Using Word Dependent Transition Models in HMM Based Word Alignment for Statistical Machine Translation

Xiaodong He, Microsoft Research

ACL 07 2nd SMT workshop, 2007

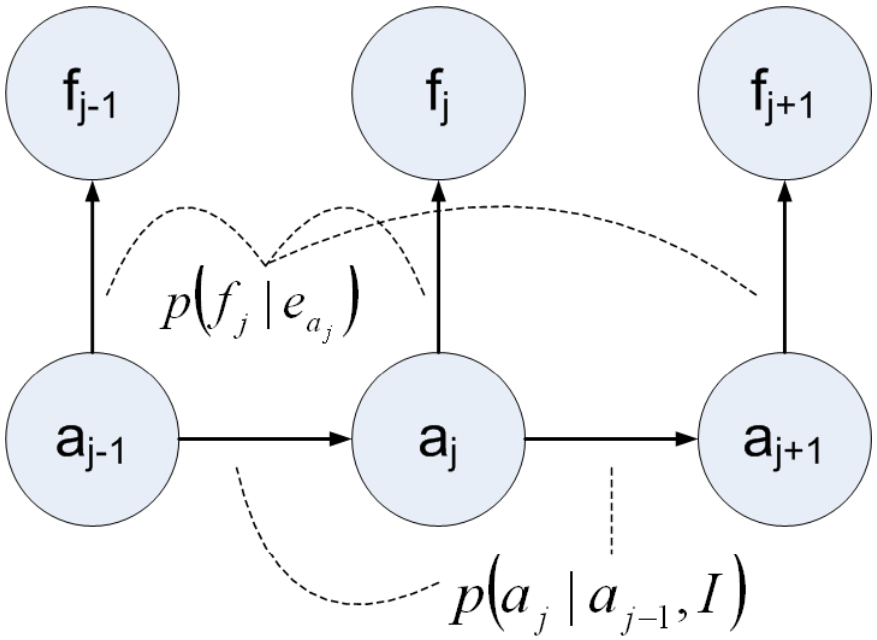
Presented by Mei Yang, University of Washington

February 20th, 2008

Goal and Approach

- Improve the transition models for HMM alignment
- Introduce word-dependence into the translation models for better modeling
- Use maximum a posteriori (MAP) estimation to address data sparseness

Conventional HMM Alignment

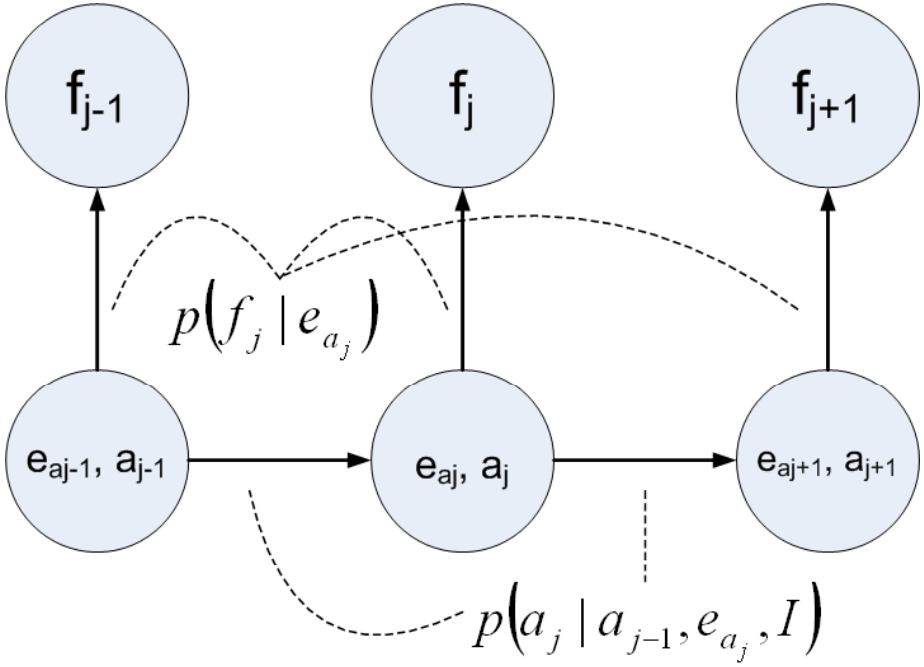


$$p(f_1^J \mid e_1^I) = \sum_{a_1^J} \prod_{j=1}^{J} \left[p(a_j \mid a_{j-1}, I)\, p(f_j \mid e_{a_j}) \right]$$

Transition models are context independent!

Can depend on word classes

Word-dependent HMM Alignment



$$p(f_1^J \mid e_1^I) = \sum_{a_1^J} \prod_{j=1}^{J} \left[p(a_j \mid a_{j-1}, e_{a_j}, I)\, p(f_j \mid e_{a_j}) \right]$$

Maximum Likelihood Estimation

$$p_{ML}\left(i \mid i^{'}, e, I\right) \propto c\left(i - i^{'}; e\right)$$

$$c(d;e) = \sum_{j=1}^{J-1} \sum_{i=1}^{I} \delta\left(e_{a_j}, e\right) p\left(a_j = i, a_{j+1} = i + d \mid f_1^J, e_1^I, \Lambda^{'}\right)$$

- Expected counts: *c(d;e)*
- Over-fitting for infrequent words *e*

MAP Estimation

- See page 83 for the equations (7) – (11)

- A Dirichlet prior over the parameters of transition models
- Hyper-parameters are chosen proportional to the word-independent transition model

- Back-off the word-dependent model to the word-independent model
- "Count" merging

Experiment #1: AER

- English-French Hansards corpus
- Training data: 500K sentence pairs
- Test data: 447 manually aligned sentence pairs
  - *sure* and *possible* alignments
- IBM model 1, HMM model (word-independent or word-independent), and IBM model 4
  - both directions combined

- Figure 1 and Table 1~3 show that WDHMM outperforms the baseline HMM and IBM model 4

Experiment #2: BLEU

- English-to-French track of NAACL 2006 Europarl evaluation workshop
- Training set: 688K sentence pairs
- Devset set: 500 sentence pairs, MERT
- Devtest set: 2000 sentence pairs, the prior parameter
- Test set: 2000 sentences
- NC-test set: 1064 out-of-domain sentences

- Figure 2, 3 and Table 4,5 show that WDHMM gives significant improvement on BLEU over the baseline and IBM model 4

Others

- WDHMM runs as fast as the conventional HMM
- WDHMM need extra memory that is proportional to the vocabulary size of the source language
  - Bucketed distortion

Conclusion

- Pros
  - Relatively simple technique
  - Consistent improvement over the baseline HMM and IBM model 4 alignments
  - …
- Cons
  - An extra prior parameter
  - …
- Questions?