Paraphrasing as Zero-shot Translation with Feature-guided Diversity Enhancement

Ziyue Yan[1,2], Hongying Zan[1], Xinglin Lyu[1], Hongfei Xu[1]
[1]Zhengzhou University, Henan, China
[2]Dawning Information Industry Co.,Ltd., Henan, China
zyyannlp@foxmail.com, {iehyzan, xllyu}@zzu.edu.cn
Correspondence: Hongfei Xu hfxunlp@foxmail.com

Abstract

Paraphrasing uses different words, sentence structures, or expressions to convey similar semantics. It is an effective training data augmentation method to improve low-resource Natural Language Processing (NLP) tasks. Existing studies normally leverage parallel corpora to construct parabanks, regarding the Machine Translation (MT) results of source sentences as the paraphrases of the corresponding target sentences. As MT models are usually trained on the same parallel corpus, translation of the training set may suffer from overfitting, which leads to less diverse paraphrases. Training paraphrasers on the parabank generated via MT may also suffer from the information loss issue, as the parabank is derived from the parallel corpora, and the knowledge inside the parabank is a subset of that inside the parallel corpora. In this paper, we train bidirectional Multilingual Neural Machine Translation (MNMT) on the bi-directional bilingual parallel corpus, and use the MNMT model directly as a paraphrasing model by asking it to generate "translations" of the input language. As some source tokens also appear in the translation in the parallel corpus, we introduce "copy"/"not-copy" tags to indicate the existence/non-existence of source tokens in the target translation during training, and use the "not-copy" tag to encourage paraphrasing during inference. Manual and automatic evaluation results show that our ParaMNMT method can generate paraphrases of higher semantic consistency, literal fluency and sentential diversity compared to existing parabanks and LLMs. Our data augmentation experiments verify the effectiveness of ParaMNMT on improving low-resource NLP tasks.

1 Introduction

Paraphrase restates the same meaning with different expressions (Bhagat and Hovy, 2013). Paraphrasing is an effective data augmentation method for many Natural Language Processing (NLP) tasks, such as low-resource Machine Translation (MT) (Khayrallah et al., 2020), automatic MT evaluation (Thompson and Post, 2020a), creative generation (Tian et al., 2021), improving model robustness (Huang and Chang, 2021), among others.

The training of high-quality neural paraphrasers relies on a large amount of paraphrases (Li et al., 2019; Kumar et al., 2020). As manual annotations are expensive and difficult to scale (Madnani et al., 2012; Iyer et al., 2017b), most studies explore the possibility of automatic paraphrase generation. One common choice is back-translation (Wieting and Gimpel, 2017; Hu et al., 2019a,b), which generates paraphrases of target sentences by translating the corresponding source sentences in the parallel corpora. As MT models are usually trained on the same parallel corpus, translations of the training set may suffer from overfitting and are less diverse. The training of paraphrasers based on the machine translated parabank may also suffer from the information loss issue, as the parabank is derived from the parallel corpora, and training on the parabank may only gain a subset of the knowledge in the parallel corpora.

In this paper, we leverage the bidirectional Multilingual Neural Machine translation (MNMT) model trained on the bilingual parallel data and to facilitate paraphrasing via zero-shot translation (Thompson and Post, 2020b). Specifically, we add the language token at the beginning of sentences to indicate the translation direction during MNMT modeling, so the model can generate translations under the instruction of the language token. During paraphrasing, we use the language token of the input source sentence and ask the model to "translate" into the same language. To encourage the model using diverse expressions, we propose to assign each token with a feature tag for its existence in the target translation during MNMT training, "copy" for its existence in the target translation, and "not-copy" for the opposite. Despite that the source and

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 17211–17223

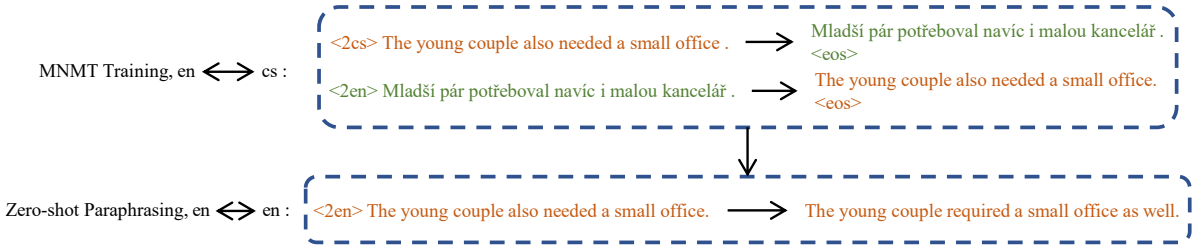

Figure 1: MNMT training and zero-shot paraphrasing.

target sides are in different languages in the parallel corpora, they still share some tokens especially when shared byte-pair-encoding is applied. During paraphrasing, we assign high-frequency tokens with the "not-copy" tag as instruction, and increase the paraphrase diversity by assigning more tokens with the "not-copy" tag.

Our main contributions are as follows:

- We facilitate paraphrasing by zero-shot translation with MNMT models, using the source language token to ask the model to "translate" into the input language. Compared to training the paraphraser on the parabank constructed by pairing the machine translations of source sentences with the corresponding target sentences of the parallel corpora, our single-stage method does not suffer from information loss.
- We assign "copy" and "not-copy" feature tags to each token, indicating its existence in the target sentence, and enhance the paraphrasing diversity using more "not-copy" tags.
- Manual and automatic evaluations show that our method generates paraphrases of higher semantic consistency, literal fluency, and sentential diversity than existing parabanks.
- Paraphrasing with our method leads to significant improvements on all low-resource GLUE tasks, showing the effectiveness of data augmentation with our paraphrasing approach.

