A Dual Reinforcement Method for Data Augmentation using Middle Sentences for Machine Translation

Wenyi Tang tang@akane.waseda.jp
Yves Lepage yves.lepage@waseda.jp
Graduate School of IPS, Waseda University, Kitakyushu, Japan

Abstract

This paper presents an approach to enhance the quality of machine translation by leveraging middle sentences as pivot points and employing dual reinforcement learning. Conventional methods for generating parallel sentence pairs for machine translation rely on parallel corpora, which may be scarce, resulting in limitations in translation quality. In contrast, our proposed method entails training two machine translation models in opposite directions, utilizing the middle sentence as a bridge for a virtuous feedback loop between the two models. This feedback loop resembles reinforcement learning, facilitating the models to make informed decisions based on mutual feedback. Experimental results substantiate that our proposed method significantly improves machine translation quality.

1 Introduction

The accuracy of neural machine translation is limited by the quantity of available training data (Wang et al., 2022; Sennrich et al., 2016), leading to the development of various techniques for data augmentation. In this paper, we propose a novel method that leverages middle sentences (Wang et al., 2021) as pivot points and uses dual reinforcement learning (Zhou et al., 2019) for data augmentation in machine translation.

Dual learning (He et al., 2016; Yi et al., 2017; Zhou et al., 2019) entails the concurrent training of two neural networks, to enhance translation accuracy by leveraging the reconstruction model's ability to generate synthetic parallel sentence pairs. Data augmentation involves artificially augmenting the size of the training data by generating additional sentence pairs through diverse techniques, such as back-translation (Brislin, 1970; Douglas and Craig, 2007; Edunov et al., 2018). These techniques offer potential solutions to mitigate the scarcity of parallel corpora and improve the quality of machine translation models by providing supplementary training data.

In our proposal, we aim to combine the strengths of dual learning and data augmentation with the use of middle sentences as pivot points to reinforce the training process and further enhance the accuracy of the machine translation model. We start by presenting our dual reinforcement method in Section 2. We present our experiment setup in Section 3 and results in Section 4.

Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track, pages 48–58
September 4–8, 2023, Macau SAR, China.

2 Methods

Our method combines the use of a dual learning framework with data augmentation techniques, leveraging the middle sentences of parallel sentence pairs as pivot points. The general process involves generating additional parallel sentence pairs through middle sentence generation, using the middle sentences to create new sentence pairs and refining the translations using a machine translation model. This process is repeated iteratively, forming a reinforcement loop that enhances the quality of the translation model through synthetic data, i.e., middle sentences. In the following subsections, we provide a detailed explanation of each step in our method.

