Homework2 Write-up

Group members:

Jingyi Wu, Yirui Wang, Huanyu Yang

Algorithm Description:

In this homework, we have implemented IBM Model 1 and IBM model 2. Both use EM algorithm.

For IBM Model 1, first, we train the data with stemming. We use a API named Snowball Steamer. Then we need to realize the EM algorithm. In the E step, we uniformly set the probability of translation at 1 / length of source language. Then count how many time a source language word is translated into a target language word, weighed by the translation probability. In the M step, we estimate the new probability of translation based on the counts of source language. Finally, we align the word by the largest probability of translation and write it to the output file.

In this model, we realize both French to English and English to French. After comparing the result, we choose French to English as the basement of IBM Model 2.

For IBM Model 2, we use the probability of translation from IBM Model 1 but not uniform value. We also stem the training data and realize the EM algorithm. In the E step, we set alignment probability uniformly at 1 / (length of French + 1). Then count how many time a French word is translated into an English word, weighed by translation probability multiple alignment probability. In the M step, we estimate the new probability of translation and alignment based on the counts of French. Finally, we align the word by the largest translation probability multiple alignment probability and write it to the output file.

Experiment Analyze:

We have implement both French to English and English to French. For the experiment, we test how many iterations can make the result best and set the iterations is 10.

(The result is based on 10W sentences)

	Precision	Recall	AER
IBM Model 1 e2f	0.611650	0.804734	0.326723

IBM Model 1 f2e	0.652042	0.781065	0.304304
IBM Model 2	0.747352	0.860947	0.214214

From the result we can see, f2e is a little better than e2f. And we also test 1000 and 1w sentences. We find that with the size of training data increasing, ARE will reduce. And IBM Model 2 is much better than Model 1. Because we don't set the probability of translation uniformly but use the result of Model 1.