

Programming Assignment 5: Machine Translation

CSE 256: Statistical NLP: Spring 2022

IBM Model 1

Description of IBM Model 1 - What are IBM Models used for? What are the limitations?

IBM models are used for translation and alignment, and they have been the majority part of the statistical machine translation systems for a long time. It is also a noisy channel model. For the limitation of the model:

- Each word in the target sentence can at most generated by one word from the source sentence, which does not include the fact that multiple source words would be mapped to a single target word
- Model 1 does not consider the word position factor and each word in a sentence are independent, assuming the distortion will be $1/(l+1)$ all the time

Description of EM Algorithm - What are the strengths and weaknesses?

Expectation-Maximization Algorithm is to find the maximum-likelihood estimates of models' parameters, where its strength is able to estimate even though only part of the data is observed, while at the same time the possibilities of the result will increase along with each iteration until converged. Its weakness includes slow convergence and only is able to converge to local optimal only.

Method Overview - Provide a high-level description of your implementation

My implementation of the IBM 1 include three main steps. First, when getting the corpus, each English sentence will added a NULL word at the beginning, and each word alignment will be initialized as $1/n(e)$ where $n(e)$ is the number of different words that occur in any translation of a sentence containing English word e . After the initialization, the algorithm will running 5 iteration of EM algorithm against the corpus. For each pair of sentences, we loop thru each foreign word in the sentence, and gather the possible delta value of each pair of word t value divided by the sum of pair of the foreign word and the whole English sentence's words t value. With this delta, we calculate the count accordingly and update t value in each iteration. After each iteration, we use dev test set against our model by finding the optimal alignment and calculate the F1 Score.

Results and Discussion - Report your F1 score and Show how F1 score changes after each iteration

Iteration 1 Result:

Type	Total	Precision	Recall	F1-Score
total	5920	0.211	0.217	0.214

Iteration 2 Result:

Type	Total	Precision	Recall	F1-Score
total	5920	0.374	0.386	0.380

Iteration 3 Result:

Type	Total	Precision	Recall	F1-Score
total	5920	0.402	0.415	0.409

Iteration 4 Result:

Type	Total	Precision	Recall	F1-Score
total	5920	0.410	0.423	0.416

Iteration 5 Result:

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

The final result of F1-Score is 0.420, which match the expected value. The F1-Score, based on the observation from the above data is increasing over time. We initialize model parameter with length of sentences which is expected that initially not a well behaving model, but with each iteration of EM algorithm, it is getting better with training on parallel corpus.

IBM Model 2

**Description of IBM Model 2 - Why is it better than IBM Model 1?
What are the limitations?**

IBM Model 2 introduce the distortion parameter, which considers the position of words in the target sentence and the source sentences. An example of such relationship would be that the source word and target word alignment should be within a relative index instead of far apart. It also has limitations:

- Its convergence would largely depend on the initial value. Different values will end up into different local optimal. In our example, we use pre-trained t value from IBM Model 1 when we doing Model 2 to avoid unstable result
- The same as Model 1, Model 2 does not map relation of multiple source word to a single target word

Method Overview - Provide a brief, high-level description of your implementation

In my implementation, I make use of the part of IBM 1 model implementation to handle the text, and for the initialization part, t value is initialized to the value of IBM 1 model 5th iteration, and the distortion value is initialized as $1/(l+1)$, where l is the sentence lengths of English sentences without NULL word. EM algorithm also includes new parameters calculation, addition to existing t value updates, and each time I build the alignment pair (j, i, l, m) and update the distortion value accordingly.

Results and Discussion - Report your F1 score

Iteration 1 Result:

Type	Total	Precision	Recall	F1-Score
total	5920	0.432	0.446	0.438

Iteration 2 Result:

Type	Total	Precision	Recall	F1-Score
total	5920	0.436	0.450	0.442

Iteration 3 Result:

Type	Total	Precision	Recall	F1-Score
total	5920	0.439	0.453	0.446

Iteration 4 Result:

Type	Total	Precision	Recall	F1-Score
total	5920	0.439	0.453	0.446

Iteration 5 Result:

Type	Total	Precision	Recall	F1-Score
------	-------	-----------	--------	----------

total	5920	0.442	0.456	0.449
-------	------	-------	-------	-------

The final result of the F1-Score is **0.449** which match the expected value. The same as previous observation, the model's performance is improving after each iteration, although it seems to be reaching a threshold.

Alignment Examples

Missing alignment marked italicized, and **W** means wrong alignment

Correct Alignment

Missing one alignment

	thank	you	for	your	statement	,	commissioner	.
gracias	X	X						
por			X					
sus				X				
palabras					X			
,						X		
señor							X	
comisario							X	
.								X

Misalignment

	if	not	,	when	will	it	be	in	operation	?
si	X									
no		X								
es						W				
así								W		
,			X							
¿										X
cuándo				X				W		
estará					X	X	X	W		
en								W	X	
marcha								X	X	
?										X

The misalignment example exposes the weakness of IBM 2 that two or more English word will not be able to properly map to a single foreign word, while in the case no such situation happened, the model behaves much better.