ParaZh-22M: a Large-Scale Chinese Parabank via Machine Translation

Wenjie Hao[1], **Hongfei Xu**[1], **Deyi Xiong**[2], **Hongying Zan**[1,3], **Lingling Mu**[1*]
[1] Zhengzhou University, [2] Tianjin University, [3] Peng Cheng Laboratory
haowj9977@163.com, hfxunlp@foxmail.com, dyxiong@tju.edu.cn
{iehyzan,iellmu}@zzu.edu.cn

Abstract

Paraphrasing, i.e., restating the same meaning in different ways, is an important data augmentation approach for natural language processing (NLP). Zhang et al. (2019b) propose to extract sentence-level paraphrases from multiple Chinese translations of the same source texts, and construct the PKU Paraphrase Bank of 0.5M sentence pairs. However, despite being the largest Chinese parabank to date, the size of PKU parabank is limited by the availability of one-to-many sentence translation data, and cannot well support the training of large Chinese paraphrasers. In this paper, we relieve the restriction with one-to-many sentence translation data, and construct ParaZh-22M, a larger Chinese parabank that is composed of 22M sentence pairs, based on one-to-one bilingual sentence translation data and machine translation (MT). In our data augmentation experiments, we show that paraphrasing based on ParaZh-22M can bring about consistent and significant improvements over several strong baselines on a wide range of Chinese NLP tasks, including a number of Chinese natural language understanding benchmarks (CLUE) and low-resource machine translation. [1]

1 Introduction

A paraphrase is a restatement of meaning with different expressions (Bhagat and Hovy, 2013). Paraphrasing has been proven to be an effective data augmentation approach for many NLP tasks, ranging from linguistically controlled paraphrase generation (Iyyer et al., 2018; Chen et al., 2019; Li et al., 2019; Sun et al., 2021), style transfer (Krishna et al., 2020), to applications like low-resource machine translation (Khayrallah et al., 2020) and automatic MT evaluation (Thompson and Post, 2020; Bawden et al., 2020).

Zhang et al. (2019b) extract sentence-level paraphrases from multiple Chinese translations of the same source texts, and create the largest Chinese paraphrase bank (PKU Parabank) to date, which contains 509,832 pairs of paraphrased sentences. However, the amount of one-to-many sentence translation data constrains the size of PKU parabank, and it cannot meet the requirement to train large Chinese paraphrasers.

Inspired by Wieting and Gimpel (2018) and Hu et al. (2019a,b), we propose to relax the restriction that requires one-to-many translation data on the construction of large-scale Chinese parabanks, by utilizing bilingual one-to-one translation data of larger scales and MT, and construct ParaZh-22M. Specifically, we leverage the huge Chinese-English machine translation data from WMT 2021 (Akhbardeh et al., 2021) of 30.4M sentence pairs, apply **strict rules** to ensure the data quality, and translate the English side of the parallel corpus to Chinese with the cutting-edge deep Transformers and **several approaches** to ensure the translation quality. We pick the machine translated Chinese sentences considering both diversity and semantic consistency, and pair with the corresponding original Chinese references to form paraphrase pairs. Compared to the PKU parabank, ParaZh-22M is $\sim$40 times as large, involves a broader range of paraphrase phenomena and domains, and can support the training of large Chinese paraphrasers.

Our main contributions are as follows:

- We propose to relieve the need of one-to-many translation data for the construction of Chinese parabank, and construct a Chinese parabank of 22M sentence pairs based on one-to-one sentence translation data and advanced MT models, which involves many domains and is $\sim$40 times as large as the previous largest PKU Chinese parabank;
- We test the effects of data augmentation via

[*] Corresponding author.

[1] We opensource our dataset at https://github.com/haowj9977/parazh-22M.

Proceedings of the 29th International Conference on Computational Linguistics, pages 3885–3897
October 12–17, 2022.

paraphrasing based on our parabank on a wide range of Chinese NLP tasks, including short/long text classification, natural language inference, keyword recognition, and low-resource machine translation, and show that paraphrasing based on our parabank is able to achieve consistent and significant improvements over several baselines.

2 Construction of ParaZh-22M

Zhang et al. (2019b) extract sentence-level paraphrases from multiple Chinese translations of the same source texts. Their approach requires one-to-many translation data, which is hard to collect. Instead, we try to relieve this restriction in Chinese parabank construction, and build the parabank based on one-to-one sentence-level translation data. Specifically, we translate the translation of Chinese sentences in the parallel data back to Chinese with the cutting-edge Neural Machine Translation (NMT) technology, and construct the parabank by pairing the machine translated Chinese sentences with the corresponding original Chinese sentences.

We suggest that the semantic consistency and quality of the parallel bank is ensured by the cutting-edge NMT algorithm, as the translation quality of advanced NMT methods is already close to that of translation agencies in high resource scenarios (Akhbardeh et al., 2021).

