EM on IBM model 1 (brute force)

October 25, 2017

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

The pairs are

<table>
<thead>
<tr>
<th>$s^1$</th>
<th>la maison</th>
<th>$s^2$</th>
<th>la fleur</th>
</tr>
</thead>
<tbody>
<tr>
<td>$o^1$</td>
<td>the house</td>
<td>$o^2$</td>
<td>the flower</td>
</tr>
</tbody>
</table>

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 $o$ with NULL, there are $3^2 = 9$ possibilities. To save a little in the pencil-and-paper calculations, we will consider a version which does not allow aligning positions in $o$ with NULL. In this case, there are $2^2 = 4$ possibilities:

<table>
<thead>
<tr>
<th>la</th>
<th>ma</th>
<th>la</th>
<th>ma</th>
<th>la</th>
<th>ma</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>ho</td>
<td>the</td>
<td>ho</td>
<td>the</td>
<td>ho</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>la</th>
<th>fl</th>
<th>la</th>
<th>fl</th>
<th>la</th>
<th>fl</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>flo</td>
<td>the</td>
<td>flo</td>
<td>the</td>
<td>flo</td>
</tr>
</tbody>
</table>

use $a_1^1 \ldots a_4^1$ for the 4 possible alignments between $o^1$ and $s^1$
use $a_1^2 \ldots a_4^2$ for the 4 possible alignments between $o^2$ and $s^2$

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

$$tr(o|s) \quad \text{the} \quad \frac{1}{3} \quad \frac{1}{3} \quad \frac{1}{3}$$

$$\quad \text{ho} \quad \frac{1}{3} \quad \frac{1}{3} \quad \frac{1}{3}$$

$$\quad \text{flo} \quad \frac{1}{3} \quad \frac{1}{3} \quad \frac{1}{3}$$

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

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

(1)

and in the derivation $\frac{1}{(\ell_{o}+1)^{n}}$ terms cancelled. In the corresponding derivation disallowing alignments to NULL, there will instead be a cancellation of $\frac{1}{(\ell_{o})^{n}}$ 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 $p(a|o, s)$ 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 $a_1^n$, calculate $num(a_1^n)$:

\[
\begin{align*}
num(a_1^1) & = \frac{1}{3} \\
num(a_1^2) & = \text{ditto} \\
num(a_1^3) & = \text{ditto} \\
num(a_1^4) & = \text{ditto}
\end{align*}
\]

sim. for each $a_2^n$ calculate $num(a_2^n)$. At this stage, these all work out as $\frac{1}{9}$.

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

\[
P(a_d^n|o^d, s^d) = \frac{num(a_d^n)}{\sum_{a} num(a^n)}
\]

\[
\begin{align*}
P(a_1^1|o^1, s^1) & = \frac{1}{9} \\
P(a_1^2|o^1, s^1) & = \text{ditto} \\
P(a_1^3|o^1, s^1) & = \text{ditto} \\
P(a_1^4|o^1, s^1) & = \text{ditto}
\end{align*}
\]

\[
\begin{align*}
P(a_2^1|o^2, s^2) & = \frac{1}{9} \\
P(a_2^2|o^2, s^2) & = \text{ditto} \\
P(a_2^3|o^2, s^2) & = \text{ditto} \\
P(a_2^4|o^2, s^2) & = \text{ditto}
\end{align*}
\]

Notice these numbers make intuitive sense: with all $tr(o|s)$ set equal, all alignments should be equally probable, giving a value of $\frac{1}{9}$ 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:\footnote{to read this table the (ho,la) entry has 1:1 3 to indicate in first pair $(o^1, s^1)$, the (ho,la) pairing occurs in alignments $a_1^1$ and $a_1^3$, and the pairing never occurs in the alignments for the second pair}

