Gappy Phrasal Alignment by Agreement

Microsoft® Research

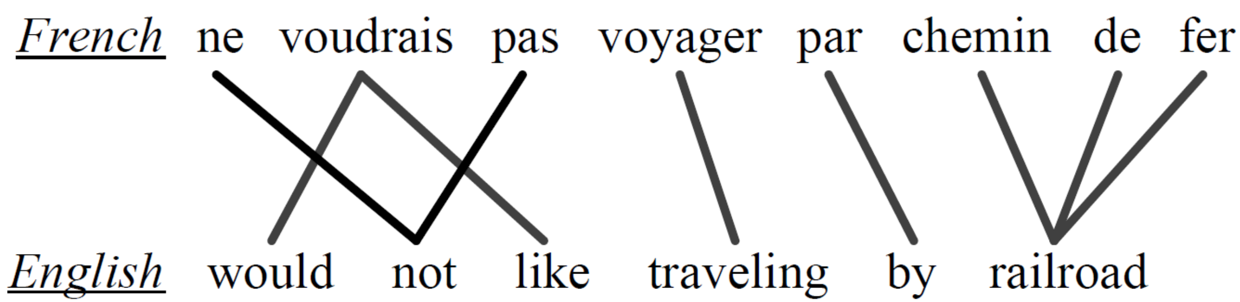
Mohit Bansal, Chris Quirk, and Robert C. Moore
Summer 2010, NLP Group

Introduction

- High level motivation:
  - Word alignment is a pervasive problem
    - Crucial component in MT systems
      - To build phrase tables
      - To extract synchronous syntactic rules
    - Also used in other NLP problems:
      - entailment
      - paraphrase
      - question answering
      - summarization
      - spell correction, etc.

Introduction

- The limitation to ***words*** is obviously wrong



- People have tried to correct this for a while now
  - Phrase-based alignment
  - Pseudo-words
- Our contribution: clean, fast phrasal alignment model

hidden semi-markov model	(observations can be phrases…)
for phrase-to-phrase alignment	(…and states…)
using alignment by agreement	(…***meaningful*** states…no phrase penalty)
allowing subsequences	(…finally, with ***ne .. pas*** !)

Related work

Two major influences :

1) Conditional phrase-based alignment models

– Word-to-phrase HMM is one approach (Deng & Byrne'05)
  - model subsequent words from the same state using a bigram model
  - change only the parameterization and not set of possible alignments

– Phrase-to-phrase alignments (Daume & Marcu'04; DeNero et al.'06)
  - unconstrained model may overfit using unusual segmentations

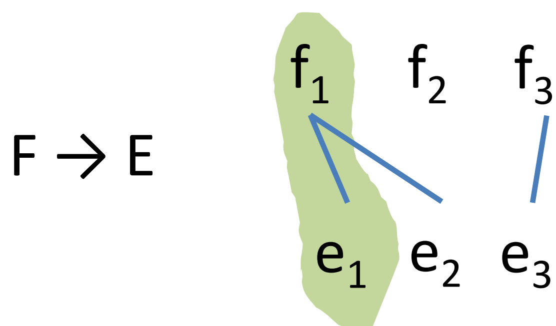
– Phrase-based hidden semi-markov model (Ferrer & Juan'09)
  - interpolates with Model 1 and monotonic (no reordering)

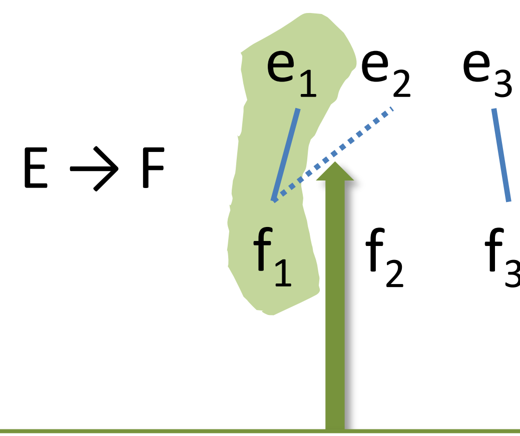
Microsoft Research

Related work

2) Alignment by agreement (Liang et al. 2006)

- soft intersection cleans up and symmetrizes word HMM alignments
- symmetric portion of HMM space is only word-to-word

