**CMPT 825: Natural Language Processing** **Spring 2006**

Lecture 5 — Jan 25, 2006

*Lecturer: Anoop Sarkar* *Scribe: Maxim Roy*

## 5.1 Required Reading

- Read section 31-33 of the workbook “A Statistical MT tutorial Workbook” written by Kevin Knight.
- Read the pages 23-34 of the paper “The Mathematics of Statistical Machine Translation: Parameter Estimation” written by Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra and Robert L. Mercer.

## 5.2 Word-alignments

Good alignment doesn’t necessarily guaranty good translation. Producing good alignments depend on aligned sentence pairs. We can have good alignment on training data and eventually end up with bad translation. There are many papers that just focus on word-alignments not word-translation. People think that there is a hope that better alignment might lead to better translations.

## 5.3 Model 1 Training

In last lecture we discussed on how to compute the probability of alignment between sentence pair e(English sentence) and f(French sentence) which is:

$$P(a|e,f) = \frac{P(a,f|e)}{P(f|e)} = \frac{P(a,f|e)}{\sum_{\mathrm{a}} P(a,f|e)} \tag{5.1}$$

We can compute the numerator of the above equation using Model 1 or Model 3. Model 1 uses only one parameter $t$. Model 3 can also be used to compute the numerator which uses 4 parameters instead of just $t$, which are $n,d,t,p$ where $n$ denotes how many words in French are generated by each word in English, $d$ is distortion parameter which denotes the position of a French word in the French translation generated given the position of the English word, $t$ denotes which French word to generate depending on the English word and $p$ denotes how to generate spurious French words. Using Model 1 we can write the denominator of *equation 5.1* as

$$\sum_{\mathrm{a}} P(a, f|e) = \sum_{\mathrm{a}} \prod_{\mathrm{j=1}}^{\mathrm{m}} t(fj|eaj) \tag{5.2}$$

we can factor that as

$$P(a|e, f) = \prod_{\mathrm{j=1}}^{\mathrm{m}} \sum_{\mathrm{i=0}}^{l} t(fj|ei) \tag{5.3}$$

Here for *equation 5.2* we need $m{*}(l{+}1)^m$ arithmetic operation where as for *equation 5.3* we need $(l+1) * m$ arithmetic operations which is much more efficient way to compute the denominator. So we can efficiently compute $P(f|e)$ for Model 1.

One interesting feature of Model 1 is that it has a special form that ensures that EM training is guaranteed to converge to a global maximum. The *figure 5.1* represents it.

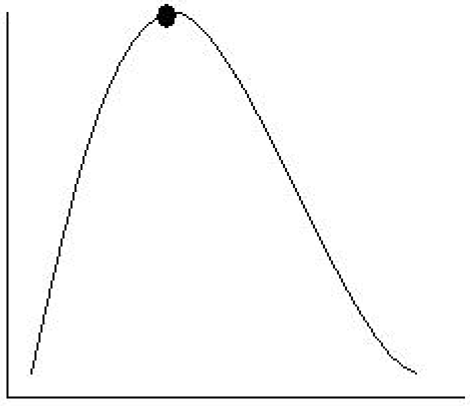

**Figure 5.1.** EM converges to global maximum in Model 1

### 5.3.1 Generative Story of Model 1

The generative story for Model 1 is that suppose we have a French sentence $f_1, f_2, ......f_m$ and English sentence $e_0, e_1, e_2, .....e_l$ and for any French word $f_j$ we can pick any of the $e_0, e_1, e_2, ...e_l$. So there is a chance that all $f_1, f_2, f_3, ......f_m$ might be generated by a single $e_i$ so $t(f_1|e_i), t(f_2|e_i), ....(f_m|e_i)$. The below figure also illustrates the same thing that every $f_j$ can be replaced by the same $e_i$ .

