Winter Project Poster

Zhenhua Kong
kon.gzhen@163.com

Abstract

This report briefly describes the work during the winter vacation internship. It mainly includes project introduction, data construction, model training, model evaluation, etc. With the continuous expansion of the application scope of machine translation, the demand for web translation is also expanding, but it is difficult for existing models to retain tags. Here, a generalization method is proposed. By first generalizing the specimen of the analyst, Then, the method of tag restoration for disputes is to achieve the purpose of retaining tags, thereby improving the quality of web page translation. The method proposed here can use the Fairseq framework to achieve a tag retention rate of more than 90% with 10,000 tagged Chinese to English training data.

1 Introduction

The application of neural machine translation models is becoming more and more widespread, and is also used in web translation in translation companies. However, the existing models do not have a good protection effect on HTML tags in web translation. This allows tags to be preserved during translation.

The system is able to successfully translate text paragraphs containing HTML tags without modification or with only a few manual modifications.

We mainly use Facebook's open source Fairseq framework, which successfully applies convolution to machine translation, and is superior to traditional RNN-based models in both effect and efficiency.

Based on this, we will train multiple models and finally choose the one with the best translation effect based on the evaluation method. Therefore, we constructed three models and three data processing methods, combined with each other to evaluate the best solution.

Three data processing methods:

1. Do nothing with the label
2. Protect tags during word segmentation
3. Protect tags during both word segmentation and BPE algorithms

Then evaluate the model with Precision, Recall, F1, BLEU values, etc. Find the one that works best, and generate a model.

2 Data set construction

2.1 Overview

For the data set, we got 100,000 Chinese-English bilingual corpora from the translation company, it is not enough compared to the trained model, so we have to construct train data.

First, we analyze the syntax of html. In HTML, markup elements in markup are enclosed in angle brackets, and elements with slashes indicate the end of the markup description; most markups must be used in pairs to indicate the beginning and end of the effect.

Type	L1	L2	L3
Frequent location	Both ends of the sentence	The inside of the sentence	The inside of the sentence
Appearance	Appear in pairs	Appear in pairs	Appear alone
Example	<a> <p> <q> <h1>	<b> <small> <mark> <span>	<br/> …

	<title> …	<strong> …	

Table 1: HTML tag classification

In addition, some tags appear only in pairs on both sides of the complete sentence, and we name them L1. Some tags appear only in pairs inside the complete sentence, and we name them L2. And there are some tags that appear only in pairs and we name them L3.

Then we find relatively high quality bilingual corpora(wmt2018, 8 million Chinese-English bilingual corpora, with the source language being Chinese and the target language being English).

For Chinese corpus, we use jieba word segmentation, and for English corpus, we use the script in moses. Then, we use fast_align to get the word alignment information.

Then we construct the training data set and test data set. Based on the word alignment information, we randomly insert L1 tags at both ends of the sentence and randomly insert L2 and L3 tags inside the sentence. Then add corresponding tags to the corresponding target language according to the word alignment information. We use it as training data set, and the test data set uses the data given by the company.

2.2 Structural target

Our goal is to construct a large number of labeled bilingual parallel corpus. Tags mainly come from tags in HTML syntax, and the constructed data should conform to HTML syntax (including tag location and tag closure, etc.).

2.3 Data Sources

- Unlabeled bilingual corpus based on wmt2018, 8 million Chinese-English bilingual corpora, with the source language being Chinese and the target language being English.

- The tags inserted mainly come from the tags that appear in the HTML grammar and the high-frequency tags provided by the translation company.

2.4 Construction method

2.4.1 Process

1. Perform data preprocessing, including Chinese and English word segmentation, lowercase, clean, and use fast_align to construct word alignment information.
2. Read single sentence Chinese corpus, corresponding English corpus and word alignment information.
3. Randomly decide (50%) whether to insert L1 type tags at both ends.
4. Randomly decide how many L2(and L3) tags to insert in the sentence.
5. For each L2 (and L3) tag, randomly decide where to insert.
6. Repeatedly generate positions until no interleaving with the previous label.

2.4.2 Framework

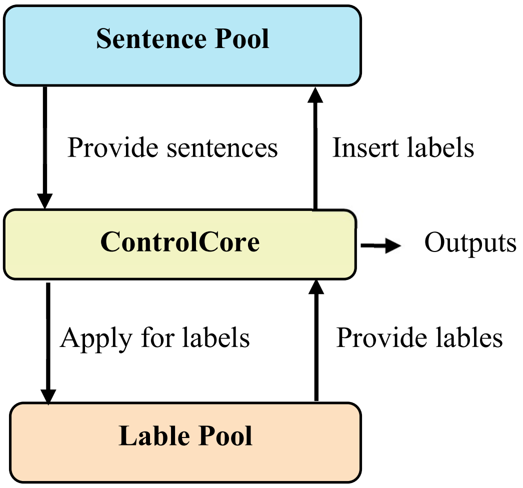

Figure1: Our data generator framework

Lable Pool: This type of polling provides the name of the label to be inserted.

Sentence Pool: This type of operation is for sentences. It can provide sentences and support the insertion of tags at any position in the sentence. At the end, the completed sentence is stitched and returned.

