelli et al., 2014), where the evaluation metric is Spearman’s rank correlation coefficient (ρ) between estimated similarities and human labels. In all experiments, we report average scores of 5 training and evaluation trials with random seeds.

5.2. Data Augmentation

We generated paraphrases to use as positive pairs for SimCSE with the raw corpus instead of generating them by dropouts as the original SimCSE does. For SimCSE with the NLI corpus, we paraphrased pairs in the NLI corpus and added them to the original corpus, which doubles the size of the training corpus. Specifically, a sentence is paired with positive and negative samples in the NLI corpus, the three of which were paraphrased and paired again.

We assumed paraphrases with high semantic similarity are appropriate because SimCSE aims to improve sentence embeddings for better represent-

[9] https://github.com/bsolomon1124/pycld3

[10] https://huggingface.co/bert-base-uncased

[11] https://github.com/princeton-nlp/SimCSE

	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg
SimCSE on Raw corpus								
(Gao et al., 2021)	67.19	**81.13**	**73.13**	80.51	77.72	76.08	70.66	75.20
RTT	63.58	73.62	66.48	75.39	73.82	69.83	66.56	69.90
⟨BLEU0_5⟩	69.10	77.17	71.21	80.66	79.28	77.82	71.78	75.29
⟨BLEU20⟩	**69.89**	78.69	72.25	**80.90**	**80.02**	**78.39**	**72.20**	**76.05**
⟨BLEU40⟩	68.91	78.99	72.05	80.67	79.63	77.65	72.13	75.72
SimCSE on NLI corpus								
(Gao et al., 2021)	75.32	84.81	80.30	85.58	81.05	84.39	80.42	81.70
RTT	76.32	83.86	80.65	85.88	81.68	84.65	80.34	81.91
⟨BLEU0_5⟩	**76.82**	84.84	**80.76**	**86.31**	**81.72**	**85.03**	**80.63**	**82.30**
⟨BLEU20⟩	76.42	**85.22**	80.73	86.06	81.53	84.90	80.43	82.18
⟨BLEU40⟩	76.19	84.96	80.57	85.87	81.58	84.76	80.61	82.07

Table 2: Spearman's rank correlation coefficients ($\rho \times 100$) measured on the test sets (**Bold** font indicates the highest scores.)

ing semantic similarity. In data augmentation, we fixed the semantic similarity as ⟨SIM95⟩ and varied lexical similarities by combining tags of ⟨BLEU0_5⟩, ⟨BLEU20⟩, and ⟨BLEU40⟩, respectively. We also compared data augmentation by round-trip translation (denoted as 'RTT' hereafter) as a baseline. For RTT, we used the same machine translation models that generated our training corpus (Section 3.3). We used a beam search of size 5[12] and constrained the output length to be 0.75 to 1.5 times the input length. This setting is identical to the decoding method we adapted for our model.

5.3. Overall Results

Table 2 shows the results when trained SimCSE with the raw and NLI corpora, respectively. The first rows are the original SimCSE without data augmentation reproduced by us. For SimCSE with the raw corpus, our model achieved the best average score when using the lexical similarity of ⟨BLEU20⟩. Conversely, RTT largely degraded the score of the original SimCSE. We conjecture this may be because lexical similarities between paraphrases by RTT are too high, as shown in Figure 1(a).

For SimCSE with the NLI corpus, our model again achieved the highest average scores with the lowest lexical similarity of ⟨BLEU0_5⟩. These results imply that the appropriate similarities are task-dependent, which we dig into in Section 6.

5.4. Effects of Augmentation Scale

Next, we investigate the effects of the scale of data augmentation. Specifically, we scale up the original corpus by adding paraphrased pairs on SimCSE with the NLI corpus; i.e., we generate multiple paraphrases of a sentence by obtaining N-best hypotheses with beam search.

Figure 4 shows the trends; the X-axis represents the multiplier of the augmented corpus size relative to the original size. The Y-axis shows the average Spearman's rank correlation coefficients measured on the STS tasks after SimCSE training of BERT. We started data augmentation from the half-sized (used the first half of the corpus) and the full-sized NLI corpus, which are indicated by dashed and solid lines, respectively. We also evaluated the setting that merges paraphrases generated using different lexical similarity tags as one corpus, denoted as 'Merge' with yellow lines. Overall, our paraphrase generation model consistently outperforms RTT. It is remarkable that on the half-sized setting, SimCSE with the double-sized augmented corpus by ⟨BLEU0_5⟩ is competitive to the original NLI corpus.

