applied sciences

MDPI

Article

A Mongolian–Chinese Neural Machine Translation Method Based on Semantic-Context Data Augmentation

Huinuan Zhang, Yatu Ji *, Nier Wu and Min Lu

School of Information Engineering, Inner Mongolia University of Technology, Hohhot 010051, China; 20211800101@imut.edu.cn (H.Z.); wunier04@imut.edu.cn (N.W.); cslumin@imut.edu.cn (M.L.)
* Correspondence: mljyt@imut.edu.cn

Abstract: Neural machine translation (NMT) typically relies on a substantial number of bilingual parallel corpora for effective training. Mongolian, as a low-resource language, has relatively few parallel corpora, resulting in poor translation performance. Data augmentation (DA) is a practical and promising method to solve problems related to data sparsity and single semantic structure by expanding the size and structure of available data. In order to address the issues of data sparsity and semantic inconsistency in Mongolian–Chinese NMT processes, this paper proposes a new semantic-context DA method. This method adds an additional semantic encoder based on the original translation model, which utilizes both source and target sentences to generate different semantic vectors to enhance each training instance. The results show that this method significantly improves the quality of Mongolian–Chinese NMT tasks, with an increase of approximately 2.5 BLEU values compared to the basic Transformer model. Compared to the basic model, this method can achieve the same translation results with about half of the data, greatly improving translation efficiency.

Keywords: Mongolian neural machine translation; data sparseness; semantic-context data augmentation

Citation: Zhang, H.; Ji, Y.; Wu, N.; Lu, M. A Mongolian–Chinese Neural Machine Translation Method Based on Semantic-Context Data Augmentation. *Appl. Sci.* **2024**, *14*, 3442. https://doi.org/10.3390/app14083442

Academic Editors: José Ramón Méndez Reboredo and David Ruano-Ordás

Received: 20 March 2024
Revised: 15 April 2024
Accepted: 16 April 2024
Published: 19 April 2024

1. Introduction

Data augmentation (DA) is a technology that employs relevant methods to expand data when the training data are limited. It has shown remarkable effects in various deep learning and machine learning tasks. While extensively and successfully applied in the field of computer vision [1], there is no in-depth research on natural language processing (NLP) and there has been limited success in achieving performance improvements. The existing data augmentation methods can generally be divided into two categories: one is word-replacement methods, and the other is relying on translation models for DA.

The first method is mainly to replace the words in the training data to achieve the effect of DA. Fadaee et al. [2] improved the quality of sentence translation by replacing high-frequency words with low-frequency words in language models. Xia et al. [3] proposed a universal framework for implementing data augmentation, which can obtain data from rich to scarce languages. Zhu et al. [4] proposed a method for enhancing "soft" contextual data by randomly selecting words and then obtaining "soft words" closely related to the context of the word through a language model. Zhou et al. [5] reordered the target sentence to match the order of the source language and used it as an additional source for training-time supervision. Liu et al. [6] proposed a causal relationship based on language models and phrase alignment, which constructs a pseudo-parallel corpus by generating counterfactual aligned phrases.

For the second method, the DA effect is mainly achieved by generating pseudo-parallel corpora through translation models. Sennrich et al. [7] were the first to propose the use of monolingual corpora for data augmentation. Fadaee [2] and Sugiyama [8] used back-translation methods to reverse-translate monolingual data from the target language, generating more source-language-enhanced versions. Isaac et al. [9] proposed a

reverse-translation method with labels, which labels real data and pseudo-data differently, so that the model can effectively distinguish these two types of data. Edunov et al. [10] demonstrated through experiments that adding a certain proportion of noise to the generated pseudo-data can effectively improve translation quality. Wu et al. [11] proposed a bilingual data extraction method called "Extract Edit" for generating high-quality bilingual data through substitution and back-translation. Jiao et al. [12] proposed a method of alternating training with real data and comprehensive data, which significantly improved the performance of translation. Wu et al. [13] used reverse translation to generate pseudo-bilingual data by separately generating source and target languages from third-party single-axis languages. Zhang et al. [14] proposed using self-learning algorithms to generate pseudo-parallel corpora from monolingual data, and also using two NMT models in a multitasking learning framework to predict translations and reorder source sentences. Abdulsumin et al. [15] proposed an improvement on self-learning methods, namely iterative self training methods. Hoang et al. [16] improved the performance of NMT by iterating reverse translation in high- and low-resource situations. Zhen et al. [17] established adversarial networks to improve the performance of NMT models. Zhang et al. [18] introduced a generative adversarial network into unsupervised machine translation, mapping the distribution of source and target words into a shared semantic space. Wei et al. [19] proposed a new data augmentation model—continuous semantic augmentation—to improve translation quality.

However, there are few studies on DA methods in Mongolian neural machine translation (MNMT). The structure of Mongolian sentences takes the subject–object–predicate form, which is different from most target languages in vocabulary, grammar and rhetorical devices [20]. Therefore, the development of MNMT has been restricted by the scarcity of parallel corpora for a long time. In MNMT, some simple DA techniques, such as insertion, deletion, modification, and other methods to expand the parallel corpus, often result in a large semantic difference between the generated enhanced sentences and the original sentences, and the diversity of the generated training samples is limited. Ji et al. [21] adopted a new preprocessing technique of generative adversarial networks and mixed morphological noise for the problem of unknown words in Mongolian–Chinese NMT. Hei et al. [22] proposed a word segmentation method based on BERT DA, significantly improving the performance of Mongolian–Chinese NMT. Although some studies have been devoted to improving the quality of Mongolian–Chinese NMT, there are few studies on the DA of Mongolian–Chinese materials, especially on the semantic DA in Mongolian–Chinese NMT.