2 Our Method

2.1 Paraphrasing based on MNMT

Regular paraphrase generation usually uses the translation model to translate the source sentence into the target language, then pairs the translation with the corresponding target reference to create a parabank using parallel corpus. The paraphraser is obtained by training a paraphrase model on the parabank. The multi-stage method may suffer from information loss for: 1) the MT model trained on the same parallel corpus is likely to generate translations similar to the reference due to overfitting, and 2) the training of the paraphraser only utilizes the parabank, which only contains a subset of the knowledge of the parallel corpus.

To avoid potential information loss of the multi-stage method, we employ the bidirectional Multilingual Neural Machine Translation (MNNT) model (Firat et al., 2016) directly as a paraphraser. We add the target language token at the beginning of sentences during MNMT training, and paraphrasing is accomplished by zero-shot "translating" the input sentence to the same language using the source language token. As shown in Figure 1.

2.2 Diversity Control with Copy Tags

Zero-shot translation may lead to outputs similar to the inputs, we enhance the diversity of "translation" with "copy" and "not-copy" feature tags, as shown in Figure 2. During MNMT training, we check the existence of source tokens in the target translation. Despite that the source and target sentences are in different languages in the parallel corpora, they may still share some tokens, especially when shared byte-pair-encoding is applied. We assign the "copy" tag to source tokens which also appear in the target sentences, and the "not-copy" tag to the other tokens. In the MNMT modeling, we retrieve embeddings for the "copy" or "not-copy" tags at each position like embedding lookup for tokens, and add the feature embeddings to token embeddings together with positional embeddings. The model is encouraged to re-generate the input tokens with the "copy" tag and to transform the input tokens with the "not-copy" tag during training. In our train data (Section 2.3), 18.7% of source tokens

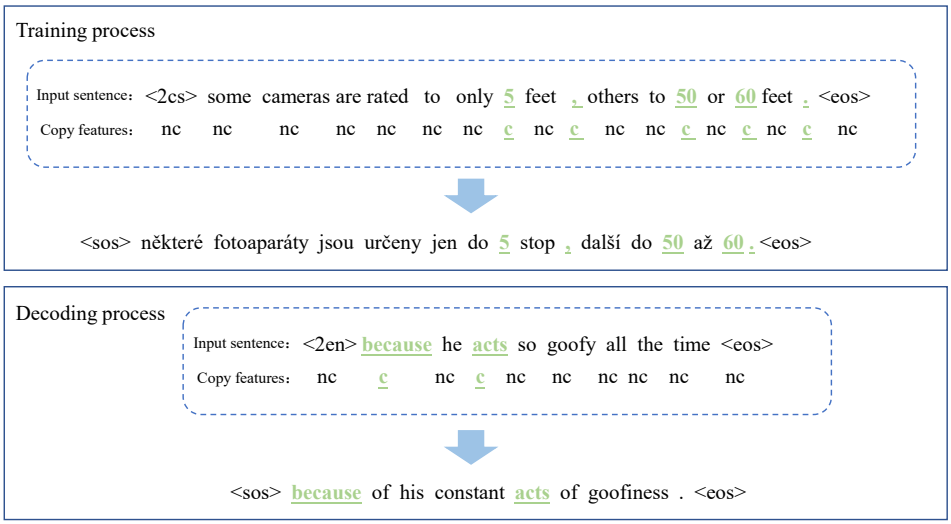

Figure 2: ParaMNMT with copy features. "c" and "nc" stand for the "copy" and the "not-copy" tags respectively.

are assigned with the "copy" tag.

During paraphrasing, we increase the diversity by assigning more tokens with the "not-copy" tag, asking the model to transform these tokens. We investigate two strategies for the choosing of "not-copy" tokens: 1) high-frequency. We assign the "not-copy" tag to tokens of higher frequencies and the "copy" tag to the other tokens, because high-frequency words are more likely to be function words which can often be replaced via syntactic paraphrasing. 2) high-entropy. We mask individual word each at a time and feed the masked sentence into BERT to get the prediction probability distribution of the masked word, we calculate the entropy of the masked word based on the probability distribution and assign the "not-copy" tag to the tokens of higher entropy, because the model is more uncertain for the prediction of the high-entropy words, which may also be easier to be replaced by the other words.

In practice, most of the time the model can generate a legal sentence according to the given copy/not-copy tags by adjusting the sentence structure or using words of similar meanings during inference. In a few cases (7.17%), the model may ignore the "not-copy" tag and use the original tokens.

2.3 Training of MNMT Paraphraser

For fair comparison, we implemented our method on the English-Czech parallel corpora following previous work (Hu et al., 2019a,b). We used the English-Czech parallel corpus of the WMT 2016 news translation task (Bojar et al., 2016a) for our experiments, containing Europarl v7 (Koehn, 2005), Common Crawl, News Commentary v11 and CzEng 1.6pre (Bojar et al., 2016b).

We swapped the source sentences and target translations of the original parallel corpus to obtain the parallel corpus for the reverse translation direction. We added the language tokens of the target language at the beginning of each sentence of both the original parallel corpus and the reversed parallel corpus, and concatenated them as the bidirectional training data for MNMT. Byte Pair Encoding (BPE) (Sennrich et al., 2016) was applied to address the unknown word issue.

