

Lecture 3 — 4, Jan 17-19, 2006

Lecturer: Anoop Sarkar

Scribe: Gholamreza Haffari

3.1 Required reading

- Read sections 23 - 29 of the workbook "A Statistical MT Tutorial Workbook", written by Kevin Knight.
- Brown et al. "The Mathematics of Statistical Machine Translation: Parameter Estimation", 1993.

3.2 IBM Translation Models

We want to describe different models for hypothetical translation mechanisms from English to French. Suppose we have an English sentence \mathbf{e} and its French translation \mathbf{f} . Let the size of $\mathbf{e} = e_1, \dots, e_l$ to be l where e_i is its i th word. Similarly, Let the size of $\mathbf{f} = f_1, \dots, f_m$ to be m where f_j is its j th word.

It is assumed that each French word f_j is translated from *one* (unknown) English word e_i . As an example let "A B C" be our English sentence and "X Y Z" be our French sentence. It might be the case that "X" is translated from "B", "Y" is translated from "A", and "Z" is translated from "C" (see figure 3.1). Another way to express this is to write (2,1,3) below the French sentence, which for each word position in \mathbf{f} shows the word position in \mathbf{e} that has produced it. This sequence of positions is called the *alignment* and denoted by \mathbf{a} .

We also assume that there is a special position e_0 which can spuriously produce some French words. At this time, there is a constraint that each word in a position in the \mathbf{f} can only be produced by only one word in a position in the \mathbf{e} . In other words, for each position in the \mathbf{f} there are $l + 1$ positions in \mathbf{e} which can be responsible for producing (the word in) it, i.e. the number of alignments = $(l + 1) \times \dots \times (l + 1) = (l + 1)^m$. Denote the set of all alignments between \mathbf{e} and \mathbf{f} by $\mathcal{A}(\mathbf{e}, \mathbf{f})$. Therefore the probability that

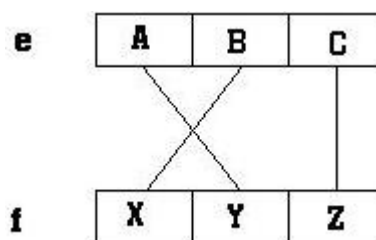


Figure 3.1. Alignment.

a given French sentence \mathbf{f} to be the translation of an English sentence \mathbf{e} is:

$$P(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a} \in \mathcal{A}(\mathbf{e}, \mathbf{f})} P(\mathbf{f}, \mathbf{a}|\mathbf{e})$$

In the general case, the following expression can be written for the probability that a French sentence \mathbf{f} has been produced based on the English sentence \mathbf{e} via an alignment \mathbf{a} :

$$P(\mathbf{f}, \mathbf{a}|\mathbf{e}) = P(m|\mathbf{e}) \prod_{j=1}^m P(a_j | a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) \cdot P(f_j | a_1^j, f_1^{j-1}, m, \mathbf{e}) \quad (3.1)$$

where a_j means the alignment for the j th position, $a_1^j = a_1, \dots, a_j$ means the alignments for the positions 1 to j , and the same thing for f_j and f_1^j . Moreover the term $P(m|\mathbf{e})$ refers to the probability of a particular length for the French sentence given the English sentence.

3.2.1 Model 1

In Model 1, several assumptions are made to simplify the expression (3.1). In particular the assumptions are:

- $P(m|\mathbf{e})$ is a fixed number ϵ
- $P(a_j | a_1^{j-1}, f_1^{j-1}, m, \mathbf{e})$ only depends on the length of English sentence and is equal to $\frac{1}{i+1}$.
- $P(f_j | a_1^j, f_1^{j-1}, m, \mathbf{e})$ only depends on the English word in the alignment corresponding to this French word. In other words it is denoted by

$t(f_j|e_{a_j})$ and called the *translation probability*. $t(\cdot|e)$ shows the probability of generating different French words based on a particular English word e .

Consequently, the expression (3.1) is simplified to:

$$P(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^m t(f_j|e_{a_j}) \quad (3.2)$$

Now suppose we are given an English sentence \mathbf{e} , how do we generate its French translation based on Model 1? Simply! At first a length for the translation \mathbf{f} is chosen based on ϵ . Afterwards, for each position j in \mathbf{f} , a number from zero to l is selected which specifies the English word in \mathbf{e} that should produce this French word, and then the French word is generated based on the parameters $t(f|e)$. This is called the *generating story*. Later we will see other generating stories corresponding to other translation models.

