Machine Translation: Summer Term 2021

Ex3: IBM Models 1, 2 and 3; Expectation Maximisation (EM)

Sangeet Sagar (7009050)
sasa00001@stud.uni-saarland.de

July 5, 2021

1. **Why is the following idea of estimating translation probabilities $p(e|f)$, where e is an English sentence and f is a foreign sentence, not a good idea:**

$$p(e|f) =_{MLE} \frac{\text{number of times } f \text{ translates as } e}{\text{number of times } f \text{ translates into anything}} \tag{1}$$

   - Even if we have million pair of parallel data, there are lots of sentences we never see in the data
   - There are many sentences we see only once, and if see such sentences only once it is not suitable for estimating probability $p(e|f)$. For any sentence that is absent in the training data, we would not have any probability.

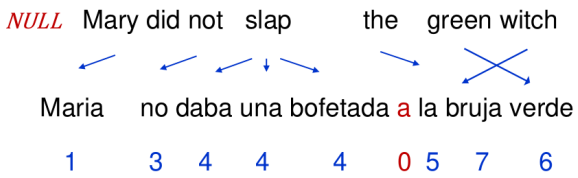
2. **What strategy can you apply to do better?**
   We apply the generative approach to model how e gets turned into f in the noisy channel Bayesian inversion in a number of small steps. We parametrise the model i.e. break it into small steps and estimate parameters for the small steps. This way we can compute $p(e|f)$ for any new sentences that we have not seen before.

3. **Given the following two sentences and the alignment vector, what does the alignment vector say?**

   - ***Mary did not slap the green witch .***
   - ***Maria no daba una bofetada a la bruia verde.***
   - **1,2,4,4,4,0,5,7,6,8**

   The alignment vector encodes information about the alignment order for the target sentence from the source sentence. It means that the *Maria* in Spanish is aligned to 1st word i.e. *Maria* in English, *no* is aligned with word 3 in English, *daba* is connected with word 4 and so on. We also have a *NULL* (hidden) element at the beginning of the English sentence and a is connected with this 0 word or the *NULL* word.

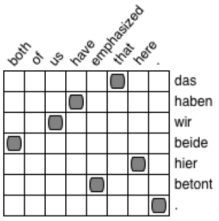
4. **Draw the alignment in 3.**

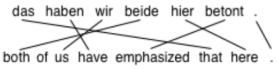


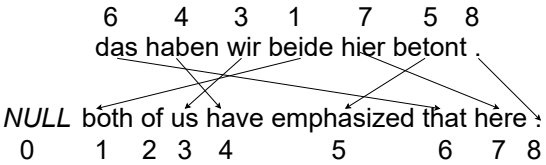
5. **Given the alignments in 3 and 4, in your own words explain the null element, fertility, reordering and translation parameters.**
   When aligning a source and target sentence, certain words appear in the target sentence that basically do not align with any word in the source sentence. Such words are aligned with *NULL* word positioned at 0 in the source sentence. This *NULL* element is responsible for anything that is idiosyncratic to the target language in the translation.

Fertility is the number of English words generated by a foreign word [Source :`https://hpi.de/fileadmin/user_upload/fachgebiete/plattner/teaching/MachineTranslation/MT2015/MT05_IBMModels.pdf`Mariana Neves]. The fertility parameter $n(\varphi_i|e_i)$ denotes the probability that e is aligned with φ words.
The translation parameter $t(f_j|e_{a_j})$ denotes the probability of the French word f given the English word e.
Distortion or reordering parameter $d(j|a_j, l, m)$ tells us when the position of a word in the source and its translation are not in the same place. It denotes probability that a_jth French word is connected to jth English word, given sentence lengths of e and f are l and m respectively.

6. **Express the alignment drawn in the pictures below as an alignment vector:**







Hence the alignment is: $< 6, 4, 3, 1, 7, 5, 8 >$

7. **Draw the alignment in 3 and 4 above as a two-dimensional grid.**

maria	did	not	slap	the	green	witch	
*							Maira
		*					no
			*				daba
			*				una
			*				bofetada
							a
				*			la
						*	bruja
					*		verde

8. **Given that your source string has l words and your target string has m words, how many alignments can you have between the source and the target string? (Remember to include the null element).**
$(l+1)^m$

9. **In what sense are alignments a the hidden structure of translation. In what sense is a a latent variable in IBM models 1 (and the others). Explain in your own words.**
The alignments a are not given in the parallel training data but they are part of how we "tell our story" of which French word comes from a English word. Therefore, a is a latent variable modeling hidden structure.