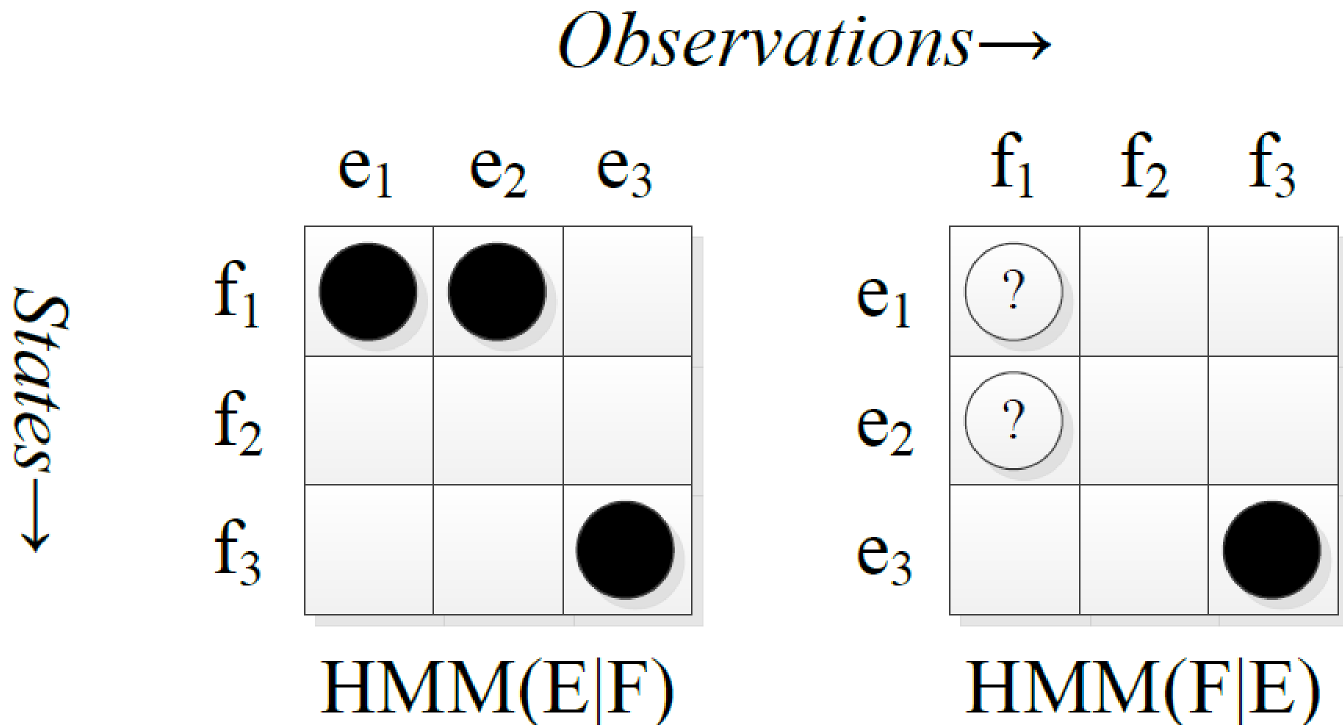




Here we can't capture $f_1 \sim e_1 e_2$ after agreement because 2 states e_1 and e_2 cannot align to the same observation f_1 in the E → F direction

Microsoft® Research

Word-to-word constraint

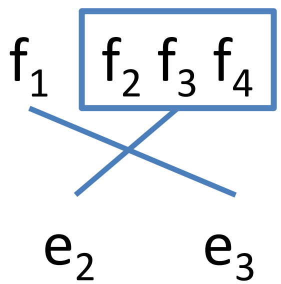


Model of *F* given *E* can't represent phrasal alignment {e1,e2} ~ {f1}: probability mass is distributed between {f1} ~ {e1} and {f1} ~ {e2}. Agreement of forward and backward HMM alignments places less mass on phrasal links and more mass on word-to-word links.

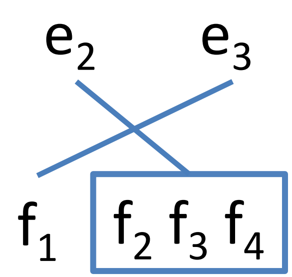
Contributions

- Unite phrasal alignment and alignment by agreement
  - Allow phrases at both state and observation side

F → E



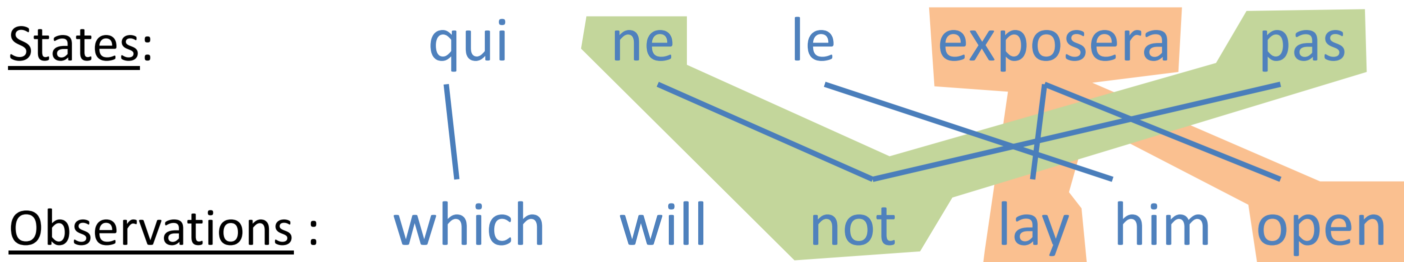
E → F



  - Agreement favors alignments meaningful in both directions
    - With word alignment, agreement removes phrases ☹
    - With phrase-to-phrase alignment, agreement reinforces meaningful phrases – avoids overfitting ☺

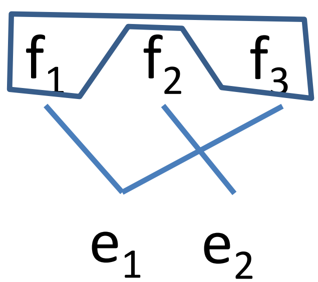
Contributions

- Furthermore, we can allow subsequences

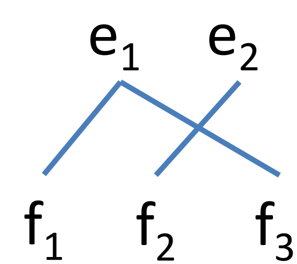


  - State space extended to include gappy phrases
  - Gappy obsrv phrases approximated as multiple obsrv words emitting from a single state