In addition, we have the following observations: (1) Mixing paraphrases with different lexical similarities is harmful. We had an assumption that combining paraphrases with various lexical similarities further improves the performance. However, the assumption does not hold for most settings. We conjecture that lexically diverse paraphrases benefit SimCSE on NLI corpus while the diversity (similarity) level should be consistent in a corpus. Further investigation constitutes our future work. (2) Improvement gets saturated regarding the scale of augmentation, which is consistent with previous studies (Fadaee et al., 2017). Notably, the peak size of the augmented corpus depends on the size of the original corpus; double on the half-sized setting (except ⟨BLEU40⟩) while quintuple on the full-sized setting regardless of the lexical similarities. Nonetheless, the performance tends to drop faster

[12] We did not use the greedy decoding with sampling as it generates lexically diverse but semantically dissimilar sentences, which was empirically confirmed inferior to beam search on data augmentation.

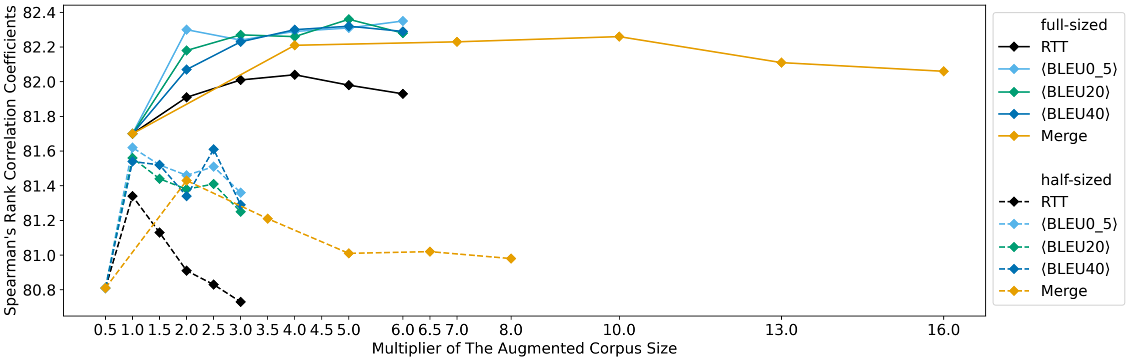


Figure 4: Effects of the scale of data augmentation

	Average STS score	Sentence BLEU	vocabulary size	perplexity
(Gao et al., 2021)	81.70	14.21	**96,721**	**247.36**
RTT	81.90	14.12	75,631	184.53
⟨BLEU0_5⟩	**82.37**	**11.25**	91,318	135.58
⟨BLEU20⟩	82.16	11.85	88,032	155.73
⟨BLEU40⟩	82.15	12.62	85,597	178.39

Table 3: Linguistic diversities in the NLI corpus and its paraphrases, and average STS scores

on the half-sized setting when adding more paraphrases.

5.5. Critical Features for Data Augmentation by Paraphrasing

In this section, we further investigate what features are crucial for effective data augmentation by lexically diverse paraphrasing for SimCSE. We measured linguistic diversities in the original NLI corpus and their paraphrases generated by RTT and our model with different tags, respectively. Specifically, we use the average sentence BLEU between pairs, vocabulary size, and the perplexity computed by pre-trained GPT-2 (Radford et al., 2019)[13]. While a lower sentence BLEU means paraphrase *pairs* are lexically more diverse, a lower perplexity indicates the *corpus* is less diverse (more uniform) (Moore and Lewis, 2010) as a whole.

Table 3 shows the results. The first row corresponds to the SimCSE trained with the original NLI corpus, while others are trained with only the paraphrased corpora of the same size. Remarkably, all SimCSE models trained by our paraphrases, ⟨BLEU0_5⟩, ⟨BLEU20⟩, and ⟨BLEU40⟩, outperformed the original SimCSE despite that they are trained only on synthetic sentences. They also outperformed paraphrasing by RTT. Table 3 reveals that our paraphrases have lower sentence BLEU scores and perplexities than the ones generated by RTT, while their vocabulary sizes are larger. These results indicate that for improving SimCSE by data augmentation, the lexical diversity between sentence pairs and uniformity as a corpus is important.

6. Effects on Pre-Fine-Tuning

We evaluate our paraphrase generation model on data augmentation for pre-fine-tuning a pre-trained language model. Specifically, we apply our model to Supplementary Training on Intermediate Labeled-data Tasks (STILTs) (Phang et al., 2018).

6.1. Preliminary: STILTs

Pre-fine-tuning improves the performance of the pretrained language model on downstream tasks by conducting additional training before fine-tuning. Phang et al. (2018) showed that pre-fine-tuning on BERT-large with the MNLI corpus is the best-performing combination in STILTs.