Training

We are given as a sample a pair of English sentence and its French translation $\{(\mathbf{e}, \mathbf{f})\}$. Suppose we want to build a model based on Model 1, how do we find the translation probabilities $t(f|e)$? By counting! Simply we look at all of the alignments and count in how many of them f is the translation of e and then we normalize the counts:

$$t(f|e) = \frac{\sum_{\mathbf{a}} \#tc(f, e; \mathbf{a})}{\sum_{f'} \sum_{\mathbf{a}} \#tc(f', e; \mathbf{a})}$$

where we have assumed that all of the alignment have equal probability. Now suppose the alignments do not have equal probabilities, we simply change the counts to the *weighted* counts, aka fractional counts:

$$t(f|e) = \frac{\sum_{\mathbf{a}} \#tc(f, e; \mathbf{a})P(\mathbf{a})}{\sum_{f'} \sum_{\mathbf{a}} \#tc(f', e; \mathbf{a})P(\mathbf{a})} \quad (3.3)$$

Expression (3.2) can be used to find the probability of each alignment $P(\mathbf{a}|\mathbf{e}, \mathbf{f}) = \frac{\mathbf{f}, \mathbf{a}|\mathbf{e}}{P(\mathbf{f}|\mathbf{e})}$ and since $P(\mathbf{a})$ appears in both numerator and denominator of the expression (3.3), the term $P(\mathbf{a})$ can be ignored. Notice that in calculating the probability of alignments we need the translation probabilities! What

do we do? We start from uniform values for translation probabilities. Then, the alignment probabilities are computed and translation probabilities are updated based on the fractional counts. Based on these updated translation probabilities, the learning process continues until it converges to a fixed point.

The other issue is the time needed to calculate the numerator and denominator of the expression (3.3). As we saw, the number of alignments is exponential in the length of the French sentence, how can we calculate the fractional counts over all possible alignments *efficiently*? Look at the following table:

$$\begin{pmatrix} P(f_1|e_0) & P(f_2|e_0) & \dots & P(f_m|e_0) \\ P(f_1|e_1) & P(f_2|e_1) & \dots & P(f_m|e_1) \\ P(f_1|e_2) & P(f_2|e_2) & \dots & P(f_m|e_2) \\ \vdots & \vdots & \vdots & \vdots \\ P(f_1|e_l) & P(f_2|e_l) & \dots & P(f_m|e_l) \end{pmatrix}$$

Starts from an arbitrary point in the first column, then go to an arbitrary point in the second column until reach the last column. This path corresponds to one possible way (alignment) for generating the French sentence from English sentence, and its probability is the product of the numbers along the path (call it the value of path)! The sum of values of all paths can be seen to be the sum of the numbers in the first column times the sum of the numbers in the second columns until the last column. In other words it is:

$$\sum_{i=0}^l P(f_1|e_i) \times \sum_{i=0}^l P(f_2|e_i) \dots \sum_{i=0}^l P(f_m|e_i) = \prod_{j=1}^m \sum_{i=0}^l P(f_j|e_i)$$

Question. What would be the sum of probabilities of all alignments in which the first word in the French sentence f_1 is translated from the first word in the English sentence e_1 ?

The Best Alignment

Suppose we are given a pair of English and French sentences, and asked to find the alignment for it based on Model 1. Having the translation probabilities $t(f|e)$, the procedure is as follows: for each position j in the French sentence, we choose $\arg \max_{0 \leq i < l} t(f_j|e_i)$.

3.2.2 Model 2

In the Model 1 we assumed that a French word in a position j can be connected to all the English words in positions zero to l equally likely. In the Model 2, this assumption is changed to that we consider for each position in the French sentence the probabilities like $P(i|j, m, l)$ such that $\sum_i P(i|j, m, l) = 1$. In model 2 all of the assumptions are the same as Model 1 except that here we have *alignment probabilities* $a(a_j|j, m, l) = P(a_j|a_1^{j-1}, f_1^{j-1}, m, \mathbf{e})$. Therefore, the probability expression (3.1) is simplified to:

$$P(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \epsilon \prod_{j=1}^m a(a_j|j, m, l) \cdot t(f_j|e_{a_j}) \quad (3.4)$$

The generating story is as follows. At first a length for the translation \mathbf{f} is chosen based on ϵ . Afterwards, for each position j in \mathbf{f} , a number from zero to l is selected based on the alignment probabilities which specifies the English word in \mathbf{e} that should produce this French word, and then the French word is generated based on the parameters $t(f|e)$.