The construction of ParaZh-22M can be divided into 4 steps: 1) data collection, 2) data processing and cleaning, 3) training of NMT models, and 4) paraphrase generation.

2.1 Data Collection

We leverage bilingual parallel sentence data for the construction of the Chinese parabank and the training of NMT models, and monolingual data to further boost the performance of NMT via back-translation (Sennrich et al., 2016a). We select several datasets from WMT 2021 Chinese-English news translation task (Akhbardeh et al., 2021), and statistics are shown in Table 1. Even though the data is collected for the news translation task in WMT, they indeed involve many domains, e.g., the ParaCrawl corpus is the extraction of parallel sentences from the web regardless of their domains.

	Dataset	Size
bilingual	United Nations Parallel Corpus	15.9M
	ParaCrawl	14.2M
	News Commentary v16	0.3M
	total	30.4M
monolingual	News Crawl	10.6M

Table 1: Statistics of the bilingual and monolingual data. Size: the number of sentence pairs (for bilingual data) / sentences (for monolingual data). The monolingual data contain 10.6M sentences per language.

Bilingual parallel corpus To ensure the quality of the training NMT models and the parabank, we manually check the quality of each dataset provided by WMT 2021 for the Chinese-English news translation task, and take three datasets into consideration: the United Nations Parallel Corpus v1.0 (Ziemski et al., 2016), News Commentary v16, and ParaCrawl dataset (Bañón et al., 2020).

The United Nations Parallel Corpus (Ziemski et al., 2016) contains over 15.9M English-Chinese sentence pairs, which is composed of official records and other parliamentary documents of the United Nations that are in the public domain. The current version of the corpus contains content that was produced and manually translated between 1990 and 2014.

The News Commentary dataset is a collection of news about general politics, economics and science. Its English-Chinese section has about 0.3M sentence pairs.

The ParaCrawl dataset (Bañón et al., 2020) contains about 14.2M English-Chinese parallel sentences, constructed through web crawling software. Although its quality is slightly worse than the other two datasets in our manual evaluation, the ParaCrawl dataset involves many domains and provides a large number of training samples. In our experiments on the Zh→En task with base Transformers, using ParaCrawl dataset can bring about +5.4 and +3.8 BLEU improvements on the newstest 2020 and newstest 2021 test sets respectively.

Monolingual corpus Back translation is a simple and effective approach to improve the performance of MT with monolingual data (Sennrich et al., 2016a; Fadaee and Monz, 2018; Edunov et al., 2018; Wang et al., 2019b; Dou et al., 2020; Wei et al., 2020a; Marie et al., 2020). To further boost the performance of our NMT models (§ 2.3), and to obtain more accurate probability estimation in dual scoring (§ 2.4), we collect monolingual data of both languages, and augment the parallel data with the back translated monolingual data for the training of NMT models. Using back-translation data for NMT models' training also helps improve the translation diversity and alleviate the overfitting issue on the parallel data. The back translated data are not used for the construction of the parabank, to avoid introducing back-translation noise into the parabank and to ensure the quality of the parabank.

Specially, we extract ∼10.6M sentences for both English and Chinese from the monolingual News Crawl dataset, which provides article texts from various online news.

2.2 Data Processing and Cleaning

The quality of the dataset affects the performance of NMT and decides the quality of the parabank.

As many data are crawled from the web, we first standardize the texts with the following pipeline:

1. removing sentences with encoding errors;
2. replacing full-width characters with their corresponding half-width characters;
3. normalizing punctuation;
4. converting all named and numeric character HTML references (e.g., &gt;, &#62;, &#x3e) to their corresponding Unicode characters;
5. converting Traditional Chinese to Simplified Chinese through OpenCC. [2]

For the training of NMT models, we tokenize and truecase the English part with Moses (Koehn et al., 2007), and segment Chinese sentences into words using jieba. [3]

To clean the English-Chinese parallel corpus: 1) we only retain the most frequent instance when the source sentence has multiple translations in the data, 2) we remove the training instances where low frequency tokens take a large part of the sentence pairs, and 3) we remove sentence pairs with abnormally large source-vs-target length ratios. After the data cleaning, around 26.4M sentence pairs are left for the training of NMT models and the construction of the parabank.

We perform independent Byte Pair Encoding (BPE) (Sennrich et al., 2016b) for English and Chinese corpus with 32k merge operations to address the unknown word issue.

2.3 Training of NMT models

To construct the parabank, we only need to translate the English sentences into Chinese with NMT, but we have trained two NMT models for the forward and reverse translation directions for back translation (Sennrich et al., 2016a) and dual scoring (§ 2.4).