The performance of pre-fine-tuned BERT was evaluated on tasks from the GLUE benchmark (Wang et al., 2018): CoLA (Warstadt et al., 2019) for linguistic acceptability estimation, SST-2 (Socher et al., 2013) for binary sentiment classification, MRPC (Dolan and Brockett, 2005) and QQP[14] for paraphrase recognition, STS-B for semantic textual similarity estimation, and MNLI, QNLI (Rajpurkar et al., 2016)[15], and RTE (Bentivogli et al.,

[13] https://huggingface.co/gpt2

[14] https://www.quora.com/q/quoradata/

[15] Phang et al. (2018) used the older QNLIv1, whereas we used the newer QNLIv2.

	CoLA	SST-2	MRPC	QQP	STS-B	MNLI-m/mm	QNLI	RTE
BERT	**58.5**	94.3	88.3	**72.4**	86.8	**86.5** / 85.6	92.7	69.0
STILTs	57.0	94.2	89.0	71.7	**88.9**	– / –	92.5	79.4
RTT	56.4	94.2	88.6	71.6	88.2	86.3 / 86.1	92.4	79.6
Ours	55.7 ⟨SIM70⟩ ⟨BLEU20⟩	**94.9** ⟨SIM70⟩ ⟨BLEU40⟩	**89.2** ⟨SIM95⟩ ⟨BLEU0_5⟩	71.7 ⟨SIM70⟩ ⟨BLEU20⟩	88.7 ⟨SIM80⟩ ⟨BLEU40⟩	**86.5** / **86.2** ⟨SIM95⟩ / ⟨SIM95⟩ ⟨BLEU20⟩ / ⟨BLEU0_5⟩	**93.0** ⟨SIM95⟩ ⟨BLEU0_5⟩	**80.0** ⟨SIM95⟩ ⟨BLEU0_5⟩

Table 4: Test set scores computed in the GLUE benchmark server (**Bold** font indicates the highest scores.)

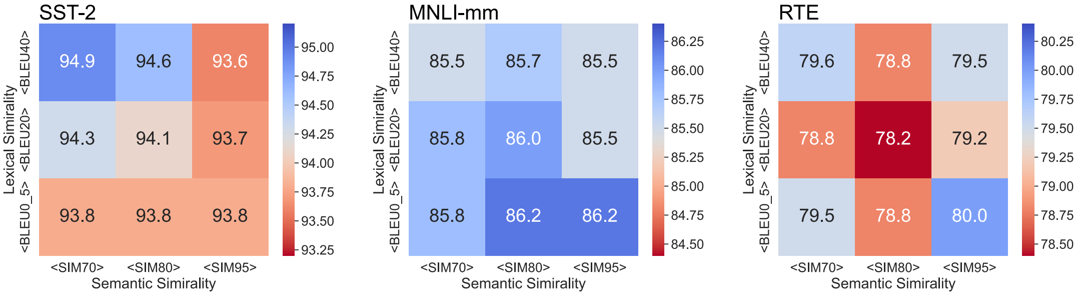


Figure 5: Heatmaps of the performance of models trained with augmented corpora by our paraphrases with different semantic and lexical similarities compared to STILTs with the MNLI corpus

2009) for natural language inference.[16] After fine-tuning using the training sets of these tasks, the test set scores were computed on the GLUE benchmark server[17]. An exception was MNLI, whose training set was used for pre-fine-tuning, and thus fine-tuning was skipped. For more details on the fine-tuning settings, please refer to Appendix C.

6.2. Data Augmentation

We expanded the training set of the MNLI corpus using our lexically diverse paraphrase generation model and conducted pre-fine-tuning under the same settings with STILTs.[18] Paraphrases were generated in the same manner as the NLI corpus in Section 5.2 using three semantic similarity tags of ⟨SIM70⟩, ⟨SIM80⟩, and ⟨SIM95⟩ and three lexical similarity tags of ⟨BLEU0_5⟩, ⟨BLEU20⟩, and ⟨BLEU40⟩, which produces nine tag combinations in total. Again, we compared the performance to data augmentation by round-trip translation (denoted as 'RTT' hereafter).

