IBM Model 1: clarified pseudocode

1 Pseudocode

The lecture notes by Collins (2011) give a thorough introduction to the well-known IBM models 1 and 2, which were originally presented by Brown et al. (1993). However, the pseudocode for IBM model 1 (Figure 4) in the lecture notes by Collins is a bit confused and there are some parts that can be removed. Here is a new description that includes only the things that you need to implement IBM model 1. We use the same notation as Collins.

Input: A training corpus $(f^{(k)}, e^{(k)})$ for $k = 1..n$,
where $f^{(k)} = f_1^{(k)}, \ldots, f_{m_k}^{(k)}$ and $e^{(k)} = e_1^{(k)}, \ldots, e_{l_k}^{(k)}$
Initialization: Initialize $t(f|e)$ parameters (e.g. to random values).
Algorithm:

for $t = 1, \ldots, T$: `# For each EM iteration`
  set all counts $c(f, e)$ and $c(e)$ to 0
  for $k = 1, \ldots, n$: `# For each sentence pair`
    for $i = 1, \ldots, m_k$: `# For each French word`
      for $j = 0, \ldots, l_k$: `# For each English word including the NULL word`

$$\delta(k,i,j) = \frac{t(f_i^{(k)}|e_j^{(k)})}{\sum_{j'=0}^{l_k} t(f_i^{(k)}|e_{j'}^{(k)})}$$ `# Compute alignment prob.`

$$c(e_j^{(k)}, f_i^{(k)}) = c(e_j^{(k)}, f_i^{(k)}) + \delta(k,i,j)$$ `# Update pseudocount`

$$c(e_j^{(k)}) = c(e_j^{(k)}) + \delta(k,i,j)$$ `# Update pseudocount`

  set $t(f|e) = \frac{c(e,f)}{c(e)}$ `# Reestimate probabilities`

Output: parameters $t(f|e)$

2 Toy example walkthrough

To make the pseudocode presented above a bit more concrete, let's compute the alignment probabilities for a toy example. For instance, let's assume that sentence pair k in our English–French training data is *the black cat / le chat noir*. For each French word $f_i^{(k)}$, we want to compute a probability $\delta(k,i,j)$ that this French word is aligned with English word $e_j^{(k)}$. This alignment probability will then be used to update the pseudocounts.

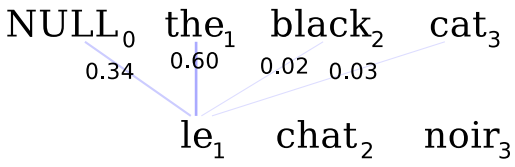
We'll start with the French word *le* at position 1 and compute the alignment probabilities for each English word including the dummy null word. To do this, we use our current estimates of the translation probabilities. For instance, let's hypothetically assume that we have

$$t(le|\text{NULL}) = 0.2 \quad t(le|the) = 0.35$$
$$t(le|black) = 0.01 \quad t(le|cat) = 0.02$$

What is the probability that *le* is aligned with *the*, as we should expect? We get

$$\delta(k,1,1) = \frac{0.35}{0.2 + 0.35 + 0.01 + 0.02} = 0.60$$

We compute all alignment probabilities for *le* and we get this result:



Now, we can update the pseudocounts $c(e,f)$: we increase $c(NULL,le)$ by 0.34, $c(the,le)$ by 0.60, $c(black,le)$ by 0.02, and $c(cat,le)$ by 0.03.

We continue and consider the alignment probabilities for *chat*. We might have the following hypothetical translation probabilites:

$$t(chat|\text{NULL}) = 0.005 \quad t(chat|the) = 0.01$$
$$t(chat|black) = 0.05 \quad t(chat|cat) = 0.7$$

This gives us the following alignment probabilities for *chat* in this sentence. As expected, *cat* is the one that seems most likely to be aligned with *chat*.

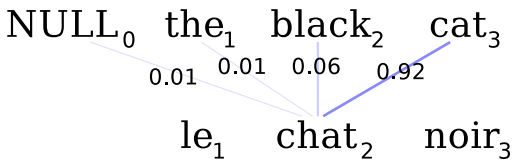


Figure 1: Finding the alignment probabilities for the French word *chat*.

We then do the same thing for *noir*, and probably *black* will be the one that has the highest alignment probability.

References

[Brown et al.1993] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

[Collins2011] Michael Collins. 2011. Statistical Machine Translation: IBM Models 1 and 2. Lecture notes, Columbia University.