Therefore, in response to the scarcity of corpora and semantic inconsistency in Mongolian–Chinese NMT, we propose a method based on semantic-context DA, which does not generate explicit training samples, but uses semantic vector fusion to directly integrate enhancement into the training process. This method can significantly improve the quality of MNMT. Specifically, we initially trained a semantic encoder using contrastive learning to map source and target sentences into the same semantic space. Post training optimization, we sampled K source sentence vectors from the semantic space where source and target sentences intersect, and then fused them with the vectors output by the original encoder. Finally, the fused semantic rich vector was input into the decoder for decoding operation. We evaluated our framework on Mongolian–Chinese translation tasks, which is a much better method than using the baseline model.

The main contributions of this paper are as follows:

(1) Aiming to address the problem of data sparsity and missing semantics in MNMT, this paper proposes a semantic-context DA technique and use different DA methods to generate pseudo-parallel corpora.

(2) We train an additional semantic encoder using the method of contrastive learning to extract the semantic information of the source and target in the Mongolian–Chinese NMT process.

(3) The model proposed in this paper shows significant performance on both the Mongolian and Chinese low-resource datasets as well as the datasets generated through DA.

2. Semantic-Context Data Augmentation

Semantic-context DA is a technique that merges source and target data for DA. It can simultaneously utilize the semantic information of the source and target during training to generate diverse semantic information-rich vectors. These vectors can be fused with the vector generated by the encoder and then translated to improve translation performance. This method augments the original sentences without requiring additional training samples. A shared semantic space exists between the source and target languages of NMT, established by a semantic encoder, which contains many vector variants with the same semantics as the original training sample. This method can alleviate the problems of data sparsity and semantic inconsistency in Mongolian–Chinese NMT.

Therefore, for data augmentation within the semantic space, this paper utilizes Transformer as the primary network and the pre-trained model XLM-R (Cross-lingual Language Model with RoBERTa architecture) [23] as an additional semantic encoder. The XLM-R model is based on the RoBERTa (Robustly optimized BERT approach) architecture, but in essence, the XLM-R model is a deep learning model based on Transformer's self-attention mechanism. By pre-training the model on large-scale text data, the semantic representation and features learned by the XLM-R model can capture the semantic information and language structure in the text. This general semantic representation is beneficial for various downstream natural language-processing tasks, because it provides a general representation that can be applied to different tasks and fields. A specific structure diagram of XLM-R is shown in Figure 1. XLM-R is used as a semantic encoder, which can dynamically allocate attention weights according to different positions and semantic relationships in the input sequence. At the same time, the model can better understand the semantic content of the input text, and take into account the relationships between and importance of words in the text when encoding, resulting in more accurate and rich semantic coding.

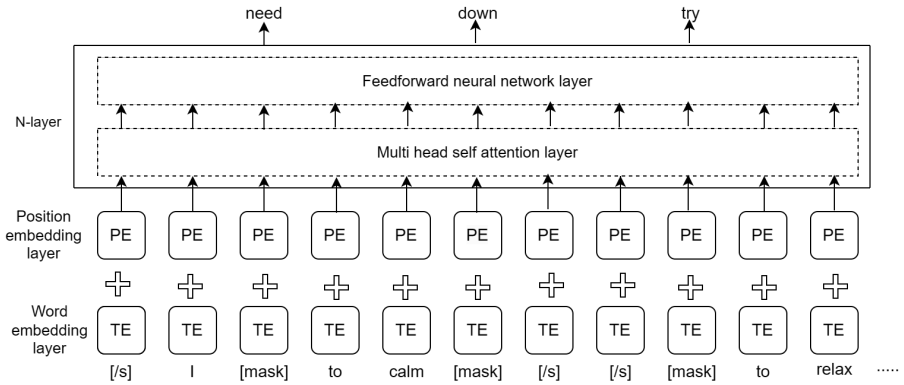


Figure 1. XLM-R structure diagram.

The pre-trained semantic encoder is proficient in capturing both grammar and semantic information within the text. Within the semantic encoder, a forward function $\phi(. : \theta^{'})$ parameterized by $\theta^{'}$ is defined, enabling the mapping of discrete sentences into continuous vector representations [19]. This semantic encoder aims to convert the source sentence x and the target sentence y into real-valued vectors, $r_x = \phi(x : \theta^{'})$ and $r_y = \phi(y : \theta^{'})$, respectively. By mapping the semantic information of Mongolian and Chinese into the same vector space, the encoder achieves semantic representation and comprehension of these two languages. For the generation of semantic regions, the following method is used. Assuming $(x^{(i)}, y^{(i)})$ and $(x^{(j)}, y^{(j)})$ are two randomly selected training instances from the training corpus, the adjacent semantic region $v(r_{x^{(i)}}, r_{y^{(i)}})$ is defined as the union of two closed balls centered on $r_{x^{(i)}}$ and $r_{y^{(i)}}$, respectively. The radius of the two balls is $d = ||r_{x^{(i)}} - r_{y^{(i)}}||_2$, which is the Euclidean distance between $r_{x^{(i)}}$ and $r_{y^{(i)}}$. d is also considered a relaxation variable for determining semantic equivalence, meaning that a vector whose distance from $r_{x^{(i)}}$ or $r_{y^{(i)}}$ does not exceed d is semantically equivalent to $r_{x^{(i)}}$ or $r_{y^{(i)}}$.