6.3. Results and Discussion

Table 4 shows the overall results; for our paraphrase generation model, only the scores of the best-performing tag combinations are listed.[19] The first row is direct fine-tuning of BERT, and the second row is STILTs with the MNLI corpus. The results confirm that our model outperforms STILTs in 5 tasks and data augmentation by RTT in 8 tasks, respectively. Figure 5 visualises the performance variations of our model depending on the tag combinations on tasks of SST-2, MNLI-mm, and RTE, compared to the original STILTs (see Appendix C for other tasks). In these heatmaps, blue indicates improvement, red indicates deterioration, and grey represents comparable scores to the original STILTs. These results clearly show that the appropriate tag settings are task-dependent. Therefore, the controllability of semantic and lexical similarities in paraphrase generation is crucial. Practically, users may explore appropriate tag settings using a small development set.

When we investigate the performance of our paraphrase generation model on each task, it is particularly effective on tasks like NLI and with a smaller training corpus, such as MRPC and RTE. This trend is consistent with the findings by Arase and Tsujii (2019). While SST-2 does not satisfy these characteristics, our model still improves its performance. Considering that the lower semantic similarities are preferred on this task, i.e., ⟨SIM70⟩ and ⟨SIM80⟩, BERT may have enhanced the robustness for sentiment analysis through pre-fine-tuning with the semantically more diverse augmented corpus. In contrast, STILTs did not contribute to QQP and CoLA, even if we applied data augmentation. We conjecture this is because QQP have sufficiently large fine-tuning corpora, and CoLA is too distant from the pre-fine-tuning task.

[16] WNLI was excluded because of the known problem: https://gluebenchmark.com/faq

[17] https://gluebenchmark.com/leaderboard

[18] Only the batch size was expanded from 24 to 32 on all models compared for training efficiency.

[19] When multiple tag combinations have the same best score, only one sample is shown due to space limitation.

7. Summary and Future Work

We developed a paraphrase generation model with controllability of semantic and lexical similarities. Extensive experiments confirmed the effectiveness of our model. Furthermore, the results revealed that appropriate levels of these similarities depend on downstream tasks while mixing paraphrases of different semantic and lexical similarities is harmful to data augmentation.

In future work, we will further investigate the relationship between semantic and lexical similarities and the effects of data augmentation. We will also apply our model to data augmentation for paraphrasing tasks with scarce resources, such as text simplification (Sun et al., 2023) and style transfer (Kajiwara et al., 2020).

Acknowledgements

This work was supported by LINE Corporation, Japan.

8. Bibliographical References

Alham Fikri Aji, Radityo Eko Prasojo Tirana Noor Fatyanosa, Philip Arthur, Suci Fitriany, Salma Qonitah, Nadhifa Zulfa, Tomi Santoso, and Mahendra Data. 2021. ParaCotta: Synthetic multilingual paraphrase corpora from the most diverse translation sample pair. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 533–542.

Yuki Arase and Jun'ichi Tsujii. 2019. Transfer Fine-Tuning: A BERT Case Study. In *Proceedings of Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5393–5404.

Elron Bandel, Ranit Aharonov, Michal Shmueli-Scheuer, Ilya Shnayderman, Noam Slonim, and Liat Ein-Dor. 2022. Quality Controlled Paraphrase Generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 596–609.

Yu Bao, Hao Zhou, Shujian Huang, Lei Li, Lili Mou, Olga Vechtomova, Xin-yu Dai, and Jiajun Chen. 2019. Generating Sentences from Disentangled Syntactic and Semantic Spaces. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6008–6019.

Yue Cao and Xiaojun Wan. 2020. DivGAN: Towards diverse paraphrase generation via diversified generative adversarial network. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2411–2421.

David Chen and William Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 190–200.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. Controllable Paraphrase Generation with a Syntactic Exemplar. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5972–5984.

Yi Chen, Haiyun Jiang, Lemao Liu, Rui Wang, Shuming Shi, and Ruifeng Xu. 2022. MCPG: A flexible multi-level controllable framework for unsupervised paraphrase generation. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 5948–5958.

Jishnu Ray Chowdhury, Yong Zhuang, and Shuyi Wang. 2022. Novelty controlled paraphrase generation with retrieval augmented conditional prompt tuning. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 36(10):10535–10544.

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 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.

Johanes Effendi, Sakriani Sakti, Katsuhito Sudoh, and Satoshi Nakamura. 2018. Multi-paraphrase Augmentation to Leverage Neural Caption Translation. In *Proceedings of the International Conference on Spoken Language Translation (IWSLT)*, pages 181–188.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 567–573.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical Neural Story Generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 889–898.

Silin Gao, Yichi Zhang, Zhijian Ou, and Zhou Yu. 2020. Paraphrase Augmented Task-Oriented Dialog Generation. In *Proceedings of the Annual*

Meeting of the Association for Computational Linguistics (ACL), pages 639–649.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6894–6910.

