Artificial Life and Robotics
https://doi.org/10.1007/s10015-024-00964-5

ORIGINAL ARTICLE

Research on English–Chinese machine translation shift based on word vector similarity

Qingqing Ma[1]

Received: 9 July 2024 / Accepted: 13 August 2024

Abstract
In English–Chinese machine translation shift, the processing of out-of-vocabulary (OOV) words has a great impact on translation quality. Aiming at OOV, this paper proposed a method based on word vector similarity, calculated the word vector similarity based on the Skip-gram model, used the most similar words to replace OOV in the source sentences, and used the replaced corpus to train the Transformer model. It was found that when the original corpus was used for training, the bilingual evaluation understudy-4 (BLEU-4) of the Transformer model on NIST2006 and NIST2008 was 37.29 and 30.73, respectively. However, when the word vector similarity was used for processing and low-frequency OOV words were retained, the BLEU-4 of the Transformer model on NIST2006 and NIST2008 was improved to 37.36 and 30.78 respectively, showing an increase. Moreover, the translation quality obtained by retaining low-frequency OOV words was better than that obtained by removing low-frequency OOV words. The experimental results prove that the English–Chinese machine translation shift method based on word vector similarity is reliable and can be applied in practice.

Keywords Word vector similarity · English–Chinese translation · Machine translation

1 Introduction

Under the influence of globalization, international communication becomes more and more frequent, and people have more and more opportunities to come into contact with non-native languages in life and work, which leads to a more urgent need for translation. Machine translation refers to the process of converting one language into another language with the assistance of computer technology [1]. In the current research, how to obtain a higher quality of machine translation has become a key and difficult issue [2]. Xiang et al. [3] put forward the method of integrating language differentiation features to improve the quality of Chinese–Vietnamese machine translation, designed a two-way long short-term memory model, and found that the proposed method can effectively enhance translation quality through the experiment on large-scale bilingual data. Lee et al. [4] proposed an attention mechanism based on reinforced learning to solve the delay problem of machine translation model in online scenes. Through experiments, it was found that the model has better translation quality and comparable delay compared to other models. For the machine translation of Dayak language, Khaikal et al. [5] extracted data from web pages to build a corpus based on statistical methods and found that the highest accuracy rate reached 49.15%, which was about 3% higher than the other machine translations. Pandey et al. [6] designed a translation system from Hindi to Chhattisgarh based on the open source software Moses and found that the accuracy rate of the system reached 75% through testing on 1000 sentences. English and Chinese are widely used in various real-life situations. With the impact of cross-cultural communication, there is an increasing need for translation between English and Chinese, which raises higher demands on current English–Chinese machine translation shift. The growing number of out-of-vocabulary (OOV) words results in a decrease in translation quality, hindering effective communication support. Therefore, in order to further improve the translation quality of English–Chinese machine translation shift, this paper designed a method based on word vector similarity that replaces OOV words with similar words. Experimental analysis verified the effectiveness of this method. The study

✉ Qingqing Ma
mqq_qing@hotmail.com

[1] Nanchang Institute of Technology, Changbei Economic Development Zone, No. 901, Hero Avenue, Nanchang City 330044, Jiangxi, China

Published online: 16 September 2024

in this article not only provides a higher quality method for English–Chinese machine translation shift, but also offers some new insights into the deep research of Transformer models in machine translation. It demonstrates the impact of OOV word processing methods on translation quality and provides some references for OOV word processing in machine translation research for other languages, which is beneficial for further enhancing the quality of machine translation and promoting its better application in practice.

2 Translation model of english-chinese machine translation

The transformer model has a good application in many languages [7]. The transformer model includes six encoders and six decoders [8]. In the encoder, input text is first converted into input vector, and then based on position encoding, it is passed into the encoder layer.

