Higher IBM Models: Complementary Material

Jakub Waszczuk

Decemer 2018

1 Composing conditional distributions

The goal of this section is to show that the composition of two conditional distributions $P(X \mid Y)$ and $P(Y \mid Z)$ via simple multiplication leads to a valid conditional distribution $P(X, Y \mid Z)$.

Remark. *We will be only concerned with one property of conditional distribution $P(X \mid Y)$ – namely, that for any $y \in \mathrm{Val}(Y)$:*

$$\sum_{x \in \mathrm{Val}(x)} P(X = x \mid Y = y) = 1 \tag{1}$$

where $\mathrm{Val}(X)$ represents X's codomain (the set of values that X can take).

Remark. *In the following, we rely on simplified notation and write $P(x)$ to denote $P(X = x)$, $P(x \mid y)$ to denote $P(X = x \mid Y = y)$, etc., as long as the corresponding variables are clear from the context.*

Proposition 1. *Let X, Y, and Z be three random variables, and $P(X \mid Y)$, $P(Y \mid Z)$ be two conditional distributions. Then, $P(X, Y \mid Z)$ defined as:*

$$P(X, Y \mid Z) = P(X \mid Y) \times P(Y \mid Z) \tag{2}$$

which basically means:

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Y = y) \times P(Y = y \mid Z = z) \tag{3}$$

is also a valid conditional distribution. In particular, for each $z \in \mathrm{Val}(Z)$:

$$\sum_{x \in \mathrm{Val}(X), y \in \mathrm{Val}(Y)} P(x, y \mid z) = 1 \tag{4}$$

Proof. First of all, $\sum_{x \in \mathrm{Val}(X), y \in \mathrm{Val}(Y)}$ means that we sum over all possible pairs of values (x, y) (cartesian product of $\mathrm{Val}(X)$ and $\mathrm{Val}(Y)$). This is equivalent to summing over (i) all possible values of Y and, for each such $y \in \mathrm{Val}(Y)$, (ii) all possible values of X. Hence, the LHS of Eq. 4 can be rewritten as:

$$\sum_{y \in \mathrm{Val}(Y)} \sum_{x \in \mathrm{Val}(X)} P(x, y \mid z)$$

By definition (i.e., Eq. 3), we can split $P(x, y \mid z)$ as $P(x \mid y) \times P(y \mid z)$:

$$\sum_{y \in \mathrm{Val}(Y)} \sum_{x \in \mathrm{Val}(X)} P(x \mid y) \times P(y \mid z)$$

Since $P(y \mid z)$ does not depend on x, we can extract it from the inner sum:

$$\sum_{y \in \mathrm{Val}(Y)} P(y \mid z) \left(\sum_{x \in \mathrm{Val}(X)} P(x \mid y) \right)$$

Since $P(X \mid Y)$ is a conditional distribution, $\sum_{x \in \mathrm{Val}(X)} P(x \mid y) = 1$. Hence:

$$\sum_{y \in \mathrm{Val}(Y)} P(y \mid z) \times 1 = \sum_{y \in \mathrm{Val}(Y)} P(y \mid z)$$

But $P(Y \mid Z)$ is also a conditional distribution, and therefore:

$$\sum_{y \in \mathrm{Val}(Y)} P(y \mid z) = 1$$

$\square$

2 Number of tableaux

Proposition 2. *Given input* $\boldsymbol{f}$*, output* $\boldsymbol{e}$*, and alignment* $a \in A(m, n)$*, there are*

$$\binom{m - \phi_0}{\phi_0} \times \prod_{i=1}^{n} \phi_i! \tag{5}$$

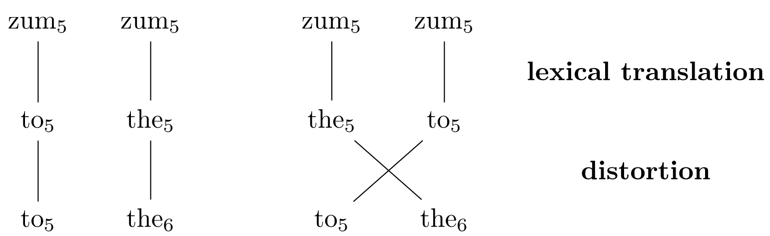
different tableaux $t \in \mathcal{T}_{\boldsymbol{e},\boldsymbol{f}}(a)$ *consistent with alignment* a*.*

2.1 Fertility

First of all, let's show the reason for the factor $\prod_{i=1}^{n} \phi_i!$. For the moment, let's focus on the example from the lecture and the word *zum* with fertility 2. There are two ($2! = 1 \cdot 2$) possible ways of translating *zum* to *to the*:

- *zum* is lexically translated to *to the* and kept intact in the distrotion step
- *zum* is lexically translated to *the to* and reordered in the distrotion step

Both options are represented on the tableau below.



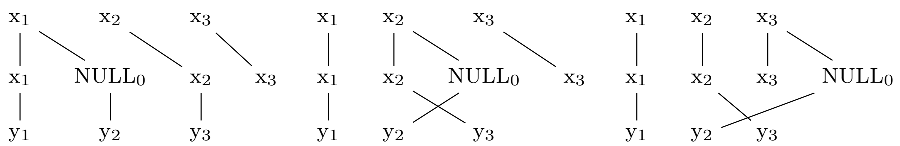
In general, for a word on position $i \in \{1, \ldots, n\}$ with fertility ϕ_i, the corresponding translations (all consistent with the same alignment a) can be generated in any order and, then, reordered in the distortion step. Therefore, all the $\phi_i!$ permutations have to be considered.