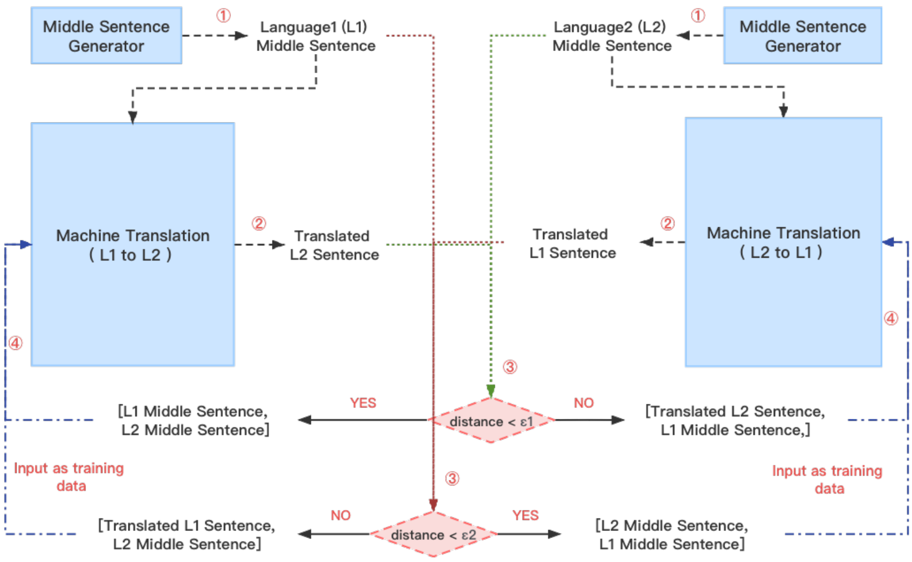


Figure 1: Framework of dual reinforcement method

2.1 Middle Sentence Generation

A middle sentence refers to a sentence that is generated or identified as an intermediate sentence between two given sentences, namely the start sentence and the end sentence (Wang et al., 2021). They suggest computing the middle sentences using Formula 1.

$$m = \frac{1}{2} \times (s + e) \tag{1}$$

Our method uses the semantic representations of the input sentences, i.e., their embedding vectors obtained using a pre-trained language model. Specifically, we use the following formula to calculate the embedding vector of the middle sentence:

$$m = \frac{1}{2} \times \frac{\|s\| + \|e\|}{\|s + e\|}(s + e) \tag{2}$$

where s and e represent the embedding vectors of the start and end sentences, respectively. The resulting embedding vector m represents the semantic midpoint between the two input sentences.

The inclusion of normalization terms in the Formula 2 takes into account the lengths of the input vectors. This ensures that the resulting midpoint vector has a relatively similar length as the input vectors, regardless of their initial lengths. By considering the magnitudes of the vectors, the equation provides a better suited representation of the semantic center between the start and end sentences.

Once the embedding vector of the middle sentence is obtained, we utilize it as input to a decoder model to generate an actual sentence.

By using the aforesaid technique, we create middle sentences for two languages, L_1 and L_2, by entering two parallel sentence pairs in each language. The problem is to check whether this pair of middle sentences is parallel and suitable for use as training data to enhance machine translation quality.

Let us take Chinese and English as examples. We randomly select a pair of start and end sentences in Chinese, such as '我爱吃苹果' (I love eating apples) and '我想学习' (I want to study). The generated intermediate sentence is '我爱学习' (I love study). Similarly, in English, we generated 'i like study' as the middle sentence.

2.2 Generation of Corresponding Translations

Once the middle sentences in two languages are generated, they can be used as input to their respective machine translation models to obtain corresponding translations. For instance, the middle sentences of L_1 can be fed into the machine translation model for translation in the direction L_1 to L_2, resulting in the translated sentences in L_2. And similarity for sentences in L_2, resulting in translations in L_1.

For the same example as above, we can translate the Chinese middle sentence '我爱学习' into English as 'I love study,' and the translation of the English middle sentence 'i like study' would be '我喜欢学习' in Chinese.

2.3 Selection of Sentence Pairs

We begin by measuring the distance between the L_1 middle sentence and the translated L_1 sentence obtained through the L_2 to L_1 machine translation model using the $L2$ middle sentence. For that, we use euclidean distance with a pre-set threshold. If the L_1 middle sentence bears significant resemblance to the translated L_1 sentence, indicating that the middle sentence in L_1 aligns closely with the L_1 sentence obtained through machine translation of the L_2 middle sentence, then we consider the L_1 middle sentence to be both middle and parallel to the L_2 middle sentence. They can be regarded as a pair of parallel sentences and utilized as training data for machine translation. Similarly, in the other direction with L_2 and L_1.

If the L_1 middle sentence and the translated L_1 sentence exceed the distance threshold, then we consider the L_1 middle sentence and the L_2 middle sentence to be middle but not parallel. As we aim to have parallel sentences that can improve machine translation model accuracy, we treat the L_1 middle sentence and its L_2 translation obtained through machine translation as a pair of parallel sentences. These parallel sentences can be utilized for training the L_2 to L_1 machine translation model. Similarly, in the other direction.

We continue the aforementioned process and calculate the distance between the Chinese middle sentence '我爱学习' and the translation of the English middle sentence, '我喜欢学习' It is evident that these two sentences are very similar, indicating that we can determine that the Chinese middle sentence '我爱学习' and the English middle sentence 'i like study' is a pair of parallel middle sentences. The same applies to the English middle sentence in the other translation direction.