We employ the Transformer translation model (Vaswani et al., 2017) for NMT, as it has achieved the state-of-the-art performance in MT evaluations (Akhbardeh et al., 2021). We first use the parallel data to train 2 NMT models. Then we use greedy decoding to construct synthetic parallel data by back-translating the monolingual data (Sennrich et al., 2016a; Edunov et al., 2018). The back-translation data is then mixed with the original parallel data (the monolingual sentences at the target side and the greedy decoding texts at the source side). We fine-tune the NMT models trained in the first step on the mixed data for another $300k$ steps for improved performance.

To obtain good translation quality, we adopted the Transformer Big setting with 1024 and 4096 as the embedding dimension and the number of hidden units of the feed-forward layer respectively, together with a 12-layer deep encoder (Bapna et al., 2018; Wang et al., 2019a; Wu et al., 2019; Wei et al., 2020b; Zhang et al., 2019a; Xu et al., 2020a; Li et al., 2020; Huang et al., 2020; Xiong et al., 2020; Mehta et al., 2021; Li et al., 2021; Xu et al., 2021b). Parameters were initialized under the Lipschitz constraint (Xu et al., 2020a) to ensure the convergence of deep encoders. Since these NMT models are used to translate tens of millions of sentences (monolingual data for back-translation and the MT training set for the construction of the parabank), we used a 6-layer decoder instead of a deeper one to preserve the decoding efficiency (Kasai et al., 2021; Xu et al., 2021a). The number of warm-up steps was set to $8k$. We used a batch size of around $25k$ target tokens achieved by gradient accumulation (Xu et al., 2020b), and trained the models for $300k$ steps, which takes about 50 hours to train a model on 4 Nvidia A100 GPUs. We averaged the last 20 checkpoints saved with an interval of $1,500$ training steps.

The newstest 2019 was used as the development set, and newstest 2020 and newstest 2021 as the test set. The beam size of the decoder was set to 4, and translation quality was evaluated by case-sensitive BLEU (Papineni et al., 2002) with the SacreBLEU

[2] https://github.com/BYVoid/OpenCC

[3] https://github.com/fxsjy/jieba

Model	Zh→En		En→Zh	
	newstest20	newstest21	newstest20	newstest21
NMT	30.53	24.74	42.38	32.85
BT-NMT	30.97	25.18	52.42	42.99

Table 2: BLEU scores of our NMT models on the WMT 20 and WMT 21 news translation test sets.

toolkit (Post, 2018; Bawden et al., 2020). Results are shown in Table 2.

Table 2 shows that our BT-NMT models can obtain comparably strong translation performance.

2.4 Paraphrase Generation

Language filtering We find that there are some English words in the Chinese part of the parallel data, which may affect the quality of the constructed parabank. To address this issue, we remove sentence pairs where a large percentage of English words appear in its Chinese sentence. Specifically, we check the percentage of English characters in Chinese sentences, and sentence pairs will be dropped if the proportion is larger than 60%.

Generating paraphrase candidates The English sentences are semantically consistent with the corresponding Chinese sentences in the parallel data. So we can obtain paraphrases of the original Chinese sentences by translating the English sentences into Chinese with MT.

We use the En→Zh BT-NMT model for the translation. For each English sentence, we use a beam size of 15 and collect all Chinese beam search candidates. Then, we pair each MT candidate with the corresponding original Chinese sentence to get a candidate Chinese paraphrase pair.

We find that En→Zh translation is more challenging than that for the construction of English parabanks although the NMT model obtains a high BLEU score for character-level evaluation (used for Chinese translations), and approaches like sampling/constrained-decoding (Post and Vilar, 2018) further drop the performance (by $\sim$5 BLEU), causing semantic changes. Hence, we put a higher priority on translation quality to ensure the semantic consistency without using diversity-oriented approaches, such as sampling and constrained decoding. We suggest that our work provides a valuable reference for the construction of many other languages' parabanks with MT when ensuring MT quality is a problem.

Edit-distance ratio filtering To effectively ensure the diversity of the parabank, we compute the edit-distance ratio (the edit distance divided by the length) between the beam search candidate and the corresponding original Chinese sentence, and use a minimum edit-distance ratio of 12% to filter the paraphrase pairs. We note that, as the parallel data are large, it is easy to further filter out a large subset with an edit-distance threshold larger than ours.

Dual scoring filtering The En→Zh model may leave some source tokens untranslated, leading to the under-translation issue (Tu et al., 2016). Measuring the round-trip translation consistency has been proven to be an effective way to address this and to improve the translation quality (Goto and Tanaka, 2017). Instead of selecting the beam search candidate with the highest decoding probability ($p_{forward}$), we also take the force decoding probability of the reverse model (Zh→En) $p_{reverse}$ into consideration. We re-rank the beam search candidates by summing the forward and reverse probabilities.

$$p_{dual} = p_{forward} + p_{reverse} \tag{1}$$

During filtering, we first select the candidate with the highest p_{dual} from beam search results for each remaining Chinese sentence. Then we derive the per-token probability of all instances of the dataset based on p_{dual}, and only retain $\sim$22M sentence pairs with the highest per-token probability to further ensure the quality, obtaining the final Chinese parabank, ParaZh-22M.

