Chapter 16

Word Alignment

16.1 Problem

A *parallel text* is a corpus of text that expresses the same meaning in two (or more) different languages. Usually we assume that a parallel text is already *sentence-aligned*, that is, it consists of *sentence pairs*, each of which expresses the same meaning in two languages. Conventionally, following Brown et al. (1993), the two languages are referred to as English and French even when other languages are possible. Our example uses English and Spanish.

Here is an example parallel text:

1. garcia and associates
   garcia y asociados

2. his associates are not strong
   sus asociados no son fuertes

The *word alignment* problem is to figure out which French (Spanish) words correspond to which English words. This would be the correct word alignment for our example:

1. garcia and associates
   | | |
   garcia y asociados

2. his associates are not strong
   | | \ | |
   sus asociados no son fuertes

More formally: let

- $\mathbf{f} = f_1 \cdots f_m$ range over French sentences
- $\mathbf{e} = e_1 \cdots e_\ell$ range over English sentences

- $\mathbf{a} = (a_1, \ldots, a_m)$ range over possible many-to-one alignments, where $a_j = i$ means that French word j is aligned to English word i, and $a_j = \text{NULL}$ means that French word j is unaligned. Thus the alignment for sentence (2) above is $(1, 2, 4, 3, 5)$.

We are given a sequence of $(\mathbf{f}, \mathbf{e})$ pairs. We are going to define a model of $P(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$ and our job is to estimate the parameters of the model to maximize the likelihood $P(\mathbf{f} \mid \mathbf{e})$.

16.2 Model 1

IBM Model 1 (Brown et al., 1993) is the first in a series of five seminal models for statistical word alignment. The basic generative story goes like this:

1. Choose m with uniform probability $\epsilon = \frac{1}{M}$, where M is the maximum length of any French sentence in the corpus.
2. Generate an alignment $a_1, \ldots, a_m$, again with uniform probability.
3. Generate French words $f_1, \ldots, f_m$, each with probability $t(f_j \mid e_{a_j})$ or $t(f_j \mid \text{NULL})$.

The model is:

$$P(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = P(m) \prod_{j=1}^{m} P(a_j) P(f_j \mid a_j) \tag{16.1}$$

$$P(m) = \frac{1}{M} \tag{16.2}$$

$$P(a_j) = \frac{1}{\ell + 1} \tag{16.3}$$

$$P(f_j \mid a_j) = t(f_j \mid e_{a_j}) \tag{16.4}$$

or, putting all that together:

$$P(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = \frac{1}{M} \prod_{j=1}^{m} \left(\frac{1}{\ell + 1} t\left(f_j \mid e_{a_j}\right) \right) \tag{16.5}$$

where M is the maximum length of any French sentence (say, 100). In practice, it doesn't actually matter what number you choose.

The parameters of the model are the word-translation probabilities $t(f \mid e)$, which can be estimated using Expectation-Maximization.

1. Initialize $t(\cdot \mid e)$ to uniform: $t(f \mid e) = \frac{1}{|V_f|}$, where V_f is the French vocabulary and e is any English word or NULL.
2. E-step: Reset all counts $c(f, e)$ to zero. For each sentence pair, calculate the expected number of times that word e is translated as f. Hard EM doesn't work very well here, but true EM is very easy to implement. For each i, j, the generation of f_j from e_i "competes" with the

generation of f_j from the other English words (or NULL). So we update $c(f, e)$ as follows. For each i, j:

$$c(f_j, e_i) \leftarrow c(f_j, e_i) + \frac{t(f_j \mid e_i)}{\sum_{i'} t(f_j \mid e_{i'}) + t(f_j \mid \text{NULL})}$$
$$c(f_j, \text{NULL}) \leftarrow c(f_j, \text{NULL}) + \frac{t(f_j \mid \text{NULL})}{\sum_{i'} t(f_j \mid e_{i'}) + t(f_j \mid \text{NULL})}$$

3. M-step: let $t(f \mid e) \leftarrow \frac{c(f,e)}{\sum_f c(f,e)}$, where e is any English word or NULL.

4. Go to 2.

Interestingly, for this particular model, (true) EM is guaranteed to converge to a global maximum (although the global maximum is not unique).