F → E

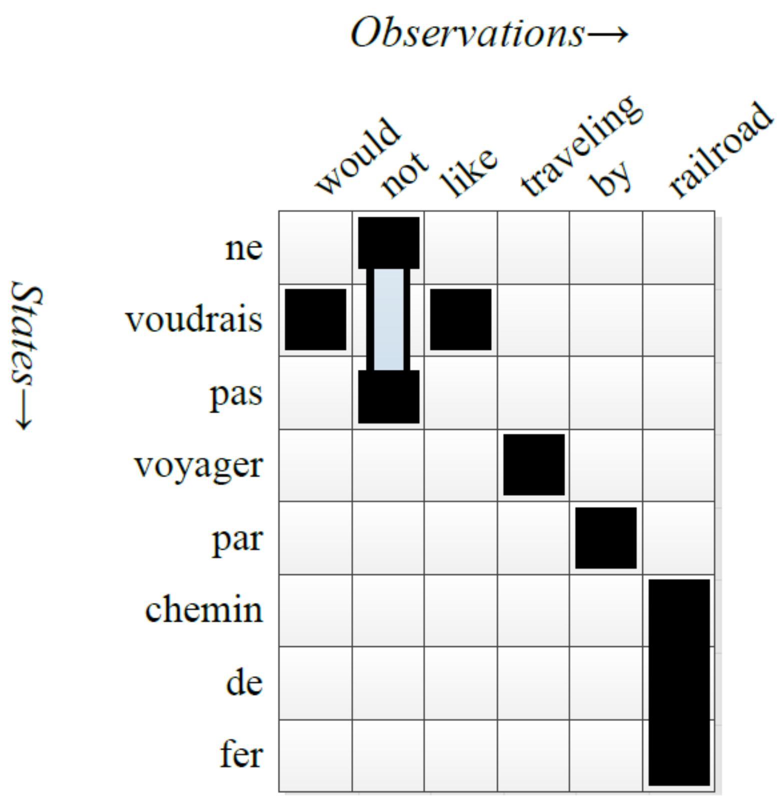


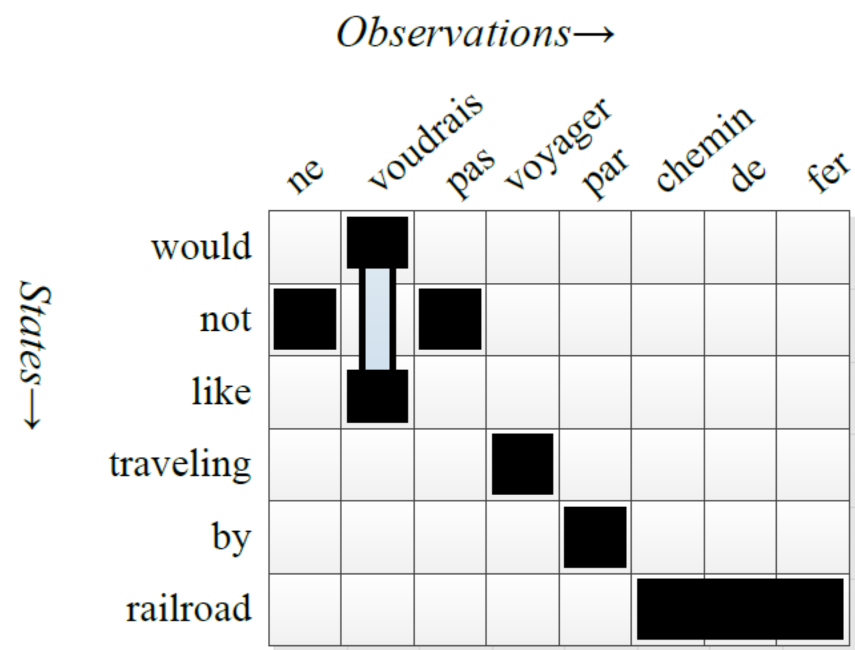
E → F



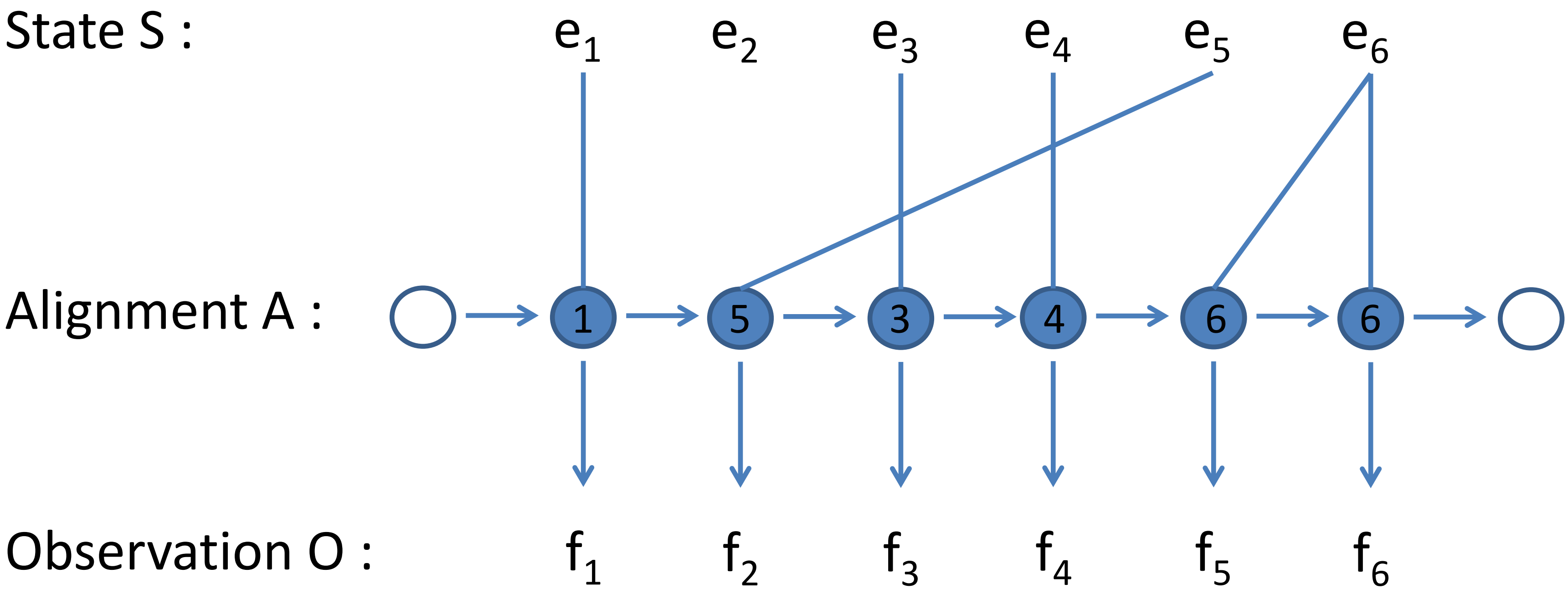
Complexity still $\sim \mathcal{O}(m^2.n)$