3 Evaluation of ParaZh-22M

We compare the constructed ParaZh-22M with two existing Chinese paraphrase datasets: PKU Paraphrase Bank (Zhang et al., 2019b) and Chinese Paraphrase from Quora (Wang et al., 2021).

PKU Paraphrase Bank Zhang et al. (2019b) construct the PKU parabank by extracting multiple

Corpus	Source materials	Size (pairs)	Length (words)	Domain
PKU Paraphrase Bank	One-to-many translation	509K	**23.05**	Literature
Chinese Paraphrase from Quora	English Quora	263K	9.80	Question
ParaZh-22M (ours)	One-to-one translation	**22M**	22.16	**General**

Table 3: Information and statistics of Chinese parabanks.

Chinese translations of the same source texts (written in English as well as other European languages). The sentence pairs are from literary work.

Chinese Paraphrase from Quora Wang et al. (2021) transfer English retelling corpus, Quora, to Chinese with machine translation engines.

3.1 Statistics

We provide the basic information of Chinese parabanks on source materials, domain, size (the number of sentence pairs), and the average sentence length (the number of Chinese words segmented by jieba) in Table 3.

Table 3 shows that: 1) ParaZh-22M is two orders of magnitude larger than the others, in terms of the number of paraphrases, it is 84 times as large as the Chinese Paraphrase from Quora (Wang et al., 2021) and 43 times as large as PKU Paraphrase Bank (Zhang et al., 2019b). 2) the average number of words of ParaZh-22M is similar to that of the PKU Paraphrase Bank, and ParaZh-22M has more words than the Chinese Paraphrase from Quora on average. And 3) as ParaZh-22M is constructed upon bilingual data which involve many domains and rich styles (for the use of 14.2M ParaCrawl data), it is more general than the other two paraphrase corpora (PKU Paraphrase Bank is constructed based on literature work while Chinese Paraphrase from Quora are translations of English Quora), and can be easily adapted to different domains.

We suggest that: 1) the large size of ParaZh-22M is crucial to support the training of large neural paraphraser models, 2) it is easy to filter out a large subset for the use of a special task given an edit-distance threshold, and 3) covering a wide range of domains makes the application of ParaZh-22M domain-agnostic, leading to robust performance.

3.2 Manual Evaluation

There lacks an ideal evaluation metric that takes both semantic consistency and diversity into account for paraphrasing. Semantic consistency, fluency and diversity are all important, while the diversity evaluation is normally against the consistency evaluation, e.g., a lower BLEU indicating higher diversity but lower semantic consistency (in MT). So we manually evaluate ParaZh-22M and PKU Paraphrase Bank (Zhang et al., 2019b) in terms of semantic consistency, literal fluency, and sentential diversity to measure their quality. We design our evaluation criteria following Wieting and Gimpel (2018); Wang et al. (2021), and specifics are shown in Table 4. For each evaluation criterion, we design 5 levels to distinguish the quality of sentence pairs.

We randomly sampled 800 sentence pairs from each dataset, and employed 8 native Chinese linguistic experts to rate them. Each sample is rated by 2 experts, and the final score is the average of their ratings. Results are shown in Table 5.

For the evaluation of ParaZh-22M, Table 5 shows that: 1) 93.9% of ParaZh-22M samples are strongly semantically consistent (with a score no less than 4, indicating that the semantic meaning of the sentence pair is nearly equivalent, or only may differ in some unimportant details). 2) 97.9% of ParaZh-22M samples are fluent (with at most one grammatical error), and 3) 97.9% of ParaZh-22M samples have at least one lexical variation.

Compared to the PKU Paraphrase Bank, ParaZh-22M achieves much higher scores in semantic consistency and literal fluency evaluation, while obtaining a slightly lower score in sentential diversity. We conjecture this might be because: 1) we pay more attention to optimizing the translation quality when constructing the parabank (§ 2), which gives the correctness (semantic consistency and fluency) a higher priority than the diversity, and 2) Zhang et al. (2019b) use one-to-many parallel data for the construction of the parabank, while we only use one-to-one translation data.

We evaluated the inter-annotator agreement with kappa (Artstein and Poesio, 2008), and obtained a kappa value of 0.87, suggesting that a high agreement is achieved with our evaluation criteria and our evaluation is reliable.

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 grammatical variation slightly.
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	The 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 4: Manual evaluation criteria of semantic consistency, literal fluency, and sentential diversity.