Tanya Goyal and Greg Durrett. 2020. Neural syntactic preordering for controlled paraphrase generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 238–252.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2018. A deep generative framework for paraphrase generation. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 32(1).

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1535–1546.

Tom Hosking and Mirella Lapata. 2021. Factorising meaning and form for intent-preserving paraphrasing. In *Proceedings of the Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 1405–1418.

J. Edward Hu, Rachel Rudinger, Matt Post, and Benjamin Van Durme. 2019. PARABANK: Monolingual Bitext Generation and Sentential Paraphrasing via Lexically-Constrained Neural Machine Translation. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 33(01):6521–6528.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1875–1885.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7943–7960.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association of Computational Linguistics (TACL)*, 5:339–351.

Shailza Jolly, Tobias Falke, Caglar Tirkaz, and Daniil Sorokin. 2020. Data-Efficient Paraphrase Generation to Bootstrap Intent Classification and Slot Labeling for New Features in Task-Oriented Dialog Systems. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 10–20.

Tomoyuki Kajiwara, Biwa Miura, and Yuki Arase. 2020. Monolingual Transfer Learning via Bilingual Translators for Style-Sensitive Paraphrase Generation. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 34(05):8042–8049.

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 Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7871–7880.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1442–1459.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pre-training approach. *arXiv*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable Text Simplification with Explicit Paraphrasing". In *Proceedings of the*

Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pages 3536–3553.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 881–893.

Yuxian Meng, Xiang Ao, Qing He, Xiaofei Sun, Qinghong Han, Fei Wu, Chun Fan, and Jiwei Li. 2021. ConRPG: Paraphrase generation using contexts as regularizer. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2551–2562.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 220–224.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 News Translation Task Submission. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*, pages 314–319.

Tong Niu, Semih Yavuz, Yingbo Zhou, Nitish Shirish Keskar, Huan Wang, and Caiming Xiong. 2021. Unsupervised Paraphrasing with Pretrained Language Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5136–5150.

Eda Okur, Saurav Sahay, and Lama Nachman. 2022. Data Augmentation with Paraphrase Generation and Entity Extraction for Multimodal Dialogue System. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 4114–4125.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 425–430.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv*.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1314–1324.

Lihua Qian, Lin Qiu, Weinan Zhang, Xin Jiang, and Yong Yu. 2019. Exploring Diverse Expressions for Paraphrase Generation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3173–3182.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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

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.

Renliang Sun, Zhixian Yang, and Xiaojun Wan. 2023. Exploiting summarization data to help text simplification. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 39–51.

Junfeng Tian, Zhiheng Zhou, Man Lan, and Yuanbin Wu. 2017. ECNU at SemEval-2017 task 1: Leverage kernel-based traditional NLP features and neural networks to build a universal model for multilingual and cross-lingual semantic textual similarity. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, pages 191–197.

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

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Haibo Zhang, Xue Zhao, Wenqing Yao, and Boxing Chen. 2022. Gcpg: A general framework for controllable paraphrase generation. In *Findings of the Association for Computational Linguistics: ACL*, pages 4035–4047.

Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. 2018. Fast and accurate reading comprehension by combining self-attention and convolution. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Daojian Zeng, Haoran Zhang, Lingyun Xiang, Jin Wang, and Guoliang Ji. 2019. User-oriented paraphrase generation with keywords controlled network. *IEEE Access*, 7:80542–80551.

Jianing Zhou and Suma Bhat. 2021. Paraphrase Generation: A Survey of the State of the Art. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5075–5086.

9. Language Resource References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, pages 252–263.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, pages 81–91.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, pages 497–511.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 Conference on Machine Translation (WMT21). In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*, pages 1–88.

Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009. The fifth PASCAL recognizing textual entailment challenge. In *Proceedings of the Text Analysis Conference (TAC)*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 632–642.

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 International Workshop on Semantic Evaluation (SemEval)*, pages 1–14.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the International Workshop on Paraphrasing (IWP)*.

Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. 2020. Wiki-40B: Multilingual language model dataset. In *Proceedings of the Lan-*

guage Resources and Evaluation Conference (LREC), pages 2440–2452.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 216–223.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2383–2392.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 1351–1361.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 353–355.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association of Computational Linguistics (TACL)*, 7:625–641.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1112–1122.

A. Implementation Details

We implemented our model using PyTorch[20] and Hugging Face Transformers[21]. For the semantic similarity estimation model, we used the cross-encoder architecture (Reimers and Gurevych, 2019). While fine-tuning used STS-B, the model performance may vary depending on the initialization seed. Therefore, we trained the model with 10 different random seeds and used the best model regarding Spearman's rank correlation coefficient measured on the validation set.

