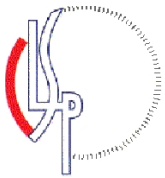
HMM Word and Phrase Alignment for Statistical Machine Translation

Yonggang Deng[1] William Byrne[1,2]

[1]Center for Language and Speech Processing,
The Johns Hopkins University,
Baltimore, MD 21218, USA

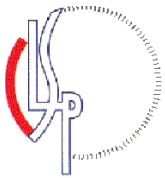
[2]Machine Intelligence Lab,
Cambridge University Engineering Department
Trumpington Street, Cambridge CB2 1PZ, UK

HLT-EMNLP Conference, 2005

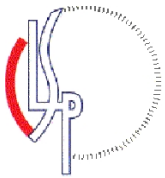


Introduction

- Goal: improve word alignments of parallel corpora for better translation
- IBM Model-4 generated by GIZA++ Toolkit (Och and Ney, '03)
  - The state of the art word alignments especially on large bitexts
- We want to propose an alternative
  - Comparable performance to Model-4
  - Efficient training, fast, with controlled memory usage
    - Can build a single model over large bitexts
  - Goal is to use the model, not just the alignments
    - Will give examples of extending phrase pairs under model-based posterior distribution
  - Present a statistical Word-to-Phrase HMM alignment model



Outline

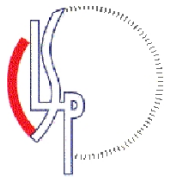


HMM-based Word Alignment Models (Vogel et al, '96)

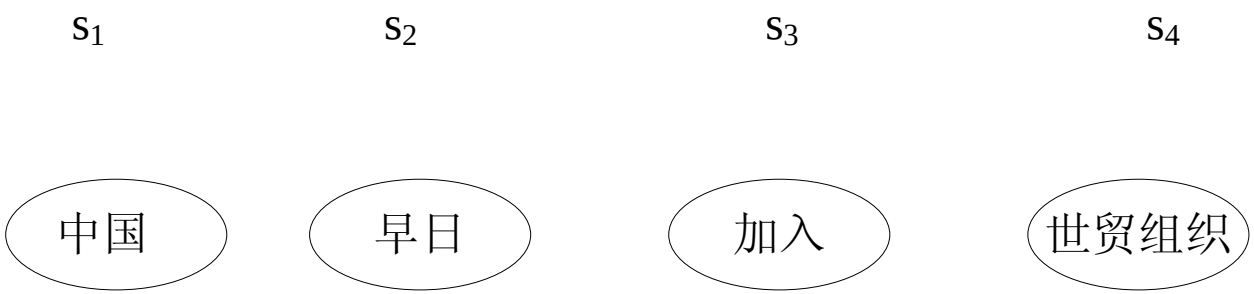
s_1	s_2	s_3	s_4
中国	早日	加入	世贸组织

china	's	accession	to	the	world	trade	organization	at	an	early	date
t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}

- State sequences $\longleftrightarrow$ word alignments
- Model parameters: transition probabilities, translation tables
- Words are generated one by one, one transition emits one target word
- Efficient Balm-Welch and Viterbi algorithms



