4W-05

Corpus Augmentation Based on Pseudo-Chinese Generation for Chinese-Japanese Neural Machine Translation

Zheng Wei[1] and Yohei Seki[2]

[1,2]*University of Tsukuba, Tsukuba, Ibaraki*

1 Introduction

With the significant development of the neural network, the neural machine translation (NMT) models have reached the state-of-the-art. However, compared with other machine translation models, the NMT model has a larger requirement for the volume of data. Koehn et al. [2] has proved that once there lacks a large-scale parallel corpus, the translation quality of NMT will be severely constrained or underperformed by other machine translation models. To deal with the situation of low resource, corpus augmentation is a powerful solution. However, the studies related to corpus augmentation for Chinese and Japanese are very scarce compared to European languages such as English and French. Therefore, we decided to adopt an approach to corpus augmentation for the Chinese-Japanese parallel corpus to improve its performance on NMT models.

In this study, we introduce a new approach called "Pseudo-Chinese Generation" (PCG) to augment the size of the Chinese-Japanese MT corpus by efficiently exploiting many types of similarities between Chinese and Japanese to "create" Chinese-Japanese parallel corpus directly from the Japanese monolingual corpus. We tested the performance of our method for boosting results under different resource settings and compared it with another method of corpus augmentation.

2 Pseudo-Chinese Generation (PCG)

One of the many similarities between Chinese and Japanese is the use of a large number of Chinese characters, known as "*Hanzi*" in Chinese and "*Kanji*" in Japanese. In 2016, there was a popular Internet slang called "Pseudo-Chinese" that was used to create new sentences by removing the *Kana* characters part of Japanese sentences so that people could understand them without the ability of understanding Japanese. Inspired by such Internet slang, we started our work of constructing corpus augmentation system using PCG. Obviously, simply removing the *Kana* characters part will greatly damage the original meaning of the sentences. To deal with such issue, we tried to find solutions to keep the original meaning. For the words that contain *Kana* characters, we tried two alternatives: replacing the original words by their synonyms that fully consist of *Kanji* characters, or make a simple dictionary to translate the words into Chinese expressions, which we called *Kanji substitution* and *Chinese expression substitution* respectively. Besides, not all sentences in the monolingual corpus are suitable to generate Pseudo-Chinese, especially the sentences that contain proper nouns in *Kana* characters. To filter them out, we also fine-tuned a filtering model based on Pretrained Japanese BERT model[1]. To summarize, the workflow of PCG is shown in Figure 1.

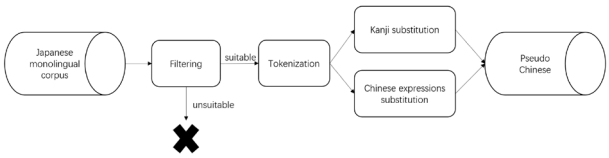

Figure 1 Workflow of Pseudo-Chinese Generation

2.1 *Kanji* substitution

Our approach is mainly based on the WordNet and Word2vec [3] model. Firstly, we detect the words fully consists of *Kanji* or special characters (including alphabet, number, etc.). Note that they can output directly without any operation. In case we have found the words including *Kana* characters, we try to substitute them with *Kanji* word. We prepared three types of candidate words for generation: (1) the *Kanji* synonym of the word that has the most similar meaning in Japanese WordNet [2] as candidate W_1; (2) the substring removed *Kana* characters from the word as candidate W_2; and (3) the representative notation of the word in (2) by Juman++ [4] as candidate W_3. Then we apply Word2vec to calculate the similarity between the candidates and original word to select the appropriate word with the most similar meanings as our target word.

2.2 Chinese expression substitution

In contrast to the words that have obvious meanings (nouns, verbs, or adjectives), it is not easy to directly

[1] https://github.com/cl-tohoku/bert-japanese

[2] https://bond-lab.github.io/wnja/index.en.html

Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.