Contributions





HMM Alignment



Model Parameters

Emission/Translation Model	*Transition/Distortion Model*
$P(O_2 = f_2 \mid S_{A_2} = e_5)$	$P(A_2 = 5 \mid A_1 = 1)$

Microsoft Research

EM Training (Baum-Welch)

≈5 iterations

E-Step

$$q(\mathbf{z};\mathbf{x}) \coloneqq p(\mathbf{z}|\mathbf{x};\theta)$$

Forward	α
Backward	β
Posteriors	γ, ξ

M-Step

$$\theta' = \operatorname{argmax}_{\theta} \sum_{\mathbf{x},\mathbf{z}} q(\mathbf{z};\mathbf{x}) \log p(\mathbf{x},\mathbf{z};\theta)$$

Translation	$p(\bar{o}|\bar{s})$
Distortion	$p(i'|i;I)$

Microsoft Research

Baum-Welch (Forward-Backward)

Forward : $\alpha_j(i)$ = probability of generating obsrv o_1 to o_j s.t. state s_i generates o_j

$$\alpha_j(i) = \left[\sum_{i'} \alpha_{j-1}(i') \,.\, p_d(i|i';I)\right] . \, p_t(o_j|\, s_i) \qquad \alpha_0(i) = 1$$

previous α

distortion model

translation model

Initialization

Backward : $\beta_j(i)$ = probability of generating obsrv o_{j+1} to o_J s.t. state s_i generates o_j

$$\beta_j(i) = \left[\sum_{i'} \beta_{j+1}(i') \,.\, p_d(i'|i;I) p_t(o_{j+1}|\, s_{i'})\right] \qquad \beta_J(i) = 1$$

next β

distortion model

translation model

Initialization

Baum-Welch

$p(O|S)$ = probability of full observation sentence $O = o_1^J$ from full state sentence $S = s_1^I$

$$p(O|S) = \sum_i \alpha_J(i) \,.\, 1 = \sum_i 1 \,.\, \beta_0(i) = \sum_i \alpha_j(i) \,.\, \beta_j(i)$$

Node Posterior : $\gamma_j(i)$ = probability, given O, that obsrv o_j generated by state s_i

$$\gamma_j(i) = \frac{\alpha_j(i).\beta_j(i)}{p(O|S)}$$

Edge Posterior : $\xi_j(i', i)$ = probability, given O, that obsrv o_j and o_{j+1} generated by states $s_{i'}$, and s_i resp.

$$\xi_j(i', i) = \frac{\alpha_j(i').p_d(i|i';I)p_t\left(o_{j+1}|\, s_i\right).\beta_{j+1}(i)}{p(O|S)}$$

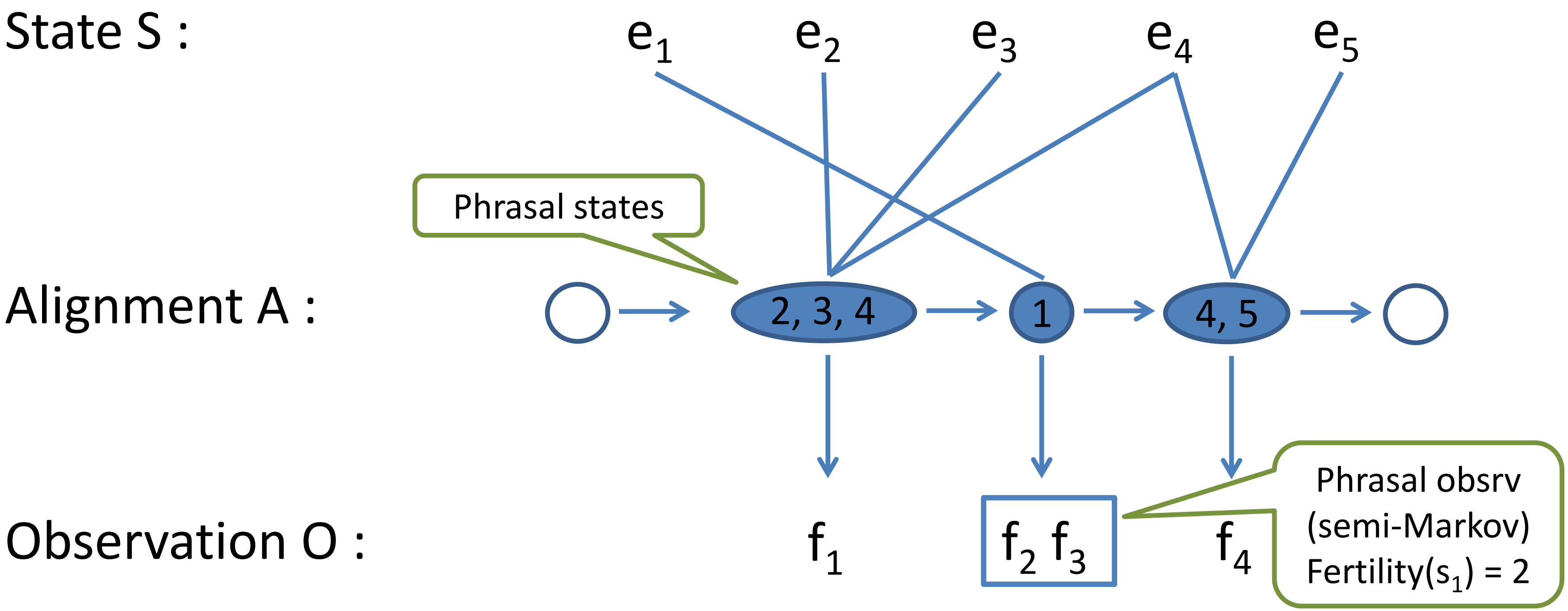
Parameter Re-estimation :

