Winter Project Proposal

Zhenhua Kong
kon.gzhen@163.com

Abstract

This proposal briefly outlines my project introduction, solutions, method evaluation, time planning, and personal goals during my winter vacation internship.

1 Introduction

NMT uses neural network-based technology to achieve more context-accurate translation, rather than translating broken sentences one word at a time. Using a large artificial neural network to calculate the probability of a sequence of words, NMT puts the complete sentence into an integrated model.

In the face of globalization, many websites need to translate their web pages into foreign languages in order to expand the international market. Based on this situation, translation companies have a need for machine translation of html text.

Although neural machine learning has shown good results when translating ordinary text, there are still many problems when facing some tagged sentences (such as html text). For example, Google Translate has the problem of misplaced labels and Youdao translation exists. The problem of missing labels. In fact, there may be cases where the content of the label itself is translated as text. These problems will cause the final web page to be displayed incorrectly. In fact, we must not translate the label, but we must retain the content and position, and then translate the remaining text to ensure that the translated html still displays the correct web page.

In summary, the goal of this system is to display the correct webpage after a section of HTML code is translated without requiring modification or only a small amount of manual modification.

2 Approaches

We decided to use the Facebook's open source fairseq framework as the basis to train the model by marking the tags in HTML so that it can memorize the tag positions and retain the tags, while replacing the translations of the corresponding positions.

Then use this model to translate labeled text, record the evaluation results, and then try different types of labeling methods to compare and get the best model.

In addition, we will also test Google Translate a lot, get inspiration by observing the translation results of some labeled sentences, and then improve our system.

2.1 Challenges

Complexity of tags-There are many tags in HTML, such as '', which are often inserted into the text, and they should be ignored during translation, and the sentences on both sides should be combined and translated. But for other tags, such as ' ', the internal text is considered as a whole. This situation causes translation difficulties.

Diversity of tags. There are a lot of tag categories in HTML, and the tags themselves often carry some other additional information, such as hyperlinks represented by '<a>'. In addition, because tags appear in pairs, Appears alone, which greatly increases the diversity of tags.

Randomness of tags. Just as we cannot predict user input, we cannot predict the possibility of HTML tags. In fact, HTML syntax is very flexible, and there may be a large number of nested relationships, which also brings a comparison to translation. Big difficulties.

Lack of computing power- In addition, considering that the laboratory's server GPU

resources are relatively lacking, and this project requires multiple methods to test and compare, it is difficult to train relatively complete models in a short time.

2.2 Our Approach

The Fairseq translation model was proposed by the Facebook AI Lab in 2017. Compared with the previous RNN-based translation models, the cnn-based model structure is adopted.

The essence of our project is to try multiple tag processing methods, find the one that works best, and generate a model. We constructed three models and three data processing methods, combined with each other to evaluate the best solution. Here are a few approaches we intend to take.

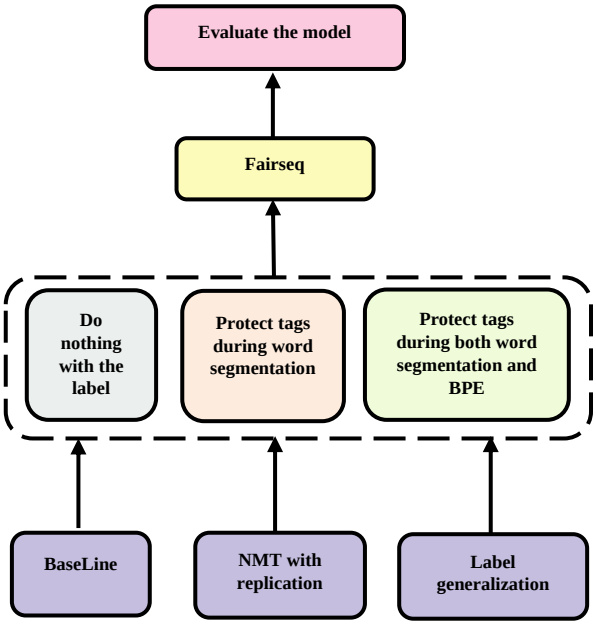

Figure1: Our solution framework

Baseline- First, 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 will also do some comparative experiments in this Baseline.

There are three main types:

- Do nothing with the label
- Protect tags during word segmentation
- Protect tags during both word segmentation and BPE algorithms

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, then take the three methods in Baseline to train separately, then test and record evaluation results.

Label generalization - 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.

Analyze Google Translate. As a leader in this industry, Google's translation of tagged text can give us some inspiration. However, its translation should be a compromise between performance and cost. In the actual use process There are also some defects found in this. We will learn from its advances in combination with our needs.

3 Data Set

For the data set, we got some data from the translation company, but it is not enough compared to the trained model, so we have to construct some data.

First we analyze the syntax of html, and analyze which tags often appear inside the text (this type is called L1), which tags often appear on both sides of the text (this type is called L2), and which tags appear unpaired (this type is called L3, and unpaired ones often appear in the middle of sentences to construct enough training data).

Then we find relatively high quality bilingual corpora. According to the word segmentation, for the source language, L1 or L3 tags are randomly inserted between words, and L2 tags are randomly inserted on both sides of the sentence. Then add corresponding tags to the corresponding target language according to the word alignment information. Considering that we have more bilingual corpora, we can construct enough training data.

In addition, there are some data given by the company, which we use as test data. Because the company's data may have some preferences.
In order to improve the translation accuracy of some special tags (such as a tag that is almost only inserted on both sides of a named entity),
Considering that the Named Entity Recognition algorithm is relatively mature, we can construct data for this property, so that the model's recognition of this label is greatly enhanced.

4 Project evaluation method

We decided to evaluate the translation effect of the model in three ways.

- Utilize indicators such as label retention accuracy, recall, and F1 value. Is to count the number of tags of the original and translation. Then calculate the retained label ratio.
- Use the BLEU value for estimation.
- Limit the evaluation conditions manually. You can manually estimate whether the surrounding tag results for a term are retained or not. In addition, you can estimate the retention of labels when they appear in the context of the specified vocabulary based on the specified labels.

5 Time Line

5.1 Week 1 (1/8 – 1/11)

Communicate the project tasks and solutions with the senior, evaluate the feasibility, and determine the follow-up work plan.

5.2 Week 2 (1/12 – 1/18)

Analyze HTML grammar and count the occurrence probability of various HTML tags of translation companies. Construct training data.

5.3 Week 3 (2/1 – 2/8)

Test the three cases of baseline and then evaluate the model.

5.4 Week 4 (2/9 – 2/15)

Training and evaluation of NMT model with replication mechanism. Training and evaluation of label generalization models

5.5 Week 5 (2/16 – 2/22)

Make a comprehensive comparison of the above models, choose the most effective model, and then write a summary report.

6 Personal goals

Complete the training of the model and complete the report. The final model can do that, for the input tagged text, it can be changed into HTML code that can directly display the web page with little or no modification.