I learned to build IBM models in this assignment. IBM is an example of a noisy channel model. In noisy channel model, suppose that f is the translation of e, then the possibility of translation f = p(f)*p(e|f), where p(f) is the possibility in the foreign language, and p(e|f) is the possibility of e given f. To model the conditional probability of pairs of english and foreign language, either p(e|f) or p(f|e), we use IBM models. Since I did my assignment with the reference of collins note, and IBM2 model is introduced at first in the note, I talk about IBM2 before IBM1 in my report.

2. IBM Model 2

2.1 Description of IBM Model 2

The definition of conditional probability in IBM2 in the writeup is: P(f1... fm; a1...am |e1 ... el;m) =product of q(ai|i; l;m)t(fi|eai) for i=1...m. Here, t(f|e) is the conditional probability of seeing foreign word for given english word e. q(j|i;l;m) is the possibility of ai = j, given the sentence length l and m. Because the most likely alignments maximizes P(a1...am |f1...fm, e1 ... el;m), for i = 1...m, we need to find ai that has the largest score of q(ai|i; l;m)t(fi|eai). By maximum estimate, t(f|e) = # of times of e aligned to f/# of times of e aligned to any foreign word, q(ai|i; l;m) = # of times fi aligned to ej for english sentence with length l, foreign sentence with length m/# of the appearance of the english sentence with length l together with foreign sentence with length m. To estimate these two parameters, we use EM, which will be introduced in 2.2.

IBM2 has q(ai|i; l;m), so it tries to model the probability distribution of fi aligned to ej, which is better than IBM1, but IBM still has the similar issue as IBM1: fertility problem, lack of context and the variability of EM convergence. The fertility problem is that an fi might be aligned to multiple or no ej by this model, but in most cases, the english and foreign words should be one-to-one aligned. Moreover, both of these two models ignore the patterns or information hidden in surrounding alignments. The EM algorithm has the convergence problem, since it could converge to different local optimum for parameter estimates with different initialization values of t and q.

2.2 Method Overview

Since delta(k,i,j) = P(ai=j |f1...fm, e1 ... el;m) = q(j|i; l;m)t(fi|ej)/sum of q(r|i; l;m)t(fi|er) for r = 0...l. Thus, the c(e, f), c(e), c(j|i, l, m) and c(i, l, m) counts are updated by the conditional probability of ai=j under this model.

The pseudocode will be:

Initialize all the t, q values

For each iteration:

    Reset all c value to be 0

    For each sentence pair (fk,ek), each foreign word in fk, and (each english word in ek union Null)

        Increment c(e, f), c(e), c(j|i, l, m) and c(i, l, m) by delta

    Get t(f|e) = c(e, f)/c(e) and q(j|i; l;m) = c(j|i, l, m) /c(i, l, m)

Return a dictionary containing all possible t(f|e) and c(j|i,l,m)

2.3 Results over 5 iterations

Type	Total	Precision	Recall	F1-Score

total	5920	0.432	0.446	0.439

==========

total	5920	0.435	0.449	0.442

==========

total	5920	0.439	0.453	0.446

==========

total	5920	0.439	0.453	0.446

==========

total	5920	0.442	0.456	0.449

2.4 Discussions

We observe that as the number of iterations increases, all of the evaluation parameters including precision, recall and f1 score increase. Unlike the IBM1 model, the f1 score of IBM2 increases with a stable speed. It is interesting that the f1 score of the fourth iteration is the same as that of the third one. This probably because in IBM1, we just initialize t(f|e) to be the uniform distribution over all foreign words that could be aligned to e in the corpus, but in IBM2, we use the t(f|e) generated by IBM1 as the start point. The start point of IBM2 therefore is closer to the local optimum. Also, the recall(ability of a classification model to identify all relevant instances) is always higher than precision(ability of a classification model to return only relevant instances) in both IBM1 and IBM2.

I calculated and sort the f1 scores for dev sentences, and got the following result (in the printed alignment, the correct alignment comes first):

The 192th sentence has the highest f-score, which is 1. There are all correctly aligned.

['one', 'issue', 'that', 'separates', 'us', 'is', 'the', 'civil', 'war', 'in', 'chechnya', '.']

['una', 'cuesti\xc3\xb3n', 'que', 'nos', 'separa', 'es', 'la', 'guerra', 'civil', 'en', 'chechenia', '.']

(['192 1 1', '192 2 2', '192 3 3', '192 4 5', '192 5 4', '192 6 6', '192 7 7', '192 8 9', '192 9 8', '192 10 10', '192 11 11', '192 12 12'], ['192 1 1', '192 2 2', '192 3 3', '192 5 4', '192 4 5', '192 6 6', '192 7 7', '192 9 8', '192 8 9', '192 10 10', '192 11 11', '192 12 12'])

The 35th sentence is the second highest, with f1 score 0.9411764705882353

['thank', 'you', 'for', 'your', 'statement', ',', 'commissioner', '.']