10. **Given that**

$$\hat{e} = \underset{e}{\operatorname{argmax}}\, p(e|f) = \underset{e}{\operatorname{argmax}}\, p(f|e) \cdot p(e)$$

which component part is modeled by IBM model 1?
We model $\text{argmax}_e\, p(f|e)$ using IBM model 1.

11. **Explain why $p(f, a|e, m) = p(a|e, m) \times p(f|a, e, m)$? What are f, e, a and m?**
This is because we apply conditional probability on $p(f, a|e, m)$ conditioned on e, m. We could simply write it as:

$$p(f, a) = p(a) \times p(f|a)$$
$$p(f|a) = \frac{p(f, a)}{p(a)}$$

- $p(f, a|e, m)$: probability of the french and a particular alignment given the English and the length of the English
- $p(a|e, m)$: probability of the alignment given the English string and the length of the French string
- $p(f|a, e, m)$: probability of the French given the alignment, the English string and length of the French string

12. **In your own words, explain why:**

$$p(f|e, m) = \sum_{a \in A} p(f, a|e, m) = \sum_{a \in A} p(a|e, m) \times p(f|a, e, m) \tag{2}$$

What is the technical term designating what is used to get rid of the a's in the right-hand side of this equation?
We have $p(f|e, m)$ i.e. probability of the French translations given the English string conditioned on the length of the French string m. We compute this by summing over all possible alignments ($\sum_{a \in A}$) A in $p(f, a|e, m)$ that given the input English string e is translating to the same French string f. This is further unpacked into product of two probabilities discussed above.
The summation term in the right hand side of the equation is used to marginalize out a by summing over all alignments in A.

13. **In your own words, explain IBM model 1:**

$$p(f, a|e, m) = p(a|e, m) \times p(f|a, e, m) = \frac{1}{(1+l)^m} \prod_{j=1}^{m} \mathbf{t}(f_j|e_{a_j})$$

$$p(f|e, m) = \sum_{a \in A} p(f, a|e, m) = \sum_{a \in A} \frac{1}{(1+l)^m} \prod_{j=1}^{m} \mathbf{t}(f_j|e_{a_j})$$

we have:
$p(a|e, m)$: probability of an alignment given an English string and length of French string. In IBM Model 1 all alignments are equally likely hence,

$$p(a|e, m) = \frac{1}{(1+l)^m}$$

Next we model $p(f|a, e, m)$ which is probability distribution over French translations given an alignment, an English sentence and length for the French sentence.

$$p(f|a, e, m) = \prod_{j=1}^{m} \mathbf{t}(f_j|e_{a_j})$$

This probability is modeled in terms of product of translation probability of the individual French words conditioned over English words. We compute conditional probability for a French word at index j and the alignment tells us the position of English word that the corresponding French word is aligned with.

The generative process in Model 1 could be summarized as:

- Pick an alignment a with probability $p(a|e, m) = \frac{1}{(1+l)^m}$

- Pick French words with probability $p(f|a,e,m) = \prod_{j=1}^{m} \mathbf{t}(f_j|e_{a_j})$

We apply conditional probability on $p(f,a|e,m)$ to simply decompose into $p(a|e,m) \times p(f|a,e,m)$. Using above equations and putting it together, we get

$$\begin{aligned} p(f,a|e,m) &= p(a|e,m) \times p(f|a,e,m) \\ &= \frac{1}{(1+l)^m} \times \prod_{j=1}^{m} \mathbf{t}(f_j|e_{a_j}) \end{aligned}$$

Now we sum over all alignments A to marginalize out a. Hence we have,

$$p(f|e,m) = \sum_{a \in A} p(f,a|e,m)$$

14. **IBM Model 2: in your own words, explain the distortion parameter**

$$\mathbf{q}(i|j,l,m) \tag{3}$$

The IBM Model 2 has an additional parameter for alignment that is not present in Model 1 i.e. distortion parameter [Source: Wiki]. The above distortion parameter is the probability that j-th French word is aligned to i-th English word given that the lengths of the English and French sentences are l and m respectively.