On fine-tuning BART to construct our lexically diverse paraphrase generation model, the batch size was set to 128. AdamW (Loshchilov and Hutter, 2019) was used as the optimiser and the learning rate was set to 5e-6 after the grid-search with 5e-6, 1e-5, and 2e-5, which showed the smallest validation loss. At the end of every epoch, the loss was computed on the validation set and the training was terminated when there was no improvement for 5 epochs.

B. Examples of Generated Paraphrases

Table 5 shows more examples of generated paraphrases. Furthermore, Table 6 demonstrates how diverse paraphrases could be when specified various tag combinations.

C. Details of Pre-Fine-Tuning Experiments

For tasks with more than $10,000$ training samples (SST, QQP, MNLI, QNLI), fine-tuning was conducted for 3 epochs. BERT-large is known to have training instability when a fine-tuning corpus is small (Devlin et al., 2019). Therefore, for tasks with smaller training sets (CoLA, MRPC, STS-B, and RTE), we fine-tuned BERT for longer epochs of 10 using 5 random seeds. From these 5 checkpoints, we selected the one with the median validation score.[22] The learning rate was set to 2e-5 and the batch size was 32.

Figure 6 shows the all heatmaps of the performance of data augmentation by our paraphrase generation model compared to STILTs without data augmentation. These heatmaps indicate that appropriate tags are strongly dependent on downstream tasks.

[20] https://pytorch.org/ (ver. $1.11.0$+cu113)

[21] https://huggingface.co/docs/transformers/

[22] Phang et al. (2018) fine-tuned BERT with 20 random seeds and selected the best model. We used the model with the median validation score for better reproducibility of our experiment results.

Source: New Broadcasting House in central London took a decade to build .

Tags	Generated paraphrases
⟨SIM95⟩ ⟨BLEU0_5⟩	The construction of the new central London Broadcasting House took ten years.
⟨SIM95⟩ ⟨BLEU40⟩	The construction of the new Broadcasting House in central London took ten years.
⟨SIM70⟩ ⟨BLEU0_5⟩	The construction of the new station in the centre of London took ten years.
⟨SIM70⟩ ⟨BLEU40⟩	The construction of the new radio station in central London took a decade.

Source: Belle Gibson was awarded Cosmo's Fun Fearless Female Award last year .

Tags	Generated paraphrases
⟨SIM95⟩ ⟨BLEU0_5⟩	Belle Gibson won last year's Cosmo Women's Fun Fearless Award.
⟨SIM95⟩ ⟨BLEU40⟩	Belle Gibson was awarded the Fun Fearless Female Award by Cosmo last year.
⟨SIM70⟩ ⟨BLEU0_5⟩	Belle Gibson won the Cosmo Women's Fun Fearless Award last year, which was awarded to the best actress in the world.
⟨SIM70⟩ ⟨BLEU40⟩	Belle Gibson won Cosmo's Fun Fearless Female Award last year for her role in the film.

Source: Rory McIlroy heads to the Masters hoping to complete a career grand slam .

Tags	Generated paraphrases
⟨SIM95⟩ ⟨BLEU0_5⟩	Rory McIlroy is heading to the Masters in the hope of ending his career as a Grand Slam winner.
⟨SIM95⟩ ⟨BLEU40⟩	Rory McIlroy is heading to the Masters hoping for a career grand slam.
⟨SIM70⟩ ⟨BLEU0_5⟩	Rory McIlroy is heading to the Masters in the hope of ending his career at the top of the leaderboard.
⟨SIM70⟩ ⟨BLEU40⟩	Rory McIlroy goes into the Masters hoping to complete a career Grand Slam with a win in the first round.

Source: The BBC is set to air a two-hour, real-time documentary following a boat .

Tags	Generated paraphrases
⟨SIM95⟩ ⟨BLEU0_5⟩	The BBC will broadcast a two-hour documentary in real time after a boat trip.
⟨SIM95⟩ ⟨BLEU40⟩	The BBC will broadcast a two-hour, real-time documentary after a boat.
⟨SIM70⟩ ⟨BLEU0_5⟩	The BBC will broadcast a two-hour documentary in real time after a boat accident.
⟨SIM70⟩ ⟨BLEU40⟩	The BBC is to broadcast a two-hour, real-time documentary about a boat crash.

Source: European pilots must fill out forms that ask about mental and physical illnesses .