As the source and target sides of the parallel corpus share some tokens especially when shared byte-pair-encoding is applied despite in different languages, we generate "copy" or "not-copy" tags for each token in the source sentence by checking its existence in the corresponding target translation. We train the MNMT model to generate the target translation given the source sentence and the corresponding copy feature sequence as input.

We used the Transformer Base (Vaswani et al., 2017) model of 6 encoder and decoder layers and an embedding dimension of 512 for MNMT modeling, and augmented the model with Language-Aware Layer Normalization (LALN) and Language-Aware Linear Transformation (LALT) for improved performance (Zhang et al., 2020).

2.4 Paraphrasing

During paraphrasing, we assign 30% of tokens in the sentence with the "not-copy" tag, and assign the other tokens with the "copy" tag. The token

Score	Semantic consistency	Literal fluency	Sentential diversity
5	Sentences have exactly the same meaning with all the same details.	The sentence pair has no grammatical error.	The sentences have more than one grammatical variation or more than two lexical variations.
4	Sentences are mostly equivalent, but some unimportant details can differ.	The sentence pair has one grammatical error.	The sentences have slight grammatical variation.
3	Sentences are roughly equivalent, with some important information missing or that differs slightly.	The sentence pair has two grammatical errors.	The sentences have unchanged grammatical structure but two lexical variations.
2	Sentences are not equivalent, even if they share slight details.	The sentence pair has three grammatical errors.	The sentences have unchanged grammatical structure but one lexical variation.
1	Sentences are totally different.	The sentence pair has more than three grammatical errors.	The sentence pair has basically unchanged grammatical structure and lexical variation.

Table 1: Manual evaluation criteria of semantic consistency, literal fluency, and sentential diversity.

frequencies were counted on the training data. In our preliminary experiments with automatic quality evaluation metrics (Section 3.2), this setting lead to a good trade-off between semantic consistency and sentential diversity.

The ParaMNMT model can paraphrase both English and Czech sentences, but we only evaluate in English, for most previous studies and evaluations are in English, and applying our method in English is friendly for comparison with previous work.

As the vocabulary of the ParaMNMT model contains tokens for all supported languages, we may prune the tokens which do not appear in the target language sentences in the parallel corpus for efficient language-specific paraphrasing. For English-only paraphrasing, we remove the corresponding rows for tokens that only appear in Czech sentences from the embedding and classifier weight matrices and the scalars in classifier bias vector. This reduces the vocabulary size by 40% and accelerates the decoding speed by 12%.

3 Quality Evaluation of ParaMNMT

We evaluated the quality of paraphrases in terms of semantic consistency, sentential diversity, and literal fluency. We compared our ParaMNMT method with existing parabanks like: ParaBank (Hu et al., 2019a), ParaBank2 (Hu et al., 2019b), ParaAMR (Huang et al., 2023), and Large Language Models (LLMs) including: LLaMa 3 8B (Dubey et al., 2024), Qwen 2.5 7B (Team, 2024) for their impressive capability (Chen et al., 2025; Zuo et al., 2025). The basic prompt for LLM paraphrasing is: *Rephrase the input, do not generate the other contents apart from the paraphrase. Input: $input_sentence*

For LLMs, we also tested the performance with chain-of-thought with Qwen 3 models (Team, 2025), a larger model (Qwen 3 14B), and a diversity-encouraging prompt (*Rephrase the input with diverse rewording, do not generate the other contents apart from the paraphrase. Input: $input_sentence*) with automatic metrics.

For LLM inference, we used the default generation parameters released together with the model files. Because we thought these generation parameters were tuned by the model authors and were suggested values. Tuning the decoding parameters such as repetition penalty or temperature may benefit the diversity, but this may require significant effort and our method can also tune these generation parameters. We suppose that tuning generation parameters may not be among the main concerns of our work.

3.1 Manual Evaluation

Manual evaluation was performed according to the criteria of Wieting and Gimpel (2017); Wang et al. (2021); Hao et al. (2022), as shown in Table 1.

We randomly sampled 800 sentence pairs for the evaluation of parabanks. For the evaluation our method and LLMs, we used the paraphraser/LLMs

	Scores	=5.0	≥4.0	≥3.0	≥2.0	Average Score
Semantic consistency	ParaBank (Hu et al., 2019a)	58.4	89.7	93.9	96.6	4.38±0.76
	ParaBank2 (Hu et al., 2019b)	69.8	93.3	95.6	98.1	4.57±0.76
	ParaAMR (Huang et al., 2023)	71.7	93.8	96.1	98.7	4.60±0.59
	Qwen 2.5 7B (Team, 2024)	53.2	61.5	74.1	89.6	3.78±0.36
	LLaMa 3 8B (Dubey et al., 2024)	55.6	64.4	78.3	91.2	3.89±0.37
	ParaMNMT (+ Copy features)	80.5	97.3	98.7	99.6	**4.76 ± 0.56**
Literal fluency	ParaBank	79.6	97.4	98.7	100.0	4.76±0.67
	ParaBank2	80.4	97.9	99.3	100.0	4.78±0.66
	ParaAMR	81.0	98.1	99.1	100.0	4.78±0.24
	Qwen 2.5 7B	78.7	89.8	97.5	100.0	4.66±0.48
	LLaMa 3 8B	79.9	90.2	99.2	100.0	4.69±0.46
	ParaMNMT (+ Copy features)	91.1	98.6	99.8	100.0	**4.89 ± 0.22**
Sentential diversity	ParaBank	62.9	70.3	81.4	91.5	4.06±2.06
	ParaBank2	69.4	79.6	88.1	97.2	4.34±1.13
	ParaAMR	70.2	80.1	87.9	97.1	4.35±1.07
	Qwen 2.5 7B	60.3	70.8	89.2	93.4	4.12±0.40
	LLaMa 3 8B	64.3	72.4	89.6	96.5	4.23±0.41
	ParaMNMT (+ Copy features)	79.8	87.5	95.2	97.9	**4.61 ± 0.89**