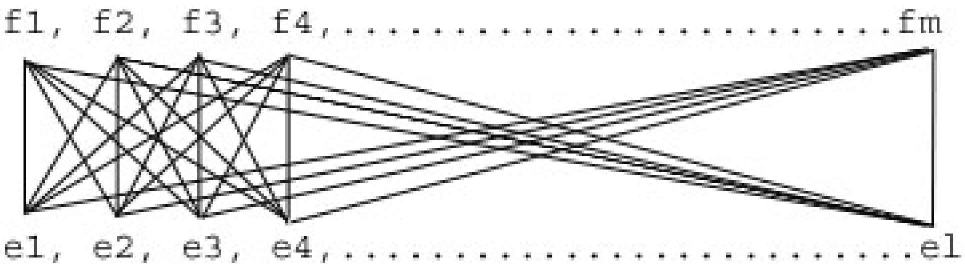
f1, f2, f3, f4,..........................fm
e1, e2, e3, e4,..........................el

**Figure 5.2.** generative story

As one English word has the possibility to generate a whole French sentence so it is not a good model for translation. But in Model 3 there is no possibility of generation of whole French sentence from one single English word.

### 5.3.2 Best Alignment for Model 1

One of the advantages of Model 1 is that it can easily compute the best alignment for a pair of sentences. Below we show the algorithm for computing the best alignment for each French word.

**Algorithm 5.1.**

```
$for\ 1 \leq j \leq m$
    $a_j = \underset{i}{argmax}\ t(f_j|e_i)$
```

The best alignment can be computed in a quadratic number of operations $(l + 1 * m)$.

## 5.4 Model 3 Training

In Model 3 we can't flip product and sum like in Model 1 as in the *equation 5.2*. So in Model 3 first we do Model 1 training then we take best values of t from Model 1 and do Model 3 training.

initially uniform t

↓

Model 1 iteration

↓↑

new t

↓

t values, uniform n ,d ,p

↓

Model 3 training ←

↓↗

(pick only best Model 1 alignment plus some more)

So instead of computing all possible alignments in Model 3 we will take the Model 1 best alignment and compute some more alignments. Now lets look at a set $A(f|e)$ which is set of all possible alignments and it's size is $(l+1)^m$, s is a subset of $A(f|e)$ : $s \in A(f|e)$. And lets say we have two alignments a and a' then,

⇒ a and a' differ by a move, if for exactly one j, $a_j \neq a'_j$

⇒ a and a' differ by a swap if $a_j = a'_j$ except at $j_1$ and $j_2$ where $a_{j_1} = a'_{j_2}$ and $a_{j_2} = a'_{j_1}$

⇒ a and a' are neighbor if they differ by a move or swap

Now set of neighbors of **a** is N(a) and let b(a) be the neighbor of **a**. If $Pr(a|f,e)$ is probability of Model 1 best alignment then $Pr(b(a)|f,e)$ is probability of neighbor of a , $Pr(b(b(a))|f,e)$ is probability of two neighbor away from a and $Pr(b^{\infty}(a)|f,e)$ is probability of infinity far away from a. Now let's say if a' is a neighbor of a which is obtained by a move of j from i to i' where $i \neq 0, i' \neq 0$, then the probability of move is:

$$Pr(a'|e,f) = Pr(a|e,f)\frac{(\phi_{i'}+1)}{\phi_i}\frac{n(\phi_{i'}+1|e_{i'})}{n(\phi_{i'}|e_{i'})}\frac{n(\phi_i-1|e_i)}{n(\phi_i|e_i)}\frac{t(f_j|e_{i'})}{t(fj|e_i)}\frac{d(j|i',m,l)}{d(j|i,m,l)}. \tag{5.4}$$

In this equation $\phi_{i'}$ is the fertility of the word in the position i' for alignment a and the fertility of this word in alignment a' is $\phi_{i'} + 1$. Similar equations can be easily derived when either i or i' is zero, or when a and a' differ by a swap.

# References

[1] P. E. Brown , S. A. Della Pietra, V. J. Della Pietra and R. L. Mercer,1993. The Mathematics of Statistical Machine Translation. Computational Linguistics, 19(2):263–311.

[2] K. Knight. 1999. A Statistical MT tutorial Workbook. JHU CLSP summer workshop.