Received: 21 Mar. 2024 | Revised: 13 Apr. 2024 | Accepted: 15 Apr. 2024 | Published online: 25 Apr. 2024
RESEARCH ARTICLE

Journal of Global Humanities and Social Sciences
2024,Vol. 5(4)156-161
DOI: 10.61360/BoniGHSS242016160404

Research on Chinese-English Patent Machine Translation Based on Fusion Strategy Model

Hua Chai[1,*]
[1]*Yulin College, China*

Abstract: In order to improve the translation quality of the machine translation model in the field of patent text, this paper proposes a patent knowledge fusion strategy. On the one hand, it integrates the structural knowledge, special phrases and professional terms in the patent text into the translation model as embedded comments, so as to improve its translation effect. On the other hand, based on XLNet, the basic translation model of Transformer is improved, and the translation performance of the model is further improved on the basis of integrating professional knowledge. The simulation results show that compared with the basic Transformer model and the Transformer model that only introduces fusion strategy, BLEU values of the proposed machine translation model combining fusion strategy and improved Transformer for the Chinese-English and English-Chinese translation tasks has increased by 3.77, 1.79, and 2.17, 0.46, respectively. It has a significant improvement in improving the quality of patent literature translation and is worthy of further research and promotion.

Keywords: machine translation; patent literature; special aspect; fusion strategy

Introduction

With the practice and development of modern agriculture, people's awareness of intellectual property rights has gradually improved, and more and more attention has been paid to the translation of patent documents and patented technology documents. However, the specialization of translation terms in specific fields has always been a major problem faced by machine translation models (Xiong, 2021). In response to this phenomenon, Cao Jingcheng et al. compared the best construction methods for multilingual patent parallel corpora in different scenarios, and found the most suitable solution to expand existing patent corpus resources, thereby improving the translation performance of multilingual neural machine translation systems (Cao et al., 2022). You Xindong et al proposed to integrate term information into Transformer patent machine translation model through replacement or addition, so as to improve the translation effect of the model on patent data sets (You et al., 2021). Although the above studies have improved the quality of translation in specific fields to a certain extent, the translation effect is limited, and there are still phenomena such as omission and over-translation of proprietary words in the text. On the basis of the above research, this paper proposes a machine translation model that combines professional knowledge fusion strategy and improved Transformer, which has certain practical value to improve the translation effect of the model in professional fields without too much modification of results.

1. Related Knowledge

1.1. Domain knowledge fusion strategy

Introducing a professional knowledge fusion strategy based on embedded annotations can improve the translation quality of machine translation models for patent related literature (Han et al., 2023). Utility model patent literature refers to the documents and materials that record a series of relevant documents in the process of application, examination and approval of utility model patents, which is usually contained in various official publications published

Corresponding Author: Hua Chai
Yulin College, China
Email: 515508552@qq.com

on a regular basis, and has the characteristics of involving many professional terms, written terms and rigorous and concise wording.

Due to the limited number of samples in the training process of such texts, which contain a large number of low-frequency words, rare words and proprietary words, the machine translation model often suffers from inaccurate translation, ignored translation or over-translation, and the translation effect is relatively poor (Du, 2023). In order to improve the translation accuracy of machine translation models on professional written texts, this paper introduced a domain knowledge fusion strategy with embedded annotations to improve the training process of the model. According to the characteristics of this type of text, the model fully learns the framework knowledge, special phrases and professional terms in the text during the training process, and replaces them into the source language sentences as embedded annotations.