Table 2: Manual evaluation results.

to paraphrase the same source sentences sampled for the evaluation of ParaAMR. We hired six language experts to score all the paraphrase pairs. Each paraphrase pair was assigned to two randomly selected experts and the Inter-Annotator-Agreement (IAA) for evaluation was 0.97 (Cohen's Kappa). The average score and the standard deviation are shown in Table 2. Table 2 shows that our method can generate paraphrases of better semantic consistency, literal fluency, and sentential diversity compared to existing parabanks and LLMs.

We provide qualitative paraphrase examples generated by different datasets, using the same source sentence as Huang et al. (2023). Table 3 shows that our ParaMNMT method with the guidance of copy feature tags can generate fluent and more diverse paraphrases while preserving the semantic meaning. Although LLMs are expected to be powerful at language modeling, We observed the lack of word diversity or semantic meaning changes in LLMs outputs. As also revealed by Wang et al. (2025b), which shows that LLM paraphrasing produces stable periodic states, such as 2-period attractor cycles, limiting linguistic diversity.

3.2 Automatic Evaluation

Due to the high cost and subjectivity of manual evaluation, we also evaluated with automatic metrics. For reproduction, we took 2000 paraphrase pairs from ParaBank (Hu et al., 2019a), ParaBank2 (Hu et al., 2019b), and ParaAMR (Huang et al., 2023) for evaluation. We used our paraphraser and LLMs to paraphrase source sentences of ParaAMR for their evaluation. We also tested the performance of constrained decoding (masking the prediction probabilities of selected not-copy tokens to zero without using copy features) and using randomly selected not-copy tokens. We used Self-BLEU (Zhu et al., 2018) to evaluate the sentential diversity. Semantic consistency was evaluated via BertScore (Zhang et al., 2019) based on vector similarity. We used the GPT-2 Perplexity (PPL) to evaluate the literal fluency. Results are shown in Table 4.

Table 4 shows that: 1) ParaMNMT without copy features can already lead to better semantic consistency and comparable literal fluency, obtaining a higher BertScore and a low PPL compared to the other parabanks and LLMs, 2) our copy feature method can effectively increase the diversity of vanilla ParaMNMT, 3) constrained decoding can obtain slightly lower self-BLEU score than assigning not-copy tags to high-frequency tokens, but the high frequency setting can obtain higher BertScore and lower PPL, and 4) the high frequency setting performs better than the random and the high entropy settings on all metrics. We assign the "not-copy" tags to high-frequency tokens by default in the other experiments based on the results.

4 Evaluation on Low-Resource Tasks

We test the effectiveness of data augmentation by paraphrasing the original English sentences of the training set on low-resource NLP tasks with various parabanks/paraphrasers/LLMs. ParaBank2 (Hu et al., 2019b) used a much larger Transformer (12 layers and an embedding dimension of 768)

Source Sentence	I know for them to approve this price, they'll need statistical documentation.
ParaBank	I know that to accept this prize, they're going to need statistical analysis. I know that in order to accept this prize, they're going to need a statistic analysis. I know that if they accept this prize, they're gonna need a statistical analysis.
ParaBank2	I know that to accept that prize, they're going to need a statistical analysis. I know that in order to accept this prize, they will require a statistical analysis. I know they'll require statistical analysis to accept that prize.
ParaAMR	I know they need statistical documentation to approve this price. There is statistic documentation I know they need to approve these prices. They need statistical documentation to approve these prices, I know.
Qwen 2.5 7B	I understand that they will require statistical documentation for them to approve this price. I understand that they will require statistical evidence for them to endorse this price. I know they'll need statistical documentation for them to approve this price.
LLaMa 3 8B	To get approval for the price, they will likely require statistical evidence. To secure their approval of this price, I'll need to provide them with detailed statistical data. To obtain approval for this price, they will require statistical data and documentation.
ParaMNMT	I realize that they'll need statistical evidence to approve this price. I recognize that approval of this price will require them to have statistical documentation. Their approval of this price hinges on having the necessary statistical documentation, I know.

Table 3: Paraphrase examples.

ID	Methods	BertScore F1↑	Self-BLEU↓	PPL↓
1	ParaBank (Hu et al., 2019a)	80.6	26.80	40.29
2	ParaBank2 (Hu et al., 2019b)	71.5	20.07	6.90
3	ParaAMR (Huang et al., 2023)	69.8	22.74	1.75
4	LLaMa 3 8B (Dubey et al., 2024)	71.3	23.75	7.92
5	Qwen 2.5 7B (Team, 2024)	80.8	30.12	10.26
6	Qwen 3 8B (CoT)	82.3	29.08	9.03
7	Qwen 3 14B (CoT, larger)	85.1	24.12	9.46
8	Qwen 3 14B (CoT, larger, diverse)	84.5	25.35	9.89
9	ParaMNMT	91.2	22.31	1.79
10	9 + Constrained decoding	87.4	**16.81**	4.51
11	9 + Copy features (random)	89.5	17.31	3.28
12	9 + Copy features (high entropy)	93.1	19.23	1.54
13	9 + Copy features (high frequency)	**94.0**	16.98	**1.43**

Table 4: Automatic evaluation results.

than ours (6 layers and an embedding dimension of 512).