We parameterize the probability of the alignment given the English string and the length of the French string in terms of product of the distortion parameters as

$$p(a|e,m) = \prod_{j=1}^{m} \mathbf{q}(a_j|j,l,m) \tag{4}$$

15. **IBM Model 2: given that**

$$p(a|e,m) = \prod_{j=1}^{m} \mathbf{q}(a_j|j,l,m) \tag{5}$$

and the following example

- $l =$ **6**
- $m =$ **7**
- $e =$ **And the program has been implemented**
- $f =$ **Le the programme a ete mis en application**
- $a = \{2,3,4,5,6,6,6\}$

what is $p(a|e,7)$

$$\begin{aligned} p(a|e,7) &= \prod_{j=1}^{m} \mathbf{q}(a_j|j,l,m) \\ &= \prod_{j=1}^{7} \mathbf{q}(a_j|j,6,7) \\ &= \mathbf{q}(2|1,6,7) \times \mathbf{q}(3|2,6,7) \times \mathbf{q}(4|3,6,7) \\ &\times \mathbf{q}(5|4,6,7) \times \mathbf{q}(6|5,6,7) \times \mathbf{q}(6|6,6,7) \\ &\times \mathbf{q}(6|7,6,7) \end{aligned}$$

16. **IBM Model 2: in your own words explain:**

$$p(f|a,e,m) = \prod_{j=1}^{m} \mathbf{t}(f_j|e_{a_j}) \tag{6}$$

In particular, what are j and a_j above?
This represents the translation probabilities of a French word given the English word which is identified by the corresponding alignment. E.g. in the above case, the English words will be identified according to the given alignment $a = \{2, 3, 4, 5, 6, 6, 6\}$.

Here j is the index of the French word and a_j is the position of the English word in the a-th alignment at index j.

17. **IBM Model 2: in your own words, explain**

$$p(f, a|e, m) = p(a|e, m) \times p(f|a, e, m) = \prod_{j=1}^{m} \mathbf{q}(a_j|j, l, m) \cdot \mathbf{t}(f_j|e_{a_j}) \tag{7}$$

In IBM Model 2:

- Pick an alignment $a = \{a_1, a_2, \cdots, a_m\}$ with probability $p(a|e, m) = \prod_{j=1}^{m} \mathbf{q}(a_j|j, l, m)$.
- Pick French words with probability $p(f|a, e, m) = \prod_{j=1}^{m} \mathbf{t}(f_j|e_{a_j})$

Putting them together we have:

$$p(f, a|e, m) = p(a|e, m) \times p(f|a, e, m) = \prod_{j=1}^{m} \mathbf{q}(a_j|j, l, m) \cdot \mathbf{t}(f_j|e_{a_j}) \tag{8}$$

18. **Express $p(f|e, m)$ in terms of the formula in 17 above, by marginalizing over the alignments.**

$$p(f|e, m) = \sum_{a \in A} p(f, a|e, m) = \sum_{a \in A} \prod_{j=1}^{m} \mathbf{q}(a_j|j, l, m) \cdot \mathbf{t}(f_j|e_{a_j})$$

19. **In your own words, please explain IBM Model 3:**

$$\begin{aligned} p(a, f|e) = &\binom{m - \varphi_0}{\varphi_0} \times p_0^{(m-2\varphi_0)} \times p_1^{\varphi_0} \\ &\times \prod_{i=1}^{l} n(\varphi_i|e_i) \times \prod_{j=1}^{m} t(f_j|e_{a_j}) \\ &\times \prod_{j:a_j \neq 0}^{m} d(j|a_j, l, m) \times \prod_{i=0}^{l} \varphi_i! \times \frac{1}{\varphi_0!} \end{aligned}$$

Recall that: $P(f|e) \sum P(a, f|e)$ **and** $P(a|e, f) = \frac{P(a,f|e)}{\sum_a P(a,f|e)}$
IBM Model 3 parameters-

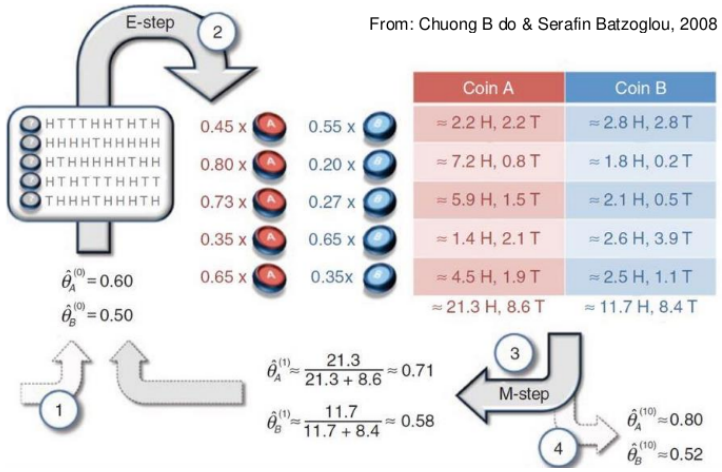
- For each word e_i in the English string, choose a fertility φ_i. This comes from the fertility distribution **n** for a particular word. The fertility choice is dependent entirely on the English word.
- For each English word e_i, we generate φ_i French words using the translation probability $t(f|e)$. The choice of French word is dependent solely on the English word that generates it and not on the English context around the English word. It is also not dependent on the other French words that have been generated from this or any other English word.
- All those French words are permuted using the distortion parameter $d(\phi_f|\phi_e, l, m)$ where each French word is assigned an absolute target "position slot".

