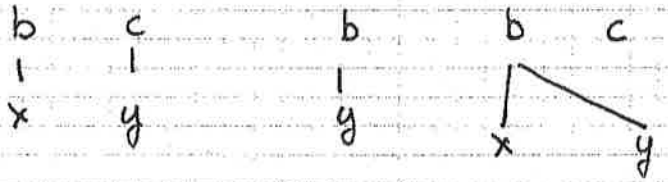
EM Training of IBM Model 1

In model 1 and Model 2 we have assumed that we have lexical translation probabilities. But dictionaries do not provide this information and we do not have corpora aligned word by word to estimate the lexical translation probabilities based on corpus counts.

So what do we do? We take our sentence-by-sentence aligned corpora and <u>pretend</u> we have corpus counts.

Let's imagine the following situation: we have a corpus of sentences aligned word by word. Then we can collect the lexical translation probabilities by just collecting the relative frequencies of translation of the words.

For example, assume the following corpus



Then, we could collect lexical translation probabilities as follows

$$t(y|b) = \frac{\text{counts}(y,b)}{\text{counts}(b)\,b)} = \frac{2}{4}$$

$$t(x|b) = \frac{\text{counts}(x,b)}{\text{counts}(\cdot,b)} = \frac{2}{4}$$

$$t(y|c) = \frac{\text{counts}(y,c)}{\text{counts}(c)} = \frac{1}{1}$$

$$t(x|c) = \frac{\text{counts}(x,c)}{\text{counts}(c)} = 0$$

Notice that these are proper probability distributions.

But let's suppose that we know that alignement ① is twice as likely as alignement ② below

Then we don't want to collect a whole unit of count for each of its occurrences. Notice that these are alignments between the same two strings, so the fact that they align differently indicates that there is uncertainty on how the two strings should be aligned.

Then, we don't want to count each word-by-word alignment as being certain, but we want to consider it a fractional count, weighed by the probability of the alignment ($p(①) = \frac{2}{3}$, $P(②) = \frac{1}{3}$, for example).

So $P(x|b) = C(x|b) \cdot P(①) = 1 \cdot \frac{2}{3}$, $P(x|b) = C(x|b) \cdot P(②) = 1 \cdot \frac{1}{3}$

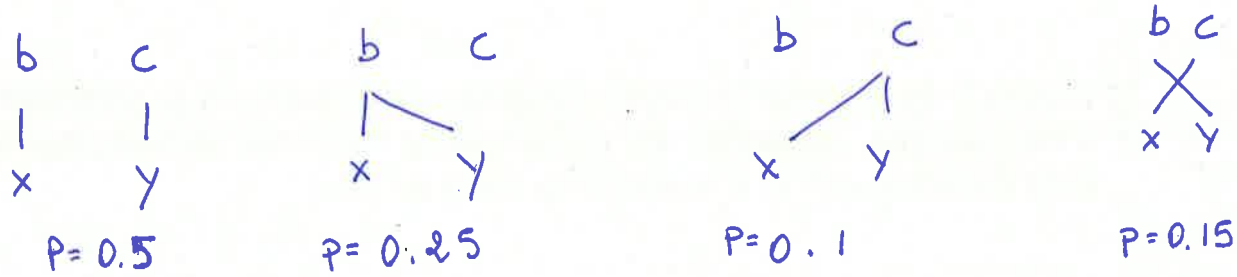
In both cases, we discount the full count by a probability coefficient. It's as if we distributed the single count of $C(x|b)$ across two alignments, mutually exclusive.

This is called collecting FRACTIONAL COUNTS

./.

We collect fractional counts to estimate the lexical translation probabilities as follows.

Let's assume the probabilities of alignment below (notice that this is a true probability distribution as it sums to 1).



FRACTIONAL COUNT

$$t(y|b) = \frac{\text{counts}(y, b)}{\text{counts}(b)} = \frac{(1 \cdot 0.25) + (1 \cdot 0.15)}{(1 \cdot 0.5) + (1 \cdot 0.25) + (1 \cdot 0.1) + (1 \cdot 0.15)}$$

$$= \frac{0.4}{1} = 0.4$$

compose to

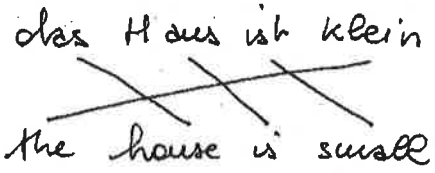
FULL COUNT

$$t(y|b) = \frac{2}{4} = 0.5$$

But where do we get the alignment probabilities?

We can write $P(a \mid e, f)$ for the probability of the alignment of a given sentence pair. How do we calculate this quantity? Consider the two following alignments

das Haus ist klein
| | | |
the house is small



which of the two alignments is more likely? You probably chose the first one because you think that

~~The pro~~ $P(\text{the} \mid \text{das}) > P(\text{the} \mid \text{klein})$

and $P(\text{house} \mid \text{Haus}) > P(\text{house} \mid \text{das})$

and $P(\text{is} \mid \text{ist}) > P(\text{is} \mid \text{Haus})$

and finally $P(\text{small} \mid \text{klein}) > P(\text{small} \mid \text{ist})$

That is: if we had word translation probabilities then we could estimate the probabilities of different alignments

An alignment has high probability if it connects words that are likely translations of each other, it has low probabilities if it connects words with very low t values.

We can show this mathematically as follows.

$$P(a \mid e, f) = \frac{P(a, e, f)}{P(e, f)} = \frac{P(a, f \mid e)\, P(e)}{P(f \mid e)\, P(e)} = \frac{P(a, f \mid e)}{P(f \mid e)}$$

The numerator of this formula is the probability we set out to calculate in Model 1 or Model 2, so we know how to calculate it, in theory, based on lexical translation probabilities and a uniform alignment (in Model 1).

$$P(f \mid e) = \sum_a P(a, f \mid e)$$ which is the same formula as above

At this point we know how to calculate alignment probabilities based on lexical translation probabilities and how to calculate lexical translation probabilities based on fractional count, namely on counts weighted by the probability of the alignment.

How do we bootstrap this process? We use EM.

- We start by assuming that lexical translation probabilities are the same for all word pairs. So $t(f|e) = \frac{1}{N}$, N the number of words that e translates into.
- Now we can ~~align~~ compute alignment probabilities for each pair of sentences, because we know that, in Model 1,

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

So we know how to calculate alignment probabilities as a normalized count of a weighted product of translation probabilities.

- Now we have alignment probabilities. So we can redo the calculation of the fractional counts for the translation probabilities weighed by the alignments. Since the alignments are based on the corpus this will give us better translation probabilities.
- We now use the new translation probabilities to compute alignment probabilities.
- We continue alternating between estimating new translation probabilities and new alignments until the probabilities converge, that is they reach a stable state.