ControlCore: This type controls the entire data construction process. Most of this type of data is randomly generated, including the position and number of insertions. For each insertion, it is compared with the previous insertion position. The insertion position is repeatedly generated until it is compared with Until the previous insertion does not interleave.

2.5 Advantages

Data constructed in this way has the following advantages.

Comprehensive: Lable Pool uses polling internally to ensure that each tag (L1, L2, L3) appears evenly.

Conform to the grammar: ControlCore performs judgment and analysis on each insertion to ensure that there is no interleaving between the tags.

Uniform label distribution. Random position insertion can ensure that the distribution of labels in the sentence is uniform throughout, that is, from a global perspective, labels have appeared in all positions, which is convenient for later training. In addition, the number of label insertions is related to the length of the sentence, so it can avoid labels Overly dense problem.

Correspond: Utilize the word alignment information and synchronize the insertion of Chinese and English corpus at the same time, which can ensure that the tag insertion position of Chinese and English corpus can correspond one to one.

3 Train

3.1 Lable generalization

We used four different methods to generalize the label, and after the experiment, we chose the method with the best retention effect.

Baseline- We will train a baseline, that is, do not perform any special processing on the label, and try to translate directly with the existing model. We use this as a baseline to compare our method for label retention accuracy and translation quality.

NMT with replication - We try to add tags at both ends of the label, in the form of '$ copy <b> $ copy', and then consider the three parts as a whole.

We marked the ends of the tag in this way to see if the neuro-machine translation model could rely on this to recognize the tag and learn to retain the contents between '$copy'.

Full generalization with pair - We improve on the basis of the previous method. We generalize each label, that is, ignore the specific content of the label, and only record its position and number (it can also record whether it is a start label or a closed label). Then we set a parameter k, that is, we only record the labels of the first k pairs of labels, such as 'lable_l1', and we consider the unpaired labels themselves as a pair, no longer distinguishing whether it is a start label or a closed label. Record them in the form of 'lable_1'. And we do not distinguish between the remaining labels, such as' $ lable '. We use the same markup in the translation to represent the remaining tags. Then we replace the content in order. Compared to fully recording the position of all tags, this method can achieve relatively good results under the current tight computing resources. This method of generalization and then restoration can effectively protect the tags in the translation, so as to improve the retention rate of the tags.

Full generalization without pair - We also tried a generalization method that does not distinguish between left and right. For all labels, we record in the form of '$lable_1'. Like the previous generalization method, we use a parameter to control record the number of subscripts. The remaining tags are generalized in the form of '$ lable'. This way, compared with the previous way to distinguish left and right generalization, the label retention effect is better.

3.2 Lable protection

In order to improve the retention effect of the label, we use three different label segmentation methods, observe them separately, and choose the best method from them.

1. Do nothing with the label. This is used as a baseline.
2. Protect tags during word segmentation. This can avoid the destruction of the tag during the word segmentation.
3. Protect tags during both word segmentation and BPE algorithms. In this way, the labels can be fully preserved, and the loss during translation is minimized.

3.3 Preprocess

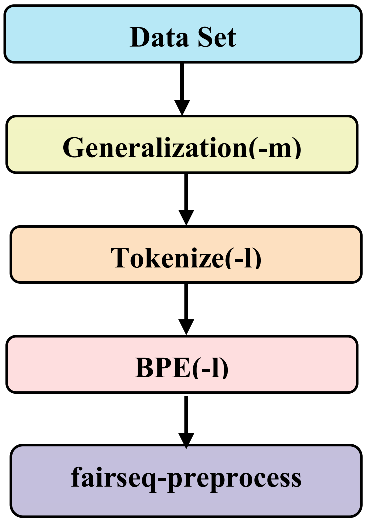

Figure 2: Our data preprocess framework

Use the **-l** and **-m** parameters to control the type of label protection and label generalization type, which is convenient for experiments.

-l can control the type of label protection, that is, 'not protected', 'protected when segmented', 'protect both segmented and bpe'.

-m can control the type of label generalization, that is, 'baseline', 'NMT with replication', 'fully generalized'.

4 Evaluation model

Since the focus of this system is translation of tagged text, measuring the retention rate of tags is an important indicator for evaluating the effectiveness of translation models.

4.1 Evaluation parameters

The evaluation model is mainly to check the retention of labels. Therefore, the evaluation is performed from four perspectives.

Precision: Precision indicates the percentage of tags in the original text that remain in the translation.

Recall: The recall rate indicates the proportion of corresponding tags in the original text found through the tags in the translation.

F1 value: The F1 value represents a comprehensive evaluation value considering accuracy and recall.

BLEU: The BLEU value represents the frequency calculation of the common words output by the model and the reference translation, which can reflect the label retention to some extent.

4.2 Calculation method

$$Precision = \frac{lable\_num\_in\_both}{lable\_num\_in\_cn}$$

$$Recall = \frac{lable\_num\_in\_both}{lable\_num\_in\_en}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

From the above three formulas, we know that the total number of labels in the original text is called **lable_num_in_cn**, the total number of labels in the translation output by the model is called **lable_num_in_en**, and the number of common parts between them is called **lable_num_in_both**. According to the formula, we can calculate Precision, Recall and F1 value.

