11-731 Machine Translation Assignment 2 Report

Ning Dong

March 24, 2017

1 Introduction

In this assignment, I created a symbolic translation model. The report is organized as follows. In section 2, I will introduce my implementation and improvement over baseline. In section 3, experiment settings and results will be presented. Section 4 is about structure of my deliverables.

2 Implementation

2.1 Alignment

I tried both IBM Model 1 and IBM Model 2 to find alignments. Both follow EM process.

2.1.1 IBM Model 1

Treat German as E, English as F, that is, switch two languages from their original role (which I saw on Piazza, would increase the performance). I tried IBM Model 1 with and without null alignment.

Parameter training: Following lecture notes Section 11.3.

Output: For each German word e, find an English word f where f=argmax(P(f|e)).

2.1.2 IBM Model 2

At parameter training phase, instead of treating P(a_j|E) as a constant like IBM Model 1 does (100), use (114) to estimate it.

Output phase is the same as IBM Model 1.

2.2 Phrase Extraction

I implement the baseline which is the same as algorithm 6 on lecture notes. When using reference alignment provided by instructor, my implementation gets identical result on valid set.

Then I explored 2 thresholds to filter some phrases. First is phrase count threshold, phrase frequency in target language(English) less than which would be discarded. Second is phrase score threshold, negative log likelihood of phrase pairs above than which would be discarded, too.

In this way, the model is more likely to be generalized, reduces variance and prevents overfitting. Moreover, since the running time of WFST compiling and decoding phase relies on number of phrases, reducing the number of it also saves time in huge amount.

2.3 WFST Construction

Construction follows chapter12 of lecture notes.

For each node, store its prefix from beginning of current pair. For each "src <eps>" or "<eps> dst", concatenate a new node after the node with its prefix. When using reference phrases provided by instructor, my implementation gets identical result as reference on valid set.

ID	Alignment	Phrase Count Threshold	Phrase Score Threshold	Valid BLEU	Test BLEU
1	IBM Model 1	0	INF	18.12	17.90
2	IBM Model 2	0	INF	18.54	18.02
3	IBM Model 1	2	3	16.55	16.52
4	IBM Model 1	2	INF	18.13	18.12
5	IBM Model 1	3	INF	17.89	18.08
6	IBM Model 1(null)	2	INF	18.13	18.20

Table 1: Results on Different Training Parameters

3 Experiment Settings and Results

For IBM Model 1, I run my program on MacBook Pro (Retina, 13-inch, Mid 2014) with 8GB memory. For IBM Model 2, because intermediate $c_{f,e,|F|,|E|}$ dictionary is huge(more than 2 million entries), memory becomes a limit, I run it on c4.4xlarge(32 GB memory) on AWS.

The strategies I use in alignment and phrase extract phase and their corresponding BLEU score on test set is reported in Table 1. Phrase Count Threshold=0 and Phrase Score Threshold=INF means they are not set at all.

From the result, we can see that IBM Model 2 only increases a little over Model 1, comparing Experiment 1 and 2.

Filtering phrases using Phrase Score Threshold unexpectedly harms the result(Experiment 3). The reason might be that the WFST could handle small probabilities(large weights on edges) while removing them reduces search space, preventing better results(although with low probability) to be found.

Filtering phrases using Phrase Count Threshold and adding null alignment improve the result a little bit(Experiment 4, 5 and 6).

4 Deliverables

My deliverables for assignment 2 include this report, code and output files.

4.1 Code

train-model1.py contains code that performs baseline IBM Model 1 training. train-model2.py contains code that performs IBM Model 2 training. phrase-extract-threshold.py contains code that performs phrase extraction with threshold to filter results. Rest of code follows name convention in run-assignment.sh given by instructor.

4.2 Output Files

Each folder (with prefix as setting ID in Table 1) contains valid and test results. Setting 1,3 and 6 contains blind result for grade.