However, if the Chinese middle sentence is '我爱学习' (I love studying), and the English middle sentence is 'i want to sleep,' which translates to '我想睡觉', it is evident that these two sentences are not similar. Therefore, the Chinese middle sentence and the English middle sentence, despite both being middle sentences, are not parallel to each other. In this case, we would replace the Chinese middle sentence with the translation of the English middle sentence and consider ‘i want to sleep’ and ‘我想睡觉’ as a pair of data to be included in the training set of the Chinese-to-English machine translation model.

2.4 Reinforcement Loop

The iterative process of utilizing dual learning and middle sentences is repeated in a reinforcement loop. The use of distance to determine sentence similarity and facilitate sentence substitution can be likened to the reward function employed in traditional reinforcement learning approaches. The refined translations from the machine translation model are used to generate additional augmented sentence pairs, which are incorporated into the training data. This loop enables continuous refinement of the model, allowing for further improvement of its accuracy over successive iterations.

3 Experimental Setup

The experimental setup for this study uses a neural machine translation (NMT) model avaiable in the OpenNMT tool (Klein et al., 2017). The selected architecture is a transformer encoder and decoder, with a word vector size of 512, 6 layers, and 8 heads, alongside an RNN size of 512. The transformer feed-forward network has a size of 2048. During training, gradients are accumulated over 8 batches, and the model is optimized using the Adam optimizer with beta1 set to 0.9, beta2 set to 0.998, and a learning rate of 0.001. Batch sizes are set to 4096, utilizing token batch type, with token normalization and a dropout rate of 0.1, while label smoothing was set to 0.1.

We employ a parallel dataset in English and Chinese extracted from Tatoeba [1]. The statistics of the dataset are presented in Table 1.

Language	Sentences	Tokens	Types	Avg. length of sentences (in char)
English	67,333	556,529	16,248	8.27
Chinese	67,333	888,743	24,864	13.20

Table 1: Statistics on Tatoeba corpus

To evaluate our system’s performance, we use three standard metrics: BLEU (Bilingual Evaluation Understudy), CHRF (CHaRacter-level F-score), and TER (Translation Error Rate). BLEU (Papineni et al., 2002) quantifies the n-gram overlap between the generated text and the reference text. CHRF (Popović, 2015) calculates the character n-gram F-score between the generated and reference text. Finally, TER (Snover et al., 2006) measures the minimum edit distance between the generated and reference text, accounting for insertions, deletions, and substitutions. Furthermore, we use SacreBLEU (Post, 2018) to conduct significance testing, and highlight the experimental outcomes that exhibited a significant improvement by bolding them.

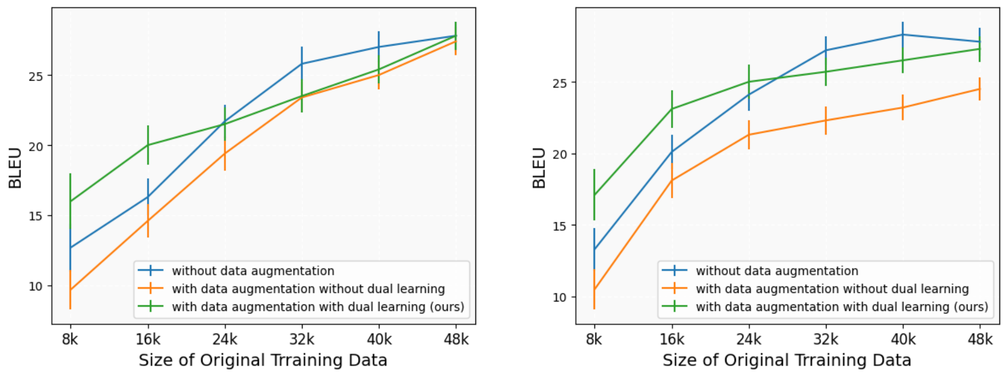
4 Results

4.1 Different Data Sizes

We conduct experiments to analyze the impact of dataset size on our results. We partition the dataset into subsets ranging from 10k to 50k, with increments of 10k. The data is then divided into training, validation, and test sets in an 8:1:1 ratio.