The BLEU value is calculated by calling the multi-bleu script provided in moses.

4.3 Effectiveness of evaluation method

First of all, the accuracy, recall rate and F1 value of the label statistics are calculated based on the number of tags in the original and translation, so it can reflect the retention of the label. If the label is damaged and lost, it can be on the F1 value intuitively.

In addition, since the BLEU value generated by the translation result that the tag successfully retains must be higher than the BLEU value that translates the tag into garbled characters, the BLEU value can also evaluate the effect of the translation model.

In summary, this set of evaluation models can well evaluate the retention effect of labels.

5 Experiment and analysis

5.1 Data Set

We carried out our experiments on Chinese-to-English translation.

The training data set and valid data set are obtained by taking 100,000 pairs of sentences from WMT18 and then processing them by the four generalization methods mentioned above.

Because most sentences in the original data provided by the translation company is too long, too many tags, has a large number of unpaired tags, and will affect the evaluation of translation quality, so we selected 2000 sentences that meet the limits as the test set.
The restrictions are as follows:

- The number of words in the sentence is in the range of [20,50]. This is to avoid translation degradation caused by too long sentences.
- The total number of words in all tags must not exceed 40% of the total number of words in the sentence. This is to prevent tag retention from affecting BLEU values too much.
- Any tag in the sentence does not exceed 10 words.

Sentences that meet the above criteria can effectively evaluate the retention rate of tags and the quality of translation.

Our final data set structure is shown in table 2.

Data set type	**Data scale**
train set	95,834
valid set	4,166
test set	2,000

Table2: Structure of our dataset

In addition, the number of our generalization records is selected to be 3, which is to record the numbering information of the first three labels, and the rest are not recorded.
For example:
Baseline:
Human <b> damages </b> <br/> environment <br/>.

NMT with replication
Human $copy <b> $copy damages $copy </b> $copy $copy <br/> $copy environment $copy <br/> $copy.

Full generalization with pair:
Human $lable_l1 damages $lable_r1 $lable_2 environment $lable_3 .

Full generalization without pair:
Human $lable_1 damages $lable_2 $lable_3 environment $lable.

5.2 Experimental Process

In the case that both the segmentation and bpe subword segmentation retain the labels, the experimental results are as follows:

Generalization type	**Precision**	**Recall**	**F1 value**
Baseline	NaN	NaN	NaN
NMT with replication	NaN	NaN	NaN
Full generalization with pair	85.2%	99.4%	91.8%
Full generalization without pair	91.1%	99.9%	95.4%

Table3: The result of label retention

The reason why Baseline and NMT With Replication do not have accuracy is because experiments have found that if the label is not generalized and the translation model is directly translated, the label will almost never be retained, but will be destroyed so that it cannot be identified tags in the translation. So it's unable to calculate precision, recall and F1 value.
It can be easily seen from the comparison that, in this process, the generalization of labels directly and then restoration can greatly improve the precision of label retention. Among them, generalization without distinguishing left and right can retain most of the labels. This is because the generalization method can avoid breaking labels during translation.
In addition, even after cleaning the test data, there are still problems with the test data having many tags, the length of a single tag, and too many non-Chinese characters (such as links, English words) in the Chinese corpus. These factors lead to deviation in BLEU. The test set contains a large number of unpaired tags, which is abnormal compared to normal HTML files. This is also a factor that affects the score.
In order to verify that the labeled training data will affect the translation quality, without changing model parameters, the following tests were performed:

The model trained using 100,000 unlabeled data has a BLEU value of 4.09 after translating two thousand unlabeled data.

The model trained using 100,000 labeled data has a BLEU value of 2.06 when translating two thousand unlabeled data.

It can be seen that the fully labeled training data will have a slight impact on the translation quality of the model, so some unlabeled training data should be added to the translation for training, so as to improve the translation quality , and unlabelled data should be mixed in proportion to labelled data.

6 Conclusion

This poster proposes a label generalization method, which achieves a good label retention effect.

It is unacceptable that translation of tagged text without translation will cause a large area of missing and errors in the tag.

Although the translation quality has decreased on the test set, the full generalization method without distinction between left and right proposed in this paper has more than 90% retention accuracy under 100,000 training data.

Although the generalization method that does not distinguish left and right can achieve better accuracy than the generalization method that distinguishes left and right, but there is still a lack of recording of the left and right information of the label. The specific method is better and requires further experiments.

In addition, the system should also add a module that can generalize labeled data. At present, the generalization of labeled data in this system is performed at the time of generation, and this aspect needs improvement.

To obtain better translation quality, also requires larger training and more suitable model parameters. When training data, labeled training data and unlabeled training data can be mixed according to a certain ratio.

References

Liang Ding, Lunjie Li, Jianghong Han, Yuqi Fan, and Donghui Hu: Detecting Domain Generation Algorithms with Bi-LSTM. Computers, Materials & Continua, vol.61, no.3, pp.1285-1304, 2019.