Research Article

Jinfeng Xue*

Machine translation of English content: A comparative study of different methods

https://doi.org/10.1515/jisys-2021-0150
received June 27, 2021; accepted August 02, 2021

Abstract: Based on neural machine translation, this article introduced the ConvS2S system and transformer system, designed a semantic sharing combined transformer system to improve translation quality, and compared the three systems on the NIST dataset. The results showed that the operation speed of the semantic sharing combined transformer system was the highest, reaching 3934.27 words per second; the BLEU value of the ConvS2S system was the smallest, followed by the transformer system and the semantic sharing combined transformer system. Taking NIST08 as an example, the BLEU values of the designed system were 4.74 and 1.49 higher than the other two systems. The analysis of examples showed that the semantic sharing combined transformer had higher translation quality. The experimental results show that the proposed system is reliable in English content translation and can be further promoted and applied in practice.

Keywords: English, machine translation, transformer system, semantic sharing, ConvS2S system

1 Introduction

Natural language processing (NLP) mainly studies [1] how to realize effective communication between humans and computers through natural language, which has been widely used in machine translation (MT) [2], public opinion monitoring [3], text classification [4], etc. The process of MT can be interpreted as decoding the source corpus and re-coding it into the target language. It is necessary to have a deep understanding of the grammar and semantics of the language to ensure high-quality MT. Neural machine translation (NMT) is one kind of MT [5]. Choi et al. [6] contextualized the word embedding vector using a nonlinear bag-of-words representation of the source sentence and used typed symbols to represent special tokens, such as numbers, proper nouns, and acronyms. Experiments on En-Fr and En-De showed that the method could significantly improve the quality of translation. Wu et al. [7] pointed out the importance of grammar knowledge for translation performance, designed a grammar-aware encoder, and incorporated it into NMT. Through experiments, they found that the method could improve the quality of translation. Lee et al. [8] introduced an NMT model, which mapped a source character sequence to a target character sequence without any segmentation. They used a character-level convolutional network with max-pooling at the encoder to reduce the length of source representation to allow the model to be trained at a speed comparable to subword-level models while capturing local regularities. The experiment found that the model showed better performance. Gu et al. [9] proposed a new MT method for languages with limited parallel data. It used the shift learning method to share multiple source languages as a target language and share the source encoder with other languages. The experiment found that the method could realize 23 BLEU on Romanian-English WMT2016 using a tiny parallel corpus of 6k sentences. Translation between English and Chinese has always been a difficult problem in MT. English belongs to the Germanic language family, which has a morphologic form and relatively

* **Corresponding author: Jinfeng Xue,** College of Petroleum Equipment and Electronic Engineering, Dongying Vocational Institute, No. 129, Dongcheng Fuqian Street, Dongying, Shandong 257091, China, e-mail: f5x7z9@126.com

fixed word order. Chinese belongs to the Sino-Tibetan language family, which expresses grammatical relations through word order and function words. To be specific, English focuses on cohesion [10] and has close sentence structure; Chinese that pays attention to coherence [11] has a relatively loose sentence structure, and it is often necessary to combine the context to understand the meaning of a sentence. In addition, there are great differences in thinking and culture between Chinese and English, which also leads to the poor quality of translation results. Therefore, this study mainly analyzed the MT method of English content. Taking NMT as the subject, this study compared the translation performance of three different NMT methods. This work makes some contributions to the realization of better translation between English and Chinese.