Using the method of comparative learning to optimize the semantic space range, each vector generates a semantically similar region in a continuous space, that is, an adjacent semantic region that contains variants of the same semantic representation as the original sentence. The adjacent semantic region $v(r_x, r_y)$ in the semantic space contains multiple semantic variants of the described sentence pairs (x, y), and contains many vector representations of similar semantics in Mongolian and Chinese. We construct negative samples in comparative learning by applying convex interpolation between the current instance and other instances in the same training batch for instance comparison. The negative sample and design are as follows:

$$r_{x'(j)} = r_{x^{(i)}} + \lambda_x \left(r_{x^{(j)}} - r_{x^{(i)}}\right), \lambda_x \in \left(\frac{d}{d'_x}, 1\right] \tag{1}$$

$$r_{y'(j)} = r_{y^{(i)}} + \lambda_y \left(r_{y^{(j)}} - r_{y^{(i)}}\right), \lambda_y \in \left(\frac{d}{d'_y}, 1\right] \tag{2}$$

Here, $d'_x = ||r_{x^{(i)}} - r_{x^{(j)}}||_2$, $d'_y = ||r_{y^{(i)}} - r_{y^{(j)}}||_2$ in the two formulas above, where d'_x and d'_y are numbers larger than d, or $r_{x'(j)} = r_{x^{(j)}}$ and $r_{y'(j)} = r_{y^{(j)}}$.

Following this design, through the interpolation of multiple instances within the same training batch, the adjacent semantic region of the i-th training instance can be comprehensively established. We dynamically adjust the values of λ_x (or λ_y) during the training process, as referenced in [24]. We optimize this vector space using contrastive learning methods and sample the source language vectors within the vector space. The sampling method employs normal distribution sampling to generate random semantic representations, which are subsequently fused with the original training samples to enhance the contextual semantic data of the training set.

When conducting comparative learning, we generate a semantic space with more similar semantic information by optimizing the loss of the difference between the semantic similarity between the negative sample and the original sentence vector and the semantic similarity between the positive sample and the original sentence vector. $s(x, y)$ represents the similarity function between the positive and negative samples, thereby deriving the loss function for contrastive learning, as depicted in Formula (3):

$$Loss_{CL}(x^{(i)}) = max(s(r_{x'(j)}, r_{x^{(i)}}) - s(r_{y^{(i)}}, r_{x^{(i)}}) + \eta, 0) \tag{3}$$

Similarly, we also expect r_{y^i} and r_{x^i} to be as similar as possible, but different from all other instances in the same training batch. Therefore, $Loss_{CL}(y^{(i)})$, as shown in Formula (4):

$$Loss_{CL}(y^{(i)}) = max(s(r_{y'(j)}, r_{y^{(i)}}) - s(r_{x^{(i)}}, r_{y^{(i)}}) + \eta, 0) \tag{4}$$

The comparative loss of training batches can be expressed as

$$Loss_{CL} = Loss_{x^{(i)}} + Loss_{y^{(i)}} \tag{5}$$

The loss function is established based on the Max-Margin Loss Function, where the constant η represents the margin. Specifically, the aim is to enhance the similarity score of positive samples while diminishing the similarity score of negative samples. In essence, η represents the maximum allowable difference between the two scores; any further increase in the difference would not yield any reward, thereby ensuring the model's generalization ability. Moreover, if the similarity score of negative examples is less than the constant η, the loss function is adjusted to bring their distance closer to η. The loss function for the translation model is the cross-entropy loss function, depicted in Formula (6):

$$Loss_{trans} = CrossEntropy(y, \hat{y}) \tag{6}$$

When integrating semantic-context DA models with translation models, the fusion loss function is

$$TotalLoss = w_1 * Loss_{CL} + w_2 * Loss_{trans} \quad (7)$$

where $w1$ and $w2$ are the weights assigned to $Loss_{CL}$ and $Loss_{trans}$, respectively. These weights can be arbitrary positive numbers, and their sum does not have to equal 1. Their values depend on the relative importance you assign to each loss function during training. To select suitable weights, it is common to conduct experiments and adjustments. One can try different weight combinations and then assess the model's performance using a validation set or cross-validation to find the optimal weight configuration. In the optimized adjacent semantic region, the normal distribution sampling method is used to sample $K(r_1, r_2 \ldots r_k)$ source language vectors in the position where the source language semantic space coincides with the target language semantic space. The vector sampled by this method contains a representation of source language semantic information and target language semantic information, which makes the semantics richer. Among them, $r_{(k)} \epsilon v(r_x, r_y)$, K are hyperparameters that determine the number of sampling vectors. The result of normal distribution sampling will yield a vector within a certain range of values, allowing the generation of new samples with a certain degree of continuity. These sampled semantic vectors are combined with the semantic vectors generated by the Transformer encoder. The resulting merged semantically enriched vectors are then fed into the decoder for decoding operations. The fusion formula is

$$r_z = (r + r_1 + r_2 + \ldots + r_k)/(k + 1) \quad (8)$$

3. Mongolian–Chinese Translation with Semantic-Context Data Augmentation Technique

This paper combines the architecture approach of the Transformer model with semantic-context data augmentation, and the model is named Sem-NMT. The specific structure is shown in Figure 2.

