EM on IBM model 1 (brute force)

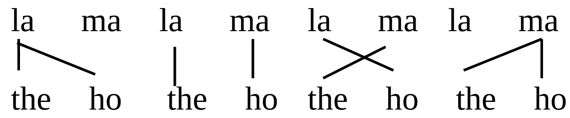
October 9, 2019

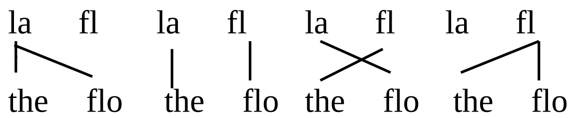
This is a worked example showing convergence of EM training with IBM model 1.

The pairs are

$\begin{vmatrix} \boldsymbol{s}^1 & \text{la maison} \\ \boldsymbol{o}^1 & \text{the house} \end{vmatrix} \quad \begin{vmatrix} \boldsymbol{s}^2 & \text{la fleur} \\ \boldsymbol{o}^2 & \text{the flower} \end{vmatrix}$

To apply the brute force EM algorithm, for each pair, each of its possible alignments has to be considered. Including the possibility of aligning positions in $\boldsymbol{o}$ with NULL, there are $3^2 = \mathbf{9}$ possibilities. To save a little in the pencil-and-paper calculations, we will consider a version which does *not* allow aligning positions in $\boldsymbol{o}$ with NULL. In this case, there are $2^2 = \mathbf{4}$ possibilities:





use $a^1_1 \dots a^1_4$ for the 4 possible alignments between $\boldsymbol{o}^1$ and $\boldsymbol{s}^1$

use $a^2_1 \dots a^2_4$ for the 4 possible alignments between $\boldsymbol{o}^2$ and $\boldsymbol{s}^2$

table shows translations probabilites, with $tr(o|s)$ shown at row o, col s, and they are all initialised to $\frac{1}{3}$

$tr(o\|s)$	la	ma	fl
the	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
ho	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
flo	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

To execute the brute force EM algorithm we need first for the pairs $\boldsymbol{o}^1, \boldsymbol{s}^1$ and $\boldsymbol{o}^2, \boldsymbol{s}^2$ to determine the conditional alignment probabilities so $p(a|\boldsymbol{o}^1, \boldsymbol{s}^1)$ and $p(a|\boldsymbol{o}^2, \boldsymbol{s}^2)$. The slides gave a derivation of the formula for $p(a|\boldsymbol{o}, \boldsymbol{s})$, it came out to be

$$p(a|\boldsymbol{o}, \boldsymbol{s}) = \frac{\prod_j [p(o_j|s_{a(j)})]}{\sum_{a'} \prod_j [p(o_j|s_{a'(j)})]} \tag{1}$$

and in the derivation $\frac{1}{(\ell_{\boldsymbol{s}}+1)^{\ell_{\boldsymbol{o}}}}$ terms cancelled. In the corresponding derivation disallowing alignments to NULL, there will instead be a cancellation of $\frac{1}{(\ell_{\boldsymbol{s}})^{\ell_{\boldsymbol{o}}}}$ terms, and exactly the same formula for the conditional alignment probability (1) will be derived.

As a name for the numerator term in (1) we will use $num(a)$

So to determine the $p(a|\boldsymbol{o}, \boldsymbol{s})$ values for each pair we need to

1. for each possible a determine $num(a)$ (ie. $\prod_j [p(o_j|s_{a(j)})]$)
2. sum these to give the denominator $\sum_a num(a)$ and then take ratios

Armed with these conditional probabilities can then compute expected counts of o, s combinations across the corpus, and from these recalculate $tr(o|s)$ probabilities.

ITERATION 1

considering the first pair, for each a_n^1 calculate $num(a_n^1)$:

$num(a_1^1)$	$num(a_2^1)$	$num(a_3^1)$	$num(a_4^1)$
$= \frac{1}{3}\frac{1}{3}$	*ditto*	*ditto*	*ditto*
$= \frac{1}{9}$			

sim. for each a_n^2 calculate $num(a_n^2)$. At this stage, these all work out as $\frac{1}{9}$.

from these to calculate the conditional probabilities $P(a_n^d|\boldsymbol{o}, \boldsymbol{s})$, need to sum the $num(a^n)$ by summing across the table and use it as denominator ie.

$$P(a_n^d|\boldsymbol{o}^d, \boldsymbol{s}^d) = \frac{num(a_n^d)}{\sum_{n'} num(a_{n'}^d)}$$

$P(a_1^1\|\boldsymbol{o}^1, \boldsymbol{s}^1)$	$P(a_2^1\|\boldsymbol{o}^1, \boldsymbol{s}^1)$	$P(a_3^1\|\boldsymbol{o}^1, \boldsymbol{s}^1)$	$P(a_4^1\|\boldsymbol{o}^1, \boldsymbol{s}^1)$
$= \frac{1/9}{4\times 1/9}$	*ditto*	*ditto*	*ditto*
$= \frac{1}{4}$			