2 Different neural MT methods

2.1 ConvS2S system

Convolutional neural network (CNN) has the characteristics of weight sharing, downsampling, etc. [12]. It has significant advantages in image processing [13] and has been widely used in the NLP field [14], such as semantic analysis [15] and language model [16]. It can also be used in MT. In the ConvS2S system, each layer in the decoder contains an attention module, and sublayers are connected by residuals. The calculation formula is as follows:

$$h^l = h^{l-1} + \text{sublayer}(h^{l-1}), \tag{1}$$

where h^l is the output of the lth sublayer, sublayer refers to the function performance of the layer, which is realized by CNN. It is assumed that the weight of every convolution kernel is w^l and the bias is b_w. The outputs of k words are merged:

$$X = \left[h^l_{i-\frac{k}{2}};\ldots,\ h^l_{i+\frac{k}{2}}\right]. \tag{2}$$

It is mapped as an output element:

$$Y = w^l X + h^l_w; \tag{3}$$

then, the output of the lth sublayer can be written as:

$$h_i^l = h_i^{l-1} + v(Y), \tag{4}$$

where $v(Y)$ refers to the function performance of convolution operation and v refers to the gated linear unit (GLU). For the input matrix

$$Y = [A; B], \tag{5}$$

its operation process is as follows:

$$V([Y])A \otimes \sigma(B), \tag{6}$$

where A and B are inputs of GLU, $\otimes$ is the multiplication of the corresponding elements of the matrix, and σ is a nonlinear activation function.

2.2 Transformer system

The transformer system [17] abandons the recurrent neural network and uses a self-attention mechanism [18], generating three vectors, q, k, and v, from the output word vector. The calculation method of self-attention mechanism is as follows:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{7}$$

where $Q, K,$ and V are matrices of three vectors and d_k is the dimension of k.

The source language sequence is set as

$$X = (x_1, x_2, \ldots, x_m), \tag{8}$$

and the target language sequence is set as

$$Y = (y_1, y_2, \ldots, y_m). \tag{9}$$

In the transformer system, the self-attention mechanism is realized by the multi-head attention module, and the calculation formulas are as follows:

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)w^O, \tag{10}$$

$$\mathrm{head}_i = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V). \tag{11}$$

The output generated from the multi-head attention module enters the forward feedback neural network (FFN) [19] to generate the output of the encoder:

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2, \tag{12}$$

where W_1 and W_2 are the weights at the first and second mappings and b_1 and b_2 are the bias at the first and second mappings.

In the system, time sequence information is obtained by position coding:

$$\mathrm{PE}_{(\mathrm{pos},2i)} = \sin(\mathrm{pos}/10{,}000^{2i/d_{\mathrm{model}}}), \tag{13}$$

$$\mathrm{PE}_{(\mathrm{pos},2i+1)} = \cos(\mathrm{pos}/10{,}000^{2i/d_{\mathrm{model}}}), \tag{14}$$

where d_{model} refers to the size of the model dimension.

The decoder part of the system is the same as the encoder, and a softmax layer is added at the end. The final output of the system is the probability distribution of candidate target words. The cross-entropy function is used for training, and the optimizer is Adam.

2.3 Transformer system combined with semantic sharing

NMT can be regarded as a model of transformation between two semantic spaces. If it can be combined with the semantic representation space of cross-language sharing, the semantic relevance of model translation results can be improved. Therefore, this article optimizes the transformer system with semantic sharing, including parameter sharing and representation sharing. First, in the process of training translation tasks, the same parameters are shared. It is assumed that the two languages to be translated are X and Y. The loss function is optimized as follows:

$$\begin{aligned}\tau^{\mathrm{total}}(\theta_{\mathrm{enc}}, \theta_{\mathrm{dec}}, \theta_X, \theta_Y) &= \tau^{X\to Y} + \tau^{Y\to X}\\ &= \tau(X, Y; \theta_{\mathrm{enc}}^{X\to Y}, \theta_{\mathrm{dec}}^{X\to Y}, \theta_{\mathrm{s}}^{X\to Y}, \theta_{\mathrm{t}}^{X\to Y})\\ &\quad + \tau(Y, X; \theta_{\mathrm{enc}}^{Y\to X}, \theta_{\mathrm{dec}}^{Y\to X}, \theta_{\mathrm{s}}^{Y\to X}, \theta_{\mathrm{t}}^{Y\to X}),\end{aligned} \tag{15}$$

$$\text{subject to } \theta_{\mathrm{enc}} = \theta_{\mathrm{enc}}^{X\to Y} + \theta_{\mathrm{enc}}^{Y\to X}, \tag{16}$$

