Parameter Estimation for IBM Model 1
CS626/CS460

Anoop Kunchukuttan
anoopk@cse.iitb.ac.in
Working under Prof. Pushpak Bhattacharyya

Training Objective

- The probability of a sentence translation is modelled in IBM Model 1 as:

$$Pr(\mathbf{f}|\mathbf{e}) = \frac{\varepsilon}{(l+1)^m} \sum_{a_1=0}^{l} \dots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_i|e_{a_j}) \tag{1}$$

- Use **Maximum Likelihood Estimation** to find the model parameters $t(f|e)$
- For a single sentence in the corpus, the objective is:

$$\max Pr(\mathbf{f}|\mathbf{e}) \tag{2}$$
$$s.t. \quad \sum_{f_i \in F} t(f_i|e) = 1 \quad \forall e \in E$$

- There will be a constraint corresponding to every word in Vocabulary(language E)

Maximizing the objective

The Lagrangian function for this objective can be written as

$$\mathcal{L}(t(f|e), \lambda_e) = \frac{\varepsilon}{(l+1)^m} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_i|e_{a_j}) - \sum_{e \in E} \lambda_e \left(\sum_{f_i \in F} t(f_i|e) - 1 \right) \quad (3)$$

Differentiating the Lagrangian w.r.t each $t(f|e)$ gives us

$$\frac{\varepsilon}{(l+1)^m} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \left(\sum_{j=1}^{m} \delta(f, f_j)\delta(e, e_{a_j}) \right) t(f|e)^{-1} \prod_{k=1}^{m} t(f_k|e_{a_k}) - \lambda_e \quad (4)$$

where δ is defined as,

$$\delta(a, b) = 1 \quad \text{if } a = b$$
$$= 0 \quad \text{if } a \neq b$$

The sum in brackets basically counts the number of times the words f and e are aligned in the sentence pair.

Maximizing the objective - 2

At the point of optimality $\frac{\partial \mathcal{L}}{\partial t(f|e)} = 0$ giving,

$$\lambda_e = t(f|e)^{-1} \sum_{a \in \mathcal{A}} Pr(f, a|e) \sum_{j=1}^{m} \delta(f, f_j)\delta(e, e_{a_j}) \quad (5)$$

$$t(f|e) = \lambda_e^{-1} \sum_{a \in \mathcal{A}} Pr(f, a|e) \sum_{j=1}^{m} \delta(f, f_j)\delta(e, e_{a_j}) \quad (6)$$

$t(f|e)$ in the form presented here is difficult to compute. As generally done in EM, we try to define $t(f|e)$ as a function of an expected quantity computed in the E-step.

Maximizing the objective - 3

- Define $c(f, e; \mathbf{f}, \mathbf{e})$ as the expected number of times word f aligns with word e in the pair of sentences $(\mathbf{f},\mathbf{e})$.

$$c(f, e; \mathbf{f}, \mathbf{e}) = \sum_{a \in \mathcal{A}} Pr(a|\mathbf{f}, \mathbf{e}) \sum_{j=1}^{m} \delta(f, f_j)\delta(e, e_{a_j}) \quad (7)$$

- Replacing the above in Equation 6,

$$t(f|e) = \lambda^{-1} Pr(f|e) c(f|e; \mathbf{f}, \mathbf{e}) \quad (8)$$

- Equation 8 gives us the formula for estimating $t(f|e)$ in the M-step.
- λ is only a normalizer as we shall see later
- You can read this equation as computing the translation probability using word alignment counts, the only difference being that these are expected counts.