$P(a_1^2\|\boldsymbol{o}^2, \boldsymbol{s}^2)$	$P(a_2^2\|\boldsymbol{o}^2, \boldsymbol{s}^2)$	$P(a_3^2\|\boldsymbol{o}^2, \boldsymbol{s}^2)$	$P(a_4^2\|\boldsymbol{o}^2, \boldsymbol{s}^2)$
$= \frac{1/9}{4\times 1/9}$	*ditto*	*ditto*	*ditto*
$= \frac{1}{4}$			

Notice these numbers make intuitive sense: with all $tr(o|s)$ set equal, all alignments should be equally probable, giving a value of $\frac{1}{4}$ for each.

Now for each possible vocabulary combination o, s combination we have to make a count by going through all the alignments and incrementing the count by how many times o is paired with s in the alignment and multiplying that by the above conditional alignment probabilities

For these short sentences the o, s count for any alignment is at most 1, and it will be handy for the calculations to note for each (o, s) the alignments where it occurs once[1]

	la	*ma*	*fl*
the	1:1 2	1:3 4	1:–
	2:1 2	2:–	2:3 4
ho	1:1 3	1:2 4	1:–
	2:–	2:–	2:–
flo	1:–	1:–	1:–
	2:1 3	2:–	2:2 4

based on this we get the following expected counts

[1]to read this table the (ho,la) entry has 1:1 3 / 2:– to indicate in first pair $(\boldsymbol{o}^1, \boldsymbol{s}^1)$, the (ho,la) pairing occurs in alignments a_1^1 and a_3^1, and the pairing never occurs in the alignments for the second pair

cnt	la	ma	fl
the	$4 \times \frac{1}{4}$	$2 \times \frac{1}{4}$	$2 \times \frac{1}{4}$
ho	$2 \times \frac{1}{4}$	$2 \times \frac{1}{4}$	0
flo	$2 \times \frac{1}{4}$	0	$2 \times \frac{1}{4}$

and for these counts get new $tr(o|s)$ by normalising by column sums

$tr(o\|s)$	la	ma	fl
the	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
ho	$\frac{1}{4}$	$\frac{1}{2}$	0
flo	$\frac{1}{4}$	0	$\frac{1}{2}$

ITERATION 2

using new $tr(o|s)$ value re-calculate for each a^1_n, $num(a^1_n)$, and for each a^2_n, $num(a^2_n)$:

$num(a^1_1)$	$num(a^1_2)$	$num(a^1_3)$	$num(a^1_4)$
$\frac{1}{2}\frac{1}{4} = \frac{1}{8}$	$\frac{1}{2}\frac{1}{2} = \frac{2}{8}$	$\frac{1}{2}\frac{1}{4} = \frac{1}{8}$	$\frac{1}{2}\frac{1}{2} = \frac{2}{8}$

$num(a^2_1)$	$num(a^2_2)$	$num(a^2_3)$	$num(a^2_4)$
$\frac{1}{2}\frac{1}{4} = \frac{1}{8}$	$\frac{1}{2}\frac{1}{2} = \frac{2}{8}$	$\frac{1}{2}\frac{1}{4} = \frac{1}{8}$	$\frac{1}{2}\frac{1}{2} = \frac{2}{8}$

then re-calculate the conditional probabilities $P(a|\boldsymbol{o}, \boldsymbol{s})$.

$P(a^1_1\|\boldsymbol{o}^1, \boldsymbol{s}^1)$	$P(a^1_2\|\boldsymbol{o}^1, \boldsymbol{s}^1)$	$P(a^1_3\|\boldsymbol{o}^1, \boldsymbol{s}^1)$	$P(a^1_4\|\boldsymbol{o}^1, \boldsymbol{s}^1)$
$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{3}$

$P(a^2_1\|\boldsymbol{o}^2, \boldsymbol{s}^2)$	$P(a^2_2\|\boldsymbol{o}^2, \boldsymbol{s}^2)$	$P(a^2_3\|\boldsymbol{o}^2, \boldsymbol{s}^2)$	$P(a^2_4\|\boldsymbol{o}^2, \boldsymbol{s}^2)$
$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{3}$

then re-calculate the expected counts of o, s combinations as shown in the table below (eg. the expected count for (the,la) is coming from a^1_1, a^1_2, a^2_1, a^2_2)