$$\theta_{\mathrm{dec}} = \theta_{\mathrm{dec}}^{X\to Y} + \theta_{\mathrm{dec}}^{Y\to X}, \tag{17}$$

$$\theta_X = \theta_{\mathrm{s}}^{X\to Y} + \theta_{\mathrm{t}}^{Y\to X}, \tag{18}$$

$$\theta_Y = \theta_{\mathrm{t}}^{X\to Y} + \theta_{\mathrm{s}}^{Y\to X}, \tag{19}$$

where $\theta^{X\rightarrow Y}$refers to the parameter of the $X \rightarrow Y$ direction model, $\theta^{Y\rightarrow X}$ refers to the parameter of the $Y \rightarrow X$ direction model, θ_{enc} refers to the parameter of the encoder, θ_{dec} refers to the parameter of the decoder, θ_{s} refers to the parameter represented by the source word, and θ_{t} refers to the parameter represented by the target word.

Based on parameter sharing, the representation generated by the model is also shared. That is to say, taking the encoder as an example, when X and Y representations are shared, the encoder can not only learn the encoding of two languages by using the same parameters but also learn the mapping from the word representation space to the hidden layer representation space of sentences. The output of the transformer system at the j moment:

$$P = (y_j|y < j, h(x)) = \text{softmax}(Wz_j + b), \tag{20}$$

where W is the transformation matrix and z_j is the hidden layer state obtained by the decoder. In the representation sharing, the above formula is modified to

$$P = (y_j|y < j, h(x)) = \text{softmax}(\theta_Y \cdot z_j). \tag{21}$$

The probability of refactoring Y is calculated as:

$$P = (y_j|y < j, h(y)) = \text{softmax}(\theta_Y \cdot z'_j). \tag{22}$$

In order to realize representation sharing, i.e., to realize reconfiguration without supervision, the loss function is optimized again:

$$\tau^{X\rightarrow Y} = \tau(X,\ Y; \theta_{\text{enc}}, \theta_{\text{dec}}, \theta_X, \theta_Y) + \alpha \cdot \tau(X,\ X; \theta_{\text{enc}}, \theta_{\text{dec}}, \theta_X, \theta_X), \tag{23}$$

where α stands for the weight of the reconstructed constraint loss.

3 Experimental analysis

3.1 Experimental setup

The ConvS2S system was implemented in the fairseq open-source tool [20]. The number of convolution layers was 16, the dimension of layers was 256, and the number of convolution kernels was 3. The optimizer nag carried by fairseq was used. The learning rate was 0.25. The parameters were the default values. The transformer system was implemented on the open-source tensor2tensor [21]. The number of network layers was 6. The dimension of different layers was 512. The multi-layer attention module used weight heads. The dimension of q, k, and v was 64. The dimension of the hidden layer was 2,048. The beam size was 8. NIST06 data sets under LDC were used for the experiment [22]. NIST06 was used as the development set, and NIST02, NIST03, NIST04, NIST05, and NIST08 were used as the test set. The translation performance of the three systems was compared.

3.2 Evaluation criteria

The results are evaluated by the BLEU value [23]. The calculation formula is

$$P_n = \frac{\sum_{c\in\{\text{candidates}\}}\sum_{n\text{-gram}\in c}\text{count}_{\text{clip}}(n\text{-gram})}{\sum_{c'\in\{\text{candidates}\}}\sum_{n\text{-gram}\in c'}\text{count}_{\text{clip}}(n\text{-gram}')}, \tag{24}$$

where P_n refers to the matching degree of the nth order, usually the fourth order, candidates refers to the candidate translation, and c refers to every sentence in the candidate translation. Finally, the calculation method of BLEU is

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{N}^{n=1} w_n \log p_n\right), \tag{25}$$

$$\text{BP} = \begin{cases} 1, c > r \\ e^{1-\frac{r}{c}}, c \leq r, \end{cases} \tag{26}$$

where BP is the length penalty factor, r and c are the length of the reference translation and candidate translation, and w_n is the weight coefficient.