[1] https://tatoeba.org

Figures 2a and 2b present the BLEU scores obtained by training on datasets of varying sizes. The general trend observed is an increase in score as the dataset size increases. When the dataset is less than 24 thousand, our proposed method outperforms the other two methods. However, as the dataset size increases, our method does not surpass the model trained on the original data. Nevertheless, our method does consistently outperform the method with data augmentation without dual learning on all dataset sizes.



(a) English to Chinese machine translation model (b) Chinese to English machine translation model

Figure 2: BLEU scores across different data sizes. The model without data augmentation uses the original data size. The models with data augmentation add up data to the original training data, three times as model data, which makes these models learn from a four times larger training data.

Considering that our experimental outcomes show superior performance when the training dataset consists of 8 thousand data points, we conduct an analysis of the original 8 thousand sentence pair data compared with the method with data augmentation without dual learning with our own data augmentation method.

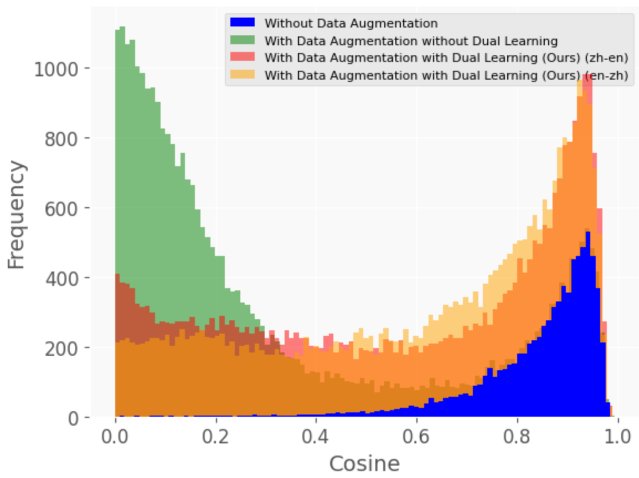


Figure 3: Distributions of training data and augmented data

The analysis of the distribution of the generated data using our method compared to the method with data augmentation without dual learning shows that our method generates data with a distribution more similar to that of the original data, as most of the generated data has a cosine similarity in the range of 0.8–1.0. In contrast, the method with data augmentation without dual learning generates data mostly in the range of 0–0.2, which may indicate lower alignment quality of the generated data. However, it is noted that our method also generates some sentence pairs with cosine similarity in the range of 0–0.4, which may explain why our method performs better with a smaller amount of raw data. It seems that when the original data is small, our method generates more high-quality sentence pairs, which can be beneficial for improving translation accuracy. However, when the dataset is large, our method may generate low-quality pairs, which potentially has a negative impact on models that have already been trained on a substantial amount of parallel data.

4.2 Impact of Parallel and Nonparallel Start-End Sentence Pairs on Machine Translation Models

To ensure the reliability and effectiveness of our proposed method, we conducted an extensive experiment to evaluate its robustness in handling both parallel and non-parallel start and end sentence pairs, which are selected at random. By examining the impact of data parallelism on the machine translation model, we aimed to investigate the performance of our proposed method under different input conditions.

Parallel	cosine similarity	Euclidean distance
Yes	0.84	0.60
No	0.08	1.79

Table 2: Similarity and distance of parallel and non-parallel sentence pairs

As observed from Figure 4, the model trained on parallel sentence pairs (dark blue bar) achieved a significantly higher BLEU score compared to the model trained on non-parallel sentence pairs (medium-dark blue bar) . This suggests that the utilization of non-parallel sentence pairs as input for machine translation models can adversely affect their accuracy. Nonetheless, it shows that our method can enhance the performance of machine translation models, even when non-parallel sentence pairs are used as input. While the use of non-parallel sentence pairs does result in a decrease in accuracy compared to parallel sentence pairs, the performance is still improved compared to the original model (medium-light blue bar) without data augmentation.