Score	Semantic Consistency		Literal Fluency		Sentential Diversity	
	Ours	PKU	Ours	PKU	Ours	PKU
= 5.0	69.1	34.3	82.4	72.0	56.9	65.3
$\geq$ 4.0	93.9	66.5	97.9	97.3	70.8	74.9
$\geq$ 3.0	98.4	87.8	99.9	99.5	84.0	87.3
$\geq$ 2.0	99.6	94.8	100.0	100.0	97.9	98.5
AVG score	**4.64±0.64**	3.89±1.11	**4.82±0.41**	4.72±0.50	4.11±1.19	**4.28±1.10**

Table 5: Manual evaluation results of our corpus and PKU Paraphrase Bank on semantic consistency, literal fluency, and sentential diversity. **Medium:** the cumulative percentages of samples with the scores. **Bottom:** the average score and the standard deviation of each criterion.

4 Using ParaZh-22M in Chinese NLP

We examine the effectiveness of data augmentation based on ParaZh-22M on a number of Chinese NLP tasks, including long/short text classification, natural language inference, keyword recognition from CLUE (a Chinese Language Understanding Evaluation benchmark) (Xu et al., 2020c) and the CCMT 2022 low-resource Chinese→Thai machine translation task, by paraphrasing the original training set.

4.1 Chinese Paraphraser

ParaZh-22M contains a large number of Chinese paraphrase examples, but cannot be directly used to augment the training sets of NLP tasks. To paraphrase arbitrary Chinese sentences, we train a Chinese paraphrase model, i.e., a Chinese paraphraser, on ParaZh-22M.

Like back translation, we use the machine translated Chinese sentences as the source input of the model, and the original Chinese sentences from the parallel data as the target when training the paraphraser on ParaZh-22M.

We used the same vocabulary and BPE as the Chinese part of NMT data (§ 2.2). We employed a base Transformer as the paraphraser. Specifically, we used 6 encoder and decoder layers, an embedding size of 512, 8 attention heads, a feed-forward layer of 2048 hidden units, and shared the encoder-decoder embeddings. The model was trained for $100k$ steps. The average of the last 5 checkpoints saved with an interval of $1,500$ training steps is served as the paraphraser.

	TNEWS					IFLYTEK				
Model	Baseline	PKU	Δ	Ours	Δ	Baseline	PKU	Δ	Ours	Δ
ALBERT-tiny	53.55	53.52	-0.03	**53.74**	**+0.19**	48.76	52.59	+3.83	**54.52**	**+5.76**
BERT-base	56.09	57.11	+1.02	**57.19**	**+1.10**	60.37	60.52	+0.15	**61.52**	**+1.15**
BERT-wwm-ext-base	56.77	57.55	+0.78	**57.69**	**+0.92**	59.88	59.75	-0.13	**61.79**	**+1.91**
avg	/	/	+0.59	/	**+0.74**	/	/	+1.28	/	**+2.94**

Table 6: Results (accuracy) on the validation sets of TNEWS and IFLYTEK tasks. "Δ" indicates the improvements over the baseline. "avg" is the average improvement of data augmentation over three baselines.

For fair comparison, we also trained a paraphraser under the same setting on the PKU parabank.

4.2 Text Classification

We conducted experiments on two text classification tasks of the CLUE benchmark (Xu et al., 2020c): TNEWS for short texts, and IFLYTEK for long texts.

The TNEWS task has 15 categories (finance, technology, sports, etc.), including 53.3k training instances and 10k validation data. The IFLYTEK task has 119 classes (food, car rental, education, etc.), with 12.1k training samples and 2.6k validation data.

We augmented the training data of these 2 tasks by paraphrasing the input sentences with our paraphraser, and constructed the synthetic data D_p by pairing paraphrases with the tag of the corresponding original sentence. We concatenated D_p with the original training set D_o as the augmented training set D_{aug}, and trained the same baseline model on the augmented training set.

We used ALBERT-tiny, BERT-base, BERT-wwm-ext-base as our baselines. ALBERT-tiny is a tiny version of ALBERT with only 4 layers and a hidden size of 312. BERT-base has 12 layers and uses a hidden size of 768. BERT-wwm-ext-base has the same configuration as BERT-base, but is pre-trained with whole word masking. We evaluated these models on the validation sets (as the test sets are not publicly available). Results are shown in Table 6.

Table 6 shows that: 1) paraphrasing based on both the PKU parabank and ParaZh-22M can lead to improvements on average, and 2) data augmentation based on ParaZh-22M leads to consistent and significant improvements over all baselines on both datasets, and brings about more accuracy improvements than based on the PKU parabank, showing the advantages of ParaZh-22M for both short and long text classification.

4.3 Natural Language Inference

We also examined the effects of paraphrasing based on ParaZh-22M on the natural language inference (NLI) task, and conducted experiments on CMNLI dataset. NLI aims to predict the relation (neutral, entailment, and contradiction) between sentence pairs. The CMNLI contains 391k training samples, and 12k validation instances.