In total, we have to individually consider all the input positions i and the $\phi_i!$ possible ways of getting them translated to the corresponding output words, hence the factor $\prod_{i=1}^{n} \phi_i!$.

2.2 NULL insertion

As described during the lecture, NULL is inserted with probability p_0 after each word generated during the fertility step. However, NULL is always inserted with index 0 and any output word on position i aligned to NULL factors in the same distortion probability $P(i \mid 0, m, n)$, regardless of where this NULL has been exactly inserted.

For instance, the following three tableaux (where x_1, x_2, x_3 result from the ferility step) all correspond to the same alignment:



In general, we need to answer the following question: what is the number of different vectors resulting from the NULL insertion step, all with the same number of NULL tokens (ϕ_0)? The answer is $\binom{m-\phi_0}{\phi_0}$, which stems from the following proposition.[1]

Proposition 3. *Let $x = (x_1, x_2, \ldots, x_n)$ be a sequence of length n and $k \in \{1, \ldots, n\}$. Then, there are $\binom{n-k}{k}$ different subsequences y of x of length k such that x_1 does not belong to y and:*

$$\forall_{i=2}^{n} \text{ either } x_{i-1} \text{ or } x_i \text{ does not belong to } y \tag{6}$$

Put differently, we are only interested in subsequences y which do not contain adjacent elements from the source sequence x and which do not contain x's first element. This corresponds to the NULL insertion step, where at most one NULL can be inserted after each word resulting form the fertility step.

Proof. We prove the above proposition by induction on n and k.

$n \geq 1, k = 1$: In this case, $\binom{n-1}{1} = n - 1$, which is correct because y contains single element which can be any x_i apart from x_1.

$n \geq 1, k > 1$: Let's consider the last elemement x_n of sequence x. We have two possibilities:

[1] We don't have to account for different permutations of output words aligned to NULL because, implicitely, IBM-3 assumes that these are generated in an ascending order. A similar assumption is adopted in IBM-4 and IBM-5 with respect to all input words, hence no need for the factor $\prod_{i=1}^{n} \phi_i!$ at all in those higher models.

1. x_n is a part of y. Then, we still need to account for subsequences of $(x_1, \ldots, x_{n-2})$ of length $k-1$.[2] From the induction hypothesis, the number of such subsequences is $\binom{n-2-(k-1)}{k-1} = \binom{n-k-1}{k-1}$.
2. x_n is not a part of y. Then, we account for the subsequences of $(x_1, \ldots, x_{n-1})$ of length k, whose number is (from the induction hypothesis) equal to $\binom{n-k-1}{k}$.

In total, this gives $\binom{n-k-1}{k-1} + \binom{n-k-1}{k}$, which (following the standard recursive calculation rule for binomial coefficients)[3] is equal to $\binom{n-k}{k}$.

□

3 Deficiency of IBM-3

Let's consider a simple case where $m = 2$, i.e., the output sentence consists of two words only. Below, all distortions possible in this case are represented, but only the first two are valid (represent permutations):

We are also given distortion probabilities, which must satisfy certain properties:[4]

- $P(1 \mid 1, 2, n) + P(2 \mid 1, 2, n) = 1$
- $P(1 \mid 2, 2, n) + P(2 \mid 1, 2, n) = 1$

Observation 1. *The total probability of all distortions in our example (including the invalid ones) is equal to* 1.

Proof. The total probability of all distortions is:

$$\begin{aligned} &P(1 \mid 1, 2, n) \cdot P(2 \mid 2, 2, n) + \\ &P(1 \mid 2, 2, n) \cdot P(2 \mid 2, 2, n) + \\ &P(1 \mid 1, 2, n) \cdot P(2 \mid 1, 2, n) + \\ &P(1 \mid 2, 2, n) \cdot P(2 \mid 1, 2, n) \end{aligned}$$

This is equal to:

$$\big(P(1 \mid 1, 2, n) + P(2 \mid 1, 2, n)\big) \times \big(P(1 \mid 2, 2, n) + P(2 \mid 2, 2, n)\big)$$

which, given that $P(1 \mid 1, 2, n) + P(2 \mid 1, 2, n) = 1$ and $P(1 \mid 2, 2, n) + P(2 \mid 2, 2, n) = 1$, is equal to 1 as well. □

The problem is that, in IBM-3, we sum over the *valid* distortions only, i.e., distortions which represent permutations. But, since the invalid distortions can get non-zero probabilities (e.g., $P(1 \mid 1, 2, n) \cdot P(2 \mid 1, 2, n)$ in the example above can be > 0), the total probability attributed to valid distortions only can be smaller than 1.

[2] Note that x_{i-1} cannot belong to y in this case because adjacent elements cannot be in y.
[3] We don't cite this recursive rule, but you can find it easilly e.g. on wikipedia.
[4] The input size n is not fixed, it depends on the fertility of words.