

Lecture 7 — Jan 31, 2006

Lecturer: Anoop Sarkar

Scribe: Akshay Gattani

1. Introduction and Required Reading

Continuing with our discussion of the different models for machine translation based on the Brown et al., 1990 paper, the lecture discusses the following 2 word alignment models which take a slightly different probabilistic approach (**Key Concept 1**: Word alignment is the correspondence between words in say an English sentence and the words of the French sentence).

- **Mixture based alignment model**
- **HMM based alignment model.**

The lecture is based on the following paper readings: [1]. HMM-Based Word Alignment in Statistical Translation. Stephan Vogel; Hermann Ney; Christoph Tillmann. COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics.

[2]. The Mathematics of Statistical Machine Translation: Parameter Estimation. Peter E Brown; Vincent J. Della Pietra; Stephen A. Della Pietra; Robert L. Mercer. Computational Linguistics, Volume 19, Number 2, June 1993

2. **Alignment with Mixture Distribution** The notations in [1] can be confusing so it's best if the notations are cleared out first. $Pr(f_1^J | e_1^I)$ is the translation probability. J and I are the length of French and English sentence respectively. $P(J|I)$ is the sentence length probability and $P(f_j | e_i)$ is the word translation probability.

First we consider the formulation of mixture-based alignment model. HMM-based alignment models build upon it later on. The mixture model decomposes the joint probability for f^J into a product over the probabilities of each word f_j as below:

$$Pr(f_1^J | e_1^I) = P(J|I) \cdot \prod_{j=1}^J p(f_j | e^I) \quad (7.1)$$

The $p(f_j|e^I)$ term can be decomposed to consider a pairwise interaction between the french word f_j and each of the english words e_i , $i = 1, \dots, I$. This decomposition is captured in the form of a mixture distribution as below.

$$Pr(f_j|e_1^I) = \sum_{i=1}^I P(i, f_j|e_1^I) = \sum_{i=1}^I P(i|j, I) \cdot p(f_j|e_i) \quad (7.2)$$

The key term here is $P(i|j, I)$, which can be interpreted as the **responsibility** that each english index i takes to explain a possible alignment with the french word at index j . (Its vital to note that $P(i|j, I)$ explains word position/index alignments and not the word translation itself). Readers familiar with the concept of '**Gaussian mixture models**' will find it analogous to '**the soft responsibility**' that each of the gaussian component k takes for 'explaining' the observation x . Mixture models come into play since we are unsure of which particular gaussian distribution component best explains our observation 'x' (in our case which english position i best aligns with the current french position j to explain the french word at position j).

Combining the above two equations, we get the following mixture-based model:

$$Pr(f_1^J|e_1^I) = P(J|I) \cdot \prod_{j=1}^J \sum_{i=1}^I P(i|j, I) \cdot p(f_j|e_i) \quad (7.3)$$

Whenever there are probabilistic 'soft assignments' involved, its generally a good rule of thumb to think that the EM algorithm might be applicable to the problem (again borrowing from the 'Gaussian mixture' analogy which i find easier to interpret and apply to any similar setting!).

Assuming uniform alignment probability $P(i|j, I) = 1/I$ gives the IBM1 model. Non-uniform alignment probabilities lead us to the IBM model 2. This model assumes that the position distance relative to the diagonal line of (j,i) plane is the dominating factor. (A good heuristic for alignment probabilities, again referred to in the HMM based model). Training and parameter learning via the EM algorithm consists of the following two iterative steps:

* Position Alignment: Given model parameters compute most likely position alignments (soft responsibilities)

* Parameter Estimation: Do maximum likelihood estimation for $P(f_j|e_i)$ going along alignment paths for all sentence pairs.

Key concept 2: In the mixture model there is no interaction between adjacent word positions, a deficiency overcome in the HMM.

3. Alignment with HMM

The motivation is that typically there is strong localization effect in aligning words in parallel texts, in other words, alignments tend to preserve the local neighborhood when going from one language to other, as is evident in the german-english sentence translation in the figure below..