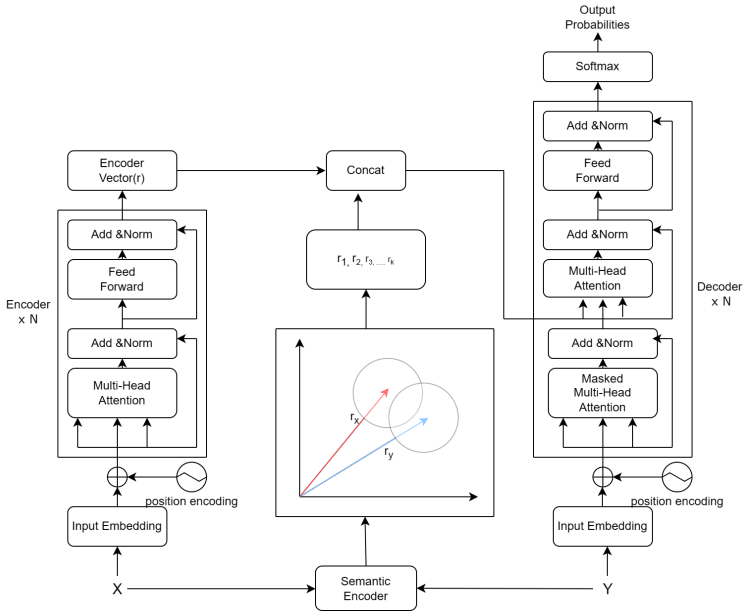


Figure 2. Semanticcontext data enhancement structure figure.

The Transformer model consists of multiple encoders and decoders. When the Transformer model is working, the encoder encodes the input sentence and transfers the encoded information to the decoder. The decoder translates Mongolian into Chinese based on encoding information. Each encoder in this model establishes communication with all decoders, allowing each decoder to access the necessary information. The internal structure of each encoder layer on the encoder side of the model remains identical, with the self-attention

layer being the most crucial component. During translation, the self-attention layer serves to guide the model on which words to prioritize, calculating the attention value for each word. The processing of the self-attention layer can better extract the features of the data. The calculation formula of the self-attention mechanism is as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^{\top}}{\sqrt{d_k}})V \tag{9}$$

where Q, K, and V are obtained by multiplying three different parameter matrices, W^Q, W^K, W^V, with the embedding vectors of the words. These three vectors describe the information contained in the vocabulary from different perspectives. In this paper, $d_k = 64$ and $\sqrt{d_k} = 8$. Both the encoder and decoder sides of this model employ a multi-head attention mechanism. After embedding the words, a word can be input into different self-attention layers, producing multiple distinct self-attention output values. Subsequently, these values are concatenated, and a parameter matrix of appropriate dimensions is used to reduce the dimension of the concatenated self-attention output. Finally, they are fed into a fully connected layer for further computation. Through this operation, we can obtain semantically rich vectors, the structure of which is shown in Figure 3.

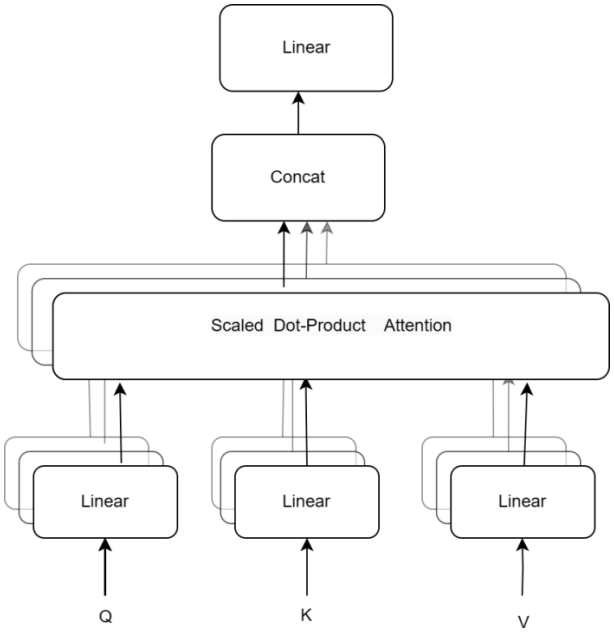


Figure 3. Multi-headed self-attention mechanism.

However, the multi-head self-attention layer alone cannot capture positional information within Mongolian sentences. This article employs positional encoding to capture structural information within sentences. The formula for positional encoding is as follows:

$$PE(POS, 2i) = sin(\frac{pos}{1000^{\frac{2i}{d_{model}}}}) \tag{10}$$

$$PE(POS, 2i+1) = cos(\frac{pos}{1000^{\frac{2i}{d_{model}}}}) \tag{11}$$

In the formula, d_{model} is the dimension of the embedding vector for Mongolian words, and i is the component index of the embedding vector for Mongolian words, with a value range of 1 to d_{model}. pos is the positional index of words in Mongolian. Formulas (10) and (11), respectively, encode the parity component index in the Mongolian word embedding vector. The utilization of trigonometric functions is attributed to their unique properties, enabling the transformation of Mongolian word encoding from the right side of the sentence to

reflect linear changes in the word encoding on the left side of the sentence. This coding method can capture the absolute position information and relative position information of Mongolian words. Figure 4 illustrates the utilization of positional encoding for Mongolian words. This involves acquiring the positional encoding of Mongolian words, adding it to the corresponding term of the word embedding vector, and subsequently inputting it into the translation model for training.

