CIS 526 : Homework 1
Alignment

Anwesha Das

February 4, 2015

1 MODELS IMPLEMENTED

The methods I experimented with are:
1. IBM Model 1
2. IBM Model 2
3. Alignment by agreement

I first implemented IBM Model 1. As part of preprocessing the text, I lowercased the data given so as to reduce sparsity. Then I added a NULL token to every source (foreign) sentence.

I ran IBM Model 1 with varying the following parameters:
a. Number of sentences being trained on
b. Threshold value

Next I tried combining Model 1 with Dice's coefficient (ACL-2006 - Blunsom, Cohn). By combining these two models, we make sure that a given alignment has the highest Dice, i.e., translation score of all the possible alignments for a given target word. Model 1 enforces competition for alignment between source and target words. By training it on 1000 lines and giving a threshold of 0.3, the combination of these two models gave an AER of 34.56, which beat the baseline.

I tried tweaking various parameters and running Model 1, with and without Dice. The results are summarized in the Table below. I observed that training our model on more number of lines and lesser threshold value gives better AER.

Next I implemented the IBM Model 2, which was seeded by the translation probabilities produced by Model 1. Running Model 2 until the EM algorithm converges, and training it on the entire dataset with a threshold value of 0.3 gave me my best AER of 27.59. As before, I observed that the AER got progressively lower by increasing the number of lines given for training.

Other than these, I experimented with word alignment by agreement (Liang

et al. 2006). I trained Model 1 in both directions, i.e., English to French and French to English. After that I took an intersection of the two alignments. The motivation was that taking an intersection would help eliminate the individual errors of the two. However, this model did not perform as well as expected by giving an AER of 50.48.

I also attempted to make another small extension to the IBM Model 1 by using a orthographic feature (ACL-2006 - Blunsom, Cohn), i.e., the absolute difference in word length. The idea behind this was that rare longer words should not be aligned with shorter and more frequent words like determiners etc. This did not give any significant improvement in the AER.

The following section summarizes the AER of all the Models implemented. The best AER (27.59) was running IBM Model 2 on the entire dataset.

2 RESULTS SUMMARY

Note: All the entries in the table below take P($English \mid French$) unless specified otherwise.

MODEL	NUMBER OF LINES	THRESHOLD	AER
MODEL 1	Full Dataset	0.4	28.76
MODEL 1 + DICE	1000	0.3	34.56
MODEL 1	1500	0.4	33.14
MODEL 1 + DICE	1500	0.4	36.23
MODEL 1	50000	0.3	28.03
MODEL 1 + Abs Difference in Word Length <5	1250	0.4	40.5
MODEL 2	Full Dataset	0.3	27.59
MODEL 2	15000	0.4	29.93
MODEL 2	15000	0.3	28.74
MODEL 2	50000	0.4	29.41
MODEL 2	50000	0.3	27.92
MODEL 2 + DICE	50000	0.3	32.21
ALIGNMENT BY AGREEMENT (MODEL 1)	50000	0.4	50.48