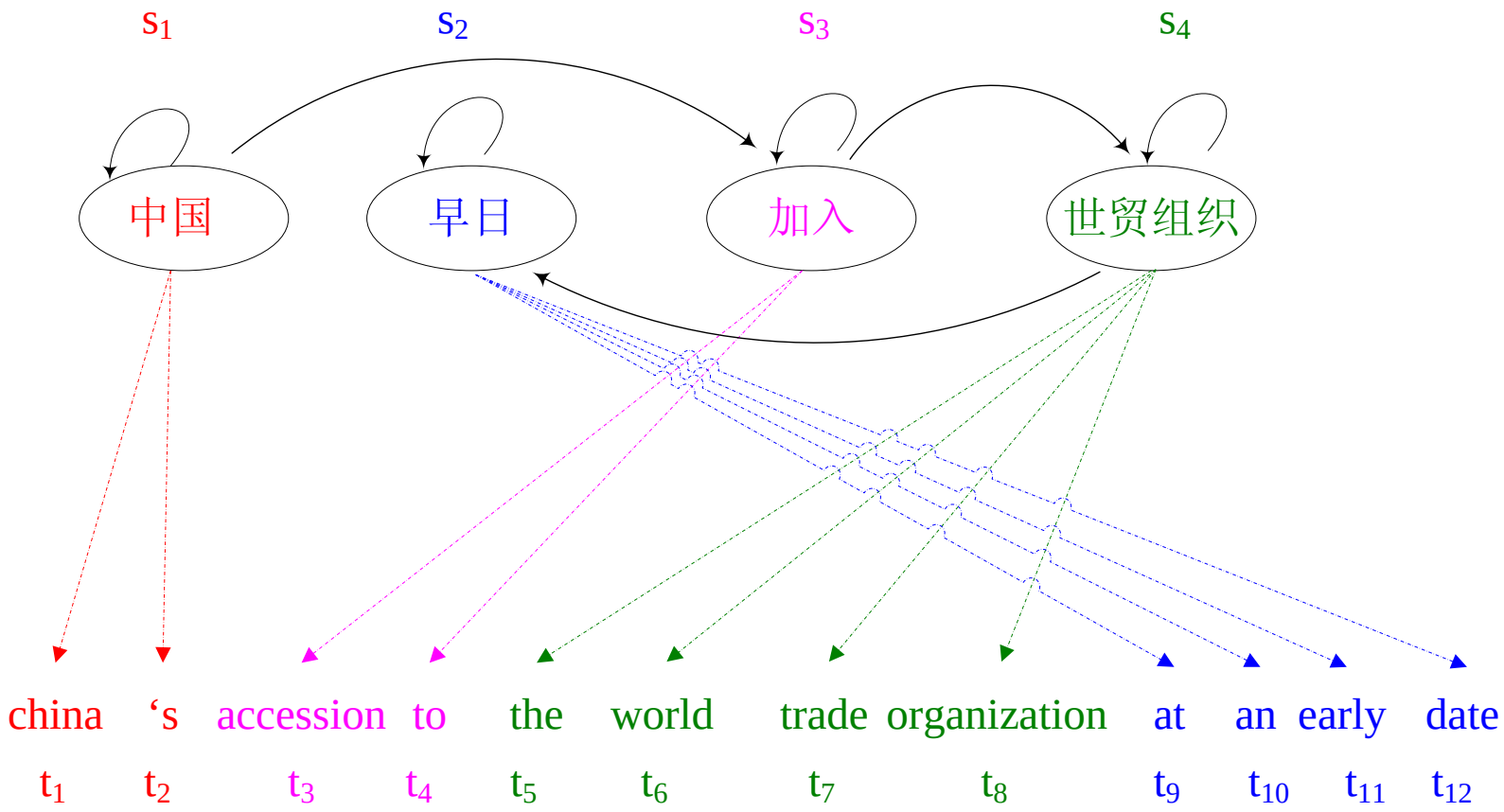
HMM-based Word Alignment Models (Vogel et al, '96)



china	's	accession	to	the	world	trade	organization	at	an	early	date
t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}

- State sequences ⟷ word alignments
- Model parameters: transition probabilities, translation tables
- Words are generated one by one, one transition emits one target word
- Efficient Balm-Welch and Viterbi algorithms

HMM-based Word Alignment Models (Vogel et al, '96)



- State sequences ⟷ word alignments
- Model parameters: transition probabilities, translation tables
- Words are generated one by one, one transition emits one target word
- Efficient Balm-Welch and Viterbi algorithms

IBM Model-4 Word Alignments (Brown et al, '93)