Note: Refer slide 7,9: `MT_6_IBM_Model_3.pdf` for more details.

Model 3 also introduces the concept of "spurious words" where for every English sentence we assume a `NULL` word at the beginning of the string that can generate spurious words in the target $t(w|NULL)$. These `NULL` words doesn't have a fertility distribution n, but a probability p_1 with which it can generate a spurious word after each properly generated word.

The generative story of Model 3. [Source: slide 12; Lecture note: `MT_6_IBM_Model_3.pdf`]

20. **In your own words, explain the main idea about Expectation Maximisation (EM):**



The expectation-maximization algorithm is an approach for performing maximum likelihood estimation in the presence of latent variables. It does this by first estimating the values for the latent variables, then optimizing the model, then repeating these two steps until convergence. [Source: EM Algorithm]

21. **Please do Expectation Maximisation (EM) to estimate $\hat{\theta_A}$ and $\hat{\theta_B}$ (the probability of A producing a head, and the probability of B producing a head) under the initial random assignments $\hat{\theta_A^{(0)}} = 0.6$ and $\hat{\theta_B^{(0)}} = 0.4$:**

Given

$$\hat{\theta_A^0} = 0.6$$
$$\hat{\theta_B^0} = 0.4$$

E-Step

1. Sequence 1: H T T T H H T H T H
   $P(A) = \frac{0.6^5 \cdot 0.4^5}{0.6^5 \cdot 0.4^5 + 0.4^5 \cdot 0.6^5} = 0.5$
   $P(B) = \frac{0.4^5 \cdot 0.6^5}{0.6^5 \cdot 0.4^5 + 0.4^5 \cdot 0.6^5} = 0.5$
   Coin A: $0.5 * 5 = 0.25H, 0.5 * 5 = 0.25T$
   Coin B: $0.5 * 5 = 0.25H, 0.5 * 5 = 0.25T$

2. Sequence 2: H H H H T H H H H H
   $P(A) = \frac{0.6^9 \cdot 0.4^1}{0.6^9 \cdot 0.4^1 + 0.4^9 \cdot 0.6^1} = 0.96$
   $P(B) = \frac{0.4^9 \cdot 0.6^1}{0.6^9 \cdot 0.4^1 + 0.4^9 \cdot 0.6^1} = 0.037$
   Coin A: $0.96 * 9 = 8.64H, 0.96 * 1 = 0.96T$
   Coin B: $0.037 * 9 = 0.333H, 0.037 * 1 = 0.037$

3. Sequence 3: H T H H H H H T H H
   $P(A) = \frac{0.6^8 \cdot 0.4^2}{0.6^8 \cdot 0.4^2 + 0.4^8 \cdot 0.6^2} = 0.91$
   $P(B) = \frac{0.4^8 \cdot 0.6^2}{0.6^8 \cdot 0.4^2 + 0.4^8 \cdot 0.6^2} = 0.08$
   Coin A: $0.91 * 8 = 7.28H, 0.91 * 2 = 1.82T$
   Coin B: $0.08 * 8 = 0.64H, 0.08 * 2 = 0.16T$

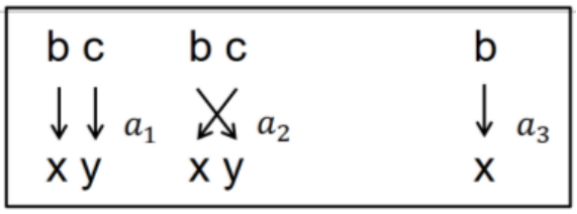
4. Sequence 3: H T H T T T H H T T
   $P(A) = \frac{0.6^4 \cdot 0.4^6}{0.6^4 \cdot 0.4^6 + 0.4^4 \cdot 0.6^6} = 0.307$
   $P(B) = \frac{0.4^4 \cdot 0.6^6}{0.6^4 \cdot 0.4^6 + 0.4^4 \cdot 0.6^6} = 0.69$
   Coin A: $0.307 * 4 = 1.228H, 0.307 * 6 = 1.842T$
   Coin B: $0.69 * 4 = 2.76H, 0.69 * 6 = 4.14T$

