Machine Translation Homework 1

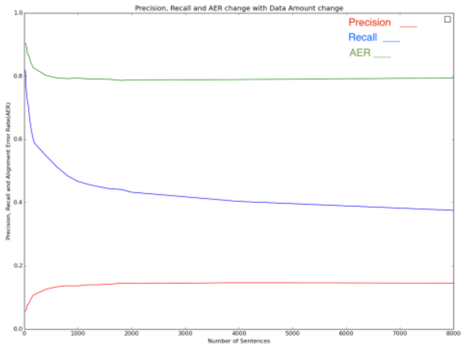
s1401192 Yunfeng Zhu s1466878 Qian Zhong

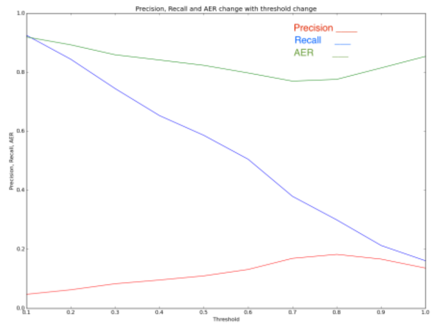
Question 1

As the figure indicates, the general trend is that the precision goes up whereas the recall and AER(alignment error rate) decreases as the sentence numbers increases. We plotted the precision, recall and AER for 20 sentences, 40 sentences so on and so forth utill it reached 200. From the figure, it could be observed that when the data amount is small, Precision, Recall and AER are more sensitive to the data amount change. The steep slope indicates a more dramatic change within this range. The impact of data amount change on Precision, Recall and AER, however, become less and less obvious when more data are being tested. When the sentence amount reached 3000, the impact of data change is only minimum. According the figure, the precision peaks around 0.145 with data amount of 2000 sentences and becomes steady thereafter. The recall is decreasing all the time, however, its decrease becomes relatively slow when the sentence amount goes to around 4000. The AER also decreases more slowly when the sentence amount comes to 2000 and maintained almost the same thereafter at around 0.78(the AER for 2000 sentences).

As demonstrated in figure 2, if we experiment with the threshold, we could observe generally the recall is negatively correlated with the threshold; in other words, recall is decreasing all the time with the increasing threshold. However, the precision is increasing until the threshold becomes 0.8, reaching a plateau at around 0.18. The AER, on the other hand, firstly decreased to around 0.76 at the threshold of 0.7 and then increased.

After studying the alignment, we could observe that the human alignment is somewhere around the diagonal. However, the machine alignment is pretty much a mess. It is relatively randomly distributed in the matrices, and if we improve the threshold, the machine would make fewer alignments. Actually, the machine aligns every words pairs as long as its dice coefficient surpasses the threshold. Judging by the dice coefficient only, the chance is high that any English and foreign word that frequently appeared together in two parallel sentences would be aligned to each other.





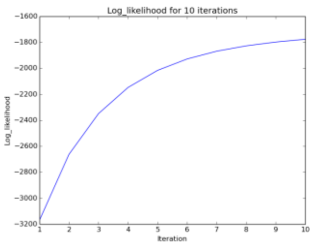
.

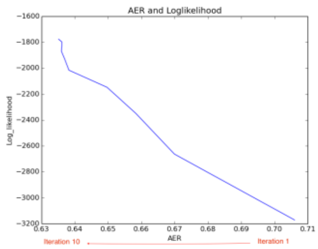
Figure 1.Precision, Recall and AER change with Data amount Change

Figure 2. Precision, Recall and AER change with the threshold change

Question 2

Generally, after we have implemented the IBM Model 1, we found the data converge after 8 iterations, as demonstrated in Figure 3. Although there is still minimum change in the likelihood, we consider this as trivial since the small amount of log likelihood change basically means very slight change in the data. We ignored this small change and we found convergence after 8 iterations. We hence plotted 10 iterations. Also, it has been observe the likelihood and AER are negatively related. After each iteration, the likelihood increases and the AER decreases, as we can see from figure 4. Figure 5 describes the Precision, Recall and AER change after each iteration, for reference only.





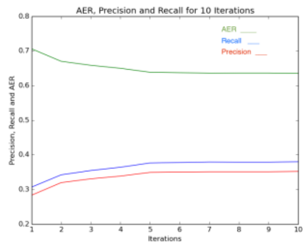


Figure 3. Log likelihood for each iteration

Figure 4. AER and Log Likelihood

Figure 5. Precision, Recall and AER for each Iteration

Question 3:

In this question, we recorded the probability of two frequent and infrequent words to better observe the translation distribution pattern. The two frequent words we choose are 'you' and 'the' and the two infrequent words we choose are 'remarkable' and 'regulations'. In the following table (figure 6), we listed the 10 most frequent alignments for these words. From the data we acquired, for frequent words, it is more likely to find one or two translations with higher probability (highlighted in Figure 6) and this probability would well surpass the probabilities of other translations for the same word. In other words, distinction is obvious between the high probability and low probability for the translations of the high frequency word. We think the probability mass is unevenly distributed among the translations for those frequent words. This, as far as we concerned, is caused by larger sample size of the frequent words and their translations. Meanwhile, on the other hand, for the infrequent words, the probability is evenly distributed among every possible translation, each with a relatively small probability. And most of these probabilities are equal to each other. This is most likely due to the small sample size of these infrequent words; each of the possible translation for these infrequent words only appeared once or twice in our dataset and thus they might have the same small translation probability in the final distribution.