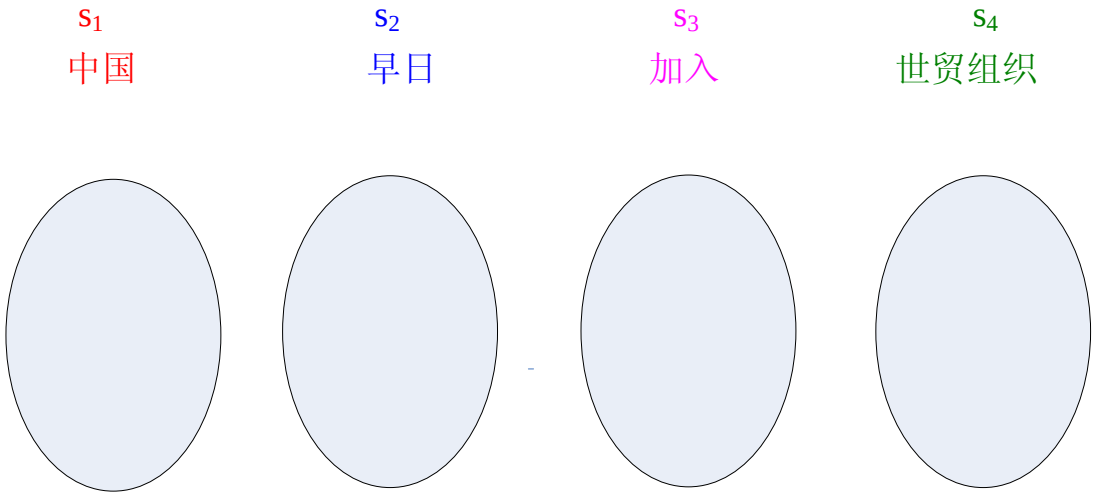
- Model-4 generated by GIZA++ Toolkit (Och and Ney, '03)

s_1	s_2	s_3	s_4
中国	早日	加入	世贸组织

- But
  - What makes the model powerful also makes computation complex
  - Exact-EM is problematic, sub-optimal estimation algorithms used
  - Difficult to compute statistics under the model

IBM Model-4 Word Alignments (Brown et al, '93)

- Model-4 generated by GIZA++ Toolkit (Och and Ney, '03)



Create a tablet for each source word

- But
  - What makes the model powerful also makes computation complex
  - Exact-EM is problematic, sub-optimal estimation algorithms used
  - Difficult to compute statistics under the model

IBM Model-4 Word Alignments (Brown et al, '93)

- Model-4 generated by GIZA++ Toolkit (Och and Ney, '03)