Based on a bilingual dictionary, this method first searches for domain knowledge in the source language and annotates the corresponding translation into the source language sentence. There are two translation annotation methods, namely concatenation and substitution. Concatenation refers to concatenation of the target phrase into the corresponding words of the source language, and annotation of the source language. Substitution refers to directly replacing the corresponding words in the source language with the translation of the target phrase. Taking "**闭锁式**水稻立体育秧技术" as an example, the labeling effects of the two methods are shown as follows:

Concatenation annotation: Stereo-structuring$_0$ Rise$_0$ Seedling$_0$ Raising$_0$ technology$_0$ in$_1$ a$_1$ Locking$_1$ Stype$_1$ Room$_1$ 闭锁式 $_2$

Substitution annotation: Stereo-structuring$_0$ Rise$_0$ Seedling$_0$ Raising$_0$ technology$_0$ 闭锁式 $_2$

Target sentence: **闭锁式**水稻立体育秧技术

In the above example sentence, the subscript "0" represents the words in the source language sentence. The subscript "1" represents the corresponding translation of the domain phrase in the target sentence in the source language sentence. The subscript "2" represents the word that appears in the target sentence after concatenation or replacement.

Different from traditional methods based on translation memory or annotation of training corpus to improve the translation model's translation effect on words in special fields, the professional knowledge fusion strategy based on embedded annotation is adopted in this paper. All the domain knowledge to be learned comes from the expansion of the training set, and the model can directly concatenate and substitute the target language on the source language, which makes the model do not need to modify the sequence-to-sequence network architecture, and can be widely used on the current machine translation model with changing architecture.

The domain knowledge in utility model patent literature can be summarized into three types: framework knowledge, normative phrases, and professional terminology. For different categories of knowledge, there are also different representations when embedding annotations. Each category of knowledge is represented as follows:

(1) Architecture knowledge representation

Architectural knowledge refers to some fixed formatting rules and organizational structure of patent texts when writing and translating, including the related words between different contents of the article, the introduction words of a patented product and components, and the introduction words of a special name of a product. Take "he patent technology relates to an intelligent toughened air flow control system based on a plant factory"as an example, which contains the architecture knowledge and the representation method as follows:

Source language: "本发明专利技术涉及一种智能钢化气流调控系统"

Target language: "The patent technology relates to an intelligent toughened air flow control system"

Its architecture is:

"本发明专利技术涉及一种（.*）系统"

"The patent technology relates to an (.*) system"

It is represented in the inline comment as follows:

Concatenation result: 本 1 发明专利技术 1 涉及 1 一种 1The2 patent2 technology2 relates2 to2 an2 智能 0 钢化 0 气流 0 调控 0 系统 1system2

Substitution result: The2 patent2 technology2 relates2 to2 an2 智能 0 钢化 0 气流 0 调控 0 system2

(2) Standardize phrase representation

Standardize phrases refer to commonly used phrases in patent literature to comply with writing standards. Take "The improvement effect of this system over the traditional plant factory wind recharge system is:" as an example, it contains standardize phrases and expression methods as follows:

Source language: "该系统相对于传统植物工厂风力补给系统的改进效果是："

Target language: "compared to the traditional plant factory wind supply system, the system has the improvement effects are that :"

The standardize phrase included in the example sentence is "改进效果/improvement effects", which is represented in the inline notes as follows:

Concatenation result: 该 0 系统 0 相对于 0 传统 0 植物 0 工厂 0 风力 0 补给 0 系统 0 的 0 改进 1 效果 1 improvement2 effects2 是 0：0

Substitution result: 该 0 系统 0 相对于 0 传统 0 植物 0 工厂 0 风力 0 补给 0 系统 0 的 0 improvement2 effects2 是 0：0

(3) Technical terms

Technical terms refer to the unified proper name for some specific matters in the field of patent in the process of forming patent literature. Taking "The patent technology relates to an plant factory system module" as an example, it includes technical terms and representation methods as follows:

Source language：“本发明设计一种植物工厂系统模块”

Target language: "The patent technology relates to an plant factory system module"

The technical term contained in the example sentence is "植物工厂/plant factory", which is represented in the embedded annotation as follows:

