Programming Assignment 4 : Statistical Machine Translation

CSE 256: Statistical NLP: Spring 2019

University of California, San Diego

Wei-Cheng Huang

1. IBM Model 1

1-1 Description of IBM Model 1

1-1-a IBM Model 1 Usage

IBM Model 1 is firstly invented for Statistical Machine Translation, while the computation is too costly since it requires calculation possibility of every foreign-english word combinations. The EM algorithm would need computation of hidden variables - probability of alignments. IBM Model 1 is then mainly used to obtain alignments.

1-1-b IBM Model 1 limitation

IBM Model 1 can only do many to one mapping, which means a function may return the same value for different input. And cannot return multiple values for one input. In our case, we can only map one English word to one Spanish word. However, real word alignments have many-to-many mappings.

1-2 Description of EM Algorithms

1-2a Description

EM algorithm could estimate missing data in the model, then estimate model parameters from completed data iteratively, until it converges.

Expectation-Steps:

- Calculate posterior over English positions $P(a_i \mid e, f)$
- Increment count of word f_i translating each word e_{a_i}

Maximization-Steps: Estimate model from data

- Renormalize counts to give probabilities

1-2b Strength and Weakness

Strength:

- The EM algorithm can be used in cases where some data values are missing.
- Easy to implement in coding.
-

Weakness:

- Highly depend on initialization, which would affect whether the model converges and find global optimal solution
- Slow convergence

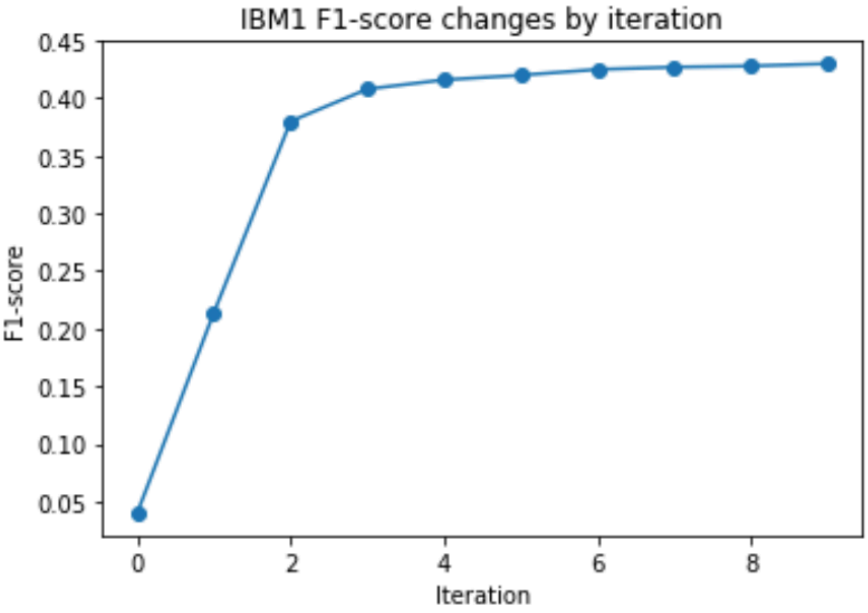
1-3 Method Overview

1. Set $n(e)$ as the number of foreign words that are possible alignments with the english word e in the training corpus
2. Initialize $t(f \mid e)$ for EM algorithm, which is the conditional probability of generating a foreign word f from an english word e as uniform distribution $1/n(e)$
3. Iterate through the english and foreign corpus number of iterations times, for each iteration, initialize $c(e, f)$for all english word and foreign word alignment as 0, initialize $c(e)$ for each english word as 0
4. For each iteration, iterate through the english sentences and foreign sentences of the corpuses and calculate $\delta(k,i,j) = \frac{t(f_i^{(k)} \mid e_j^{(k)})}{\sum_{j=0}^{l_k} t(f_i^{(k)} \mid e_j^{(k)})}$, which is the conditional probability of current iteration divided by the sum of the conditional probability, i, j are the index of the english word and foreign word of the k^{th} sentence respectfully, for each $c(e, f)$ and $c(e)$ update by adding the corresponding $\delta(k,i,j)$
5. After each iteration of running through the corpus, and each sentence, update the conditional probability $t(f \mid e)$ as $\frac{c(e,f)}{c(e)}$

1-4 Results (After 5 iterations)

Type	Total	Precision	Recall	F1-Score
total	5920	0.413	0.427	0.420

1-5 Discussions



After 4th iteration, the model almost converges and could obtain a F1-score at around 0.42. We can see from the trend that the model first predict very worst as our initial parameter were assumed, but parameters are adjusted iteratively through EM algorithm.

