Domain-Mixing for Chinese-English Neural Machine Translation

Christopher Leege

December 11, 2018

Advisor: Prof. Dragomir Radev, Dept. of Computer Science, Yale University

Abstract:

While there is a plethora of parallel corpora for languages that are similar, such as Indo-European languages, it is much harder to find corpora for language pairs that are more distance. The goal of this project was to translate Chinese fiction into English. While there are parallel corpora for Chinese-English, the majority focus on nonfiction domains, such as news and legal documents. I was unable to find any parallel corpora that dealt with modern Chinese fiction. As a result, I created my own corpus out of a popular Chinese web novel that has an official English translation. I then trained four neural network translation models and tested how they performed on both data sets. The first neural network model was trained on the nonfiction corpus, while the second model was trained on the fiction corpus. The third and fourth models were trained on a mixture of the two corpora. The third model was trained on a naïve mixture of the nonfiction and fiction corpora, while the fourth model was trained on a mixture of the nonfiction and fiction corpora with a technique called target token mixing. Target token mixing introduces a tag at the beginning of every sentence in the target training and validation sets. This tag corresponds to the domain of the sentence, and encourages the neural network to predict the domain of the sentence, adding a regularizing effect. Both the naïve mixture and the mixture with target tokens performed better than the models trained exclusively on the nonfiction or fiction corpus, which was expected. Unexpectedly, the naïve mixture performed worse than the mixture with target tokens. It is likely that more data is required for target token mixing to perform better than a naïve mixture. Another possibility is the nonfiction corpus was already too heterogenous in composition, and as a result, target token mixing did not help significantly.
1 Introduction

Chinese webnovels have wide-ranging appeal in China, with the China Internet Network Information Center reporting that 333 million Chinese, or 45.6% of China's internet user base, read webnovels in 2016. In addition, Chinese webnovels are becoming increasingly popular abroad. The most popular English language translation website receives an estimated 300,000 unique visitors daily, with the average traffic in 2018 having doubled the average traffic in 2017. The official English language website for Qidian, the largest Chinese webnovel publisher, was established in 2017. Since the establishment of their English language website, Qidian has added an average of two new English translations of their Chinese webnovels every week, published serially. In this project, I have trained a neural machine translation system for translating Chinese webnovels into English.

2 Methods

The largest challenge for creating neural machine translation systems is the data. Neural machine translation requires a massive amount of data in the form of parallel sentences. For Chinese-English translation, the requisite number of parallel sentences necessary to achieve passable translation quality is enormous. Unfortunately, Chinese-English corpora are much rarer than corpora between Indo-European languages. There is also a lack of diversity in the corpora that exist between Chinese and English. Most Chinese-English corpora are very domain specific, with news, movie subtitles, and legal proceedings being some of the common domains represented. Parallel corpora for domains such as fiction are rare if not nonexistent. Parallel corpora for Chinese webnovels do not exist.

For this project, I created my own parallel corpus of 45,402 sentences from the Chinese webnovel 天道图书馆, translated into English as Library of Heaven’s Path. The English
sentences were pulled from the official English translation of the webnovel. Unfortunately, a single corpus of 45,402 sentences is still far too small for neural machine translation.

In order to circumvent the issue of limited data, I attempted a form of data preprocessing called target token mixing. Target token mixing introduces a tag at the beginning of every sentence in the target training and validation sets. This tag corresponds to the domain of the sentence, and encourages the neural network to predict the domain of the sentence, adding a regularizing effect. Employing target token mixing, enabled me to combine multiple corpora together for training, increasing the size of the data set, giving the neural machine translation model more data to work with.

I used the nltk tokenizer and jieba, a Chinese tokenizer, to tokenize the Chinese and English corpora. I trained the neural networks using OpenNMT, an open source neural machine translation model. It employs a basic encoder-decoder model with recurrent neural networks.

3 Corpora

For the nonfiction corpus, I used the casia2015 corpus, a corpus of parallel sentences scraped from pages around the web. For the fiction corpus, I used the parallel corpus that I created.
4 Models

All four models used the same encoder-decoder structure:

The only differences in the four models were the corpora that the models were trained on. The first model was trained on the non-fiction corpus, the second model on the fiction corpus, the third model on a naïve mix of the two corpora, and the fourth model on a mix of the two corpora with target token mixing.

For this project, I trained a total of eight models. The first four models had a nonfiction corpus size of just 100,000 parallel sentences, while the second four models incorporated the entire nonfiction corpus with 1,000,000 parallel sentences.
5 Results

For the first four models, both the naïve mixture and the mixture with target token mixing performed better than the pure Casia and pure Fiction corpora. The naïve mixture performed better than the mixture with target tokens on both Casia and Fiction test data. The target token was correctly predicted 2317/2400 times on the Fiction data set, and 1044/1050 times on the Casia dataset.