$$p_t(\bar{o}|\bar{s}) = \frac{1}{Z}\sum_{O,S}\sum_{\substack{i,j \\ s_i=\bar{s} \\ o_j=\bar{o}}} \gamma_j(i)$$

$$p_d(i|i';I) = \frac{1}{Z}\sum_{\substack{O,S \\ |S|=I}}\sum_j \xi_j(i',i)$$

Microsoft Research

Hidden Semi-Markov Model



Model Parameters

Emissions : $P(O_1 = f_1 \mid S_{A_1} = e_2 e_3 e_4)$, — k → 1 alignment

$P(O_2 = f_2 f_3 \mid S_{A_2} = e_1)$ — 1 → k alignment

Transitions : $P(A_2 = 1 \mid A_1 = 2,3,4)$ — Jump (4 → 1)

Microsoft Research

Forward Algorithm (HSMM)

$\alpha_j(i,\phi)$ = probability of generating observations o_1 to o_j such that last observation-phrase $o_{j-\phi+1}^{j}$ generated by state s_i

$$\alpha_j(i,\phi) = \left[\sum_{i',\phi'} \alpha_{j-\phi}(i',\phi') \, . \, p_d(i|i';I)\right] . \, n(\phi|s_i) \, . \, \eta^{-\phi} \, . \, \eta^{-|s_i|} \, . \, p_t\left(o_{j-\phi+1}^{j} | \, s_i\right)$$

previous α | distortion model | fertility model | penalty for obsrv phrase-len | penalty for state phrase-len | translation model

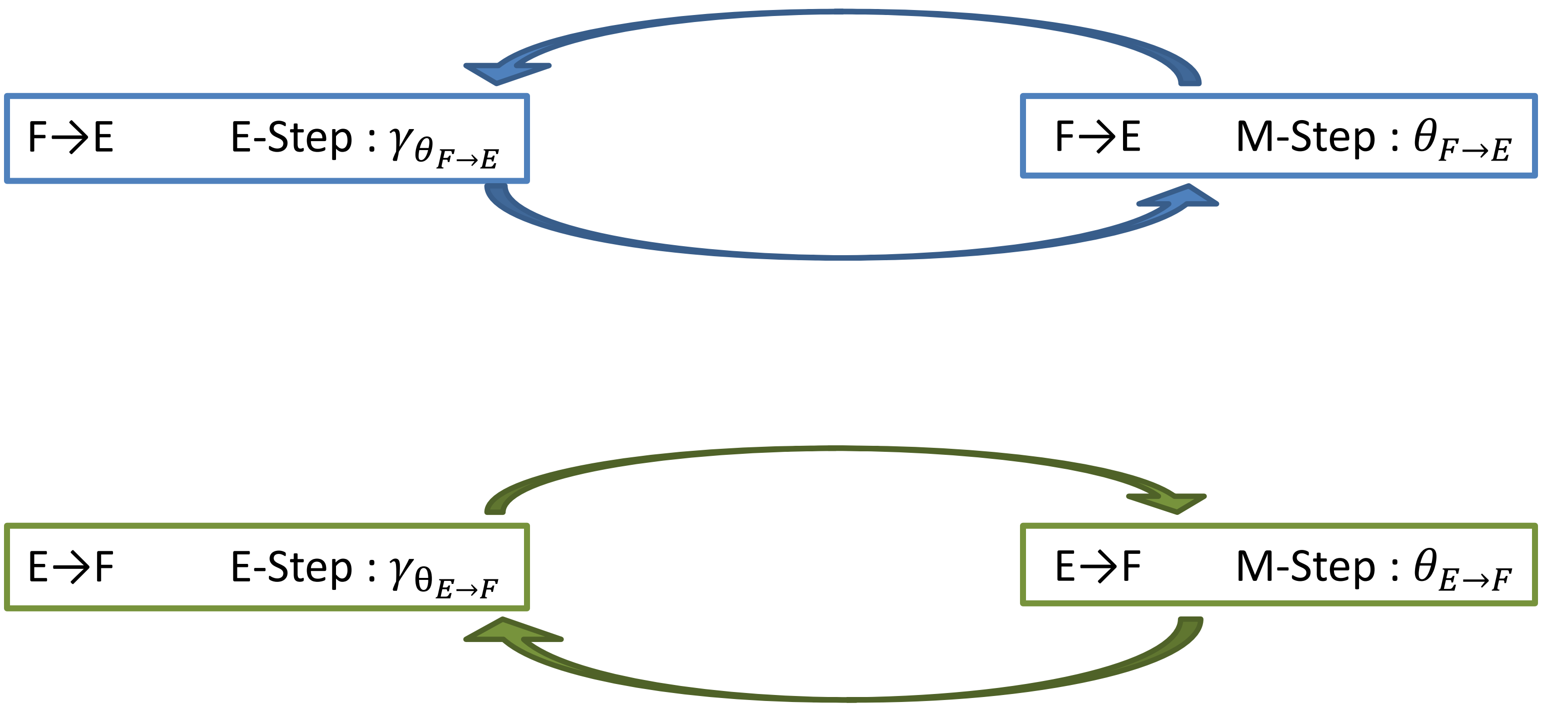
Initialize :

$$\alpha_0(i,0) = 1$$

$$\alpha_\phi(i,\phi) = p_{d_{init}}(i) \, . \, n(\phi|s_i) \, . \, \eta^{-\phi} \, . \, \eta^{-|s_i|} \, . \, p_t\left(o_1^{\phi} | \, s_i\right)$$