<table>
<thead>
<tr>
<th></th>
<th>la</th>
<th>ma</th>
<th>fl</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>1:1 2</td>
<td>1:3 4</td>
<td>1:--</td>
</tr>
<tr>
<td></td>
<td>2:1 2</td>
<td>2:--</td>
<td>2:3 4</td>
</tr>
<tr>
<td>ho</td>
<td>1:1 3</td>
<td>1:2 4</td>
<td>1:--</td>
</tr>
<tr>
<td></td>
<td>2:--</td>
<td>2:--</td>
<td>2:--</td>
</tr>
<tr>
<td>flo</td>
<td>1:--</td>
<td>1:--</td>
<td>1:--</td>
</tr>
<tr>
<td></td>
<td>2:1 3</td>
<td>2:--</td>
<td>2:2 4</td>
</tr>
</tbody>
</table>

based on this we get the following expected counts
and for these counts get new $\text{tr}(a|s)$ by normalising by column sums

\[
\begin{array}{c|ccc}
\text{tr}(a|s) & \text{la} & \text{ma} & \text{fl} \\
\hline
\text{the} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} \\
\text{ho} & \frac{1}{4} & \frac{1}{4} & 0 \\
\text{flo} & \frac{1}{4} & 0 & \frac{1}{4} \\
\end{array}
\]

ITERATION 2

using new $\text{tr}(a|s)$ value re-calculate for each $a_1^n$, $\text{num}(a_1^n)$, and for each $a_2^n$, $\text{num}(a_2^n)$:

\[
\begin{array}{cccc}
\text{num}(a_1^1) & \text{num}(a_1^2) & \text{num}(a_1^3) & \text{num}(a_1^4) \\
\frac{\frac{1}{4}}{4} & \frac{\frac{1}{4}}{4} & \frac{\frac{1}{4}}{4} & \frac{\frac{1}{4}}{4} \\
\frac{\frac{1}{4}}{8} & \frac{\frac{1}{4}}{8} & \frac{\frac{1}{4}}{8} & \frac{\frac{1}{4}}{8} \\
\text{num}(a_2^1) & \text{num}(a_2^2) & \text{num}(a_2^3) & \text{num}(a_2^4) \\
\frac{\frac{1}{4}}{8} & \frac{\frac{1}{4}}{8} & \frac{\frac{1}{4}}{8} & \frac{\frac{1}{4}}{8} \\
\end{array}
\]

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

\[
\begin{array}{cccc}
P(a_1^1|o^1, s^1) & P(a_2^1|o^1, s^1) & P(a_3^1|o^1, s^1) & P(a_4^1|o^1, s^1) \\
\frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\
\end{array}
\]

\[
\begin{array}{cccc}
P(a_1^2|o^2, s^2) & P(a_2^2|o^2, s^2) & P(a_3^2|o^2, s^2) & P(a_4^2|o^2, s^2) \\
\frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\
\end{array}
\]

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 $a_1^1$, $a_2^1$, $a_3^1$, $a_4^1$)

\[
\begin{array}{c|ccc}
\text{cnt} & \text{la} & \text{ma} & \text{fl} \\
\hline
\text{the} & \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} & \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} & \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} \\
& = \frac{4}{6} & = \frac{4}{6} & = \frac{4}{6} \\
\text{ho} & \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} & \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} & 0 \\
& = \frac{4}{6} & = \frac{4}{6} & 0 \\
\text{flo} & \frac{1}{6} + \frac{1}{6} + \frac{1}{6} & \frac{1}{6} + \frac{1}{6} + \frac{1}{6} & \frac{1}{6} + \frac{1}{6} + \frac{1}{6} \\
& = \frac{4}{6} & = \frac{4}{6} & = \frac{4}{6} \\
\end{array}
\]

and for these counts get new $\text{tr}(a|s)$ by normalising by column sums
ITERATION 3

using new \( tr(o|s) \) value re-calculate for each \( a_n^1, \text{num}(a_n^1) \) and each \( a_n^2, \text{num}(a_n^2) \):

\[
\begin{array}{c|ccc}
\text{num}(a_1^1) & \text{num}(a_2^1) & \text{num}(a_3^1) & \text{num}(a_4^1) \\
3\frac{1}{5} & 3\frac{4}{5} & 3\frac{1}{5} & 3\frac{4}{5} \\
\text{num}(a_2^2) & \text{num}(a_3^2) & \text{num}(a_4^2) & \text{num}(a_5^2) \\
3\frac{1}{5} & 3\frac{4}{5} & 3\frac{1}{5} & 3\frac{4}{5}
\end{array}
\]

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

\[
\begin{array}{cccc}
P(a_1^1|o^1, s^1) & P(a_2^1|o^1, s^1) & P(a_3^1|o^1, s^1) & P(a_4^1|o^1, s^1) \\
0.1512 & 0.4321 & 0.1080 & 0.3086 \\
P(a_3^2|o^2, s^2) & P(a_4^2|o^2, s^2) & P(a_5^2|o^2, s^2) & P(a_6^2|o^2, s^2) \\
0.1512 & 0.4321 & 0.1080 & 0.3086
\end{array}
\]