4.3 Different Euclidean Distance Threshold to Select Sentence Pairs

Given that we have a threshold for determining the degree of parallelism between the middle and translated sentences, this threshold directly impacts the quality and quantity of the training data utilized. Consequently, we perform experiments with various euclidean distance thresholds to evaluate this impact.

Figure 5 illustrates that the model reaches its best performance at a euclidean distance threshold of 0.3, after which its efficacy decreases. This observation implies that setting the threshold at 0.3 enables us to effectively eliminate non-parallel sentence pairs, while retaining an adequate number of high-quality parallel sentence pairs for training the machine translation model.

5 Conclusion

This paper presented a novel data augmentation method for enhancing machine translation performance by using middle sentences and dual learning. Our approach aims to overcome the

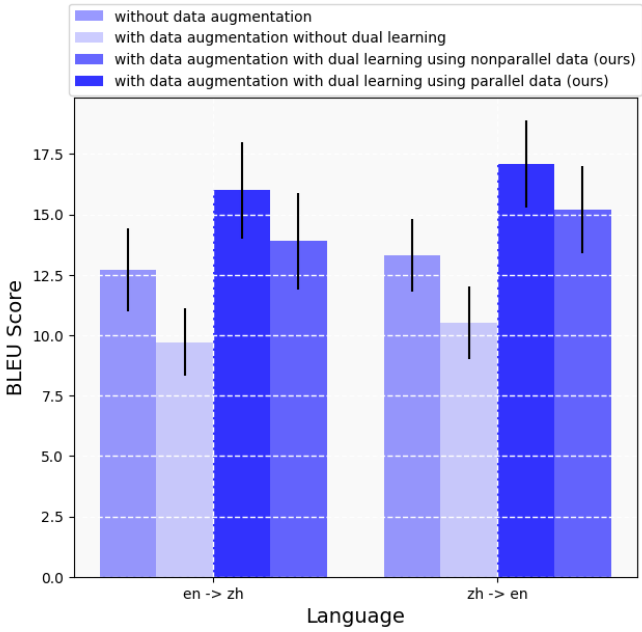
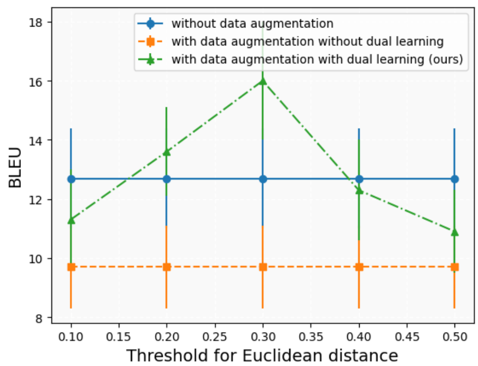
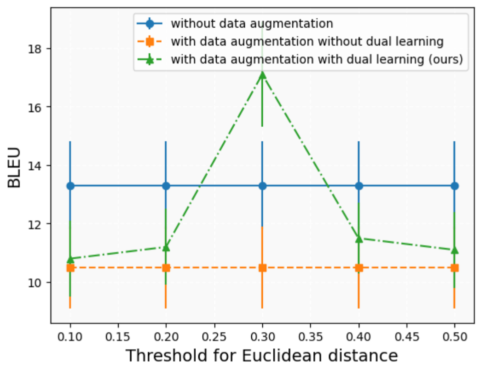


Figure 4: BLEU score across different methods



(a) en-zh machine translation model



(b) zh-en machine translation model

Figure 5: BLEU scores for various thresholds of Euclidean distance

challenge of availability and quality of parallel corpora, which can substantially impair the accuracy of machine translation systems. By utilizing middle sentences as pivot points and integrating dual learning with data augmentation techniques, we generated a considerable number of high-quality parallel sentence pairs to train machine translation models. The experimental results substantiate the superiority of our proposed method over two baseline methods.