We used the same baseline models described in § 4.2. For data augmentation, as each CMNLI training instance has a sentence pair, we investigate 4 cases: 1) augmentation by paraphrasing the first sentence ($S1$), 2) augmentation by paraphrasing the second sentence ($S2$), 3) augmentation by paraphrasing both sentences ($S12$), and 4) the combination of case 1 and case 2 ($S1 + S2$). We concatenated the paraphrased training set with the original training set. Results are shown in Table 7.

Table 7 shows that: even though paraphrasing based on the PKU parabank brings about more improvements in the S1+S2 and S2 settings with the ALBERT-tiny model than based on ParaZh-22M, data augmentation with ParaZh-22M leads to consistent and significant improvements over all baselines, and works better with larger models and stronger baselines (BERT-base and BERT-wwm-ext-base) than with the PKU parabank.

4.4 Keyword Recognition

The keyword recognition task requires the model to distinguish real keywords of paper abstracts from fake keywords. Chinese Scientific Literature (CSL) dataset (Xu et al., 2020c) contains Chinese paper abstracts and their real keywords from core journals of China, covering multiple fields of natural sciences and social sciences, with fake keywords generated through TF-IDF. CSL datasets provide 20k samples for training and 3k samples for validation.

We used the same baselines as in § 4.2. When paraphrasing the abstract, we performed beam de-

Model	Baseline	D_{aug}	S1+S2	Δ	S12	Δ	S1/S2	Δ
ALBERT-tiny	70.26	PKU	**73.67**	**+3.41**	72.01	+1.75	71.88/73.27	+1.62/+3.01
		Ours	73.07	+2.81	72.88	+2.62	72.56/72.45	+2.30/+2.19
BERT-base	79.47	PKU	79.55	+0.08	79.95	+0.48	79.90/79.81	+0.43/+0.34
		Ours	80.27	+0.80	**80.80**	**+1.33**	80.57/80.30	+1.10/+0.83
BERT-wwm-ext-base	80.92	PKU	79.98	-0.94	80.50	-0.42	80.04/80.08	-0.88/-0.84
		Ours	81.23	+0.31	81.21	+0.29	**81.28**/81.16	**+0.36**/+0.24

Table 7: Results (accuracy) on the validation set of CMNLI task.

Model	Baseline	PKU	Δ	Ours	Δ
ALBERT-tiny	74.34	76.66	+2.32	**77.20**	**+2.86**
BERT-base	79.63	79.03	-0.60	**80.90**	**+1.27**
BERT-wwm-ext-base	80.60	79.30	-1.30	**81.00**	**+0.40**
avg	/	/	+0.14	/	**+1.51**

Table 8: Results (accuracy) on the validation set of CSL task.

coding with a beam size of 15 with the paraphraser, and selected the beam search candidate that contains all corresponding keywords and has the highest decoding probability. We did not augment training instances when no beam search candidate contains the keywords. The synthetic training set D_p was then combined with the original training set D_o. Results are shown in Table 8.

Table 8 shows that: 1) data augmentation based on both the PKU parabank and ParaZh-22M can bring about improvements on average, 2) the accuracy improvements with ParaZh-22M are consistent and significant with all baselines, including in challenging cases (with BERT-base and BERT-wwm-ext-base), and are larger than with the PKU parabank, demonstrating the effectiveness of ParaZh-22M in challenging settings.

4.5 Low-Resource Machine Translation

We conducted experiments on the CCMT 2022 Chinese→Thai low-resource translation task. Its training set has 200k sentence pairs. As the evaluation does not release both the development set and the test set, we held out the last 2000 sentence pairs of the training set, and equally divided them into 2 parts for validation and test respectively. We paraphrased the Chinese sentences of the training data and paired with the corresponding Thai sentences.

We employed a 6-layer and a 12-layer Transformer as our baselines. Following Sennrich and Zhang (2019), we used an embedding dimension of 256, 4 attention heads, 1024 as the hidden dimension of the feed-forward layer, a dropout probability of 0.1, and applied 16k BPE operations enforced by sentence piece (Kudo and Richardson, 2018).

Model	Baseline	PKU	Δ	Ours	Δ
6-layer	7.15	6.35	-0.80	**7.75**	**+0.60**
12-layer	10.65	7.31	-3.34	**14.74**	**+4.09**

Table 9: Results (BLEU) on CCMT 2022 Zh→Th translation task.

We set the a beam size to 4, and evaluated translation quality via BLEU with the average of the last 5 checkpoints saved in an interval of 1,500 training steps. Results are shown in Table 9.

Table 9 shows that: 1) paraphrasing based on ParaZh-22M can lead to consistent and significant improvements in the low-resource translation task with both settings, and 2) the improvements with the 12-layer model (+4.09 BLEU) in the MT task without using pre-trained models are much larger than in CLUE tasks with pre-trained models and than with the 6-layer model with fewer parameters.