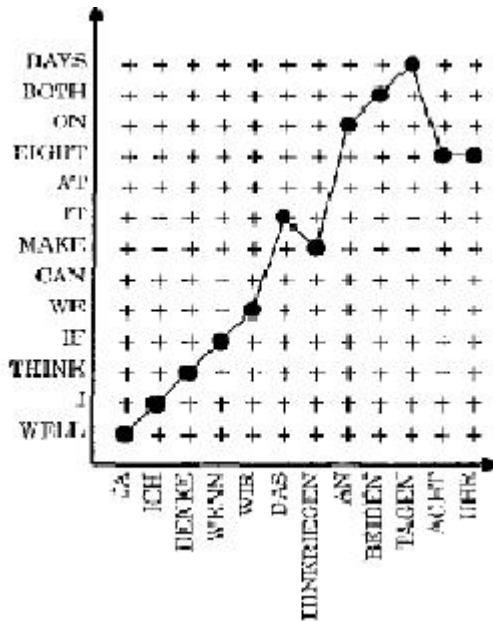


Fig 1: Showing effect of localization for a german-english sentence word alignment

Again clearing up the notation, we call the mapping $j \rightarrow^{a_j} i$, which assigns a word f_j in position j to word e_i as alignment a_j . HMM extends the notion of Mixture model by conditioning on the previous word alignment. Again referring to the mixture model analogy, **Key Concept 3:** in HMM the choice of mixture component is not selected independently but also depends on the choice of the component for the previous observation. Thus the probability of alignment $a_j = i$ should have a dependence on the previous alignment $a_{j-1} = i'$: $P(a_j|a_{j-1}, I)$.

The Graphical model for our HMM based alignment model is shown in the figure 2 below. The alignment assignments $a_j = i$ are the latent variables while translated french words f_j are the observations. I and e are non-stochastic inputs and are shown for understanding.

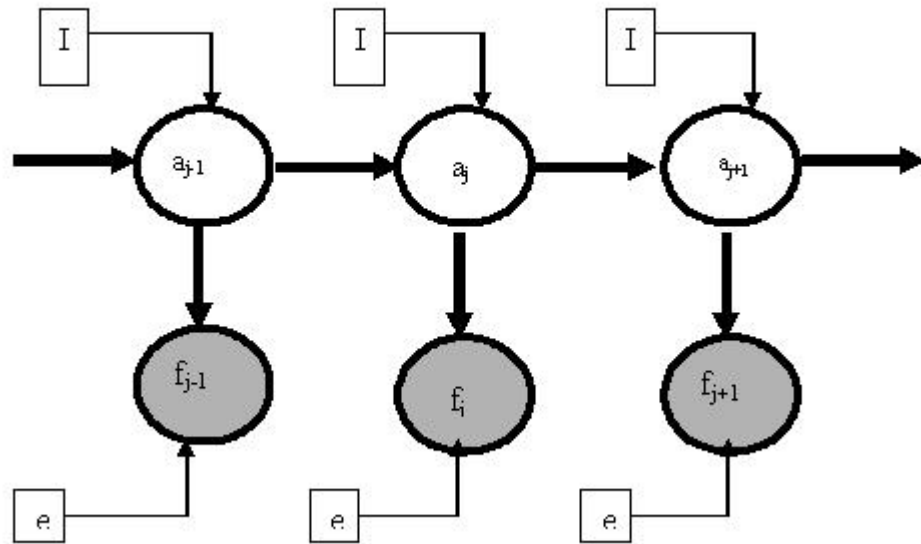


Fig 2: Graphical model for HMM based alignment model

Transition between states (alignments) is governed by the 'transition probabilities' (alignment probabilities) and the observation (translation) in the current state is based on the translation probability distribution ($P(f_j|e_{a_j})$). Combining these 2 ingredients we have the following HMM based alignment model:

$$Pr(f_1^J|e_1^J) = \sum_{a_i^J} \prod_{j=1}^J [P(a_j|a_{j-1}, I) \cdot P(f_j|e_{a_j})] \quad (7.4)$$

Additional constraints can be imposed in the HMM by making the alignment probabilities $P(a_j|a_{j-1}, I)$ dependent only on the jump width between a_j and a_{j-1} .