4.1 GLUE

We tested the effectiveness of paraphrasing on the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2019), involving nine tasks: Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2019), Stanford Sentiment Treebank (SST-2) (Socher et al., 2013), Microsoft Research Paraphrase Corpus (MRPC) (Dolan and Brockett, 2005), Quora Question Pairs (QQP) (Iyer et al., 2017a), Question-answering NLI (QNLI) (Rajpurkar et al., 2016), Recognizing Textual Entailment (RTE) (Dagan et al., 2005), Winograd NLI (WNLI) (Levesque et al., 2012), Semantic Textual Similarity Benchmark (STS-B) (Cera et al.) and Multi-Genre Natural Language Inference (MNLI) (Williams et al., 2018).

For data augmentation on sentence pair tasks, we explored four settings: S1. Paraphrasing only the first sentence; S2. Paraphrasing only the second sentence; S12. paraphrasing both sentences; and S1+S2. a combination of S1 and S2.

We tested the effectiveness over the BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) baselines implemented in the huggingface transformers library (Wolf et al., 2020) using the base setting. We evaluated the performances of fine-tuning on the original training set and the training set augmented by paraphrasing. The results of ParaTag, Taboo and DoAug are from Wang et al.

Data Augmentation over BERT		CoLA	SST-2	MRPC	QQP	QNLI	RTE	WNLI	STS-B	MNLI
None (BERT Baseline)		56.5	92.3	84.1	90.7	90.7	65.7	56.3	88.6	83.9
ParaTag (Wang et al., 2022)		-	92.3	-	90.6	90.5	65.0	46.5	88.8	84.1
ParaBank2 (Hu et al., 2019b)		56.9	92.5	-	-	91.0	70.1	59.5	-	83.9
Taboo (Cegin et al., 2024)		-	86.7	76.7	-	-	58.0	-	-	57.3
DoAug (Wang et al., 2025a)		-	88.6	81.0	-	-	56.1	-	-	59.8
LLaMa 3 8B	S1	56.7	92.3	84.9	89.5	90.4	65.6	56.3	88.4	83.1
	S2			84.3	89.6	90.5	65.7	56.4	88.2	83.5
	S12			84.4	90.5	90.8	65.9	56.4	88.5	83.7
	S1+S2			84.8	90.7	90.8	65.9	56.5	88.6	83.9
Qwen 2.5 7B	S1	56.2	92.3	84.0	89.5	90.4	65.2	56.3	88.1	83.1
	S2			84.3	89.8	90.5	65.2	56.3	88.2	83.3
	S12			84.7	90.4	90.6	65.3	56.4	88.4	83.9
	S1+S2			84.6	90.6	90.7	65.4	56.4	88.5	84.0
ParaMNMT	S1	59.1	92.9	89.1	91.1	91.1	67.9	59.9	89.3	84.2
	S2			89.0	91.1	91.2	68.7	59.6	89.0	84.2
	S12			88.9	91.3	91.3	67.7	60.0	89.4	84.3
	S1+S2			89.2	91.5	91.9	72.2	62.0	89.6	84.3
+ Copy features	S1	**60.4**	**93.2**	89.3	91.3	91.6	69.0	60.1	89.6	84.1
	S2			89.0	91.3	91.5	69.7	60.0	89.4	84.3
	S12			89.3	91.4	91.7	69.5	60.6	89.5	84.4
	S1+S2			**89.7**	**91.6**	**92.1**	**72.5**	**62.0**	**90.0**	**84.6**

Table 5: Results on GLUE tasks with BERT.

Data Augmentation over RoBERTa		CoLA	SST-2	MRPC	QQP	QNLI	RTE	WNLI	STS-B	MNLI
None (RoBERTa Baseline)		63.6	94.8	90.2	91.9	92.8	78.7	61.6	91.2	87.6
LLaMa 3 8B	S1	63.6	94.5	90.1	91.2	92.6	78.8	61.3	91.1	87.5
	S2			90.1	91.6	92.7	79.1	61.4	91.2	87.4
	S12			90.2	92.1	93.1	79.2	61.6	91.3	87.8
	S1+S2			90.5	92.2	93.4	79.3	61.8	91.6	88.0
Qwen 2.5 7B	S1	63.4	94.2	89.9	90.8	92.5	78.7	61.1	91.0	87.2
	S2			89.8	91.3	92.5	78.8	61.2	91.1	87.4
	S12			90.1	91.6	92.8	79.1	61.7	91.3	87.5
	S1+S2			90.2	92.0	93.1	79.1	61.7	91.4	87.9
ParaMNMT	S1	64.4	95.2	90.6	92.4	93.0	78.8	61.9	91.5	87.9
	S2			90.4	92.1	92.9	79.0	61.7	91.4	88.2
	S12			90.9	92.5	93.2	79.1	62.1	91.6	88.7
	S1+S2			91.1	92.8	93.8	79.4	62.4	91.8	88.9
+ Copy features	S1	**65.8**	**96.7**	90.9	92.5	93.0	79.2	62.1	91.7	88.1
	S2			90.7	92.4	93.1	79.3	61.8	91.5	88.0
	S12			91.1	92.9	93.4	79.6	62.5	91.8	88.8
	S1+S2			**91.5**	**93.1**	**93.8**	**80.0**	**62.8**	**92.1**	**89.2**

Table 6: Results on GLUE tasks with RoBERTa.