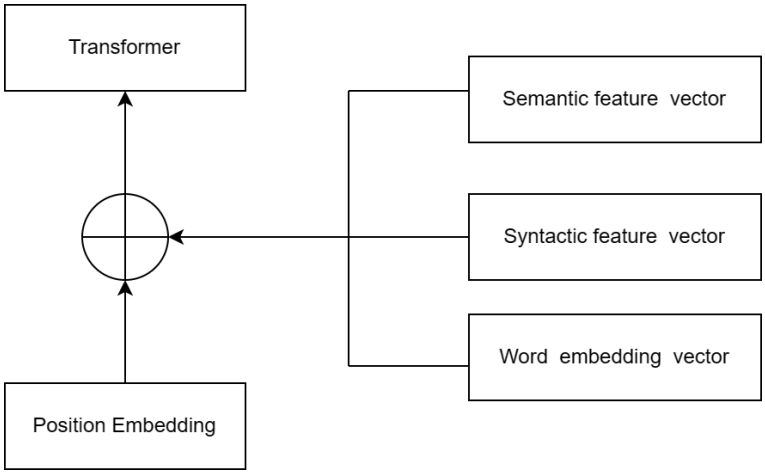


Figure 4. The application of word embedding vector and position information.

Mongolian is a low-resource language with relatively few parallel corpora. Nowadays, machine translation requires a large number of corpora to train an excellent translation model. For Mongolian, we need to use DA methods to improve the translation performance of the model. Aiming to address the problem of data sparseness and semantic inconsistency in Mongolian–Chinese neural machine translation, this paper adds a semantic encoder to extract semantic information from training samples on the basis of the Transformer translation model. Specifically, this paper maps Mongolian and Chinese to the same semantic space through a semantic encoder, and optimizes the semantic space by means of comparative learning. The vector in the semantic space is sampled and fused with the vector in the Transformer encoder and sent to the decoder for decoding operation. The fused vector simultaneously incorporates the semantic information from both the source and target. By merging multiple vectors, the diversity of semantic information is augmented, following which the enhanced vector is directed to the decoder for the decoding process. The unique architecture of the Transformer decoder allows for the extraction of Mongolian information from various perspectives. It offers diverse Mongolian information to the decoder through its distinctive interactions.

4. Experiments

This section initially presents the bilingual parallel corpus dataset utilized in this paper, along with the evaluation metrics employed in this study. Subsequently, it outlines the experimental settings where Sem-NMT is applied to the Mongolian–Chinese translation task and other low-resource languages. Finally, it provides the experimental results alongside the corresponding analysis.

4.1. Experimental Data and Evaluation Indicators

Before the translation task, we augmented the Mongolian–Chinese parallel corpus to generate a pseudo-parallel corpus. Then, we merged this pseudo-parallel corpus with the original Mongolian–Chinese corpus to conduct research on Mongolian–Chinese NMT. The dataset of Mongolian–Chinese NMT we used was a 30 w Mongolian–Chinese parallel corpus selected from the school laboratory and a 270 w Mongolian–Chinese pseudo-parallel corpus generated by a DA method. The DA methods employed in this paper included simple augmentation, back-translation, and iterative back-translation. The simple

data augmentation techniques involved random exchange, random insertion, synonym replacement, and random deletion. When conducting a simple DA, attention must be given to potential issues such as the number of operations k for random exchange, random insertion, and synonym replacement in a real Chinese sentence. Additionally, the setting of the operation word p in the random deletion operation should be considered. Here, we let $k = p * l$, where l is the length of the sentence, so that the number of operations can be dynamically adjusted according to the length of the sentence to ensure higher quality of the generated pseudo-sentences. In the Mongolian–Chinese translation task, we set the ratio of operation words to 0.1, and the relevant experiments are shown in Figure 5. The number of new corpora constructed was set to n = 1. The cosine similarity calculation method was used to calculate the similarity when the synonym is replaced, and the threshold was set to 0.8. The BPE (Byte Pair Encoding) word segmentation method dynamically generates vocabulary through character-based merging operations, thereby solving some limitations of traditional word segmentation methods such as fixed vocabulary size and sparse vocabulary. It is an important word segmentation method in low-resource-language processing. Therefore, this paper adopted the BPE word segmentation method for text. The data are shown in Table 1.

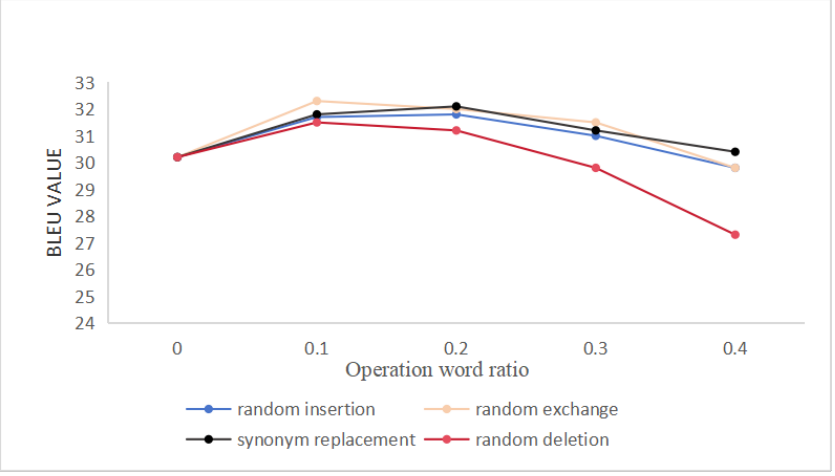


Figure 5. Determination of operation-word ratio.

Table 1. The scale of bilingual parallel corpora used in the experiment.

Corpus Type	Training Set	Validation Set	Test Set
Initial corpus	30 w	2000	2000
Easy data augmentation	60 w	4000	4000
Back-translation	60 w	4000	4000
Iterative back-translation	IT1(Zh-Mn): 60 w IT2(Mn-Zh): 90 w IT3(Zh-Mn): 120 w IT4(Mn-Zh): 150 w	5000	5000

For the evaluation index, this paper adopted the widely used automatic machine translation evaluation method and the bilingual translation quality evaluation auxiliary tool BLEU (Bilingual Evaluation Understudy). BLEU is an automated indicator for evaluating the quality of machine translation. It is evaluated by comparing the n-gram (consecutive n words) overlap between the system-generated translation and the reference translation. The higher the BLEU value, the more overlap between the translation generated by the system and the reference translation, and the better the translation quality.