['gracias', 'por', 'sus', 'palabras', ',', 'se\xc3\xb1or', 'comisario', '.']

(['35 1 1', '35 2 1', '35 3 2', '35 4 3', '35 5 4', '35 6 5', '35 7 6', '35 7 7', '35 8 8'], ['35 1 1', '35 3 2', '35 4 3', '35 5 4', '35 6 5', '35 7 6', '35 7 7', '35 8 8'])

The 89th sentence is the worst with score 0.16438356164383558

['every', 'wet', 'place', 'is', 'affected', ',', 'including', ',', 'as', 'far', 'as', 'loire-atlantique', 'is', 'concerned', ',', 'the', 'bri\xc3\xa8re', ',', 'brivet', ',', 'goulaine', ',', 'and', 'redon', 'marshes', ',', 'the', 'grand', 'lieu', 'lake', ',', 'the', 'mazerolles', 'plains', ',', 'and', 'so', 'forth', '.']

['todas', 'las', 'zonas', 'h\xc3\xbamedas', 'se', 'ven', 'afectadas', 'y', ',', 'en', 'especial', ',', 'en', 'el', 'departamento', 'de', 'loira', 'atl\xc3\xa1ntico', ',', 'los', 'pantanos', 'de', 'bri\xc3\xa8re', ',', 'brivet', ',', 'goulaine', 'y', 'redon', ',', 'el', 'lago', 'de', 'grandlieu', ',', 'las', 'llanuras', 'de', 'mazerolles', ',', 'etc.']

(['89 1 1', '89 2 4', '89 3 3', '89 4 6', '89 4 5', '89 5 7', '89 8 12', '89 12 18', '89 12 17', '89 15 19', '89 16 20', '89 17 23', '89 18 24', '89 19 25', '89 20 26', '89 21 27', '89 23 28', '89 24 29', '89 25 21', '89 26 30', '89 27 31', '89 28 34', '89 29 34', '89 30 32', '89 31 35', '89 32 36', '89 33 39', '89 34 37', '89 35 40', '89 36 41', '89 37 41', '89 38 41'], ['89 2 1', '89 3 2', '89 2 3', '89 29 4', '89 3 5', '89 21 6', '89 5 7', '89 23 8', '89 22 9', '89 11 10', '89 24 11', '89 20 12', '89 17 13', '89 27 14', '89 24 15', '89 16 16', '89 12 17', '89 12 18', '89 15 19', '89 28 20', '89 30 21', '89 27 22', '89 19 23', '89 22 24', '89 21 25', '89

22 26', '89 21 27', '89 36 28', '89 21 29', '89 31 30', '89 17 31', '89 30 32', '89 38 33', '89 34 34', '89 7 35', '89 5 36', '89 33 37', '89 38 38', '89 21 39', '89 18 40', '89 17 41'])
The 46th is the second worst with 0.17391304347826086
['10', '.', 'considerably', 'limits', 'the', 'present', 'derogations', 'for', 'cement', 'kilns', '.']
['10', '-', 'limita', 'considerablemente', 'las', 'exenciones', 'vigentes', 'para', 'las', 'instalaciones', 'en', 'cementeras', ';']
(['46 1 1', '46 3 4', '46 4 3', '46 5 5', '46 6 7', '46 7 6', '46 8 8', '46 9 12', '46 10 10', '46 11 13'], ['46 1 1', '46 10 2', '46 10 3', '46 3 4', '46 9 5', '46 9 6', '46 10 7', '46 10 8', '46 9 9', '46 9 10', '46 9 11', '46 10 12', '46 9 13'])

We observe that the best aligned sentences are very short and the positions of aligned pair of words in english and foreign sentences are adjacent, e.g. 5th word in english and 6th word in foreign sentence, or 5th - 5th in the paired sentences. In the worst aligned sentences, the english word position in a sentence is usually far, and the long length of a sentence increases the possibility that two aligned words are far apart, which increases the difficulty.

2.5 Critical Thinking

To solve the fertility problem, we need to model the probability distribution of how many english words could an input foreign word aligned to, and combine this probability in to P(f1... fm; a1...am |e1 ... el;m).

To solve the lack of context issue, we need to model the conditional probability of word classes, e.g. if the input foreign word is a verb, how likely is its english aligned word is behind an adverb? This will provide us more surrounding information, and thus the model can make a more accurate alignment.