(2022, 2025a), for they are not publicly available. Taboo and DoAug only used $1.2K$ training samples, leading to lower scores.

Following common practice, we evaluated the CoLA task by Matthews Correlation Coefficient, the STS-B task by Pearson Correlation Coefficient, the MNLI task by Matched accuracy, and the others by accuracy.

Results in Tables 5 and 6 show that: 1) S1+S2 generally lead to better performance than S1, S2 and S12 for sentence pair tasks, possibly because the S1+S2 setting augments the training set with more data, 2) our ParaMNMT method without copy features already leads to better data augmentation performance than both LLMs and existing parabanks, and 3) using copy features can bring about further improvements over the ParaMNMT method without copy features, probably due to the improved semantic consistency and diversity of paraphrases.

Methods	En→De	En→Tr	En→Fi
Transformer	27.46	15.79	22.23
ParaMNMT	27.89†	16.49‡	22.61†
+ Copy features	**28.24**‡	**16.83**‡	**22.95**‡

Table 7: Results on low-resource translation tasks. † and ‡ indicate $p < 0.1$ and $p < 0.01$ respectively in the significance test.

4.2 Machine Translation

We also tested the performance of ParaMNMT on the WMT 17 English (En) to Finnish (Fi) and Turkish (Tr) tasks, and the IWLST 14 English (En) to German (De) task. The English to Finnish, Turkish, and German tasks contain 2.6M, 0.2M, and 0.17M sentence pairs for training respectively. We adopted the 6-layer Transformer Base setting (Vaswani et al., 2017) as our baseline. We augmented the training set by paraphrasing the source English sentences and pairing with the corresponding target translations. We used a beam size of 4 for decoding with the averaged model of the last 5 checkpoints saved with an interval of $1,500$ training steps, and evaluated tokenized case-sensitive BLEU. Results are shown in Table 7.

Table 7 shows that: 1) paraphrasing without copy features can already significantly improve the performances of the low-resource translation tasks, and 2) using copy features can lead to further BLEU improvements over ParaMNMT without copy features.

4.3 Experiments over LLM Baseline

We tested the effectiveness of data augmentation with our ParaMNMT method (+ copy features) over the stronger Qwen 3 8B LLM baseline on the GLUE tasks and the translation tasks. We did not test on the STS-B task for it is a regression task. We fine-tuned Qwen 3 8B with a LoRA (Hu et al., 2022) rank of 32 for GLEU tasks and a LoRA rank of 256 for machine translation tasks. For sentence pair GLUE tasks, we used the S1+S2 setting for our method. For machine translation, we also tested the performance of the Qwen 3 8B without fine-tuning (Zhang et al., 2023). The translation prompt is: *Translate the input into [German|Turkish|Finnish], do not generate the other contents apart from the translation. Input: $input_sentence*

Results in Tables 8 and 9 validate the effectiveness of our method even with the stronger Qwen 3 8B LLM baseline.

5 Related Work

5.1 Paraphrasing

Initial paraphrasing methods are usually based on hand-crafted rules, including rule-based methods (McKeown, 1983; Zong et al., 2001), thesaurus-based methods (Kauchak and Barzilay, 2006), and lattice matching methods (Barzilay and Lee, 2003). Tiedemann and Scherrer (2019) apply multilingual neural machine translation systems for paraphrase extraction. Thompson and Post (2020b) generate paraphrases using multilingual neural machine translation models but increase the diversity by discouraging the production of n-grams presented in the input. Sjöblom et al. (2020) use paraphrase sentence pairs to train an encoder-decoder based paraphrase generation model for six languages. Deng et al. (2023) use a unified bidirectional generative framework to tackle various cross-domain ABSA tasks. Zhang et al. (2024) propose Pseudo-Inconsistent Penalization and Prior Enhanced Decoding methods to improve language consistency. Wang et al. (2025b) reveal that successive paraphrasing converges to stable periodic states, limiting linguistic diversity. Wang et al. (2025a) propose a Diversity-oriented data Augmentation (DoAug) framework to train a large language model (LLM) as a diverse paraphraser.

Diversity and Quality remain two main challenges (Pan et al., 2024; Xu and Wang, 2024; Slobodkin et al., 2024; Şahinuç et al., 2024), Chung et al. (2023) show that creating high-quality datasets with LLMs can be challenging, and increasing the output diversity of LLMs is often at the cost of data accuracy. Cegin et al. (2024) show that using taboo words can effectively improve the diversity of LLM output, but paraphrasing with hints is more effective in improving the downstream models' performance.

5.2 Parabanks

The scale of early parabanks like MRPC (Dolan et al., 2004) and Quora (Ganitkevitch et al., 2013) is normally limited, but training neural models for paraphrasing requires large parabanks. The automatic generation of large-scale corpus parabanks has become a popular research topic. ParaNMT (Wieting and Gimpel, 2017) translates one side of a parallel corpus by machine translation, pairing it with the corresponding sentence to form a paraphrase pair. Hao et al. (2022) apply the similar idea to Chinese and construct a large-scale Chinese

Methods	CoLA	SST-2	MRPC	QQP	QNLI	RTE	WNLI	MNLI
Qwen 3 8B	63.4	95.6	88.3	89.7	94.7	90.3	79.6	90.9
+ Our method	**65.5**	**97.6**	**89.7**	**91.3**	**95.8**	**91.7**	**80.6**	**92.4**

Table 8: Results on GLUE tasks with Qwen3 8B.