2. IBM Model 2

2-1 Description of IBM Model 2

IBM Model 2 not only consider the translation possibility of word-pairs (t parameters), but also consider the possibility of alignments (q parameters). Therefore, IBM Model 2 has higher F1-score and accuracy than IBM Model 1.

IBM Model 2 has same limitation as IBM 1, they can only do many-to-one mapping, which means we can only map one English word to one Spanish word in our case.

2-2 Method Overview

1. Set $n(e)$as the number of foreign words that are possible alignments with the english word e in the training corpus
2. Initialize $t(f \mid e)$, which is the conditional probability of generating a foreign word f from an english word e as uniform distribution $1/n(e)$, initialize $q(j \mid i, l, m)$, which is the alignment parameter as uniform distribution $1/(l+1)$
3. Iterate through the english and foreign corpus number of iterations times, for each iteration, initialize $c(e \mid f)$for all english word and foreign word alignment as 0, initialize c(e) for each english word as 0, initialize $c(j \mid i, l, m)$ which is the number of times we see an English sentence of length l, and a French sentence of length m, where word i in French is aligned to

word j in English as 0, initialize $c(i,l,m)$ which is the number of times we see an English sentence of length l together with a French sentence of length m as 0

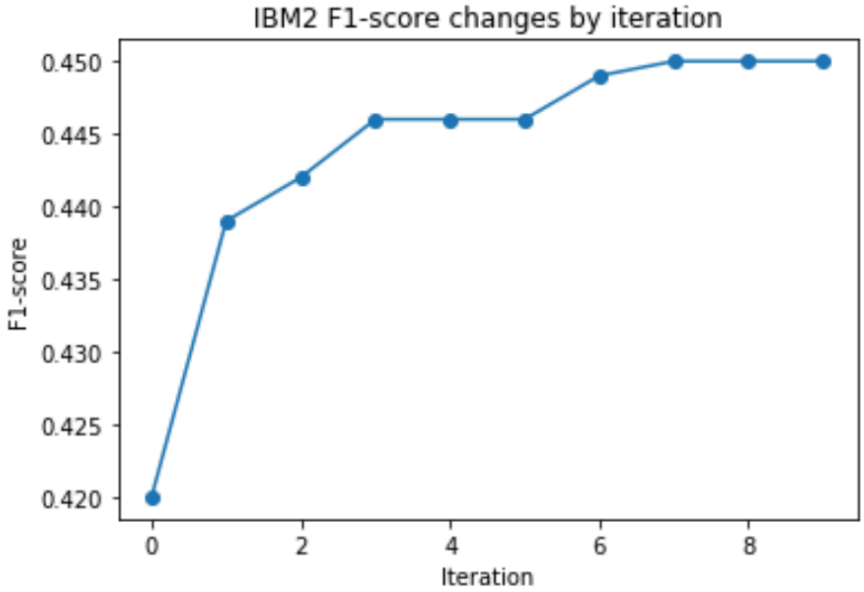
4. In each iteration, iterate through the english sentences and foreign sentences of the corpuses and calculate the $\delta(k,i,j) = \frac{q(j\mid i,l_k,m_k)t(f_i^{(k)}\mid e_j^{(k)})}{\sum_{j=0}^{l_k} q(j\mid i,l_k,m_k)t(f_i^{(k)}\mid e_j^{(k)})}$, which is the conditional probability*alignments parameter of current iteration divided by the sum of the conditional probability*alignment parameter, i, j are the index of the english word and foreign word of the k^{th}sentence respectfully, for each $c(e\mid f)$, $c(e)$, $c(j\mid i,l,m)$, and $c(i,l,m)$ update by adding the corresponding $\delta(k,i,j)$
5. After each iteration of running through the corpus, and each sentence, update the conditional probability $t(f\mid e)$ as $\frac{c(e,f)}{c(e)}$ and $q(j\mid i,l,m)$ as $\frac{c(j\mid i,l,m)}{c(i,l,m)}$

2-3 Results

Type	Total	Precision	Recall	F1-Score
total	5920	0.442	0.456	0.449

2-4 Discussions

2-4a F1-score changes after iterations, different from IBM Model 1, 0^{th} iteration utilizes t parameters from IBM Model 1, so the F1-score of 0^{th} iteration starts from around 0.42. IBM Model 2 obtains higher F1-score at 0.45 and reaches plateau after 5th iteration.