cnt	la	ma	fl
the	$\frac{1}{6} + \frac{2}{6} + \frac{1}{6} + \frac{2}{6} = \frac{6}{6}$	$\frac{1}{6} + \frac{2}{6} = \frac{3}{6}$	$\frac{1}{6} + \frac{2}{6} = \frac{3}{6}$
ho	$\frac{1}{6} + \frac{1}{6} = \frac{2}{6}$	$\frac{2}{6} + \frac{2}{6} = \frac{4}{6}$	0
flo	$\frac{1}{6} + \frac{1}{6} = \frac{2}{6}$	0	$\frac{2}{6} + \frac{2}{6} = \frac{4}{6}$

and for these counts get new $tr(o|s)$ by normalising by column sums

$tr(o\|s)$	la	ma	fl
the	$\frac{3}{5}$	$\frac{3}{7}$	$\frac{3}{7}$
ho	$\frac{1}{5}$	$\frac{4}{7}$	0
flo	$\frac{1}{5}$	0	$\frac{4}{7}$

ITERATION 3

using new $tr(o|s)$ value re-calculate for each a^1_n, $num(a^1_n)$ and each a^2_n, $num(a^2_n)$:

$num(a^1_1)$	$num(a^1_2)$	$num(a^1_3)$	$num(a^1_4)$
$\frac{3}{5}\frac{1}{5}$	$\frac{3}{5}\frac{4}{7}$	$\frac{3}{7}\frac{1}{5}$	$\frac{3}{7}\frac{4}{7}$

$num(a^2_1)$	$num(a^2_2)$	$num(a^2_3)$	$num(a^2_4)$
$\frac{3}{5}\frac{1}{5}$	$\frac{3}{5}\frac{4}{7}$	$\frac{3}{7}\frac{1}{5}$	$\frac{3}{7}\frac{4}{7}$

then re-calculate the conditional probabilities $P(a|\boldsymbol{o},\boldsymbol{s})$.

$P(a^1_1\|\boldsymbol{o}^1,\boldsymbol{s}^1)$	$P(a^1_2\|\boldsymbol{o}^1,\boldsymbol{s}^1)$	$P(a^1_3\|\boldsymbol{o}^1,\boldsymbol{s}^1)$	$P(a^1_4\|\boldsymbol{o}^1,\boldsymbol{s}^1)$
0.1512	0.4321	0.1080	0.3086

$P(a^2_1\|\boldsymbol{o}^2,\boldsymbol{s}^2)$	$P(a^2_2\|\boldsymbol{o}^2,\boldsymbol{s}^2)$	$P(a^2_3\|\boldsymbol{o}^2,\boldsymbol{s}^2)$	$P(a^2_4\|\boldsymbol{o}^2,\boldsymbol{s}^2)$
0.1512	0.4321	0.1080	0.3086

then re-calculate the expected counts of o, s combinations

cnt	la	ma	fl
the	0.1512+ 0.4321+ 0.1512+ 0.4321 = 1.167	0.1080+ 0.3086 = 0.4167	0.1080+ 0.3086 = 0.4167
ho	0.1512+ 0.1080 = 0.2593	0.4321+ 0.3086 = 0.7407	0
flo	0.1512+ 0.1080 = 0.2593	0	0.4321+ 0.3086 = 0.7407

and for these counts get new $tr(o|s)$ by normalising by column sums

$tr(o\|s)$	la	ma	fl
the	0.6923	0.36	0.36
ho	0.1538	0.64	0
flo	0.1538	0	0.64

Over the 3 iterations, $tr(the|la)$, $tr(ho|ma)$ and $tr(flo|fl)$ are steadily increasing.

If the calculations are carried on, after 10 iterations you have the following for the translation probabilities

$tr(o\|s)$	la	ma	fl
the	0.982	0.096	0.096
ho	0.009	0.904	0
flo	0.009	0	0.904

this is the history over the 10 iterations

		$o\|s$ at each iteration										
Obs	Src	0	1	2	3	4	5	6	7	8	9	10
the	la	0.33	0.5	0.6	0.69	0.77	0.84	0.89	0.93	0.95	0.97	0.98
house	la	0.33	0.25	0.2	0.15	0.11	0.081	0.056	0.037	0.024	0.015	0.009
flower	la	0.33	0.25	0.2	0.15	0.11	0.081	0.056	0.037	0.024	0.015	0.009
the	maison	0.33	0.5	0.43	0.36	0.3	0.24	0.2	0.16	0.14	0.11	0.096
house	maison	0.33	0.5	0.57	0.64	0.7	0.76	0.8	0.84	0.86	0.89	0.9
flower	maison	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
the	fleur	0.33	0.5	0.43	0.36	0.3	0.24	0.2	0.16	0.14	0.11	0.096
house	fleur	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
flower	fleur	0.33	0.5	0.57	0.64	0.7	0.76	0.8	0.84	0.86	0.89	0.9

In the end the $tr(o|s)$ table converges to:

$tr(o\|s)$	la	ma	fl
the	1	0	0
ho	0	1	0
flo	0	0	1