Tags	Generated paraphrases
⟨SIM95⟩ ⟨BLEU0_5⟩	European pilots are required to fill in forms asking for mental illness and physical illness.
⟨SIM95⟩ ⟨BLEU40⟩	European pilots must complete forms asking for mental and physical illnesses.
⟨SIM70⟩ ⟨BLEU0_5⟩	European pilots are required to fill in forms asking for mental health and physical health information, which can be submitted to the European
⟨SIM70⟩ ⟨BLEU40⟩	Pilots must fill out forms asking for mental and physical health information.

Source: Newcastle were denied a penalty against Liverpool on Monday night .

Tags	Generated paraphrases
⟨SIM95⟩ ⟨BLEU0_5⟩	Newcastle were not awarded a penalty in Monday night's game against Liverpool.
⟨SIM95⟩ ⟨BLEU40⟩	Newcastle were denied a penalty in Monday night's game against Liverpool.
⟨SIM70⟩ ⟨BLEU0_5⟩	Newcastle were not awarded a penalty in Monday night's 3-0 win over Liverpool.
⟨SIM70⟩ ⟨BLEU40⟩	Newcastle were denied a penalty in the 3-1 win over Liverpool on Monday night.

Table 5: Paraphrases of various source inputs generated by our model

Source: Maria Sharapova has been forced to withdraw with a leg injury .	
Tags	Generated paraphrases
⟨SIM95⟩ ⟨BLEU0_5⟩	Maria Sharapova withdrew with an injury to her leg.
⟨SIM95⟩ ⟨BLEU10⟩	Maria Sharapova was forced to pull out due to a leg injury.
⟨SIM95⟩ ⟨BLEU25⟩	Maria Sharapova had to withdraw due to a leg injury.
⟨SIM95⟩ ⟨BLEU40⟩	Maria Sharapova had to pull out with a leg injury.
⟨SIM90⟩ ⟨BLEU0_5⟩	Maria Sharapova withdrew from the tournament with an injury to her leg.
⟨SIM90⟩ ⟨BLEU10⟩	Maria Sharapova had to pull out of the tournament due to a leg injury.
⟨SIM90⟩ ⟨BLEU25⟩	Maria Sharapova had to pull out of the tournament with a leg injury.
⟨SIM90⟩ ⟨BLEU40⟩	Maria Sharapova has been forced to pull out of the tournament due to a leg injury.
⟨SIM80⟩ ⟨BLEU0_5⟩	Maria Sharapova withdrew from the tournament with an injury to her leg.
⟨SIM80⟩ ⟨BLEU10⟩	Maria Sharapova was forced to pull out of the tournament due to a leg injury.
⟨SIM80⟩ ⟨BLEU25⟩	Maria Sharapova was forced to pull out of the tournament with a leg injury.
⟨SIM80⟩ ⟨BLEU40⟩	Maria Sharapova was forced to retire with a leg injury.
⟨SIM70⟩ ⟨BLEU0_5⟩	Maria Sharapova withdrew from the tournament with an ankle injury.
⟨SIM70⟩ ⟨BLEU10⟩	Maria Sharapova was forced to pull out of the French Open due to a leg injury.
⟨SIM70⟩ ⟨BLEU25⟩	Maria Sharapova has been forced to pull out of the tournament due to a thigh injury.
⟨SIM70⟩ ⟨BLEU40⟩	Maria Sharapova has been forced to pull out of the French Open with a leg injury.
Source: Nike have launched the new World Cup uniforms for the US women's team .	
Tags	Generated paraphrases
⟨SIM95⟩ ⟨BLEU0_5⟩	Nike has unveiled new uniforms for the U.S. women's team at the World Cup.
⟨SIM95⟩ ⟨BLEU10⟩	Nike has unveiled the new U.S. women's World Cup uniforms.
⟨SIM95⟩ ⟨BLEU25⟩	Nike has launched new World Cup uniforms for the U.S. women's team.
⟨SIM95⟩ ⟨BLEU40⟩	Nike has unveiled the new World Cup uniforms for the U.S. women's team.
⟨SIM90⟩ ⟨BLEU0_5⟩	Nike has unveiled the new uniforms for the women's World Cup.
⟨SIM90⟩ ⟨BLEU10⟩	Nike has unveiled the new uniforms for the U.S. women's World Cup.
⟨SIM90⟩ ⟨BLEU25⟩	Nike has unveiled the new uniforms for the US women's national team at the World Cup.
⟨SIM90⟩ ⟨BLEU40⟩	Nike has unveiled the new World Cup uniforms for the U.S. women's team.
⟨SIM80⟩ ⟨BLEU0_5⟩	Nike has launched new uniforms for the women's World Cup in the United States.
⟨SIM80⟩ ⟨BLEU10⟩	Nike has unveiled the new uniforms for the U.S. women's national team.
⟨SIM80⟩ ⟨BLEU25⟩	Nike has launched the new uniforms for the women's World Cup.
⟨SIM80⟩ ⟨BLEU40⟩	Nike has launched the new World Cup uniforms for the U.S. women's national team.
⟨SIM70⟩ ⟨BLEU0_5⟩	New U.S. women's World Cup uniforms have been unveiled by Nike.
⟨SIM70⟩ ⟨BLEU10⟩	Nike has unveiled the new uniforms for the women's World Cup in Rio de Janeiro.
⟨SIM70⟩ ⟨BLEU25⟩	Nike has launched the new World Cup uniforms for the women's team of the United States, which will compete in the World Cup.
⟨SIM70⟩ ⟨BLEU40⟩	Nike has launched the new World Cup uniforms for the US women's team, which will be available in the coming weeks.