The relationship between each word and other words in the source sentence is learned through a multi-head self-attention layer. The formula is written as:

$$\text{MultiHead}(Q,K,V) = \text{concat}\left(\text{head}_1, \text{head}_2, \cdots, \text{head}_h\right)W^O, \tag{1}$$

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right), \tag{2}$$

where $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_k}$, Q, K, V indicate query, key, and value matrices, and W^O is the final context vector.

Then, the output of the multi-head self-attention layer is nonlinearly mapped through a feedforward neural network (FFN), and the formula is written as:

$$\text{FFN}(x) = \max\left(0, xw_1 + b_1\right)w_2 + b_2, \tag{3}$$

where w_1 and w_2 are weights of the first and second layers of the FFN, w_1 and w_2 are threshold values.

In the decoding layer, since the information of ungenerated words cannot be seen when decoding, the decoder uses a mask mechanism to hide the information. Finally, after all the execution of the decoder layer, for the source end sentence, the target translation is obtained according to conditional probability:

$$P(Y|X) = \prod_{i=1}^{|Y|} P\left(y_i|Y_{<i}, X\right). \tag{4}$$

where X refers to the input sentence and Y is the target translation generated by the model according to the conditional probability. At each time step i, beam search is used for decoding [9]. K best candidates are kept. The formula can be written as:

$$\log P(Y|X) = \sum_{i=1}^{|Y|} \log\left(y_i|Y_{<i}, X\right). \tag{5}$$

3 OOV word processing based on word vector similarity

In the actual English–Chinese machine translation shift, for the included OOV words, the same symbol < unk > will be used to represent them [10], which is easy to lead to ambiguity. To solve this problem, in the pre-processing stage, this paper first replaces OOV words in corpus based on word vector similarity and then puts the replaced words into the model to complete translation, so as to improve translation quality.

Distributed representation can use a fixed-dimension vector to represent words, which can avoid the problem of excessive dimension of one-hot encoding and can also reflect the semantic correlation between words. It has a good performance in natural language processing. In this paper, this method is used to find the similar words of OOV words, and the Skip-gram model in word to vector (word2vec) is used as the training tool [11].

The Skip-gram model performs one-hot encoding on the input and output and then word vector training. For the input sequence $w_1, w_2, \cdots, w_T$, the training goal is to maximize the value of the following formula:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{j \subseteq nb(t)} \log p\left(w_j|w_t\right), \tag{6}$$

where $nb(t)$ refers to the context word of w_t and $p\left(w_j|w_t\right)$ is the conditional probability value.

After obtaining the vector representation of all the words in the corpus, the similarity can be calculated according to the word vector. The similarity between OOV word vector w and its similar word vector $w\prime$ is calculated based on cosine similarity. The formula is:

$$\text{sim}(w, w\prime) = \cos(\text{vec}(w), \text{vec}(w\prime)), \tag{7}$$

$$w^* = \arg\max_{w\prime \in IV} \text{sim}(w, w\prime), \tag{8}$$

where IV refers to the common word list. The cosine similarity can be calculated based on the distance tool in word2vec. In order to further retain the meaning of the source end sentence, candidate words are screened again using the n-gram model [12]. OOV word w_i and candidate word $w\prime$ are scored based on n-gram. The formula is:

$$P(s) = \prod_{i=1}^{k} P\left(w_i | w_{i-n+1}, \cdots, w_{i-1}\right). \quad (9)$$

In this paper, the binary language model is used in calculation, i.e., $n = 2$. The appearance of a word is only related to the previous word. The score of each candidate sentence is calculated, and then the similar high-frequency words with the highest score is used to replace OOV words. The steps are as follows:

(1) The word vector is trained to find all the high-frequency words that are similar to OOV words.
(2) Alternative candidate words are scored using the n-gram model.
(3) The high-frequency words with the highest score are found out to replace OOV words.
(4) The replaced sentence is used as the source end sentence and translated using a trained transformer model.

4 Results and analysis

4.1 Experimental setup

The experiment was carried out in Linux environment. The operating system was Ubuntu 18.04 LTS. The programming language was Python 3.6, and the transformer model was implemented based on the PyTorch framework. The experimental datasets came from LDC (Table 1).