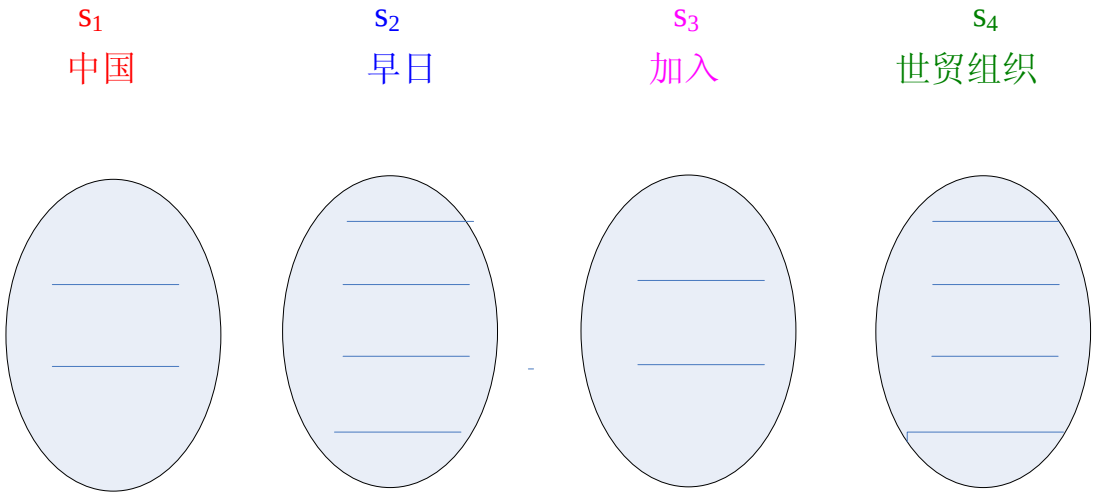
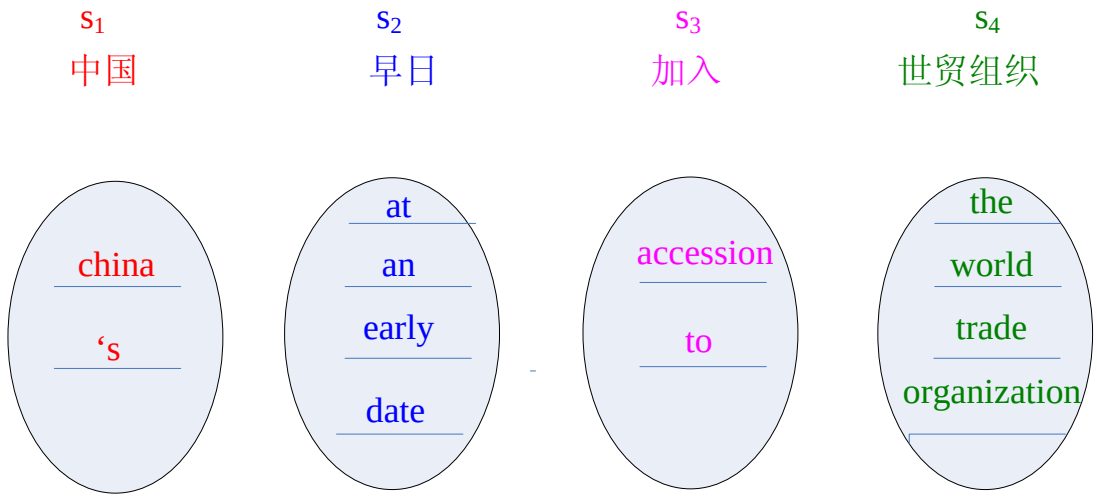


Table lookup to decide fertility: # of target words connected

- But
  - What makes the model powerful also makes computation complex
  - Exact-EM is problematic, sub-optimal estimation algorithms used
  - Difficult to compute statistics under the model

IBM Model-4 Word Alignments (Brown et al, '93)

- Model-4 generated by GIZA++ Toolkit (Och and Ney, '03)