Agreement

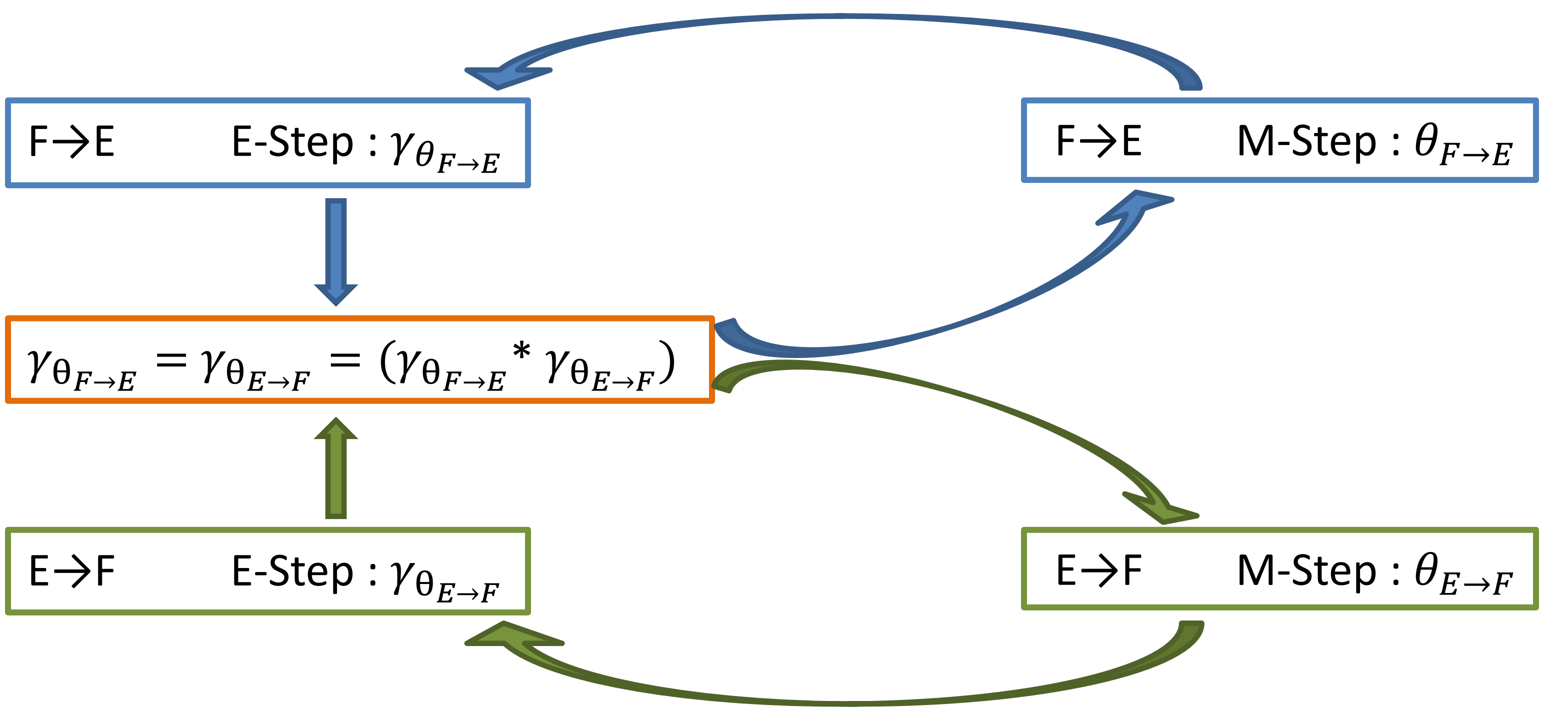
- Multiply posteriors of both directions E→F and F→E after every E-step of EM
- $q(\mathbf{z};\mathbf{x}) := \prod_{i,j} p_1\left(z_{ij}\middle|\mathbf{x};\theta_1\right).p_2\left(z_{ij}\middle|\mathbf{x};\theta_2\right)$



Microsoft Research

Agreement

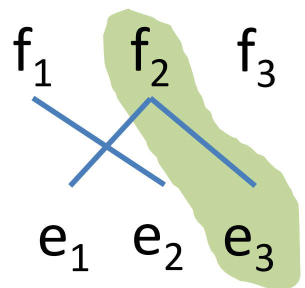
- Multiply posteriors of both directions E→F and F→E after every E-step of EM
- $q(\mathbf{z};\mathbf{x}) := \prod_{i,j} p_1(z_{ij}|\mathbf{x};\theta_1).p_2(z_{ij}|\mathbf{x};\theta_2)$



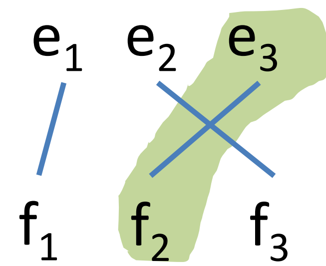
Microsoft Research

Agreement

F → E



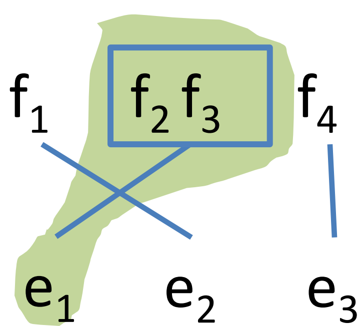
E → F



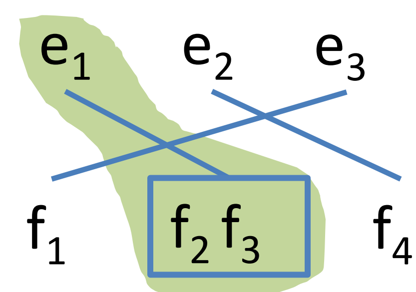
$$\gamma_{F\to E}(f_i, e_j) = \gamma_{E\to F}(e_j, f_i) = [\gamma_{F\to E}(f_i, e_j) * \gamma_{E\to F}(e_j, f_i)]$$