The Skip-gram model in the word2vec tool was used to train a 300-dimensional word vector, and the training set included 2.1 million English–Chinese parallel corpus. The forward and backward translation dictionary was obtained by GIZA + + tool [13]. Chinese was segmented using jieba [14], and English was segmented using byte pair encoding (BPE) [15]. In terms of parameter setting of the transformer model, the word embedding dimension was set to 768, the FFN dimension was 2,048, and the Adam optimizer was employed. In the Adam optimizer, $\beta_1 = \beta_2 = 0.5$, the learning rate was set as 0.5, and the dropout was set as 0.1. The 4-g BLEU [16], which is case-insensitive, was used to assess the translation quality.

Table 1 Experimental dataset

Training set	LDC2002E18, LDC2003E07, LDC2003E14, LDC2004T04, LDC2004T08, LDC2005T06, LDC2005T10
Development set	NIST 2005
Test set	NIST2006, NIST2008

4.2 Result analysis

The transformer model was employed to realize English–Chinese machine translation shift. For the corpus for training, OOV words were replaced based on word vector similarity. For low-frequency OOV words without word vector representation, two processing methods can be adopted: retain (represented by < unk >) or remove (directly delete). In order to compare the results of the English–Chinese machine translation shift method based on the word vector similarity, the comparative experiments are as follows:

(1) Experiment 1: The original corpus without OOV word processing.
(2) Experiment 2: The corpus processed based on word vector similarity, whose low-frequency OOV words are retained.
(3) Experiment 3: The corpus processed based on word vector similarity, whose low-frequency OOV words are removed.

The BLEU-4 results obtained after using two test sets are presented in Table 2.

As can be seen from Table 2, when using the original corpus, the transformer model achieved a BLEU-4 of 37.29 for IST2006 and a BLEU-4 of 30.73 for IST2008. In experiment 2, OOV words in the corpus were processed based on word vector similarity, and low-frequency OOV words were retained. As a result, the BLEU-4 of the transformer model for NIST2006 was 37.36, indicating an increase of 0.07 compared with experiment 1. The BLEU-4 for NIST2008 was 30.78, which showed an increase of 0.05 compared to experiment 1. These results proved the role of word vector similarity processing in improving the quality of English–Chinese machine translation shift. Then, in experiment 3, compared with experiment 2, low-frequency OOV words in the corpus were removed. The BLEU-4 of the transformer model for IST2006 was 37.12, which was reduced by 0.17 and 0.24 compared with experiment 1. The BLEU-4 for IST2008 was 30.33, which was reduced by 0.4 compared with experiment 1 and 0.45 compared with experiment 2. These results showed that for low-frequency OOV words, the effect of retention processing was obviously better than that of removal processing. For the direct removal of

Table 2 Translation results of English–Chinese machine translation based on word vector similarity

	NIST2006	NIST2008
Experiment 1	37.29	30.73
Experiment 2	37.36	30.78
Experiment 3	37.12	30.33

low-frequency OOV words in sentences, the sentence structure may be damaged, resulting in a significant decline in translation quality. While reserving it as < unk >, although there may be semantic loss, it ensures the integrity of sentence structure. According to the experimental results, low-frequency OOV words in English–Chinese machine translation should be reserved in addition to processing based on word vector similarity, so as to avoid a substantial decline in translation quality.

To further verify the performance of the proposed mode, it was compared with some other machine translation models, including:

the phrase-based statistical machine translation [17],
the deep neural network-based statistical machine translation [18],
the bidirectional long short-term memory-based neural machine translation [19],
the attention-based neural machine translation [20].

The results are presented in Table 3.