3.3 Comparison of results

The computing speed of the ConvS2S system, the transformer system, and the transformer system combined with semantic sharing is shown in Figure 1.

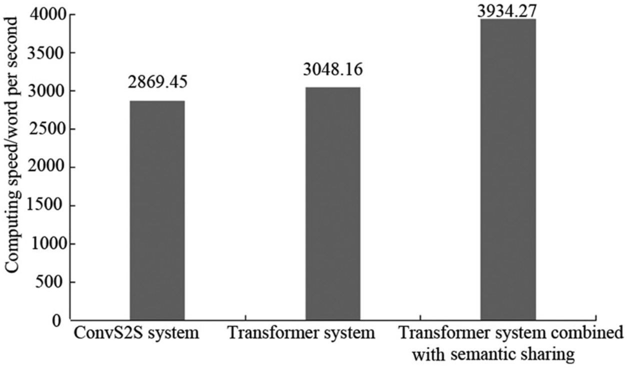


Figure 1: Comparison of the computing speed between different systems.

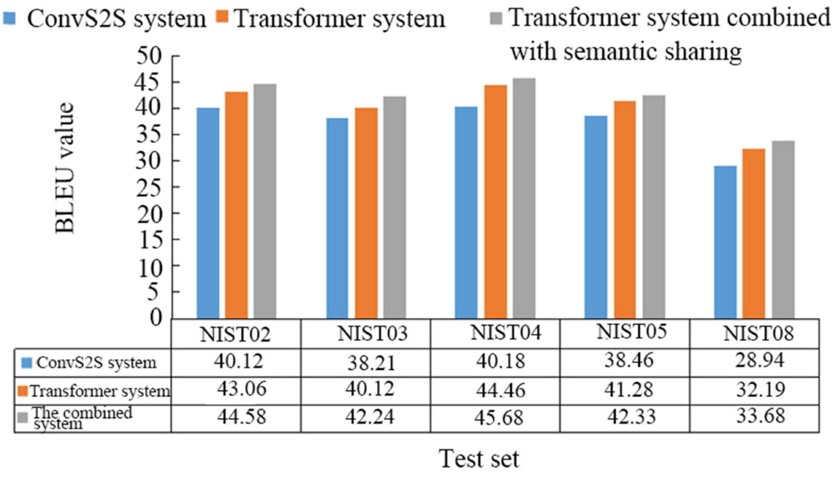
Figure 1 shows that the computing speed of the transformer system combined with semantic sharing was the fastest, reaching 3934.27 words per second, the computing speed of the ConvS2S system was 2869.45 words per second, and that of the transformer system was 3048.16 words per second. The computing speed of the transformer system combined with semantic sharing was 37.11 and 29.07% higher than that of the ConvS2S system and the transformer system.

The BLEU values of the ConvS2S system, the transformer system, and the transformer system combined with semantic sharing are shown in Figure 2.

Figure 2 shows that the BLEU value of the ConvS2S system was the smallest, followed by the transformer system and the transformer system combined with semantic sharing.

Specifically, in NIST02, the BLEU value of the transformer system combined with semantic sharing was 4.46 larger than the ConvS2S system and 1.52 larger than the transformer system; in NIST03, the BLEU value of the transformer system combined with semantic sharing was 4.03 larger than the ConvS2S system and 2.12 larger than the transformer system; in NIST04, the BLEU value of the transformer system combined with semantic sharing was 5.5 larger than the ConvS2S system and 1.22 larger than the transformer system; in NIST05, the BLEU value of the transformer system combined with semantic sharing was 3.87 larger than the ConvS2S system and 1.05 larger than the transformer system; in NIST08, the BLEU value of the transformer system combined with semantic sharing was 4.74 larger than the ConvS2S system and 1.49 larger than the transformer system. The above results verified that the transformer system combined with semantic sharing had higher quality in English content translation.