5. Sequence 4: T H H H T H H H T H
   $P(A) = \frac{0.6^7 \cdot 0.4^3}{0.6^7 \cdot 0.4^3 + 0.4^7 \cdot 0.6^3} = 0.83$
   $P(B) = \frac{0.4^7 \cdot 0.6^3}{0.6^7 \cdot 0.4^3 + 0.4^7 \cdot 0.6^3} = 0.16$
   Coin A: $0.83 * 7 = 5.81H, 0.83 * 3 = 2.49T$
   Coin B: $0.83 * 3 = 2.49H, 0.16 * 3 = 0.48T$

M-Step
Coin A: 23.208H, 7.362T
Coin B: 5.103H, 5.087T

$$\hat{\theta_A^1} = \frac{23.208}{23.208 + 7.362} = 0.76$$
$$\hat{\theta_B^2} = \frac{5.103}{5.103 + 5.087} = 0.50$$

22. **Please estimate translation parameters t using EM given the following data**



and uniform initial translation parameters:

$$t(x|b) = t(y|b) = t(x|c) = t(y|c) = \frac{1}{2} \tag{9}$$

Translation parameter(t) is given by

$$t(x|b) = \frac{\#(b \longrightarrow x)}{\#b}$$

Also,

$$P(a, f|e) = \prod_{j=1}^{2} t(f_j|e_{a_j})$$
$$P(a|f, e) = \frac{P(a, f|e)}{\sum_a P(a, f|e)}$$

EM step 1

1. $t(x|b) = t(x|c) = t(y|b) = t(y|c) = \frac{1}{2}$
2. $P(a_1, f|e) = t(f_1|e_{a_1}) \cdot t(f_2|e_{a_2}) = t(x|b) \cdot t(y|c) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$
   $P(a_2, f|e) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$
   $P(a_3, f|e) = \frac{1}{2}$
3. $P(a_1|f, e) = \frac{1/4}{1/4+1/4} = \frac{1}{2}$
   XXX Why not $1/4 + 1/4 + 1/2$
   $P(a_2|f, e) = \frac{1/4}{1/4+1/4} = \frac{1}{2}$
   $P(a_3|f, e) = \frac{1/2}{1/2} = 1$

4. Weighted by alignment probabilities
   $\#t(x|b) = 1/2$
   $\#t(y|b) = 1/2 + 1 = 3/2$
   $\#t(x|c) = 1/2$
   $\#t(y|b) = 1/2$

5. Normalized fractional counts
   $t(x|b) = \frac{1/2}{4/2} = 1/4$
   $t(y|b) = \frac{3/2}{4/2} = \mathbf{3/4} \uparrow$
   $t(x|c) = \frac{1/2}{1} = 1/2$
   $t(y|b) = \mathbf{1/2}$

EM step 2

1. $P(a_1, f|e) = t(x|b) \cdot t(y|c) = \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8}$
   $P(a_2, f|e) = \frac{1}{2} \cdot \frac{3}{4} = \frac{3}{8}$
   $P(a_3, f|e) = \frac{3}{4}$
2. $P(a_1|f, e) = \frac{1/8}{1/8+3/8} = \frac{1}{4}$
   XXX Why $1/8 + 3/8$ and why not $1/8 + 3/8 + 3/8$
   $P(a_2|f, e) = \frac{3/8}{1/8+3/8} = \frac{3}{4}$
   $P(a_3|f, e) = \frac{3/4}{3/4} = 1$

3. Weighted by alignment probabilities
   $\#t(x|b) = 1/4$
   $\#t(y|b) = 3/4 + 1 = 7/4$
   $\#t(x|c) = 3/4$
   $\#t(y|b) = 1/4$

4. Normalized fractional counts
   $t(x|b) = \frac{1/4}{8/4} = 1/8$
   $t(y|b) = \frac{7/4}{8/4} = \mathbf{7/8} \uparrow$
   $t(x|c) = \frac{3/4}{1} = 3/4$
   $t(y|b) = \mathbf{1/4} \downarrow$

References