Methods	En-De	En-Tr	En-Fi
Prompt	27.27	20.03	17.58
LoRA	30.49	22.61	18.95
+ Our method	**31.30**‡	**23.54**‡	**19.63**‡

Table 9: Results on low-resource translation tasks with Qwen3 8B.

parabank. Most studies explore decoding methods to improve the paraphrasing diversity, including lexical constraints (Hu et al., 2019a), cluster-based constrained sampling (Hu et al., 2019b), latent space sampling (Roy and Grangier, 2019; Cao and Wan, 2020), word order controlling (Goyal and Durrett, 2020) and syntax-based methods (Huang et al., 2022; Lee et al., 2022). Huang et al. (2023) create syntactically diverse paraphrase datasets using abstract meaning representation.

6 Conclusion

In this paper, we facilitate paraphrasing via zero-shot translation with MNMT models. Compared to constructing paraphrases by pairing the translation of source sentences with the corresponding target sentences of the parallel corpora and then training the paraphraser, our single-stage method avoids the information loss issue with the multi-stage method. In order to control the diversity of paraphrases, we assign "copy" and "not-copy" feature tags to each token during MNMT training, indicating their existence in the target translation, teaching the model to transform the tokens assigned with the "not-copy" feature tag. We encourage the model to generate diverse paraphrases with more "not-copy" tags.

Both manual and automatic evaluations show that our method leads to better paraphrases than existing parabanks in semantic consistency, literal fluency, and the copy feature method can effectively improve the sentential diversity. Data augmentation experiment results show that ParaMNMT can effectively improve the performance of both natural language understanding tasks (GLUE) and low-resource MT tasks.

Limitations

We only adopted a simple and small MNMT model for our study, scaling or adopting advanced MNMT models may lead to better performance.

Acknowledgments

We thank anonymous reviewers for their insightful comments. Ziyue Yan and Hongfei Xu are partially supported by the National Natural Science Foundation of China (Grant No. 62306284), China Postdoctoral Science Foundation (Grant No. 2023M743189), and the Natural Science Foundation of Henan Province (Grant No. 232300421386). Ziyue Yan and Hongying Zan acknowledge the support of the National Natural Science Foundation of China (Grant No. U23A20316).

References

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 16–23.

Rahul Bhagat and Eduard H. Hovy. 2013. Squibs: What is a paraphrase? *Computational Linguistics*, 39:463–472.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016a. Findings of the 2016 conference on machine translation (wmt16). In *First conference on machine translation*, pages 131–198. Association for Computational Linguistics.

Ondřej Bojar, Ondřej Dušek, Tom Kocmi, Jindřich Libovický, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. 2016b. CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing Tools Dockered. In *Text, Speech and Dialogue: 19th International Conference, TSD 2016, Brno, Czech Republic, September 12-16, 2016, Proceedings*. Springer Verlag. In press.

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.

Jan Cegin, Branislav Pecher, Jakub Simko, Ivan Srba, Maria Bielikova, and Peter Brusilovsky. 2024. Effects of diversity incentives on sample diversity and downstream model performance in LLM-based text augmentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13148–13171, Bangkok, Thailand. Association for Computational Linguistics.

Daniel Cera, Mona Diabb, Eneko Agirrec, Inigo Lopez-Gazpioc, Lucia Speciad, and Basque Country Donostia. Semeval-2017 task 1: Semantic textual similarity multilingual and cross-lingual focused evaluation.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2025. Benchmarking LLMs for translating classical Chinese poetry: Evaluating adequacy, fluency, and elegance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33019–33036, Suzhou, China. Association for Computational Linguistics.

John Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, Toronto, Canada. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2023. Bidirectional generative framework for cross-domain aspect-based sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12272–12285, Toronto, Canada. Association for Computational Linguistics.

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.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third international workshop on paraphrasing (IWP2005)*.

William B. Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *International Conference on Computational Linguistics*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *North American Chapter of the Association for Computational Linguistics*.

Tanya Goyal and Greg Durrett. 2020. Neural syntactic preordering for controlled paraphrase generation. In *Annual Meeting of the Association for Computational Linguistics*.

Wenjie Hao, Hongfei Xu, Deyi Xiong, Hongying Zan, and Lingling Mu. 2022. Parazh-22m: A large-scale chinese parabank via machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3885–3897.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

J. Edward Hu, Rachel Rudinger, Matt Post, and Benjamin Van Durme. 2019a. Parabank: Monolingual bitext generation and sentential paraphrasing via lexically-constrained neural machine translation. In *AAAI Conference on Artificial Intelligence*.

J. Edward Hu, Abhinav Singh, Nils Holzenberger, Matt Post, and Benjamin Van Durme. 2019b. Large-scale, diverse, paraphrastic bitexts via sampling and clustering. In *Conference on Computational Natural Language Learning*.

Kuan-Hao Huang and Kai-Wei Chang. 2021. Generating syntactically controlled paraphrases without using annotated parallel pairs. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1022–1033.

Kuan-Hao Huang, Varun Iyer, I-Hung Hsu, Anoop Kumar, Kai-Wei Chang, and Aram Galstyan. 2023. ParaAMR: A large-scale syntactically diverse paraphrase dataset by AMR back-translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8047–8061, Toronto, Canada. Association for Computational Linguistics.