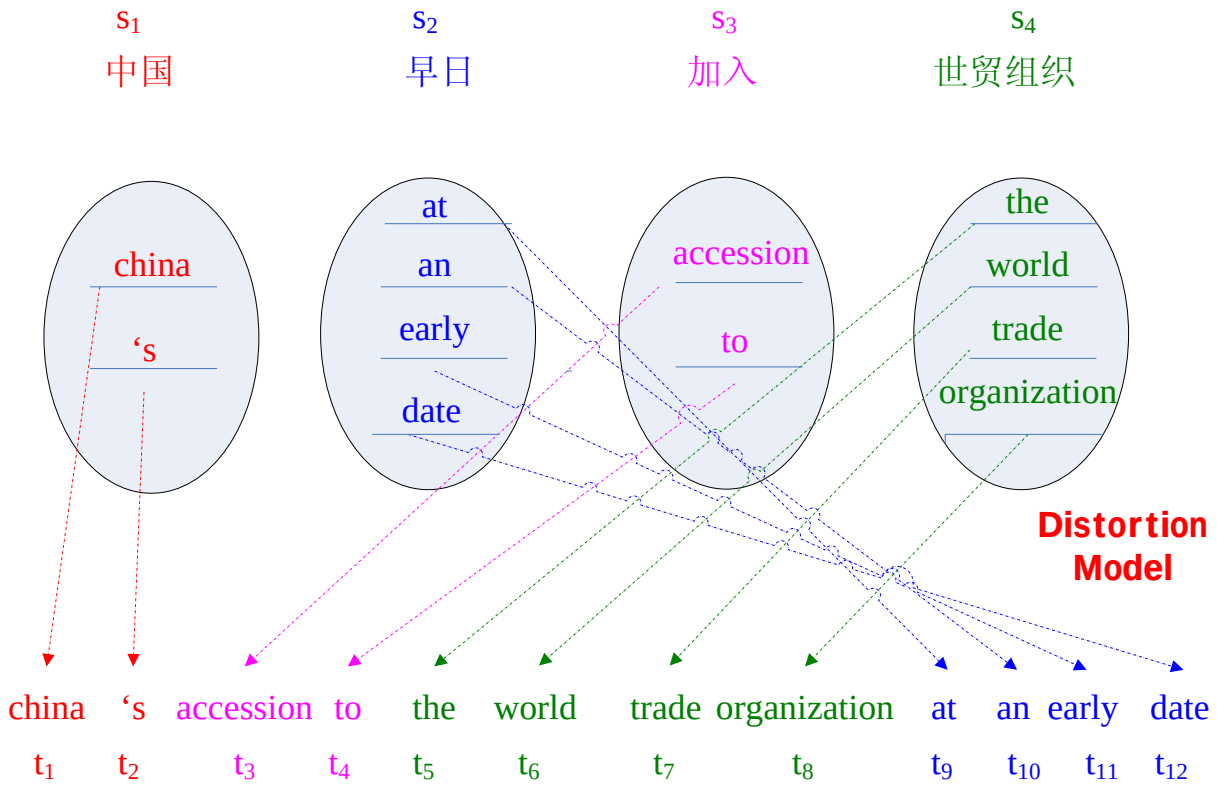


Sample target words from translation table i.i.d.

- But
  - What makes the model powerful also makes computation complex
  - Exact-EM is problematic, sub-optimal estimation algorithms used
  - Difficult to compute statistics under the model

IBM Model-4 Word Alignments (Brown et al, '93)

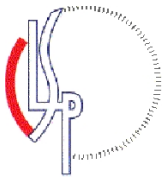
- Model-4 generated by GIZA++ Toolkit (Och and Ney, '03)



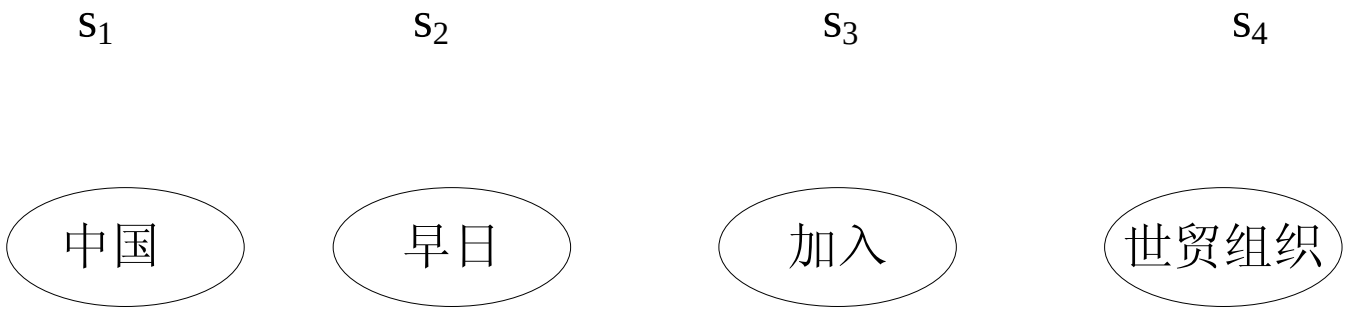
- But
  - What makes the model powerful also makes computation complex
  - Exact-EM is problematic, sub-optimal estimation algorithms used
  - Difficult to compute statistics under the model

Goal

- To find replacement for Model-4 word-to-word Alignment with
  - parameter estimation as efficient as HMM
    - retain Markov property, omit distortion
  - comparable performance to Model-4
    - incorporate fertility
  - efficient statistics computation
    - want to use models, not just the alignments