4.2. Experimental Settings

We divided the training process into two parts. The aim of the first part was to train a single semantic encoder for semantic-context data augmentation. The aim of the second part was to train a Transformer NMT model with semantic data augmentation.

The purpose of the first part was to use the XLM-R pre-training model as our semantic encoder. XLM-R is a language model based on the RoBERTa architecture, which has strong semantic coding ability. By pre-training the model on a large-scale multilingual corpus, XLM-R can learn common semantic representations and features. As a semantic encoder, XLM-R can convert the input text into a semantically rich vector representation, which contains the semantic information and context of the input text. In our experiment, we chose the XLM-R model as our semantic encoder and evaluated its performance on our task. It can map the text of different languages into a shared semantic space. Here, we used the semantic encoder to map Mongolian and Chinese into a semantic space and used the method of comparative learning to optimize the semantic space. The parameters were set using the parameters of the XLM-R Base. The number of encoder layers was 12, the number of attention heads was 12, and the number of hidden units was 768.

The purpose of the second part was to train an NMT model with DA. Sampling semantic vectors from the optimized semantic space and merging these vectors with the semantic vectors output by the Transformer encoder, the fused semantic vectors were then fed into the decoder for decoding operations. The translation model uses the basic Transformer architecture [25]. The basic Transformer architecture is a deep learning model based on a self-attention mechanism for sequence-to-sequence learning tasks. It consists of an encoder and a decoder, where the encoder is responsible for encoding the input sequence into a context-sensitive representation. The self-attention mechanism of the Transformer architecture can capture the semantic information and context relationship in the input sequence, resulting in rich semantic coding. Specifically, both the encoder and decoder are composed of six blocks. The dimensions of the embedded sub-layer and the feedforward sub-layer were set to 512 and 1024, respectively. The number of attention heads was set to 4. The default learning rate was 0.003.

4.3. Experimental Results and Analysis

As shown in Table 2, the table presents the results of various methods in the Mongolian–Chinese translation task. From the data in the table, it can be concluded that the semantic-context data enhancement method proposed by us has a better translation effect on multiple different datasets than the existing DA strategies, such as back-translation [7], soft context data enhancement [4], SwitchOut [26], etc.

Table 2. BLEU of translation results.

Model	Initial Corpus	Easy Data Augmentation	Back-Translation	Iterative Back-Translation
Transformer	30.2	30.8	31.1	31.6
SwitchOut	31.3	31.6	31.9	32.4
SemAug	30.78	31.16	31.62	31.56
SCA	31.02	31.17	32.03	32.17
Sem-NMT (ours)	**32.73**	**33.15**	**33.57**	**34.55**

The experimental results also show the importance of large-scale training data for MNMT. We observed that using simple DA, back-translation, and iterative back-translation DA methods and our proposed semantic-context DA method significantly improved translation quality. However, the method proposed in this paper generally achieved the best results and outperformed the other DA methods in the testing set. Because our method uses both the source and target sentence vectors in the translation process, we can obtain a

vector representation with rich semantic information. By using our method and the existing method in the pseudo-parallel corpus generated by different DA methods, it is obvious that with the expansion of the corpus scale, the translation effect is constantly improving. Due to its direct operation of sentences, simple data augmentation easily leads to issues of missing semantics. Compared with the back-translation and iterative back-translation methods, the translation results on the pseudo-parallel corpus generated by simple DA are poor. In general, the expansion of corpus size can effectively improve the quality of Mongolian–Chinese translation tasks.

The main feature of Mongolian syntax is that the subject and object come before the predicate verb, and their order is not fixed. This syntactic structure often leads to structural confusion when translating Mongolian into other languages. As shown in Figure 6, several typical representative sentences were selected to translate Mongolian into Chinese using the Sem-NMT model. From the translation result of the first sentence, it can be seen that when translating ordinary sentences, the translated sentences are roughly the same as the reference translation. From the second translation result, it can be seen that passive sentences can easily be translated into content that is contrary to the original semantics in MNMT. This is because Mongolian grammar takes the subject–object–verb structure, while Chinese takes the subject–verb–object structure, which is not very conducive to translating complex sentences. In future research, the translation model also needs to be further improved to account for the corresponding grammatical structure of Mongolian and Chinese. Through the analysis of the translation results of Mongolian and Chinese, the translation model can effectively alleviate the problem of unknown words, and the translation effect of some place names and specific nouns is particularly prominent.

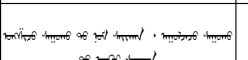
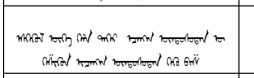
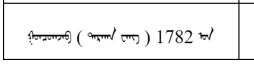
Mongolian sentences	Translate sentences	Reference translation
	Books are good when you sit and read; friends are good when you meet and get together.	Books are good when you're sitting down to read; friends are good when you're meeting them.
	Members of linguistic minorities are members of minorities	Members of linguistic minorities are not considered to be minorities
	Gongqiqzab (King of Daiqing) 1782	Gongqiqzab (King of Daiqing) 1782

Figure 6. Translation sample.

