Assignment2: Phrase-Based Neural Machine Translation Model

Yinzhi Yu

March 21, 2017

1 Introduction

Phrase-based machine translation is to build a model which memorizes multi-symbol strings, and translates this string as a single segment. One target is to build a WFST. In order to get the WFST, a previous step should be taken which is phrase extraction. In order to get those German-English phrase pairs, we need to first get the alignments for each translation instance. To get the alignment, ibm model is one of the appropriate choices. Therefore, it becomes clear that this assignment can be divided into 3 main tasks: ibm model training, phrase extraction and finite state machine creation.

2 Implementation

2.1 IBM Model

Specifically, IBM Model 1 is used in this part. The main idea is based on the EM algorithm. Specifically, E step here is to get the C_e and C_{ef} based on current θ. And M step here is to recompute the θ value for each e,f pairs.

In the training part, θ value for each pair is initialized as 1.0/source language vocabulary size. Then in each iteration, the pairs involved will not change. I leave θ values for uninvolved pairs zero and keep updating those involved according to EM algorithm. The maximal iteration time I set is 20. However, it seems that starting from around 13, 14 iteration, the log likelihood value changes only a little.

In the alignment part, I go through all the German words in a sentence, and looked for the English word in the target sentence which has the highest probability to match to this German word.

In IBM model, I also tried to add null alignment symbol. At first, I only add null symbol in English, which means there can be a German word which cannot be mapped to any one of the words in the original English sentence. Then in the training part, things stay the same. In the alignment part, I go through each German word in a sentence, and if one word mapped to Null Symbol according to the θ values I trained, then the alignment of this word won't be output. It has to be mentioned that, this method has a rather lower score than not adding Null Symbol for English. So I tried to add Null symbol for German and for English both. Training part stay the same, and the only difference in the alignment part is that I only go through the first (n-1) words in German and skip the last Null Symbol and the reason is very clear.

2.2 Phrase Extraction

Phrase Extraction is based on the previous IBM model, since we want to extract phrases that are consistent with word alignments. The phrase extraction algorithm mainly focuses on 2 constraints: one is to ensure that phrases should at least contain one corresponding aligned word and the other one is to ensure phrases which only have part of the necessary content are not included.

I set the maximal length of the phrase as 3. My idea is to go through each German word in a sentence and set them as the start, then try the following 2 words and the start word itself as the end. By doing this, I can have a phrase in German. Targeting at this German phrase, I got their aligning English words. The min and max value of the English word's index can form a range. If no word in this range maps to other German word, then it's a concrete phrase. In addition, expansion is also included. But the constraint that the maximal phrase length is 3 goes through the whole prgram.

Running on the given alignment file, I can get the same phrase as given phrase file. So I think the code in this part has no problem.

2.3 Phrase to FST

In this part, I keep a global dictionary saving the states id. For each phrase pair, starting from the first word in source language, I keep track of the whole previous words, and use this as a unique state for this current word. If this combination is not in the dictionary, then add it. Also, when going through the source language, I output "id1 id2 token <eps>", if going through target language, I output "id1 id2 <eps> token". And at last of each pair, I output "idLast idInit <eps> <eps> p".

3 Result

Under the setting that the maximal phrase length = 3 and no null symbol added to either language, I trained my model on the training dataset and got BLEU score = 18.06 on validation set and BLEU score = 18.19 on the test set. I also tried to let the maximal phrase length increase to 4, I didn't get improvements.

After adding null symbol in English, I got a BLEU score = 16.44 on the validation set. It's really bad. I also tried on adding null symbol for both langauges, but I haven't got the result yet. So currently, the best I can got is from the first model I tried, with BLEU score as 18.19 on the test set.

4 Problems and Future Improvements

The first problem I met is when I tried to add null symbol in the IBM model. At first, I think I need to add null symbol in English, but the result is not good. I guess it's because I switched English and German in the training part.(Since doing this can have better performance than the original version in the text book) Then I add null symbol for both 2 languages. But I haven't got the results yet.

The second problem I met is in the fst creation part. Since I ran it on the given phrase file but cannot get the same result as the given fst, so I guessed there could be some problems. I found that only keep the concatenated words both in German and English can be ambiguous. So I add a mark for each word before the word to tell the difference between the words with same spelling from source and target languages, beacause there can be some words have the same spelling in both of the languages.

For further improvement, I think there are more improvement space in the alignment part than the other 2 parts. I don't think I get a kind of accurate alignment based on my IBM model. Both the training part and how to deal with the null symbol need to be further examined. Apart from this, the phrase penalty can also be considered.