Computing in the E-step

- $c(f|e)$ needs to be computed in the E-step.
- Equation 7 requires iterating over all alignments - computationally intractable
- Solution exists for Model 1. c can be computed without enumeration of alignments
- Key idea is to rewrite objective function as **Product of Sum**

$$\sum_{a_1=0}^{l} \dots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j|e_{a_j}) = \prod_{j=1}^{m} \sum_{i=1}^{l} t(f_j|e_i) \tag{9}$$

As an example,

$$t_{11}t_{21} + t_{11}t_{22} + t_{12}t_{21} + t_{12}t_{22} = (t_{11} + t_{12})(t_{21} + t_{22})$$

Notation: $t_{ji} = t(f_j|e_i)$

Computing in the E-step - 2

The Lagrangian function in terms of this revised objective function

$$\mathcal{L}(t, \lambda) = \frac{\epsilon}{(l+1)} \prod_{j=1}^{m} \sum_{i=1}^{l} t(f_j|e_i) - \sum_{einE} \lambda_e \left(\sum_{f_i \in F} t(f_i|e) - 1 \right) \quad (10)$$

On differentiating w.r.t $t(f|e)$

$$\frac{\partial \mathcal{L}}{\partial t(f|e)} = \frac{\sum_j \delta(f_j, j) \sum_i \delta(e_j, e)}{\sum_k t(f|e_k)} Pr(\mathbf{f}|\mathbf{e}) - \lambda_e$$

How did we get this? See the next slide.

Why? - 2 cases I

1. t_{11} and t_{21} may actually refer to the same word pair: Here two positions in the F language sentence may have the same word. If we differentiate,

$$(t_{11} + t_{12} + t_{13})(t_{21} + t_{22} + t_{23})(t_{31} + t_{32} + t_{33})$$

w.r.t t_{11}, the first term would go to 1. However, if $t_{11} = t_{21}$ the differentiation gives the result,

$$((t_{11} + t_{12} + t_{13}) + (t_{21} + t_{22} + t_{23}))\,(t_{31} + t_{32} + t_{33})$$

Note that the two sum terms are identical, since the F-language words are the same. So, counting the number of words f we could write the above as,

$$\frac{2}{(t_{11} + t_{12} + t_{13})}\,((t_{11} + t_{12} + t_{13})(t_{21} + t_{22} + t_{23})(t_{31} + t_{32} + t_{33}))$$

Why? - 2 cases II

In general, we can write the above as,

$$\frac{\sum_j \delta(f_j, j)}{\sum_k t(f|e_k)} \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j|e_i) \qquad (11)$$

2. t_{11} and t_{12} may actually refer to the same word pair: Hence two position in the E language sentence may have the same word. So if we differentiate,

$$(t_{11} + t_{12} + t_{13})(t_{21} + t_{22} + t_{23})(t_{31} + t_{32} + t_{33})$$

w.r.t t_{11}, we would get,

$$\frac{2}{(t_{11} + t_{12} + t_{13})}((t_{11} + t_{12} + t_{13})(t_{21} + t_{22} + t_{23})(t_{31} + t_{32} + t_{33}))$$

Getting these two cases together and setting the derivative to 0 will give you:

$$\frac{\sum_j \delta(f_j, j) \sum_i \delta(e_j, e)}{\sum_k t(f|e_k)} Pr(\mathbf{f}|\mathbf{e}) - \lambda_e \quad (12)$$

$$\lambda_e = \frac{\sum_j \delta(f_j, j) \sum_i \delta(e_j, e)}{\sum_k t(f|e_k)} Pr(\mathbf{f}|\mathbf{e}) \quad (13)$$

Substituting for λ from Equation 13 into Equation 8 gives us,

$$c(f|e; \mathbf{f}, \mathbf{e}) = \frac{t(f|e)}{\sum_k t(f|e_k)} \sum_i \delta(e_i, e) \sum_j \delta(f_j, f) \quad (14)$$

$c(f_k|e; \mathbf{f}, \mathbf{e})$ can be interpreted as weighing the max number of possible alignments by the translation probability
λ_e can be computed using the constraint $\sum_k t(f_k|e) = 1$ giving,

$$\lambda_e = \sum_k Pr(\mathbf{f}|\mathbf{e}) \quad (15)$$

Q.E.D