Let's see how this works on our toy example. The initial model is uniform:

	f						
e	asociados	fuertes	garcia	no	son	sus	y
NULL	1/7	1/7	1/7	1/7	1/7	1/7	1/7
and	1/7	1/7	1/7	1/7	1/7	1/7	1/7
are	1/7	1/7	1/7	1/7	1/7	1/7	1/7
associates	1/7	1/7	1/7	1/7	1/7	1/7	1/7
garcia	1/7	1/7	1/7	1/7	1/7	1/7	1/7
his	1/7	1/7	1/7	1/7	1/7	1/7	1/7
not	1/7	1/7	1/7	1/7	1/7	1/7	1/7
strong	1/7	1/7	1/7	1/7	1/7	1/7	1/7

(Note that each row sums to one.)

E step: We want to know the *expected* (fractional) number of times that each French word is translated from each English word. Consider sentence (1). First, for each English position i and French position j, compute the total probability of all alignments that translate f_j from e_i, that is, $P(a_j = i, f_j \mid \mathbf{e})$. Because the choice of a_j and f_j is independent of all the other alignments and French words, this works out to be $\frac{1}{M} \frac{1}{\ell+1} t(f_j \mid e_i)$. The first two terms make $\frac{1}{400}$, and the $t(f_j \mid e_i)$ term is:

	f_j		
e_i	garcia	y	asociados
NULL	1/7	1/7	1/7
garcia	1/7	1/7	1/7
and	1/7	1/7	1/7
associates	1/7	1/7	1/7

But we want the fractional count of times that f_j is translated from e_i (that is, $P(a_j = i \mid \mathbf{f}, \mathbf{e})$. We get this by renormalizing each *column*:

	f_j		
e_i	garcia	y	asociados
NULL	1/4	1/4	1/4
garcia	1/4	1/4	1/4
and	1/4	1/4	1/4
associates	1/4	1/4	1/4

This is because each French word occurs exactly once, and in each "possible world" is aligned to exactly one English word (or NULL). So for each French word, the fractional counts of which English word it's translated from should sum to one.

Similarly for sentence (2):

	f_j				
e_i	sus	asociados	no	son	fuertes
NULL	1/6	1/6	1/6	1/6	1/6
his	1/6	1/6	1/6	1/6	1/6
associates	1/6	1/6	1/6	1/6	1/6
are	1/6	1/6	1/6	1/6	1/6
not	1/6	1/6	1/6	1/6	1/6
strong	1/6	1/6	1/6	1/6	1/6

To complete the E step, we count up how many times each French word was translated from each English word:

	f						
e	asociados	fuertes	garcia	no	son	sus	y
NULL	5/12	1/6	1/4	1/6	1/6	1/6	1/4
and	1/4	0	1/4	0	0	0	1/4
are	1/6	1/6	0	1/6	1/6	1/6	0
associates	5/12	1/6	1/4	1/6	1/6	1/6	1/4
garcia	1/4	0	1/4	0	0	0	1/6
his	1/6	1/6	0	1/6	1/6	1/6	0
not	1/6	1/6	0	1/6	1/6	1/6	0
strong	1/6	1/6	0	1/6	1/6	1/6	0

In the M step, we renormalize for each e (that is, for each row) to obtain probabilities $t(f \mid e)$:

	f						
e	asociados	fuertes	garcia	no	son	sus	y
NULL	5/19	2/19	3/19	2/19	2/19	2/19	3/19
and	1/3	0	1/3	0	0	0	1/3
are	1/5	1/5	0	1/5	1/5	1/5	0
associates	5/19	2/19	3/19	2/19	2/19	2/19	3/19
garcia	1/3	0	1/3	0	0	0	1/3
his	1/5	1/5	0	1/5	1/5	1/5	0
not	1/5	1/5	0	1/5	1/5	1/5	0
strong	1/5	1/5	0	1/5	1/5	1/5	0

After one iteration of EM, we can start to see what the model is learning. It correctly learns that *associates* most likely translates to *asociados*. It knows that *and* should translate to one of *garcia*, *y*, or *asociados*, but can't decide which; at the next iteration, it will start to learn that *and* does not translate to *asociados*, but it will never learn to distinguish *garcia* and *y*, because the model is too weak (it doesn't know anything about word order) and/or there isn't enough data (to observe sentences where *garcia* occurs without *y* or vice versa).

References

Brown, Peter F. et al. (1993). "The Mathematics of Statistical Machine Translation: Parameter Estimation". In: *Computational Linguistics* 19, pp. 263–311.