5 Related Work

Data augmentation via paraphrasing is beneficial for many NLP tasks, such as question answering (Dong et al., 2017), semantic parsing (Berant and Liang, 2014; Su and Yan, 2017) and machine translation (Cho et al., 2014; Khayrallah et al., 2020), especially in low-resource scenarios. Paraphrasing relies heavily on large scale paraphrase datasets.

Construction of English paraphrase data Most paraphrase corpus construction studies are for English (Suzuki et al., 2017; Mallinson et al., 2017). Given the development of NMT, Wieting and Gimpel (2018) leverage large amounts of bilingual parallel data to generate paraphrases via MT. Hu et al. (2019a) add lexical constraints during NMT decoding to enrich the diversity. Hu et al. (2019b) cluster over constrained sampling decoding candidates to generate diverse paraphrases. Compared to Wieting and Gimpel (2018) and Hu et al. (2019a,b), we assign a higher priority to the quality of machine translation than the diversity to ensure the translation correctness and the semantic consistency, without using constrained decoding or sampling that hampers the translation quality.

Chinese paraphrase corpus To our knowledge, existing Chinese parabanks are much smaller than large scale English parabanks. Zhang et al. (2019b) extract sentence-level paraphrases from multiple Chinese translations of the same source text, obtaining the PKU Paraphrase Bank of 509,832 paraphrase pairs. Wang et al. (2021) translate the question retelling Quora corpus into Chinese with multiple MT engines, and construct a Chinese parabank of 263,729 sentence pairs. Compared to their work, ParaZh-22M is much larger and involves many domains.

6 Conclusion

In this paper, we relieve the requirement of one-to-many translation data for the construction of Chinese parabank, and construct a Chinese parabank of 22M sentence pairs, ParaZh-22M, utilizing one-to-one sentence-level parallel data and MT technology. ParaZh-22M involves many domains and is over 40 times as large as the previous largest PKU Chinese Paraphrase Bank. Human evaluation on semantic consistency, fluency and sentential diversity shows the good quality of ParaZh-22M.

We test the effects of data augmentation via paraphrasing based on ParaZh-22M on a wide range of Chinese NLP tasks, including short/long text classification, natural language inference, keyword recognition, and low-resource machine translation. Our experiment results show that paraphrasing based on ParaZh-22M is able to achieve consistent and significant improvements over several baselines in all evaluations, demonstrating the contribution of ParaZh-22M to Chinese NLP tasks.

Acknowledgements

We thank anonymous reviewers for their insightful comments. Wenjie Hao, Hongfei Xu, Hongyin Zan and Lingling Mu acknowledge the support of the National Social Science Fund of China (Grant No. 17ZDA138 and Grant No. 14BYY096). Deyi Xiong is partially supported by the Key Research and Development Program of Yunnan Province (Grant No. 202203AA080004) and Zhejiang Lab (Grant No. 2022KH0AB01).

References

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 Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Ankur Bapna, Mia Chen, Orhan Firat, Yuan Cao, and Yonghui Wu. 2018. Training deeper neural machine translation models with transparent attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3028–3033. Association for Computational Linguistics.

Rachel Bawden, Biao Zhang, Lisa Yankovskaya, Andre Tättar, and Matt Post. 2020. A study in improving BLEU reference coverage with diverse automatic paraphrasing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 918–932, Online. Association for Computational Linguistics.

Jonathan Berant and Percy Liang. 2014. Semantic parsing via paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 1415–1425, Baltimore, Maryland. Association for Computational Linguistics.

Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. Controllable paraphrase generation with a syntactic exemplar. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5972–5984, Florence, Italy. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 875–886, Copenhagen, Denmark. Association for Computational Linguistics.

Zi-Yi Dou, Antonios Anastasopoulos, and Graham Neubig. 2020. Dynamic data selection and weighting for iterative back-translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5894–5904, Online. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Marzieh Fadaee and Christof Monz. 2018. Back-translation sampling by targeting difficult words in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 436–446, Brussels, Belgium. Association for Computational Linguistics.

Isao Goto and Hideki Tanaka. 2017. Detecting untranslated content for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 47–55, Vancouver. Association for Computational Linguistics.

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 *Proceedings of AAAI*, volume 33, pages 6521–6528, Hawaii, USA.

J. Edward Hu, Abhinav Singh, Nils Holzenberger, Matt Post, and Benjamin Van Durme. 2019b. Large-scale, diverse, paraphrastic bitexts via sampling and clustering. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 44–54, Hong Kong, China. Association for Computational Linguistics.

Xiao Shi Huang, Felipe Perez, Jimmy Ba, and Maksims Volkovs. 2020. Improving transformer optimization through better initialization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4475–4483. PMLR.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, and Noah Smith. 2021. Deep encoder, shallow decoder: Reevaluating non-autoregressive machine translation. In *International Conference on Learning Representations*.