Kuan-Hao Huang, Varun Iyer, Anoop Kumar, Sriram Venkatapathy, Kai-Wei Chang, and A. G. Galstyan. 2022. Unsupervised syntactically controlled paraphrase generation with abstract meaning representations. In *Conference on Empirical Methods in Natural Language Processing*.

Shankar Iyer, Nikhil Dandekar, Kornél Csernai, et al. 2017a. First quora dataset release: Question pairs. *data. quora. com.*

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. 2017b. First quora dataset release: Question pairs.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for automatic evaluation. In *North American Chapter of the Association for Computational Linguistics*.

Huda Khayrallah, Brian Thompson, Matt Post, and Philipp Koehn. 2020. Simulated multiple reference training improves low-resource machine translation. In *Conference on Empirical Methods in Natural Language Processing*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Machine Translation Summit*.

Ashutosh Kumar, Kabir Ahuja, Raghuram Vadapalli, and Partha Pratim Talukdar. 2020. Syntax-guided controlled generation of paraphrases. *Transactions of the Association for Computational Linguistics*, 8:330–345.

Fei-Tzin Lee, Miguel Ballesteros, Feng Nan, and Kathleen McKeown. 2022. Using structured content plans for fine-grained syntactic control in pretrained language model generation. In *International Conference on Computational Linguistics*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Zichao Li, Xin Jiang, Lifeng Shang, and Qun Liu. 2019. Decomposable neural paraphrase generation. In *Annual Meeting of the Association for Computational Linguistics*.

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 pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nitin Madnani, Joel R. Tetreault, and Martin Chodorow. 2012. Re-examining machine translation metrics for paraphrase identification. In *North American Chapter of the Association for Computational Linguistics*.

Kathleen McKeown. 1983. Paraphrasing questions using given and new information. *Am. J. Comput. Linguistics*, 9:1–10.

Xingyuan Pan, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, and Shanbo Cheng. 2024. G-DIG: Towards gradient-based DIverse and hiGh-quality instruction data selection for machine translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15395–15406, Bangkok, Thailand. Association for Computational Linguistics.

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

Aurko Roy and David Grangier. 2019. Unsupervised paraphrasing without translation. In *Annual Meeting of the Association for Computational Linguistics*.

Furkan Şahinuç, Ilia Kuznetsov, Yufang Hou, and Iryna Gurevych. 2024. Systematic task exploration with LLMs: A study in citation text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4832–4855, Bangkok, Thailand. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Eetu Sjöblom, Mathias Creutz, and Yves Scherrer. 2020. Paraphrase generation and evaluation on colloquial-style sentences. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1814–1822.

Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. 2024. Attribute first, then generate: Locally-attributable grounded text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3344, Bangkok, Thailand. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Brian Thompson and Matt Post. 2020a. 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.

Brian Thompson and Matt Post. 2020b. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation*, pages 561–570.

Yufei Tian, Arvind krishna Sridhar, and Nanyun Peng. 2021. Hypogen: Hyperbole generation with commonsense and counterfactual knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1583–1593.

Jörg Tiedemann and Yves Scherrer. 2019. Measuring semantic abstraction of multilingual nmt with paraphrase recognition and generation tasks. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 35–42.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*.

Shuohang Wang, Ruochen Xu, Yang Liu, Chenguang Zhu, and Michael Zeng. 2022. Paratag: A dataset of paraphrase tagging for fine-grained labels, nlg evaluation, and data augmentation. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 7111–7122.

Yasong Wang, Mingtong Liu, Yujie Zhang, Yufeng Chen, et al. 2021. Research on the construction and application of paraphrase parallel corpus. *Beijing Da Xue Xue Bao*, 57(1):68–74.

Zaitian Wang, Jinghan Zhang, Xinhao Zhang, Kunpeng Liu, Pengfei Wang, and Yuanchun Zhou. 2025a. Diversity-oriented data augmentation with large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22265–22283, Vienna, Austria. Association for Computational Linguistics.

Zhilin Wang, Yafu Li, Jianhao Yan, Yu Cheng, and Yue Zhang. 2025b. Unveiling attractor cycles in large language models: A dynamical systems view of successive paraphrasing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12740–12755, Vienna, Austria. Association for Computational Linguistics.

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

John Wieting and Kevin Gimpel. 2017. Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Annual Meeting of the Association for Computational Linguistics*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL-HLT*, pages 1112–1122.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ziyao Xu and Houfeng Wang. 2024. SPOR: A comprehensive and practical evaluation method for compositional generalization in data-to-text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 604–621, Bangkok, Thailand. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Liang Zhang, Qin Jin, Haoyang Huang, Dongdong Zhang, and Furu Wei. 2024. Respond in my language: Mitigating language inconsistency in response generation based on large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4177–4192, Bangkok, Thailand. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference*

on research & development in information retrieval, pages 1097–1100.

Chengqing Zong, Yujie Zhang, Kazuhide Yamamoto, Masashi Sakamoto, and Satoshi Shirai. 2001. Approach to spoken chinese paraphrasing based on feature extraction. In *Natural Language Processing Pacific Rim Symposium*.

Fei Zuo, Kehai Chen, Yu Zhang, Zhengshan Xue, and Min Zhang. 2025. InImageTrans: Multimodal LLM-based text image machine translation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20256–20277, Vienna, Austria. Association for Computational Linguistics.