From Table 3, it can be observed that both statistical machine translation methods had relatively low BLEU-4 on NIST2006 and NIST2008. Compared to the statistical machine translation models, neural machine translation methods performed better and achieve higher BLEU-4. However, the overall comparison showed that the method proposed in this paper had more advantages. It improved the BLEU-4 by 1.72, 1.52, 0.59, and 0.24, respectively for NIST2006 compared to the other methods, and improved the BLEU-4 by 2.16, 1.81, 1.23, and 0.54, respectively for NIST2008 compared to the other methods. These results proved the effectiveness of the proposed method in enhancing translation quality.

Two translation examples were analyzed as follows:

Example 1: Source end Sentence: 奥勃说, 所有三位大使都对菲律宾政府加强安全措施表示满意。

The reference result of manual translation: Ople said that all the three ambassadors expressed their satisfaction with the strengthened measures taken by the Philippines government.

Table 3 Results of comparisons with the other machine translation models

	NIST2006	NIST2008
Literature [17]	35.64	28.62
Literature [18]	35.84	28.97
Literature [19]	36.77	29.55
Literature [20]	37.12	30.24
The method proposed in this paper	37.36	30.78

The transformer model: He said all three ambassadors expressed satisfaction with the Philippines government's increased security measures.

The transformer model based on word vector similarity: According to < unk >, all three ambassadors were satisfied with strengthened security measures taken by the Philippines government.

Example 2: Source end Sentence: 仲介商今天说, 总部设在香港的地产集团 “华人置业” (Chinese Estates) 以2亿8000万英镑, 在伦敦买下高盛 (Goldman Sachs) 投资银行欧洲总部所在的大楼。

The reference result of manual translation: An intermediary said today that Hong Kong-based property group “Chinese Estates” has bought a building in London which houses the European headquarters of investment bank Goldman Sachs for 280 million British pounds.

The transformer model: Chinese Estates, the Hong Kong-based property group, has bought the European headquarters of Goldman Sachs’ investment bank in London for £280 m, agents said today.

The transformer model based on word vector similarity: The intermediary said today that Hong Kong-based property group “Chinese Estates” has bought a building in London which houses the European headquarters of investment bank Goldman Sachs for 280 million British pounds.

In example 1, "奥勃" in the source sentence is a person name and also a low-frequency OOV word, and no similar word can be replaced. In the transformer model trained with original corpus, the word was translated into "He", while in the Transformer model based on word vector similarity, the word was replaced by < unk >, maintaining the integrity of the sentence structure.

In Example 2, the word "仲介商" in the source end sentence as an OOV word was replaced by "中介商". The translated result had a good similarity with the reference result of manual translation, while in the original corpus training, the word is translated as "agents", which is somewhat different from the actual semantics.

From the analysis of translation examples, it can be found that the Transformer model based on word vector similarity obtained better translation results in English–Chinese machine translation and can be applied in practice.

5 Conclusion

Aiming at the OOV problem in English–Chinese machine translation, this paper proposed a processing method based on word vector similarity and conducted an experimental analysis with the transformer model. The results showed that after OOV word processing based on word vector similarity, the translation quality of the transformer model improved to some extent. In addition, for low-frequency OOV words, the

effect of retention processing was better than that of removal processing. The results of the proposed method were more similar to the result of manual translation. Therefore, the proposed method can be further applied in practice. However, there are some limitations in this study. For example, only one method, word2vec, was considered in the training of word vectors. Additionally, no novel improvements were made to the design of the transformer model for machine translation. Therefore, future work will explore the application of more advanced word vector techniques and consider optimizing the transformer model to further enhance the quality of English-to-Chinese machine translation. Furthermore, assessing the applicability of the proposed approach in machine translation for other languages will also be considered.

Data availability The data used and analyzed in the paper are available from corresponding author upon reasonable requests.