The translation results of two sentences were analyzed, as shown in Table 1.



	NIST02	NIST03	NIST04	NIST05	NIST08
ConvS2S system	40.12	38.21	40.18	38.46	28.94
Transformer system	43.06	40.12	44.46	41.28	32.19
The combined system	44.58	42.24	45.68	42.33	33.68

Figure 2: Comparison of BLEU values between different systems.

Table 1: Examples of translation results

Source corpus	Look, man, you don't get to do anything.
Reference translation	兄弟，你什么都不需要做。
Translation of the CnvS2S system	你看，男人，你不应该尽。
Translation of the transformer system	男人，你不需要做任何事。
Translation of the transformer system combined with semantic sharing	听着，老兄，你什么都不用做。
Source corpus	This one means a lot to me.
Reference translation	这对我来说意义重大。
Translation of the CnvS2S system	这其中意味着很多给我。
Translation of the transformer system	这个对我来说意味着很多。
Translation of the transformer system combined with semantic sharing	这意味着对我很重要。

Table 1 shows that there were some differences between the translation of the ConvS2S system and the transformer system, and the reference translation in translating English content; from the semantic perspective, the differences were large. The translation results of the transformer system combined with semantic sharing were very similar to the reference translation, with stronger readability and higher translation quality.

4 Discussion

NMT maps sentences of the source language directly to those of the target language through an end-to-end method [24], which is significantly better than statistical MT in the case of sufficient parallel corpus. It does not require separate modules for word alignment and tuning order but outputs translation results directly through a neural network. It not only has a wide range of applications in practice, but also has a very important research value in the field of translation [25].

This study mainly compared two NMT methods, the ConvS2S system and the transformer system, and improved the transformer system by semantic sharing. Experiments were carried out with the NIST dataset as an example. First, in terms of computing speed, the ConvS2S system had the slowest computing speed, 2869.45 words per second, showing high computational complexity, while the computing speed of the transformer system was 3048.16 words per second, showing an increase of 6.23% compared to the ConvS2S system. The computing speed of the improved system reached 3934.27 words per second, which was significantly higher than the first two systems, i.e., the improved system had an advantage in computational efficiency. The comparison of BLEU values showed that all three systems showed similar results on different data sets, i.e., the ConvS2S system < the transformer system < the improved system. The average BLEU values of the three systems were 37.18, 40.22, and 41.7, respectively, and the BLEU value of the transformer system was 3.04 larger than the ConvS2S system, while the BLEU value of the improved system was 4.52 larger than the ConvS2S system and 1.48 larger than the transformer system. The above results revealed that the improved system had better performance both in terms of computational efficiency and translation performance. Finally, the comparison of translation results showed that the translation results of both ConvS2S system and transformer system had different degrees of semantic differences in the translation of English example sentences, which did not fully express the meaning of the source sentences and had shortcomings in the completeness and readability, but the improved system designed achieved better translation of English example sentences and more accurate results, showing higher translation quality.

Some results have been achieved in the comparison of MT methods for English content in this study; however, there are still some shortcomings. In future research, more NMT methods can be improved and compared, and experiments can be conducted on more datasets to further improve the efficiency and quality of English translation.

5 Conclusion

This study introduced two NMT methods for English content translation, the ConvS2S system and the transformer system, and designed a transformer system combined with semantic sharing to improve translation quality. The experiment on the NIST data set showed that the transformer system combined with semantic sharing had a better performance in computing speed and BLEU value, showing reliability in improving the efficiency and quality of English content translation. The designed system can be further promoted and applied in practice.

Conflict of interest: The author state no conflict of interest.

References

[1] Lee L. Book reviews: foundations of statistical natural language processing. Microbiology. 2015;144 (pt 4)(3).

[2] He H. The parallel corpus for information extraction based on natural language processing and machine translation. Expert Syst. 2018;36:e12349.

[3] Zhang Y, Chen J, Liu B, Yang Y, Li H, Zheng X, et al. COVID-19 public opinion and emotion monitoring system based on time series thermal new word mining. Comput Mater Con. 2020;64:1415–34.