Concatenation result: 本 0 发明设计 0 一种 0 植物 1 工厂 1plant2 factory2 系统 0 模块 0

Substitution result: 本 0 发明设计 0 一种 0 plant2 factory2 系统 0 模块 0

(4) Fusion embedding

According to the characteristics of different types of domain knowledge, different methods are used to integrate and embed them in the model. Among them, the architectural knowledge is merged by "substitution", while the standardize phrases and technical terms are merged by "concatenation". In the machine translation model taking decoder-encoder as framework, the encoder usually maps the input sequence $(x_1, x_2, ..., x_n)$ to a series of high-dimensional vector $\tilde{z} = (z_1, z_2, ..., z_n)$ representations, and then combines the context $\tilde{z}$ information to generate a new output sequence $(y_1, y_2, ..., y_n)$ by the decoder. Therefore, this paper summarizes the knowledge in the field of patents and concatenates them as embedded annotations to replace them in the training corpus. In the training process, the model will automatically encode these domain knowledge, and gradually learn to use them in the iterative process.

1.2. Improved transformer machine translation model

The Encoder-Decoder general neural machine translation model framework is used to construct the Chinese-English patent machine translation model in this paper. The proposed Chinese-English patent machine translation model is implemented by improving the structure of the traditional Transformer neural machine network. Where, the encoder is replaced by an XLNet pre-trained model from the six layer stack of the Transformer encoder, and the decoder continues to use Transformer decoder (Chao et al., 2022).

XLNet model is a pre-training language model proposed by Google in recent years by combining the advantages of auto-regressive model and self-coding model (Li et al., 2022). Compared with the single auto-encoding model and the auto-regressive model, XLNet model can use both contextual and future information by arranging the pre-training method of the language model. In addition, XLNet model also

introduces a dual stream attention mechanism in the multi-head self-attention mechanism of each layer, improving the ability of traditional Transformer models to capture long sequence information. Compared to Transformer encoder, it can more fully grasp the feature information of the source language, thereby further improving the translation quality of the model (Zhao et al., 2024).

Suppose the input sequence is $x=(x_1,x_2,...,x_n)$. Where, n and x_i represent the total length of the sequence and the i th element in the sequence, respectively. The dynamic adaptive input sentence length algorithm is used to truncate and fill the input source language sentences, and the processed sequences are input to the embedding layer of XLNet, which converts it into a word vector matrix, and then encoded by XLNet. Moreover, XLNet uses a dual-stream attention mechanism and a feedforward neural network at each layer to operate on the left and right contexts respectively, and then normalize the obtained results through LayerNorm to complete the encoding of the source language.

The decoding end processing is implemented by Transformer decoder. Before decoding, it is necessary to perform vector fusion on each position in the input sequence. The output vector $H_{1,i}^{(l)}$ and $H_{2,i}^{(l)}$ of XLNet are spliced, and the obtained vector h_i is the final vector representation after fusion. Finally, at the decoding end, each time step predicts the next target language word based on the vector representation at the encoder end and the output from the previous step, and outputs the generation probability of the next word.

2. Chinese-English Patent Machine Translation Model Combining Fusion Strategy and Improved Transformer

A Chinese-English patent machine translation model is established based on the proposed patent knowledge fusion strategy using embedded annotations and XLNet-Transformer encoder-decoder framework. The model mainly includes three parts: preprocessing module, encoder module, and decoder module. In the preprocessing section, the model first focuses on the knowledge fusion strategy, and uses a sequence annotation model to extract patent terms based on sub word segmentation. The extracted knowledge of different categories of terms is fused through embedded annotations and input into XLNet-Transformer encoder-decoder. In the encoding part, XLNet preprocessing model integrates the semantic information and the left and right contextual information through the dual-stream self-attention mechanism and feedforward neural network. Finally, in the decoding part, Transformer decoder obtains the vector representation of the word information in the sentence by multi-head self-attention mechanism, predicts the next target language word according to the source language and generated target language context information, and outputs the probability distribution of the word.

3. Simulation Experiment

Based on Ubuntu 18.04.5 LTS operating system and PyTorch deep learning framework, a Chinese-English patent machine translation model combining the proposed fusion strategy and improved Transformer is constructed and simulated. The system is equipped with Intel(R) Xeon(R) Gold 6152 CPU @ 2.10GHz CPU and NVIDIA Tesla V100 graphics card, and the video memory is 32GB. In the training process of Chinese-English patent machine translation model, the dimension of the feedforward fully connected layer of XLNet encoder is set to 2048, the batch size is 64, and the maximum length of the sentence input is dynamically processed using AISL adaptive processing. The learning rate and random deactivation probability are $3.0\times10-4$ and 0.1, respectively. The cluster search size of Transformer decoder is set to 4.