Figure 7 show the changes in the loss and accuracy of the Sem-NMT model during training. From this graph, it can be seen that in the early stages of model training, the loss function decreases rapidly, and the accuracy of translation improves rapidly. With the increase in training time, the change in loss also increases, but the increase is small, and finally tends to stabilize. As the training time increases, the accuracy of model translation also continuously improves and tends to stabilize when reaching a certain value. Based on the changing trends of these two graphs, it can be seen that our proposed method is stable and effective in overall training. Compared to the basic Transformer model, the Sem-NMT model has improved convergence speed and accuracy to a certain extent, indicating that the Sem-NMT data augmentation model proposed in this paper can significantly improve the quality of Mongolian–Chinese neural machine translation.

In order to explore the translation performance of this method in other low-resource languages, we obtained publicly available datasets from the internet. The Vietnamese–Chinese dataset was selected from CWMT2020, and 30 w sentences were extracted for training, while the Tibetan–Chinese dataset was selected from the Opus corpus [27] and pre-processed using the method of OpusFilter [28]. Then, 30 w sentences of the Tibetan–Chinese parallel corpus were extracted as the training set. Table 3 shows the specific translation results.

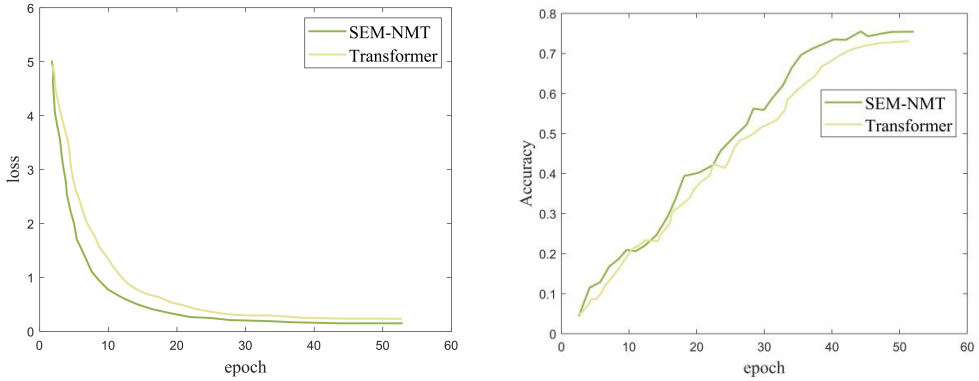


Figure 7. Changes in loss and accuracy during the translation process.

Table 3. Translation results of Sem-NMT on different datasets.

Model	Mn-Ch	Vi-Ch	Ti-Ch
Transformer	30.26	31.4	19.98
SwitchOut	31.32	29.38	20.34
SemAug	30.78	29.14	20.51
SCA	31.02	28.77	20.87
Sem-NMT (ours)	32.73	32.15	21.78

From the above translation results, it can be seen that the Sem-NMT data augmentation model proposed in this paper can not only achieve good translation results in Mongolian–Chinese neural machine translation tasks, but also increase the BLEU values by 0.8 and 1.8 in low-resource Vietnamese–Chinese and Tibetan–Chinese translation tasks, respectively. So, the Sem-NMT data augmentation model proposed in this paper can effectively improve the quality of low-resource machine translation.

In summary, the semantic-context DA method we proposed can achieve good results in the Mongolian–Chinese NMT task. It improves the BLEU value by 2.5 for the basic Transformer model. Compared to the basic model, this method can achieve the same translation results with about half of the data, greatly improving translation efficiency. In addition, by combining simple DA, backtracking, and iterative backtracking methods on this basis, the translation quality has also been significantly improved.

5. Conclusions

In the context of the robust evolution of data augmentation methods in NMT, this study explores the data augmentation challenges within MNMT. Aiming to solve the problem of data sparseness and semantic inconsistency in MNMT, this paper proposes an MNMT method based on semantic-context DA. This method adds an additional semantic encoder to the Transformer translation model to extract the semantic information of the source and target languages. This study maps Mongolian and Chinese into a shared semantic space by semantic encoding, and applies comparative learning to optimize the space that contains rich semantic vector variants. Subsequently, a semantic vector is sampled from this space, fused with the vector generated by the encoder, and then fed into the decoder for translation. This DA method can use the semantic information of both the source and target to improve translation performance during the training process. Compared with the baseline model, the data augmentation method proposed in this paper significantly improved the translation results, especially in the translation of low-frequency words such as place names and names, and its effectiveness has also been verified in other low-resource datasets.

While the use of semantic-context data augmentation methods to generate diverse training samples alleviates the issue of resource scarcity in MNMT, there still exists a discernible gap between the generated samples and the original training data. In future endeavors, we aim to further explore methods to optimize data augmentation techniques, striving to enhance the quality of MNMT, specifically focusing on the issues of grammar and semantic alignment between Mongolian and Chinese.

Author Contributions: Conceptualization, N.W. and M.L.; Methodology, H.Z., Y.J. and N.W.; Validation, H.Z.; Resources, N.W.; Writing—original draft, H.Z.; Writing—review & editing, Y.J. and M.L.; Supervision, Y.J. and M.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China grant number 62066035, 62206138 and Universities Directly Under the Autonomous Region Funded by the Fundamental Research Fund Project grant number JY20220122, JY20220089, RZ2300001739, RZ2300001743, JY20220186 and Research program of science and technology at Universities of Inner Mongolia Autonomous Region grant number NJZZ22251, NJZZ23081 and Science Research Foundation of Inner Mongolia University of Technology grant number BS2021079, ZZ202118, DC2300001261, DC2300001258, DC2300001262. The APC was funded by Universities Directly Under the Autonomous Region Funded by the Fundamental Research Fund Project grant number JY20220186.