Table 2 BLEU score of each model

Model	Augmented Corpus	Filtering	20k	50k	100k	200k
Baseline	N/A	N/A	19.37	33.10	39.69	41.42
Proposed Method (PCG)	263k	**Yes**	**31.46** (+12.09)	**36.58** (+3.48)	40.42 (+0.73)	42.16 (+0.74)
Proposed Method (PCG)	608k	No	**30.60** (+11.23)	35.78 (+2.68)	40.33 (+0.64)	42.40 (+0.98)
Comparison Method (BT)	608k	No	28.60 (+9.23)	36.04 (+2.94)	41.11 (+2.42)	42.40 (+0.98)

generate *Kanji* synonyms over the remaining words, especially for the four POS categories: adverb, particles, conjunction, and suffix. Fortunately, these words tend not to be deformed, so we can directly translate these words by doing a direct translation of them. However, even though they are limited, the amount of these words prevents us from translating all of them.

To keep as much information as possible while doing the Chinese expression substitution, we have counted the frequency of occurrence of these four types of words in the part of the monolingual corpus. Then, we selected the top five frequently occurring words in these four POS categories and substituted them with the corresponding Chinese words. We also defined substitution rules for particles and pronouns because they have a much lower amount of words. In Table 1, we give an example of the Pseudo-Chinese.

Table 1 Example of Pseudo-Chinese

Japanese	しかし紙の黄変と繊維の損傷が起こった。
Pseudo-Chinese	但紙之黄変和繊維之損傷発生。

2.3 Filtering noise from monolingual corpus

In this module, we prevent excessive noise from being mixed into the model to improve the generation quality as much as possible. At first, we extracted a part of sentences from the target monolingual corpus and generated the Pseudo-Chinese by our method. Then, we annotated whether the generated Pseudo-Chinese sentences have same meaning to the original sentences. Finally, we trained a text classification model by using the BERT [1] to filter out the noisy sentences for PCG in the huge monolingual corpus.

3 Experiments

3.1 Method

We built our translation system using the open-source NLP tool Fairseq [3]. We used Transformer as our baseline model. To evaluate the quality of the translation, we used the BLEU score. For the original parallel corpus, we used the ASPEC-JC corpus[4]. For the monolingual Japanese corpus used to augment the original corpus, we employed Wikipedia JP[5], which consists of 608k Japanese sentences (and of 263K after applying filtering module in Section 2.3). In order to verify the effect of PCG on the translation performance of the model under different resource settings, we randomly selected 20k, 50k, 100k, and 200k Chinese and Japanese utterance pairs from the original parallel corpus to construct the baseline model respectively. As a comparison, we also tested the improvement of back translation (BT) and unfiltered Pseudo-Chinese on the original model under the same resource settings.

3.2 Results

The results are shown in Table 2. From the results, we can see that under low-resource settings, our proposed method has better performance than back translation. Also, we have also proved the effectiveness of the filtering module described in Section 2.3.

Acknowledgements

This work was partially supported by the Japanese Society for the Promotion of Science Grant-in-Aid for Challenging Exploratory Research (#22K19822), Grant-in-Aid for Scientific Research (B) (#19H04420), and by ROIS NII Open Collaborative Research 2022 (22S0103).

Reference

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL-HLT*, pages 4171-4186, 2019.

[2] Philipp Koehn and Rebecca Knowles. Six Challenges for Neural Machine Translation. In *Proc. of* the *First Workshop on Neural Machine Translation*, pages 28-39, 2017.

[3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Proc. of NIPS 2013*, pages 3111-3119, 2013.

[4] Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. Juman++: A Morphological Analysis Toolkit for Scriptio Continua. In *Proc. of EMNLP 2018: System Demonstrations*, pages 54-59, 2018.

[3] https://github.com/facebookresearch/fairseq
[4] https://jipsti.jst.go.jp/aspec/
[5] https://kairozu.github.io/updates/japanese-wiki-corpus

Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.