[4] Bo T, Kay S, He H. Toward optimal feature selection in Naive Bayes for text categorization. IEEE T Knowl Data En. 2016;28:2508–21.

[5] Castilho S, Moorkens J, Gaspari F, Calixto I, Tinsley J, Way A. Is neural machine translation the new state of the art? Prague Bull Math Ling. 2017;108:109–20.

[6] Choi H, Cho K, Bengio Y. Context-dependent word representation for neural machine translation. Comput Speech Lang. 2017;45:149–60.

[7] Wu S, Zhang D, Zhang Z, Yang N, Li M, Zhou M. Dependency-to-dependency neural machine translation. IEEE/ACM T Audio Spe. 2018;26:2132–41.

[8] Lee J, Cho K, Hofmann T. Fully character-level neural machine translation without explicit segmentation. Trans Assoc Comput Ling. 2017;5:365–78.

[9] Gu JT, Hassan H, Devlin J, Li V. Universal neural machine translation for extremely low resource languages. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2018. p. 344–54.

[10] Tejada MAZ, Gallardo CN, Ferradá MCM, López MIC. 2L English texts and cohesion in upper CEFR levels: a corpus-based approach. Proc Soc Behav Sci. 2015;212:192–7.

[11] Simpson A, Wu Z, Li Y. Grammatical roles, coherence relations, and the interpretation of pronouns in Chinese. Ling Sin. 2016;2:1–20.

[12] Yamaguchi T, Ikehara M. Multi-stage dense CNN demosaicking with downsampling and re-indexing structure. IEEE Access. 2020;8:175160–68.

[13] Omer K, Caucci L, Kupinski M. CNN performance dependence on linear image processing. Electr Imag. 2020;310:1–7.

[14] He Y, Yu LC, Lai KR, Liu WY. YZU-NLP at EmoInt-2017: determining emotion intensity using a bi-directional LSTM-CNN model. Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis; 2017. p. 238–42.

[15] Rosewelt A, Renjit A. Semantic analysis-based relevant data retrieval model using feature selection, summarization and CNN. Soft Comput. 2020;24:16983–7000.

[16] Wang ZR, Du J, Wang JM. Writer-aware CNN for parsimonious HMM-based offline handwritten Chinese text recognition. Pattern Recogn. 2018;100:107102.

[17] Zhang Y, Shi XY, Mi SY, Yang X. Image captioning with transformer and knowledge graph. Pattern Recogn Lett. 2021;143:43–9.

[18] Wang D, Hu H, Chen D. Transformer with sparse self-attention mechanism for image captioning. Electron Lett. 2020;56:764–6.

[19] Pan Y, Yu H. Biomimetic hybrid feedback feedforward neural-network learning control. IEEE T Neur Net Lear. 2017;28:1481–7.

[20] Ott M, Edunov S, Baevski A, Fan A, Gross S, Ng N, et al. Fairseq: a fast, extensible toolkit for sequence modeling. Proceedings of the 2019 Conference of the North; 2019.

[21] Vaswani A, Bengio S, Brevdo E, Chollet F, Gomez A, Gouws S, et al. Tensor2Tensor for neural machine translation; 2018.

[22] Zhang J, Wei XL, Zheng CH, Wang B, Wang F, Chen P. Compound identification using random projection for gas chromatography-mass spectrometry data. Int J Mass Spectrom. 2016;407:16–21.

[23] Luong MT, Sutskever I, Le QV, Vinyals O, Zaremba W. Addressing the rare word problem in neural machine translation. Bull UASVM Vet Med. 2015;27:82–6.

[24] Wu SZ, Zhang DD, Yang N, Li M, Zhou M. Sequence-to-dependency neural machine translation. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics; 2017. p. 698–707.

[25] Chen HD, Huang SJ, Chiang D, Chen JJ. Improved neural machine translation with a syntax-aware encoder and decoder. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics; 2017. p. 1936–45.