1. IBM Model 1

1.1 Description of IBM Model 1

The IBM model is used for estimating the conditional probability such as p(e|f) or p(f|e). Also, since we don't know what initialization values are appropriate for t in IBM2 and the EM algorithm is sensitive to initial values, we can use IBM1 to evaluate t at first and use the result of t as the initialization value in IBM2. In IBM1, we suppose the fi is uniformly aligned to ej, j=0...l, i = 1...m where e0 = NULL, english sentence has length l and foreign sentence has length m. Thus q(j|i; l;m) = 1/(1+l), which causes P(f1... fm; a1...am |e1 ... el;m) becomes 1/(l+1)^m*product of t(fi|eai) for i = 1...m, and delta(k,i,j) becomes t(fi|ej)/sum of t(fi|er) for r = 0...l.

It is almost impossible that fi is uniformly aligned to ej, so the assumption of IBM1 is not very valid. IBM1 also shares the fertility and lack of context problem in IBM2 that I have mentioned above.

1.2 Description of EM Algorithm

Iteratively increment the value of c(e,f) and c(e) by delta through all of the (f,e) pairs of the parallel corpus, and get the updated t(f|e) = c(e,f)/c(e) after each iteration through the parallel corpus.

Strength: EM can estimate the maximum likelihood estimate value even when data is partially observed. Here since the value of ai is not revealed in the training data, EM would be a good tool for us to get parameter estimate.

Weakness: The results of EM might have a large variability depending on the start points, since it finds the local optimum instead of global optimum for the parameter, and we need to make sure that this local optimum is equal or very close to the global optimum.

1.3 Method Overview

The pseudocode will be:

Initialize all the t, q values

For each iteration:

    Reset all c value to be 0

    For each sentence pair (fk,ek), each foreign word in fk, and (each english word in ek union Null)

        Increment c(e, f) and c(e) by delta

    Get t(f|e) = c(e, f)/c(e)

Return a dictionary containing all possible t(f|e)

1.4 Results

Type	Total	Precision	Recall	F1-Score
total	5920	0.211	0.217	0.214
total	5920	0.375	0.387	0.380
total	5920	0.402	0.415	0.408
total	5920	0.410	0.423	0.416
total	5920	0.413	0.427	0.420

1.5 Discussions

We observe that as the number of iterations increases, all of the evaluation parameters including precision, recall and f1 score increase. Moreover, the speed of improvement decreases. For example, from the first to the second iteration, the difference of precision, recall and f1 score are 0.165, 0.170, and 0.166; but from the fourth to the last iteration, the difference of precision, recall and f1 score are 0.003,0.004, and 0.004. This is because EM gradually converge as iterations goes up.

3. Growing Alignments

3.1 Method Overview

We use IBM2 to estimate p(e|f) and p(f|e), make alignment guess on dev corpus, and gain the intersection and union of predicted alignments with p(f|e) and p(e|f). We use intersection as a starting point, and then add alignment one by one. From the union of the alignments, if there is an alignment point adjacent to point(e.g. The 5th word is close to the 6th word) that has existed in the intersection and align a word that has no alignment in the intersection, we add it into our alignment. Repeat this step until there is no alignment could be added.

3.2 Results and Discussions

I tried to calculate p(e|f) but my f1 score result is not good. It might be the issue of my code, but I went through it and could not find the mistake. Based on my current p(e|f), I doubt if it will improve the f-score much by growing alignments, since the intersection of alignments with p(f|e) and p(e|f) should be just a small percentage, and 0.1 is really far from 0.45.

IBM1:

Type	Total	Precision	Recall	F1-Score
total	5920	0.050	0.049	0.050
Type	Total	Precision	Recall	F1-Score
total	5920	0.095	0.092	0.093
Type	Total	Precision	Recall	F1-Score
total	5920	0.100	0.098	0.099
Type	Total	Precision	Recall	F1-Score
total	5920	0.101	0.098	0.099
Type	Total	Precision	Recall	F1-Score
total	5920	0.101	0.099	0.100

IBM2:

Type	Total	Precision	Recall	F1-Score
total	5920	0.112	0.109	0.111
Type	Total	Precision	Recall	F1-Score
total	5920	0.113	0.111	0.112
Type	Total	Precision	Recall	F1-Score
total	5920	0.112	0.109	0.110
Type	Total	Precision	Recall	F1-Score
total	5920	0.113	0.110	0.112
Type	Total	Precision	Recall	F1-Score
total	5920	0.114	0.111	0.112

3.3 Critical Thinking

Since with this heuristic, the alignment could be one to many, or many to one. To improve the accuracy, we should model the probability distribution of one to many and many to one, and take this into consideration when we add new alignments points from union.