2-4b Highly Correct examples:

Notation: (applied to all the following tables)

● : is the alignment predicted and correct

● : the correct alignments but not predicted

● (false): misaligned predicted alignments

The IBM Model 2 only predicted 1 sentence pair in dev set completely correct, so the second correct example I use is nearly correct sentence pair with short sentence length.

Idx: 191 (error rate: 0 %)

	una	cuestión	que	nos	separa	Es	La	Guerra	civil	En	Chechenia	.
One	●											
Issue		●										
that			●									
Separates					●							
Us				●								
Is						●						
The							●					
Civil									●			
War								●				
In										●		
Chechnya											●	

	una	cuestión	que	nos	separa	Es	La	Guerra	civil	En	Chechenia	.
.												●

Idx : 46 (error rate: 1.28 %)

	La	unión	es	Y	Debe	Seguir	Siendo	Una	Associati on	De	Estad os	.
the	●											
union		●										
Is			●									
,												
and				●								
shoul d					●							
Remai n						●	●		● (false)			
,												
a								●				
Union									●			
Of										●		
States											●	
.												●

2-4c Misaligned Examples

Idx: 55 (error rate: 25.0 %)

	no	Hay	Estadiscas	.
No	●			
Statistical		● (false)	●	
Data			●	

	no	Hay	Estadiscas	.
Exists		●	● (false)	
.				●

Idx:196 (error rate : 14.58 %)

	Le	Desco	El	Mayor	De	Los	éxitos	.
I								
Wish		●				● (false)		
You	●							
Every			●	●	●			
Success	● (false)		● (false)	● (false)			●	
.								●

2-4d Discussion of result

From my observation, the accuracy would be higher under two conditions:

1. Word alignment points are diagonal
2. Lengths of English and Foreign sentences are almost the same

On the other hand, the model performs worst when:

1. Lengths of English and Foreign sentences are relatively different
2. Correct word alignment points are disorder.

2-5 Critical Thinking: Possible methods to improving IBM2

1. Running more iterations : The result above shows that more iteration could possibly increase the F1-score up to the plateau.

2. Implementing the heuristics: The result in the next part suggests it would solve the main problem of IBM Models, which is many-to-one mapping. With Heuristics methods, we can do many-to-many alignments, which is more applicable to the real case.

6. Growing Alignments

3-1 Method Overview

1. We use the model p(e|f) and p(f|e) to obtain union and intersection of the alignments
2. Starting from Intersected alignments, if the adjacent points are union, then include them as final alignments.
3. Finally , add other points if there is no corresponding alignments for any of English or foreign words. (However, I found that without using 3th step could obtain higher F1-score, so the result following are the one without step 3)

3-2 Results and Discussion

3-2a Result : F1-score

Type	Total	Precision	Recall	F1-Score
total	5920	0.706	0.426	0.531

3-2b Highly correct example with one-to-many alignment.

Most of the alignments results are better than IBM models by the Heuristics Methods, since it is generated bilaterally, the model is more robust and could generate many-to-many alignments.

Idx: 34 (error_rate: 1.5%)

	Gracias	Por	Sus	Palabras	,	señor	Comisario	.
Thank	●							
You	●							
For		●						
Your	● (false)		●					
Statement				●				

	Gracias	Por	Sus	Palabras	,	señor	Comisario	.
,					●			
Commissioner						●	●	
.								●

3-2c Example failing due to the alignment heuristics

Some of examples in the heuristics methods would cause even worst performance than in IBM Model 2, when wrong union points are coincidentally adjacent of intersection points.

Idx: 10 (error rate: 7.5 %)

	Siempre	He	Tenido	Conianza	.
I		● (false)			
Have		●			
Always	●				
Had			●		
Confidence				●	
In			● (false)		
This				● (false)	
.					●

3-3 Critical Thinking

Develop a way to choose other points that are not aligned by any of English words or Spanish words.