Data Availability Statement: Restrictions apply to the availability of these data. Data were obtained from the Artificial Intelligence Laboratory, School of Information Engineering, Inner Mongolia University of Technology, and are available from the authors with permission from the Artificial Intelligence Laboratory, School of Information Engineering, Inner Mongolia University of Technology.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mikołajczyk, A.; Grochowski, M. Data augmentation for improving deep learning in image classification problem. In Proceedings of the 2018 International Interdisciplinary PhD Workshop (IIPhDW), Swinoujscie, Poland, 9–12 May 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 117–122.
2. Fadaee, M.; Bisazza, A.; Monz, C. Data augmentation for low-resource neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vancouver, BC, Canada, 30 July–4 August 2017; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 567–573.
3. Xia, M.; Kong, X.; Anastasopoulos, A.; Neubig, G. Generalized data augmentation for low-resource translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 5786–5796.
4. Gao, F.; Zhu, J.; Wu, L.; Xia, Y.; Qin, T.; Cheng, X.; Zhou, W.; Liu, T.Y. Soft contextual data augmentation for neural machine translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 5539–5544.
5. Zhou, C.; Ma, X.; Hu, J.; Neubig, G. Handling syntactic divergence in low-resource machine translation. 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), Hong Kong, China, 3–7 November 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 1388–1394.
6. Liu, Q.; Kusner, M.; Blunsom, P. Counterfactual data augmentation for neural machine translation. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021; pp. 187–197.
7. Sennrich, R.; Haddow, B.; Birch, A. 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), Berlin, Germany, 7–12 August 2016; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 86–96.
8. Sugiyama, A.; Yoshinaga, N. Data augmentation using back-translation for context-aware neural machine translation. In Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019), Hong Kong, China, 3 November 2019; pp. 35–44.
9. Caswell, I.; Chelba, C.; Grangier, D. Tagged back-translation. In Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers), Florence, Italy, 1–2 August 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 53–63.
10. Edunov, S.; Ott, M.; Auli, M.; Grangier, D. Understanding back-translation at scale. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 489–500.

11. Wu, J.; Wang, X.; Wang, W.Y. Extract and edit: An alternative to back-translation for unsupervised neural machine translation. 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), Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 1173–1183.
12. Jiao, R.; Yang, Z.; Sun, M.; Liu, Y. Alternated training with synthetic and authentic data for neural machine translation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online Event, 1–6 August 2021; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 1828–1834.
13. Wu, L.; Wang, Y.; Xia, Y.; Qin, T.; Lai, J.; Liu, T.Y. Exploiting Monolingual Data at Scale for Neural Machine Translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Hong Kong, China, 3–7 November 2019; pp. 4207–4216.
14. Zhang, J.; Zong, C. Exploiting source-side monolingual data in neural machine translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 1535–1545.
15. Abdulmumin, I.; Galadanci, B.S.; Isa, A. Enhanced back-translation for low resource neural machine translation using self-training. In Proceedings of the Information and Communication Technology and Applications: Third International Conference, ICTA 2020, Minna, Nigeria, 24–27 November 2020; Revised Selected Papers 3; Springer: Berlin/Heidelberg, Germany, 2021; pp. 355–371.
16. Hoang, C.D.V.; Koehn, P.; Haffari, G.; Cohn, T. Iterative back-translation for neural machine translation. In Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, Melbourne, Australia, 20 July 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 18–24.
17. Yang, Z.; Chen, W.; Wang, F.; Xu, B. Improving neural machine translation with conditional sequence generative adversarial nets. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, LA, USA, 1–6 June 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 1346–1355.
18. Zhang, D.; Luo, M.; He, F. Reconstructed similarity for faster GANs-based word translation to mitigate hubness. *Neurocomputing* **2019**, *362*, 83–93. [CrossRef]
19. Wei, X.; Yu, H.; Hu, Y.; Weng, R.; Luo, W.; Xie, J.; Jin, R. Learning to generalize to more: Continuous semantic augmentation for neural machine translation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 22–27 May 2022; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 7930–7944.
20. Cao, Y.; Gao, Y.; Li, M.; Feng, T.; Jingru, W.; Sha, F. Research on Mongolian Chinese neural machine translation based on monolingual corpus and word vector alignment. *J. Chin. Inf. Process.* **2020**, *34*, 27–32.
21. Ji, Y.; Hou, H.; Chen, J.; Wu, N. Adversarial training for unknown word problems in neural machine translation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process. (TALLIP)* **2019**, *19*, 1–12. [CrossRef]
22. He, W.; Xiu, Z.; Bao, J.; Chen, M.; Wangsi, R. The Mongolian-to-Chinese neural machine translation system, incorporating BERT-based data augmentation for word segmentation. *J. Xiamen Univ.* **2022**, *12*, 061.
23. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 8440–8451.
24. Wei, X.; Weng, R.; Hu, Y.; Xing, L.; Yu, H.; Luo, W. On learning universal representations across languages. *arXiv* **2020**, arXiv:2007.15960.
25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
26. Wang, X.; Pham, H.; Dai, Z.; Neubig, G. SwitchOut: an efficient data augmentation algorithm for neural machine translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 856–861.
27. Tiedemann, J. Parallel data, tools and interfaces in OPUS. In Proceedings of the Lrec. Citeseer, Istanbul, Turkey, 23–25 May 2012; Volume 2012, pp. 2214–2218.
28. Aulamo, M.; Virpioja, S.; Tiedemann, J. OpusFilter: A configurable parallel corpus filtering toolbox. In Proceedings of the 2020 Annual Conference of the Association for Computational Linguistics, Online, 5–10 July 2020; The Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 150–156.