Assignment 2: Phrase-based Statistical Machine Translation

Chunting Zhou
Andrew ID: chuntinz

March 23, 2017

Abstract

In this report, we detail the implementation of a phrase-base machine translation system. We give some hacks and observations that is useful for performance improvement. We achieve a BLEU score of 18.53 on the test dataset and 18.78 on the validation dataset of IWSLT2016.

1 Introduction

Statistical machine translation learns alignments between parallel languages from data statistics, usually an EM algorithm is performed to maximize the log likelihood of observed data pairs. Afterwards, pairwise phrases are extracted based on the alignment table obtained from the last step. Finally, weighted finite state transducers are created to compose the target translation sentence.

2 System Description

In this section, we describe the implementation details in this assignment. In particular, we need to first implement IBM model 1 and obtain the alignment for parallel sentences, then phrase extraction is performed based on the alignment table, lastly we need to create the WFST for the phrases.

2.1 IBM Model 1

The main idea of symbolic machine translation is the noisy-channel model where the translation model of $P(F|E)$ and the language model of $P(E)$ are separated. With the Bayes' rule, we aim to restore the original source given the observed output. IBM model 1 imposes oversimplified independent assumptions on the latent variables and assumes the latent alignment is generated from an uniform distribution.

For IBM Model 1, we test three types of translation models, which are $p(f|e), p(e|f)$ and the intersection model respectively. Among these models, we find that $\mathbf{p}(\mathbf{f}|\mathbf{e})$ performs best over the other choices in case that we are translating German into English in

Settings	BLEU (Test)	BLEU (Dev)
Baseline	30.29	30.43
Best Model (Max_PL=4, Min_PF=2, EM_Iter=30)	18.53	18.78
Others (Max_PL=3, Min_PF=2, EM_Iter=30)	18.45	18.72
Others (Max_PL=5, Min_PF=2, EM_Iter=30)	18.43	18.70
Intersection Model (Max_PL=4, Min_PF=2, EM_Iter=30)	18.03	18.23

Table 1: BLEU score on test and dev data under different hyperparameter settings. Max_PL, Min_PF denotes the maximum length of extracted phrases and the minimum occurrence frequency of phrase pairs respectively.

the test phase. We conjecture that this is due to the small size of the training data which results in a lot of NULL alignments when using the intersection model. However, the intersection model should lead to better alignment accuracy.

2.2 Phrase Extraction

We implement the Algorithm 6 in chapter 13 for phrase extraction. We find that **limiting the occurrence frequency of phrase pairs** can not only help accelerate the decoding speed but also improves the BLEU score on the test data. Because this significantly reduces the size of resulting phrase table as well as the FST table. Meanwhile, we also **vary the maximum possible length of extracted phrases** from 3 to 5 and find that phrases with maximum length of 4 performs best.

2.3 Creating FST tables

In the last step, we transform the phrase pair table along alone the negative log probability of $p(fp|ep)$ into the weighted finite state transducers. In the meantime, we have the n-gram language model for the target language and create a FST for this too, The two FST are composed to form the final noisy channel model to translate German into English.

3 Experiments

3.1 Datasets and Setups

We use the provided IWSLT 2016 datasets for training, validation and testing. Specifically, we use the filtered and lowercased version of this data set. We set the number of iterations of the EM algorithm for IBM model 1 to be 30.

3.2 Results

We present the results in Table 1. To conclude, we find the pruning of phrases is very effective regarding the decoding speed and the translation performance. We assume that these rare phrases are unhelpful in decoding but increase the search space of FST. We achieve a BLEU score of 18.53 on the test data.