<table>
<thead>
<tr>
<th>Smaller</th>
<th>Casia2015</th>
<th>Fiction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Casia2015</td>
<td>8.79</td>
<td>2.01</td>
</tr>
<tr>
<td>Fiction</td>
<td>1.12</td>
<td>11.79</td>
</tr>
<tr>
<td>Mix</td>
<td>9.44</td>
<td>13.67</td>
</tr>
<tr>
<td>Mix+Token</td>
<td>9.37</td>
<td>13.01</td>
</tr>
</tbody>
</table>

Bleu Scores for the four models trained with the smaller portion of the casia2015 corpus.

For the second four models, the pure Fiction dataset performed better than either of the mixed datasets on Fiction. The naïve mixture performed best on the Casia test data, and the pure Casia and the mixture with target tokens performed equally well on Casia test data.
<table>
<thead>
<tr>
<th>Larger</th>
<th>Casia2015</th>
<th>Fiction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Casia2015</td>
<td>14.38</td>
<td>3.42</td>
</tr>
<tr>
<td>Fiction</td>
<td>0.74</td>
<td><strong>10.60</strong></td>
</tr>
<tr>
<td>Mix</td>
<td><strong>14.43</strong></td>
<td>10.50</td>
</tr>
<tr>
<td>Mix+Token</td>
<td>14.38</td>
<td>10.03</td>
</tr>
</tbody>
</table>

Bleu scores for the four models trained with the full casia2015 corpus.

Here we compare the model trained exclusively on casia2015 with the naïve mixed model:

**Model 3 predicting Fiction:**

The vast range of the mountains, the tree vegetation covering it, could it be possible for a formation to hide in the formation... Why would you find it?
```
``Teacher ...``
```
The crowd had no choice but to look around, but they could only focus on Zhang Xuan's body.
``This ...``
Zhang Xuan scratched his head. ``Let's see whether I am useless or not. I have the the the the front of my family and I don't know where...``
Halfway through his words, his eyes suddenly lit up, ``Right, I thought that a solution is way above that of mine. There's no problem with it!``
``Is there a way to do so?``
Everyone was taken aback, and they stared at Zhang Xuan with glowing eyes.
Mo Yu was also curious.
Model 1 predicting Fiction:

The mountain of the mountain, trees covered with trees and could still hide hide... How do we find?
`` The teacher...``
The people looked round at all, and there was no way to put the eyes on the 悬.
``...``

张悬 挠 was n't 挠. `` I have no the 故居, `` the keeper of her family said ...

As the words said, suddenly half a sudden of light: `` Yes, I think a way, I absolutely think of a way, no problem! ``
`` Is there a way to do? ``

All the people had a second look at the same character as a man's eye.
It is also very curious to see that bell is also curious.

Gold Standard:

The vast mountain range was filled with plants, and on top of that, the old residence was likely to be hidden by a formation ... How could they find it under such overwhelmingly unlikely odds?
`` Teacher ...''

After taking a look around, the crowd eventually gave up. Thus, all eyes gathered on Zhang Xuan.
`` This ...''

Zhang Xuan shook his head. `` There's no use asking me, I've never been to the Qu Clan's old residence before ...''
Halfway through his words, Zhang Xuan’s eyes suddenly lit up, and he said, "Wait, I just thought of a good idea!"

``You have an idea?"

Everyone froze for a moment before turning to look at Zhang Xuan with glowing eyes.

Mo Yu was also curious.

Model 3 is significantly better than Model 1. One of model 1’s problems is not knowing how to translate names. Model 3 seems to have learned how to translate names to some extent.
6 Conclusion

It is unexpected that the naïve mixing of the Casia and Fiction data sets resulted in better BLEU scores than the mixing with target tokens. Two possible explanations are as follows. One, the Casia data set may be too mixed already. The Casia data set was pulled from various places around the web, and while there are no domain tags, it does seem that the number of domains covered are fairly broad. It is possible that the heterogeneity of the corpus had already led to a degradation of translation quality. If possible, I would redo the experiment with a more homogenous corpus as the base, then try naïve mixing and mixing with target tokens. A second possibility is simply that more data is required. It is possible that the naïve mixing improved scores in both categories because the model was still being refined, and having a larger amount of data was more important than having domain specific data.
References:


   
   https://doi.org/10.18653/v1/P17-4012