- Phrase Agreement : ***need phrases on both observation and state sides***

F → E



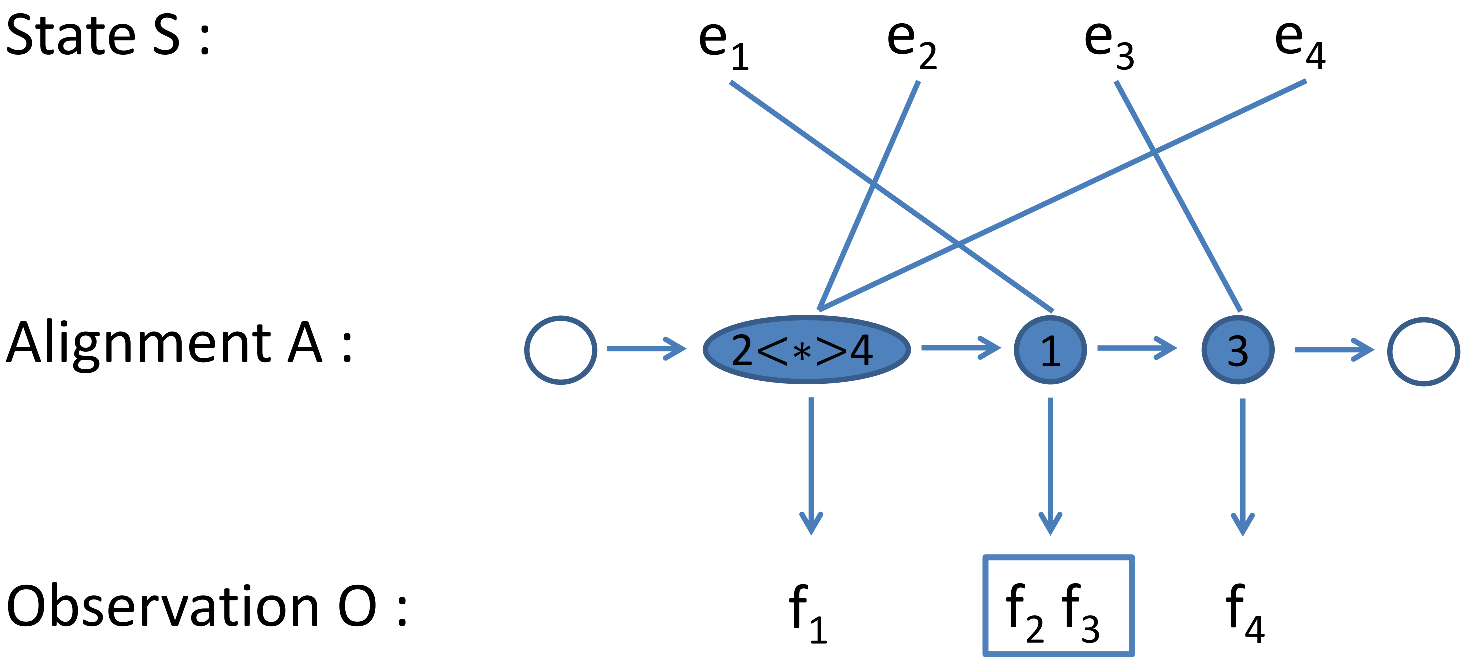
E → F



$$\gamma_{F\to E}(f_a^b, e_k) = \gamma_{E\to F}(e_k, f_a^b) = [\gamma_{F\to E}(f_a^b, e_k) * \gamma_{E\to F}(e_k, f_a^b)]$$

Microsoft Research

Gappy Agreement

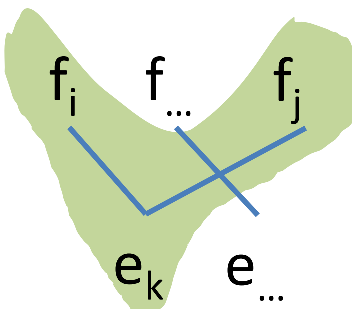


- ***Asymmetry***
  - Computing posterior of gappy observation phrase is inefficient
  - Hence, approximate posterior of $e_k \sim \{f_i <*> f_j\}$ using posteriors of $e_k \sim \{f_i\}$ and $e_k \sim \{f_j\}$

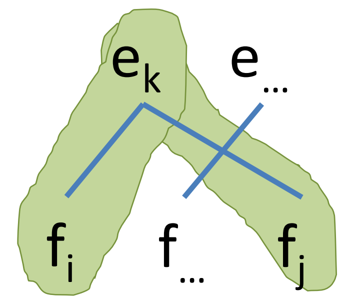
Gappy Agreement

- Modified agreement, with approximate computation of posterior
  - Reverse direction of gapped state corresponds to a revisited state emitting two discontiguous observations

F → E



E → F



$$\gamma_{F\to E}\left(f_i <*> f_j, e_k\right) \ *= \ \min^{\ddagger}\left\{\gamma_{E\to F}(e_k, f_i), \gamma_{E\to F}\left(e_k, f_j\right)\right\}$$

$$\gamma_{E\to F}(e_k, f_i) \ *= \gamma_{F\to E}(f_i, e_k) + \sum_{h<i<j}\left\{\gamma_{F\to E}(f_h <*> f_i, e_k) + \gamma_{F\to E}\left(f_i <*> f_j, e_k\right)\right\}$$