Training

The training procedure is similar to the training of Model 1. For translation probabilities, we use the expression (3.3) except that in this case the alignment probabilities are calculated based on (3.4). The trick used for efficient computation of it is also similar to the previous:

$$\begin{aligned} \sum_{i=0}^l P(f_1|e_i)P(i|1, m, l) &\times \sum_{i=0}^l P(f_2|e_i)P(i|2, m, l) \dots \sum_{i=0}^l P(f_m|e_i)P(i|m, m, l) \\ &= \prod_{j=1}^m \sum_{i=0}^l P(f_j|e_i)P(i|j, m, l) \end{aligned}$$

For estimating the alignment probabilities, a similar method is used. Suppose the i th position in the English sentence is connected to the j th position in the French sentence:

$$a(i|i, m, l) = \frac{t(f_j|e_i)}{\sum_{i'=0}^l t(f_j|e_{i'})a(i'|j, m, l)}$$

The Best Alignment

Suppose we are given a pair of English and French sentences, and asked to find the alignment for it based on Model 2. Having the translation probabilities $t(f|e)$ and alignment probabilities $a(i|j, m, l)$, the procedure is as follows: for each position j in the French sentence, we choose $\arg \max_{0 \leq i \leq l} t(f_j|e_i)a(i|j, m, l)$.

3.3 Model 3

Model 3 is very different from the previous models, and there are some new parameters in this model which did not exist in the previous ones. There is a parameter called *fertility* $n(\phi|e)$ which shows what is the probability of generating ϕ French words based on the English word e . The other parameter is called *distortion* $d(j|i, m, l)$ which shows the probability of placing the French word corresponding to the English word in the position i th of \mathbf{e} into position j th of the \mathbf{f} where the size of English and French sentences is l and m respectively. We have also a coin with probability p_H of coming Head: At some points we want to probabilistically decide whether to produce a spurious word or not; we simply toss the coin if Head comes out the word is produced, otherwise it is not produced.

Now we introduce the Model 3 by mentioning its generating story:

1. For each word e_i ($1 \leq i \leq l$) choose its fertility with probability $n(\phi_i|e_i)$.
2. Add all of fertilities $s = \sum_{i=1}^l \phi_i$. Toss the coin s times, count the number of times Head comes out and call it ϕ_0 . In fact ϕ_0 is the fertility of e_0 . The length of the French sentence will be $m = s + \phi_0$.
3. For each $i = 0, 1, \dots, l$ and each $k = 1, \dots, \phi_i$ choose a French word τ_{ik} with probability $t(\tau_{ik}|e_i)$.
4. For each $i = 1, \dots, l$ and each $k = 1, \dots, \phi_i$ choose target French position π_{ik} with probability $d(\pi_{ik}|i, l, m)$. After this phase, s positions in the set of positions $1, 2, \dots, m$ are occupied.
5. Consider the set of free positions after the previous step, and choose the position of the first spurious word π_{01} uniformly at random from them ($\#$ possibilities = ϕ_0). Then choose the position of the next spurious word π_{02} from the remaining positions ($\#$ possibilities = $\phi_0 - 1$), and

so on. The number of possible assignments of positions to $\pi_{01}, \dots, \pi_{0\phi_0}$ would be $\phi_0 \times (\phi_0 - 1) \dots \times 1 = \phi_0!$

6. Output the French sentence with words τ_{ik} in positions π_{ik} ($0 \leq i \leq l, 1 \leq k \leq \phi_i$).

Let us consider an example. Assume the English sentence is just one word "A", fertility is selected to be $\phi_1 = 2$, then $[\tau_{11} = X, \tau_{12} = Y]$, and finally $[\pi_{11} = 1, \pi_{12} = 2]$. The resulting French sentence is "XY". We can also show this generation process incompletely by using alignment (see figure 3.2). Now consider these values $[\pi_{11} = 2, \pi_{12} = 1]$ and $[\tau_{11} = Y, \tau_{12} = X]$. Again the resulting French sentence is "XY", and the alignment is what we saw before in the figure 3.2. Therefore we can see that there may be more than one way to reach to a sentence **f** from a sentence **e** based on our generating story. All of these different ways of translating **e** to **f** have the same alignment representation. That is the reason why *using the alignment is an incomplete way of representing what really happens in Model 3*.

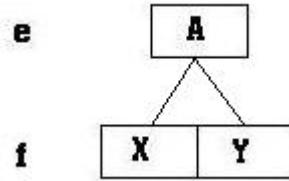


Figure 3.2. Alignment.

Suppose we are given a pair of sentences (**e**,**f**), what is the probability of generating **f** from **e** based on the generating story of the Model 3? Denote $(\tau_{i1}, \dots, \tau_{i\phi_i})$ by τ_i , and similarly $(\pi_{i1}, \dots, \pi_{i\phi_i})$ by π_i . Suppose a specific values for $(\tau_0, \dots, \tau_l, \pi_0, \dots, \pi_l)$ which transform **e** to **f**; we compute its probability. Then we add the probability of all possible ways of converting **e** to **f** to find $P(\mathbf{f}|\mathbf{e})$. What is $P(\tau_0, \dots, \tau_l, \pi_0, \dots, \pi_l|\mathbf{e})$? To answer this, first the fertilities $(\phi_0, \phi_1, \dots, \phi_l)$ must be chosen with the probability:

$$\prod_{i=0}^l P(\phi_i) = \binom{s}{\phi_0} (p_H)^{\phi_0} (1 - p_H)^{s-\phi_0} \cdot \prod_{i=1}^l n(\phi_i|e_i) \quad (3.5)$$

where the first term in the right hand side is the probability of ϕ_0 times coming Head in $s = \sum_{i=1}^l \phi_i$ experiments. After specifying the fertilities,

the French words have to be generated. The probability of generating the French word τ_{ik} only depends on the English word in the position i , hence its probability is $t(\tau_{ik}|e_i)$. Since the words are generated independently, the probability of generating the French words (τ_0, \dots, τ_l) is:

$$\prod_{i=0}^l \left(\prod_{k=1}^{\phi_i} t(\tau_{ik}|e_i) \right) = \prod_{j=1}^m t(f_j|e_{a_j}) \quad (3.6)$$

where we have written the right hand side by noting the definition of alignment¹. Now we have to specify the probability of a permutation (π_0, \dots, π_l) . For the probability of positions corresponding to (π_1, \dots, π_l) we use the distortion probabilities. For π_0 there is not any distortion probabilities, instead recall that there are ϕ_0 vacant positions and assigning each $\phi_0!$ possible configuration to π_0 is equally likely. Therefore, the probability of a specific permutation² is:

$$\frac{1}{\phi_0!} \cdot \prod_{i=1}^l \left(\prod_{k=1}^{\phi_i} d(\pi_{ik}|i, l, m) \right) = \frac{1}{\phi_0!} \prod_{j=1, a_j \neq 0}^m d(j|a_j, l, m) \quad (3.7)$$

At this point, we have all quantities needed to calculate the probability of converting \mathbf{e} to \mathbf{f} via a specific $(\tau_0, \dots, \tau_l, \pi_0, \dots, \pi_l)$; it is the multiplication of the terms (3.5), (3.6) and (3.7). As we mentioned, one set of values for $(\tau_0, \dots, \tau_l, \pi_0, \dots, \pi_l)$ specifies an alignment. However by rearranging the values in $(\tau_0, \dots, \tau_l, \pi_0, \dots, \pi_l)$ the same alignment can be reproduced (with the same probability). To get the idea, consider τ_1 and π_1 . Suppose we swap the word-position (τ_{11}, π_{11}) with (τ_{12}, π_{12}) to get new τ_1' and π_1' . It is clear that the alignment corresponding to $(\tau_0, \tau_1', \dots, \tau_l, \pi_0, \pi_1', \dots, \pi_l)$ is the same as before. The number of possibilities of rearranging the values contained in (τ_1, π_1) such that the alignment remains unchanged is $\phi_1!$. The same fact is true for any other places $0 \leq i \leq l$. So, the number of rearranging the values of $(\tau_0, \dots, \tau_l, \pi_0, \dots, \pi_l)$ in such a way that its corresponding alignment remains unchanged is $\prod_{i=0}^l \phi_i!$. Hence, the probability of generating \mathbf{f} from

¹Note that the word in the position i of \mathbf{e} is connected to the positions $\pi_{i1}, \dots, \pi_{i\phi_i}$ in the \mathbf{f} . Therefore the value of the alignment in the position, let say, π_{ik} is i .

²Note that this permutation is not *already* a valid one, in the sense that it does not assign the same French position to two different position variables π_{ik} and $\pi_{i'k'}$. The number of such valid permutations is $m \times (m-1) \dots \times (m - \sum_{i=1}^l \phi_i + 1)$.

\mathbf{e} via a specific alignment \mathbf{a} is:

$$P(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \prod_{i=0}^l \phi_i! \cdot \binom{s}{\phi_0} (p_H)^{\phi_0} (1 - p_H)^{s - \phi_0}.$$

$$\prod_{i=1}^l n(\phi_i|e_i) \cdot \prod_{j=1}^m t(f_j|e_{a_j}) \cdot \frac{1}{\phi_0!} \prod_{j=1, a_j \neq 0}^m d(j|a_j, l, m) \quad (3.8)$$

Note that $\phi_0!$ can be removed from the above formula because it is appeared both in the numerator and denominator.

Training

The training principle is similar to the training of previous models: Counting! In other words, we start from some initial values for the parameters, then we update them based on the fractional counts. However, in this case there is not any efficient way (similar to previous models) to compute exactly the fractional counts. Instead, a subset of all good alignments is chosen and the parameters are updated based on the alignments in this set. In future we discuss it more in the class.

The Best Alignment

Suppose we are given a pair of English and French sentences, and asked to find the best alignment for it based on Model 3. Having all parameters of the model, it is very time consuming to find the best alignment. In fact, the problem is *not* decomposed to find the best choice for each individual position. Instead, all of the $(l+1)^m$ possible alignments have to be examined, and the best one is selected based on the score (3.8).