Similar to any research, there exist potential challenges and opportunities for future work. One promising direction is to examine the adaptability of our proposed method for other languages, particularly those with limited available training data. Additionally, it would be worthwhile to investigate the applicability of our method in other natural language processing tasks beyond machine translation, such as text summarization or sentiment analysis. Moreover, future research could investigate the use of more sophisticated similarity metrics to determine parallel sentence pairs.

References

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of cross-cultural psychology*, 1(3):185–216.

Douglas, S. P. and Craig, C. S. (2007). Collaborative and iterative translation: An alternative approach to back translation. *Journal of International Marketing*, 15(1):30–43.

Edunov, S., Ott, M., Auli, M., and Grangier, D. (2018). Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500.

He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T.-Y., and Ma, W.-Y. (2016). Dual learning for machine translation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 820–828.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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

Sennrich, R., Haddow, B., and Birch, A. (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.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Wang, H., Wu, H., He, Z., Huang, L., and Church, K. W. (2022). Progress in machine translation. *Engineering*, 18:143–153.

Wang, P., Wang, L., and Lepage, Y. (2021). Generating the middle sentence of two sentences using pre-trained models: a first step for text morphing. In *Proceedings of the 27th annual meeting of the Association for Natural Language Processing*, pages 1481–1485.

Yi, Z., Zhang, H., Tan, P., and Gong, M. (2017). Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857.

Zhou, J. T., Zhang, H., Jin, D., Zhu, H., Fang, M., Goh, R. S. M., and Kwok, K. (2019). Dual adversarial neural transfer for low-resource named entity recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3461–3471.

A Table of Experiment Results

A.1 Different Data Sizes

Data size	Language Pairs	Data augmentation	Dual laerning	BLEU	chrF	TER
8k	en ->zh	without	without	12.7 ± 1.7	16.1 ± 1.3	67.6 ± 1.8
		with	without	9.7 ± 1.4	12.3 ± 1.1	72.4 ± 1.6
		with	with (ours)	**16.0 ± 2.0**	**17.7 ± 1.6**	**66.5 ± 1.9**
	zh ->en	without	without	13.3 ± 1.5	28.3 ± 1.4	68.5 ± 1.9
		with	without	10.5 ± 1.4	23.5 ± 1.3	75.6 ± 1.8
		with	with (ours)	**17.1 ± 1.8**	**32.0 ± 1.7**	**62.9 ± 1.8**
16k	en ->zh	without	without	16.3 ± 1.3	19.2 ± 1.0	64.4 ± 1.4
		with	without	14.6 ± 1.2	16.2 ± 1.0	70.2 ± 1.6
		with	with (ours)	**20.0 ± 1.4**	**23.6 ± 1.2**	63.8 ± 2.3
	zh ->en	without	without	20.1 ± 1.2	35.8 ± 1.1	60.3 ± 1.3
		with	without	18.1 ± 1.2	32.3 ± 1.1	62.1 ± 1.2
		with	with (ours)	**23.1 ± 1.3**	**40.0 ± 1.3**	**56.0 ± 1.4**
24k	en ->zh	without	without	21.7 ± 1.2	22.0 ± 1.0	63.4 ± 1.2
		with	without	19.4 ± 1.2	20.7 ± 1.0	67.3 ± 1.3
		with	with (ours)	21.5 ± 1.2	22.7 ± 1.0	58.5 ± 1.2
	zh ->en	without	without	24.1 ± 1.1	37.6 ± 1.0	59.3 ± 1.1
		with	without	21.3 ± 1.0	37.2 ± 1.0	62.3 ± 1.1
		with	with (ours)	**25.0 ± 1.2**	**41.0 ± 1.0**	**55.0 ± 1.1**
32k	en ->zh	without	without	25.8 ± 1.2	28.9 ± 1.0	53.9 ± 1.1
		with	without	23.4 ± 1.1	25.0 ± 1.0	58.1 ± 1.2
		with	with (ours)	23.5 ± 1.2	27.6 ± 1.0	54.7 ± 1.1
	zh ->en	without	without	27.2 ± 1.0	44.2 ± 0.9	51.6 ± 1.0
		with	without	22.3 ± 1.0	37.4 ± 0.9	56.5 ± 0.9
		with	with (ours)	25.7 ± 1.0	40.9 ± 1.0	55.1 ± 1.0
40k	en ->zh	without	without	27.0 ± 1.1	29.4 ± 0.9	52.7 ± 1.0
		with	without	25.0 ± 1.0	26.5 ± 0.9	56.0 ± 0.9
		with	with (ours)	25.4 ± 1.0	27.3 ± 1.0	54.6 ± 1.0
	zh ->en	without	without	28.3 ± 0.9	44.8 ± 0.8	51.0 ± 0.9
		with	without	23.2 ± 0.9	40.8 ± 0.8	55.4 ± 0.9
		with	with (ours)	26.5 ± 0.9	41.9 ± 0.9	54.7 ± 0.9
48k	en ->zh	without	without	27.8 ± 1.0	30.7 ± 0.9	52.2 ± 0.9
		with	without	27.4 ± 1.0	29.2 ± 0.9	54.1 ± 0.9
		with	with (ours)	27.8 ± 1.0	30.1 ± 0.9	53.7 ± 0.9
	zh ->en	without	without	27.8 ± 1.0	45.3 ± 0.7	64.2 ± 0.9
		with	without	24.5 ± 0.8	40.7 ± 0.8	59.3 ± 1.1
		with	with (ours)	27.3 ± 0.9	42.9 ± 0.8	54.3 ± 0.9