References

1. Qiu J, Moh M, Moh TS (2021) Fast streaming translation using machine learning with transformer. Proceedings of the 2021 ACM southeast conference, USA, Apr 15–17, 2021, pp. 9–16.
2. Huang L, Chen W, Qu H (2021) Accelerating Transformer for neural machine translation. ICMLC 2021: 2021 13th International conference on machine learning and computing, Shenzhen China, Feb 26-Mar 1, 2021, pp. 191–197.
3. Zou X, Zhu JG, Gao SX, Yu ZT, Yang FA (2022) Translation quality estimation of Chinese-Vietnamese neural machine translation incorporating linguistic differentiation features. J Chin Comput Syst 43:1413–1418
4. Lee YH, Shin JH, Kim YK (2021) Simultaneous neural machine translation with a reinforced attention mechanism. ETRI J 43:775–786
5. Khaikal MF, Suryani AA (2021) Statistical machine translation Dayak Language – Indonesia Language. Informatika Mulawarman: Jurnal Ilmiah Ilmu Komputer 16:49
6. Pandey V, Padmavati DMV, Kumar R (2021) Hindi Chhattisgarhi machine translation system using statistical approach. Webology 18:208–222
7. Chen Y, Rong P (2020) Compressed-transformer: distilling knowledge from transformer for neural machine translation. NLPIR 2020: 4th international conference on natural language processing and information retrieval, Seoul Republic of Korea, Dec 18–20, 2020, pp. 131–137.
8. Dong HW, Zhou C, Berg-Kirkpatrick T, McAuley J (2022) Deep performer: score-to-audio music performance synthesis. ICASSP 2022 - 2022 IEEE international conference on acoustics, speech and signal processing (ICASSP), Singapore, Singapore, May 23–27, 2022, pp. 951–955.
9. Parreno-Torres C, Alvarez-Valdes R, Parreno F (2022) A beam search algorithm for minimizing crane times in premarshalling problems. Eur J Oper Res 302:1063–1078
10. Aqlan F, Fan X, Alqwbani A, Al-mansoub AA (2019) Arabic-Chinese neural machine translation: romanized Arabic as subword unit for Arabic-sourced translation. IEEE Access 7:133122–133135
11. Al-Saqqa S, Awajan A (2019) The use of Word2vec model in sentiment analysis: a survey. AIRC '19: 2019 international conference on artificial intelligence, robotics and control, Cairo Egypt, 2019, pp. 39–43.
12. Goel G, Bhardwaj H, Hooda I, Kumar S (2020) Optimal N-gram subset extraction for accelerating evaluation using genetic algorithm. 2020 international conference for emerging technology (INCET), Belgaum, India, Jun 05–07, 2020, pp. 1–5.
13. Andrih B, Stankovi R (2019) Extraction of bilingual terminology using graphs, dictionaries and GIZA++. Infotheca 19:119–138
14. Chen S, Zhang C, Ren P (2020) Preliminary study on exploring the trajectory of patients with COVID-19 by data mining algorithms. Chin J Med Sci Res Manag 33:E005–E005
15. Amalia A, Sitompul OS, Mantoro T, Nnababan EB (2021) Morpheme embedding for bahasa indonesia using modified byte pair encoding. IEEE Access 9:155699–155710
16. Shahnawaz MRB (2013) Rule-based approach for handling of case markers in English to Urdu/Hindi translation. Int J Knowl Eng Soft Data Parad 4(2):138–165
17. Babhulgaonkar A, Sonavane S (2022) Empirical analysis of phrase-based statistical machine translation system for English to Hindi language. Vietnam J Comput Sci 09:135–162
18. Xia Y (2020) Research on statistical machine translation model based on deep neural network. Computing 102(3):643–661
19. Paul N, Faruki I, Pranto MI, Shawon MTR, Mandal NC (2023) Bengali-English neural machine translation using deep learning techniques. 2023 international conference on electrical, computer and communication engineering (ECCE), Chittagong, Bangladesh, Feb 23–25, 2023, pp. 1–6.
20. Bakarola V, Nasriwala J (2022) Attention-based neural machine translation approach for low-resourced Indic languages—a case of Sanskrit to Hindi translation. Smart Syst: Innov Comput 235:565–572