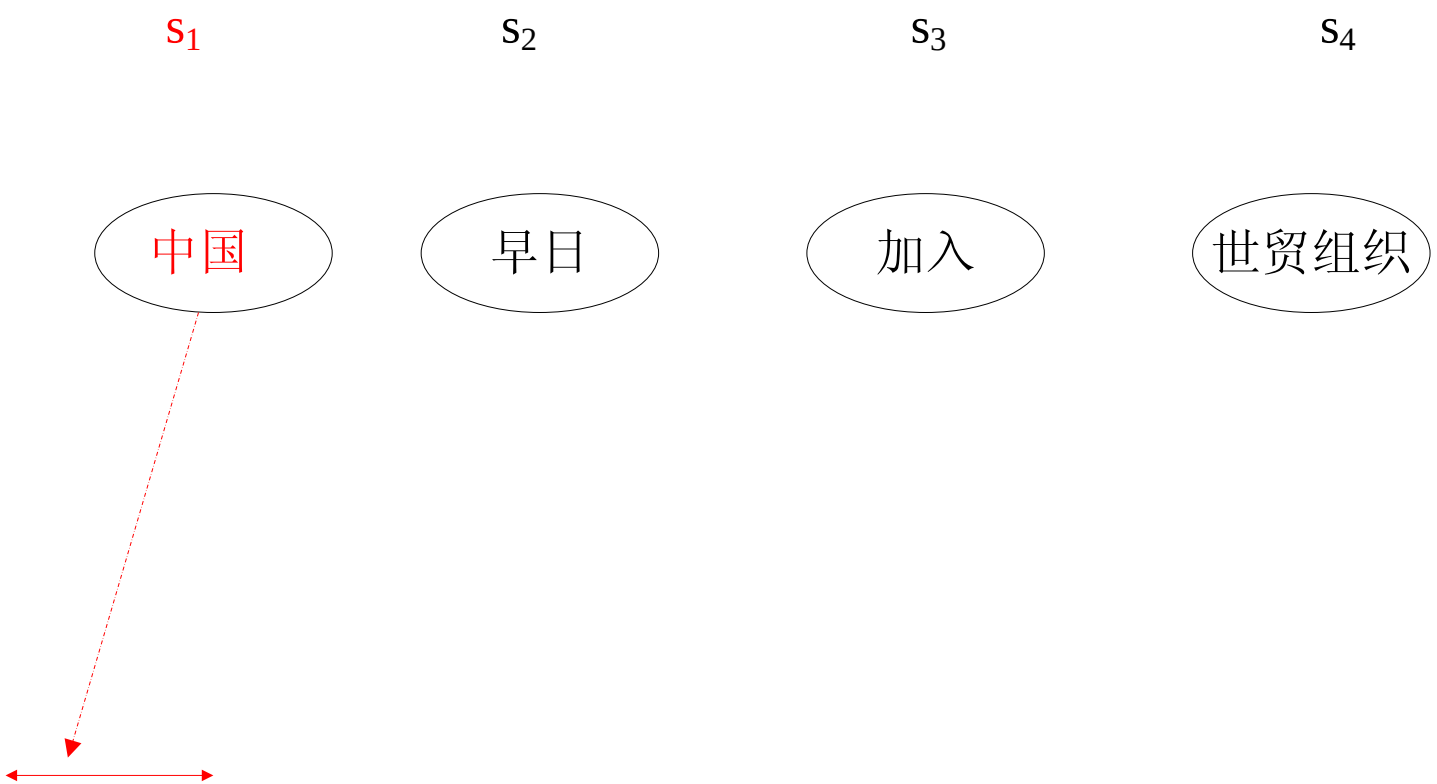
Make HMM More Powerful in Generating Observations



china	‘s	accession	to	the	world	trade	organization	at	an	early	date
t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}