Table 3: Translation results of different sizes of dataset

A.2 Different Euclidean Distance Thresholds to Select Sentence Pairs

Data augmentation	Dual learning	Euclidean dis.	Language Pairs	BLEU	chrF	TER
without	without	/	en ->zh	12.7 ± 1.7	16.1 ± 1.3	67.6 ± 1.8
			zh ->en	13.3 ± 1.5	28.3 ± 1.4	68.5 ± 1.9
with	without	/	en ->zh	9.7 ± 1.4	12.3 ± 1.1	72.4 ± 1.6
			zh ->en	10.5 ± 1.4	23.5 ± 1.3	75.6 ± 1.8
with	with (ours)	0.1	en - >zh	11.3 ± 1.7	13.2 ± 1.5	70.3 ± 1.7
			zh ->en	10.8 ± 1.3	24.0 ± 1.5	69.2 ± 1.5
		0.2	en - >zh	13.6 ± 1.5	18.3 ± 1.3	65.7 ± 1.7
			zh ->en	11.2 ± 1.3	26.7 ± 1.3	65.9 ± 1.5
		0.3	en - >zh	**16.0 ± 2.0**	**17.7 ± 1.6**	**66.5 ± 1.9**
			zh ->en	**17.1 ± 1.8**	**32.0 ± 1.7**	**62.9 ± 1.8**
		0.4	en - >zh	12.3 ± 1.7	14.8 ± 1.4	65.7 ± 1.7
			zh ->en	11.5 ± 1.2	25.1 ± 1.3	70.3 ± 1.8
		0.5	en - >zh	10.9 ± 1.4	13.9 ± 1.1	72.4 ± 1.8
			zh ->en	11.1 ± 1.3	26.1 ± 1.1	75.3 ± 2.0

Table 4: Translation results of using different euclidean distance for selecting sentence pairs

A.3 Impact of Parallel and Nonparallel Start-End Sentence Pairs on Machine Translation Models

Language Pairs	Parallel	Cos similarity	Euclidean distance	BLEU	CHRF	TER
en ->zh	Yes	0.84	0.60	**16.0 ± 2.0**	**17.7 ± 1.6**	**66.5 ± 1.9**
	No	0.08	1.79	13.9 ± 1.8	16.1 ± 1.5	70.8 ± 2.0
zh ->en	Yes	0.84	0.60	**17.1 ± 1.8**	**32.0 ± 1.7**	**62.9 ± 1.8**
	No	0.08	1.79	15.2 ± 1.7	30.1 ± 1.5	64.6 ± 1.7

Table 5: Translation results starting from parallel and nonparallel start-end sentence pairs