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

\[
\begin{array}{c|ccc}
\text{cnt} & la & ma & fl \\
the & 0.1512+ & 0.1080+ & 0.1080+ \\
 & 0.4321+ & 0.3086 & 0.3086 \\
 & 0.1512+ & = & = \\
 & 0.4321 & 0.4167 & 0.4167 \\
 & = & = & 1.167
\end{array}
\]

\[
\begin{array}{c|ccc}
ho & 0.1512+ & 0.4321+ & 0 \\
 & 0.1080 & 0.3086 & = \\
 & = & = & 0.2593 & 0.7407
\end{array}
\]

\[
\begin{array}{c|ccc}
flo & 0.1512+ & 0 & 0.4321+ \\
 & 0.1080 & 0.3086 & = \\
 & = & = & 0.2593 & 0.7407
\end{array}
\]

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

\[
\begin{array}{c|ccc}
\text{tr}(o|s) & la & ma & fl \\
the & 0.6923 & 0.36 & 0.36 \\
ho & 0.1538 & 0.64 & 0 \\
flo & 0.1538 & 0 & 0.64
\end{array}
\]

Over the 3 iterations, \( tr(\text{the}|la), tr(\text{ho}|ma) \) and \( tr(\text{flo}|fl) \) are steadily increasing.

If the calculations are carried on, after 10 iterations you have the following for the translation probabilities
\[ tr(o|s) \quad la \quad ma \quad fl \]

<table>
<thead>
<tr>
<th>Obs</th>
<th>Src</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>la</td>
<td>0.33</td>
<td>0.5</td>
<td>0.6</td>
<td>0.69</td>
<td>0.77</td>
<td>0.84</td>
<td>0.89</td>
<td>0.93</td>
<td>0.95</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>house</td>
<td>la</td>
<td>0.33</td>
<td>0.25</td>
<td>0.2</td>
<td>0.15</td>
<td>0.11</td>
<td>0.081</td>
<td>0.056</td>
<td>0.037</td>
<td>0.024</td>
<td>0.015</td>
<td>0.009</td>
</tr>
<tr>
<td>flower</td>
<td>la</td>
<td>0.33</td>
<td>0.25</td>
<td>0.2</td>
<td>0.15</td>
<td>0.11</td>
<td>0.081</td>
<td>0.056</td>
<td>0.037</td>
<td>0.024</td>
<td>0.015</td>
<td>0.009</td>
</tr>
<tr>
<td>the</td>
<td>maison</td>
<td>0.33</td>
<td>0.5</td>
<td>0.43</td>
<td>0.36</td>
<td>0.3</td>
<td>0.24</td>
<td>0.2</td>
<td>0.16</td>
<td>0.14</td>
<td>0.11</td>
<td>0.096</td>
</tr>
<tr>
<td>house</td>
<td>maison</td>
<td>0.33</td>
<td>0.5</td>
<td>0.57</td>
<td>0.64</td>
<td>0.7</td>
<td>0.76</td>
<td>0.8</td>
<td>0.84</td>
<td>0.86</td>
<td>0.89</td>
<td>0.9</td>
</tr>
<tr>
<td>flower</td>
<td>maison</td>
<td>0.33</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>the</td>
<td>fleur</td>
<td>0.33</td>
<td>0.5</td>
<td>0.43</td>
<td>0.36</td>
<td>0.3</td>
<td>0.24</td>
<td>0.2</td>
<td>0.16</td>
<td>0.14</td>
<td>0.11</td>
<td>0.096</td>
</tr>
<tr>
<td>house</td>
<td>fleur</td>
<td>0.33</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>flower</td>
<td>fleur</td>
<td>0.33</td>
<td>0.5</td>
<td>0.57</td>
<td>0.64</td>
<td>0.7</td>
<td>0.76</td>
<td>0.8</td>
<td>0.84</td>
<td>0.86</td>
<td>0.89</td>
<td>0.9</td>
</tr>
</tbody>
</table>

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

\[ tr(o|s) \quad la \quad ma \quad fl \]

<table>
<thead>
<tr>
<th>Obs</th>
<th>Src</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>la</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ho</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>flo</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>