‡ min is an upper bound on the posterior that both observations f_i and f_j are ~ e_k, since every path that passes through e_k ~ f_i & e_k ~ f_j must pass through e_k ~ f_i, therefore the posterior of e_k ~ f_i & e_k ~ f_j is less than that of e_k ~ f_i, and likewise less than that of e_k ~ f_j

Allowed Phrases

- Only allow certain 'good' phrases instead of all possible ones
- Run word-to-word HMM on full data
- Get observation phrases (contiguous and gapped) aligned to single state, i.e. $o_i^j \sim s$ for both languages/directions

- Weight the phrases o_i^j by discounted probability $\max\left(0, c(o_i^j \sim s\right) - \delta)/c(o_i^j)$ and choose top X phrases

Complexity

State length

Obsrv length

Model1 : $\prod_{i=1}^{n} p_t(o_i|s_{a_i})$ $\mathcal{O}(m\,.\,n)$

w2w HMM : $\prod_{i=1}^{n} p_d(a_i|\, a_{i-1}, m).\, p_t(o_i|s_{a_i})$ $\mathcal{O}(m^2\,.\,n)$

HSMM (phrasal obsrv of bounded length k): $\mathcal{O}(m^2\,.\,kn)$

+ Phrasal states of bounded length k : $\mathcal{O}\big((km)^2\,.\,kn\big) = \mathcal{O}(k^3m^2n)$

+ Gappy phrasal states of form w <*> w : $\mathcal{O}\big((km+m^2)^2\,.\,kn\big) = \mathcal{O}(km^4n\,)$

Still smaller than exact ITG $\mathcal{O}(n^6)$

Phrases (obsrv and states (contig and gappy))
from pruned lists : $\mathcal{O}\big((m+p)^2\,.\,(n+p)\big) \sim \mathcal{O}(m^2.n)$

Evaluation

$$Precision = \frac{|A \cap P|}{|A|} * 100\%, \qquad Recall = \frac{|A \cap S|}{|S|} * 100\%$$

where S, P and A are gold-sure, gold-possible and predicted edge-sets respectively

$$F_\beta = (1+\beta^2).\frac{Precision \; * \; Recall}{\beta^2.Precision \; + \; Recall} * 100\%$$

$$BLEU_n = \min\left(1, \frac{output-len}{ref-len}\right).\exp\sum_{i=1}^{n}\lambda_i \log p_i$$

$$p_i = \frac{\sum_{C\in\{Candidates\}}\sum_{i-gram\in C} Count_{clip}(i-gram)}{\sum_{C\in\{Candidates\}}\sum_{i-gram\in C} Count\ (i-gram)}$$

Microsoft Research

Evaluation

- Datasets
  - French-English : Hansards NAACL 2003 shared-task
    - 1.1M sentence-pairs
    - Hand-alignments from Och&Ney03
    - 137 dev-set, 347 test-set (Liang06)
  - German-English : Europarl from WMT 2010
    - 1.6M sentence-pairs
    - Hand-alignments from ChrisQ
    - 102 dev-set, 258 test-set

- Training Regimen :
  - 5 iterations of Model 1 (independent training)
  - 5 iterations of w2w HMM (independent training)
  - Initialize the p2p model using phrase-extraction from w2w Viterbi alignments
  - ***Minimality*** : Only allow 1-K or K-1 alignments, since 2-3 can be generally be decomposed into 1-1 U 1-2, etc.

Microsoft Research

Alignment F1 Results

Data	Decoding method	Word-to-word	+Contig phrases	+Gappy phrases
FE 10K	Viterbi	89.7	90.6	90.3
FE 10K	Posterior ≥ 0.1	90.1	90.4	**90.7**
FE 100K	Viterbi	93.0	93.6	93.8
FE 100K	Posterior ≥ 0.1	93.1	93.7	**93.8**
FE All	Viterbi	94.1	94.3	94.3
FE All	Posterior ≥ 0.1	94.2	94.4	**94.5**
GE 10K	Viterbi	76.2	79.6	**79.7**
GE 10K	Posterior ≥ 0.1	76.7	79.3	79.3
GE 100K	Viterbi	81.0	83.0	83.2
GE 100K	Posterior ≥ 0.1	80.7	83.1	**83.4**
GE All	Viterbi	83.0	85.2	85.6
GE All	Posterior ≥ 0.1	83.7	85.3	**85.7**

Microsoft Research

Translation BLEU Results

- Phrase-based system using only contiguous phrases consistent with the potentially gappy alignment –
  - 4 channel models, lexicalized reordering model
  - word and phrase count features, distortion penalty
  - 5-gram language model (weighted by MERT)
- Parameters tuned on dev-set BLEU using grid search
- A syntax-based or non-contiguous phrasal system (Galley and Manning, 2010) may benefit more from gappy phrases

Language pair	Word-to-word	Gappy
French-English	34.0	34.5
German-English	19.3	19.8

Conclusion

- Start with HMM alignment by agreement
- Allow phrasal observations (HSMM)
- Allow phrasal states
- Allow gappy phrasal states
- Agreement between F→E and E→F finds meaningful phrases and *makes phrase penalty almost unnecessary*
- Maintain $\sim\mathcal{O}(m^2.n)$ complexity

Microsoft Research

Future Work

- Limiting the gap length also prevents combinatorial explosion
- Translation system using discontinuous mappings at runtime (Chiang, 2007; Galley and Manning, 2010) may make better use of discontinuous alignments
- Apply model at the morpheme or character level, allowing joint inference of segmentation and alignment
- State space could be expanded and enhanced to include more possibilities: states with multiple gaps might be useful for alignment in languages with template morphology, such as Arabic or Hebrew
- A better distortion model might place a stronger distribution on the likely starting and ending points of phrases

Microsoft Research

Thank you!