In order to verify the effectiveness of patent knowledge fusion strategy with embedded annotations and XLNet-based machine translation optimization method for improving model

performance, the simulation experiment results of the constructed model are compared with the translation results of the basic Transformer model without patent knowledge and the basic Transformer model with patent knowledge. Bilingual Evaluation (BLEU), a commonly used score for translation tasks, was used as the evaluation index of the translation results of the model.

The training corpus for the model in the experiment comes from the compilation of Chinese-English corpus of utility model patent texts. 500000 Chinese English parallel sentence pairs containing domain knowledge were extracted as the training set, and 1500 and 2000 Chinese English parallel sentence pairs were extracted as the validation and testing sets. BERT+CRF+Bi-LSTM sequence labeling model is used to extract a certain amount of corpus from each data set, and the domain knowledge contained in it is extracted and labeled. The number of corpus extracted from training set, verification set and test set is 1847, 206 and 206, respectively.

4. Conclusion

Making the machine translation model generate more accurate, standardized and professional translation results for patent domain texts, this paper proposes a Chinese-English patent machine translation method combining domain knowledge fusion strategy and improved Transformer model model, and conducts translation experiments based on utility model patent texts. The results show that compared to the basic Transformer model without fusion strategy, the BLEU value of the basic Transformer model with fusion strategy is increased by 3.77 and 1.79 in Chinese-English and English-Chinese translation tasks, respectively. This indicates that the patent knowledge fusion method based on embedded annotations can effectively improve the translation performance of machine translation models. On this basis, the proposed improvement method of XLNet-based Transformer model further improves the learning ability of the model for patent text. Finally, compared with the basic Transformer model without the fusion strategy, BLEU value of the proposed Chinese-English patent machine translation model combining fusion strategy and improved Transformer for Chinese-English and English-Chinese translation tasks is increased by 5.94 and 2.25, respectively, and its translation results obtained are more accurate, standardized and professional in expression, which is more in line with the translation requirements of relevant documents in the field of utility model patents.

Conflict of Interest

The author declares that she has no conflicts of interest to this work.

References

Xiong, X. (2021). Research on errors in machine translation of patent documents from Chinese to English. *Overseas English, 2021*(11), 212–213, 218.

Cao, J., Wu, X., & Wang, Q. (2022). Research on parallel corpus construction for multi-language patent machine translation. *China Invention and Patent, 19*(06), 70–75, 80.

You, X., Yang, H., & Chen, H. (2021). Research on new energy patent machine translation by integrating terminology information. *Journal of Chinese Information Processing, 35*(12), 76–83, 93.

Han, D., Ye, N., & Zhang, G. (2023). Neural machine translation for patent text fusion domain knowledge. *Journal of Computer Applications and Software, 40*(12), 160–168.

Du, Z. (2023). Weihan neural machine translation based on transformer. *Electronic Design Engineering, 31*(22), 47–51.

Chao, Z., Ye, C., & Han, X. (2022). Research and development of Chinese-English machine translation system based on transformer. *Computer Knowledge and Technology, 18*(27), 16–17, 20.

Li, D., Shan, R., & Yin, L. (2022). Sentiment analysis of Chinese text based on XLNet. *Journal of Yanshan University, 46*(06), 547–553.

Zhao, Z., Che, J., & Lv, W. (2024). Text image generation method based on XLNet and DMGAN. *Chinese Journal of Liquid Crystals and Displays, 39*(02), 168–179.

How to Cite: Chai, H. (2024). Research on Chinese-English Patent Machine Translation Based on Fusion Strategy Model. *Journal of Global Humanities and Social Sciences*, *05*(04), 156-161.
https://doi.org/ 10.61360/BoniGHSS242016160404