Translation probability of frequent words				Translation Probability of infrequent words			
	you		the		remarkable		regulations
P(Sie|you)	0.7608847801	P(der|the)	0.2653763576	P(Bedingungen|remarkable)	0.0803773533	P(einhalten|regulations)	0.0644203544
P(,|you)	0.1208584411	P(.|the)	0.2143782005	P(verschiedensten|remarkable)	0.0803773533	P(unterbreiten|regulations)	0.0644203544
P(.|you)	0.0655336307	P(,|the)	0.1829326262	P(Multifunktionalität|remarkable)	0.0803773533	P(Rechtsvorschriften|regulations)	0.0644203544
P(ich|you)	0.0194300673	P(die|the)	0.1822837108	P(Unsere|remarkable)	0.0803773533	P(Wert|regulations)	0.0644203544
P(daß|you)	0.0045828436	P(in|the)	0.0331528384	P(ausgerichtet|remarkable)	0.0803773533	P(legen|regulations)	0.0644203544
P(hat|you)	0.0030722290	P(und|the)	0.0203739612	P(optimale|remarkable)	0.0803773533	P(großen|regulations)	0.0644169276
P(zu|you)	0.0029999021	P(den|the)	0.0165947680	P(Anbauprodukte|remarkable)	0.0803773533	P(sicher|regulations)	0.0637078421
P(und|you)	0.0027307419	P(ist|the)	0.0154460450	P(ländlichen|remarkable)	0.0803773533	P(darauf|regulations)	0.0623177307
P(der|you)	0.0021275149	P(zu|the)	0.0124781373	P(ihre|remarkable)	0.0745146927	P(bin|regulations)	0.0536674392
P(die|you)	0.0019539817	P(das|the)	0.0111399018	P(sie|remarkable)	0.0662174278	P(Frage|regulations)	0.0512217438

Figure 6. Translation Probabilities of frequent words 'you' and 'the' as well as infrequent words 'remarkable' and 'regulations'

Question 4.

The words pairs we choose are 'country' and 'countries', the latter being the plural form of the former. We tested on the first 5000 sentences to observe the translation distribution. The result show the most probable German translation for 'country' is 'Land' and the most possible translations for 'countires' are 'Länder' and 'Ländern', highlighted below (Figure 7).

Translation Probability of 'country' and 'countries'			
	country		countries
P(Land|country）	0.804906196	**P(Länder|countries）**	0.401280659
P(das|country)	0.040042196	**P(Ländern|countires）**	0.356169235
P(in|country）	0.021821581	P(in|countries)	0.049341209
P(einem|country)	0.020144076	P(die|countries)	0.029617071
P(,|country)	0.012539803	P(den|countries)	0.028701933
P(es|country）	0.01107394	P(.|countries)	0.021496504
P(Die|country）	0.01082999	P(,|countries)	0.019360353
P(und|country）	0.008122276	P(der|countries)	0.017572639
P(.|country）	0.008088487	P(zu|countries)	0.008652243
P(ist|country）	0.006054692	P(werden|countries)	0.007177821

Figure 7. The 10 most frequent German translations for ‘country’ and ‘countries’

It could be speculated from the result that German is also a morphologically rich language as ‘Länder’ and ‘Ländern’(translation for the plural form ‘countries’) are quite similar to the ‘Land’(translation for the singular form). In other words, we could treat ‘Länder’, ‘Ländern’ and ‘Land’ as different morphological forms of the same word and they might be in the same entry in a German dictionary. However, since the IBM model 1 fails to differentiate morphological forms as well as meaning issues, these morphological derivations of words and words themselves are regarded as distinctive words. Therefore, we have two completely somehow independent translation distribution for ‘country’ and ‘countries’ respectively. This might be not so reasonable, especially when it comes to translation of the morphologically rich language. It would waste a lot of unnecessary computation to handle the translation of basically the same word or words with same meaning. We might be able to preprocess our data to identify these morphological derivations of one word using finite state transducer or even using the edit distance to pre-identify these words.

Question 5

After we have studied the alignment, we find the English words are not chosen at random. It could be spotted the most human alignment is somewhere around the diagonal. However, in IBM model 1, since we consider each alignment as equally likely and the word our model chooses is somehow at random because we have not recorded any data for alignment. The only issue that matters in our current model is two words appear together in a sentence pair. The position of the word in the sentence has not been taken into consideration. This kind of method could certainly cause a high AER. And we believe that’s why our alignment in IBM model 1 is far from perfect.

Notes:

We have implemented the IBM Model 1 and IBM Model 2. The codes for the Model 2 are commented in "align.py", because we found that for the first 300 lines' experiments, the AER of Model 2 is 0.678262, which is higher than Model 1's AER(0.631880). Thus, we choose the result of Model 1 as our final result.