Table 6: Paraphrases with various tags' combinations by our model

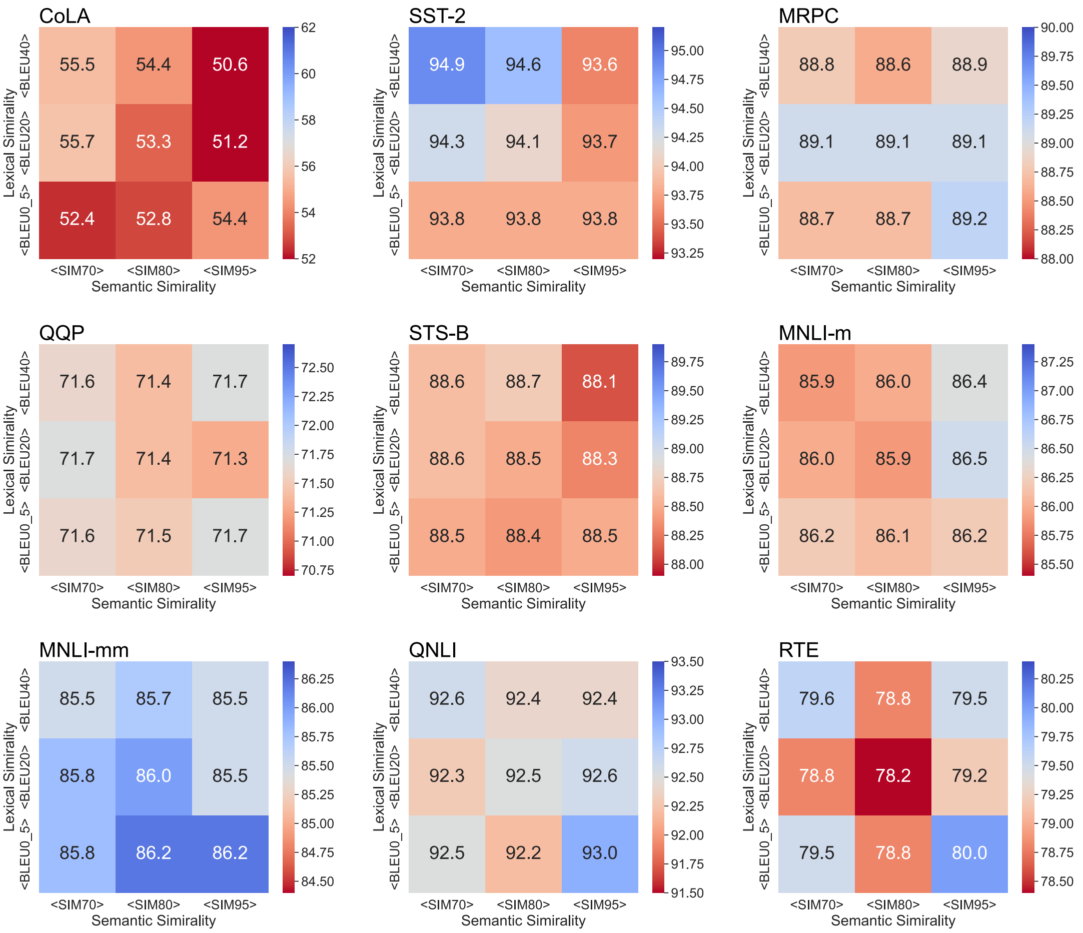


Figure 6: Heatmaps of the performance of our paraphrase generation model on all tasks, compared to STILTs using the original MNLI corpus