Huda Khayrallah, Brian Thompson, Matt Post, and Philipp Koehn. 2020. Simulated multiple reference training improves low-resource machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 82–89, Online. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Bei Li, Ziyang Wang, Hui Liu, Quan Du, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2021. Learning light-weight translation models from deep transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13217–13225.

Bei Li, Ziyang Wang, Hui Liu, Yufan Jiang, Quan Du, Tong Xiao, Huizhen Wang, and Jingbo Zhu. 2020. Shallow-to-deep training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 995–1005, Online. Association for Computational Linguistics.

Zichao Li, Xin Jiang, Lifeng Shang, and Qun Liu. 2019. Decomposable neural paraphrase generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3403–3414, Florence, Italy. Association for Computational Linguistics.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893, Valencia, Spain. Association for Computational Linguistics.

Benjamin Marie, Raphael Rubino, and Atsushi Fujita. 2020. Tagged back-translation revisited: Why does it really work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5990–5997, Online. Association for Computational Linguistics.

Sachin Mehta, Marjan Ghazvininejad, Srinivasan Iyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021. Delight: Deep and light-weight transformer. In *International Conference on Learning Representations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Yu Su and Xifeng Yan. 2017. Cross-domain semantic parsing via paraphrasing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1235–1246, Copenhagen, Denmark. Association for Computational Linguistics.

Jiao Sun, Xuezhe Ma, and Nanyun Peng. 2021. AESOP: Paraphrase generation with adaptive syntactic control. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5176–5189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yui Suzuki, Tomoyuki Kajiwara, and Mamoru Komachi. 2017. Building a non-trivial paraphrase corpus using multiple machine translation systems. In *Proceedings of ACL 2017, Student Research Workshop*, pages 36–42, Vancouver, Canada. Association for Computational Linguistics.

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

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. 2019a. Learning deep transformer models for machine translation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 1810–1822, Florence, Italy. Association for Computational Linguistics.

Shuo Wang, Yang Liu, Chao Wang, Huanbo Luan, and Maosong Sun. 2019b. Improving back-translation with uncertainty-based confidence estimation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 791–802, Hong Kong, China. Association for Computational Linguistics.

Yasong Wang, Mingtong Liu, Yujie Zhang, Jin'an Xu, and Chen Yufeng. 2021. Research on the construction and application of paraphrase parallel corpus. *Acta Scientiarum Naturalium Universitatis Pekinensis*, 57(1):68–74.

Hao-Ran Wei, Zhirui Zhang, Boxing Chen, and Weihua Luo. 2020a. Iterative domain-repaired back-translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5884–5893, Online. Association for Computational Linguistics.

Xiangpeng Wei, Heng Yu, Yue Hu, Yue Zhang, Rongxiang Weng, and Weihua Luo. 2020b. Multiscale collaborative deep models for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 414–426, Online. Association for Computational Linguistics.

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

Lijun Wu, Yiren Wang, Yingce Xia, Fei Tian, Fei Gao, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2019. Depth growing for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5558–5563, Florence, Italy. Association for Computational Linguistics.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. 2020. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10524–10533. PMLR.

Hongfei Xu, Qiuhui Liu, Josef van Genabith, Deyi Xiong, and Jingyi Zhang. 2020a. Lipschitz constrained parameter initialization for deep transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 397–402, Online. Association for Computational Linguistics.

Hongfei Xu, Josef van Genabith, Qiuhui Liu, and Deyi Xiong. 2021a. Probing word translations in the transformer and trading decoder for encoder layers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 74–85, Online. Association for Computational Linguistics.

Hongfei Xu, Josef van Genabith, Deyi Xiong, and Qiuhui Liu. 2020b. Dynamically adjusting transformer batch size by monitoring gradient direction change. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3519–3524, Online. Association for Computational Linguistics.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020c. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Peng Xu, Dhruv Kumar, Wei Yang, Wenjie Zi, Keyi Tang, Chenyang Huang, Jackie Chi Kit Cheung, Simon J.D. Prince, and Yanshuai Cao. 2021b. Optimizing deeper transformers on small datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2089–2102, Online. Association for Computational Linguistics.

Biao Zhang, Ivan Titov, and Rico Sennrich. 2019a. Improving deep transformer with depth-scaled initialization and merged attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 898–909, Hong Kong, China. Association for Computational Linguistics.

Bowei Zhang, Weiwei Sun, Xiaojun Wan, and Zongming Guo. 2019b. PKU paraphrase bank: A sentence-level paraphrase corpus for chinese. In *Natural Language Processing and Chinese Computing*, pages 814–826, Cham. Springer International Publishing.

Michal Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1.0. In *LREC 2016*.