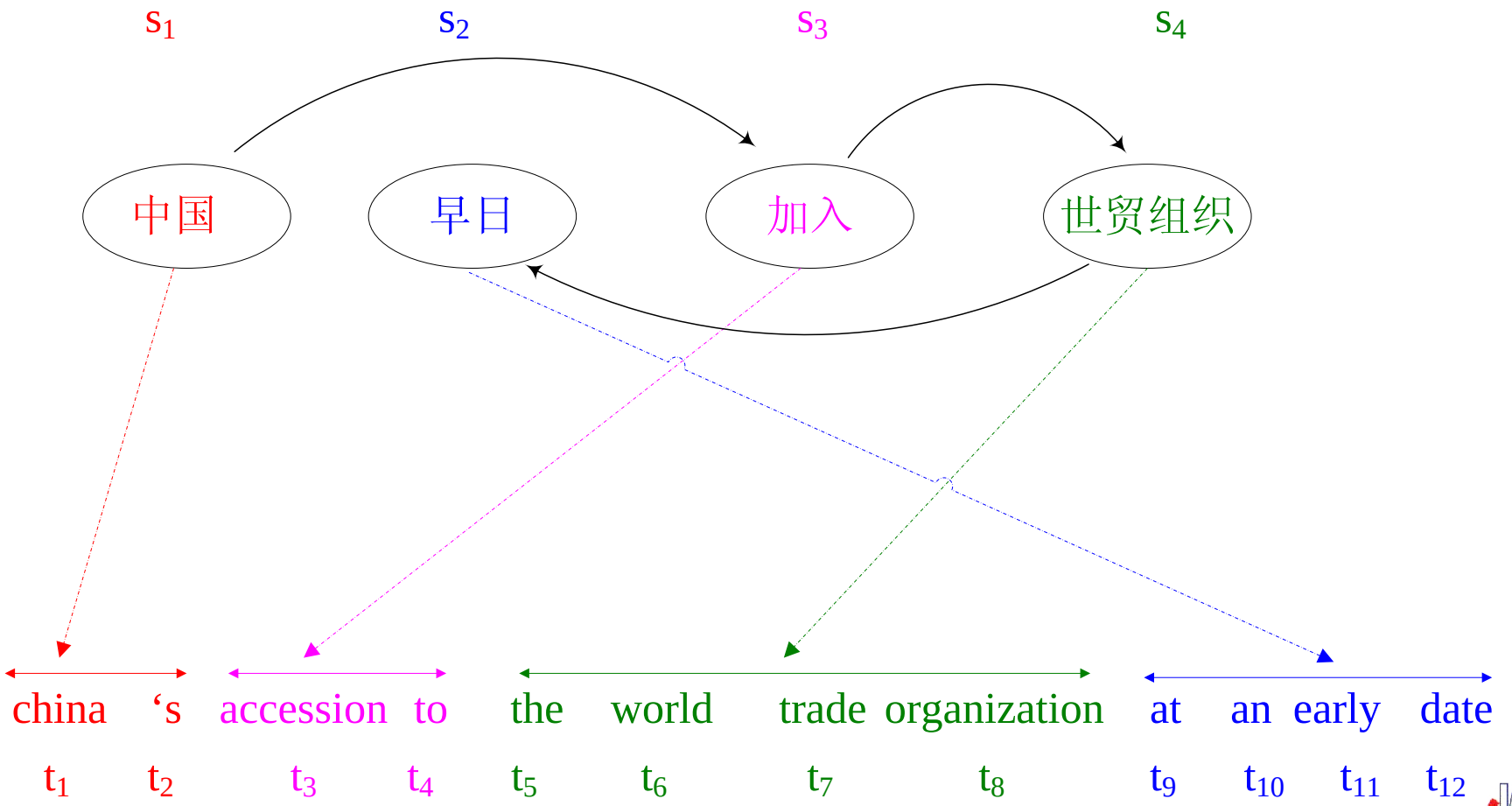
- Target phrases rather than words are emitted after jumping into a state
- Building links from source words to target phrases explicitly
- Can model dependencies between words within phrases

Make HMM More Powerful in Generating Observations



- Target phrases rather than words are emitted after jumping into a state
- Building links from source words to target phrases explicitly
- Can model dependencies between words within phrases

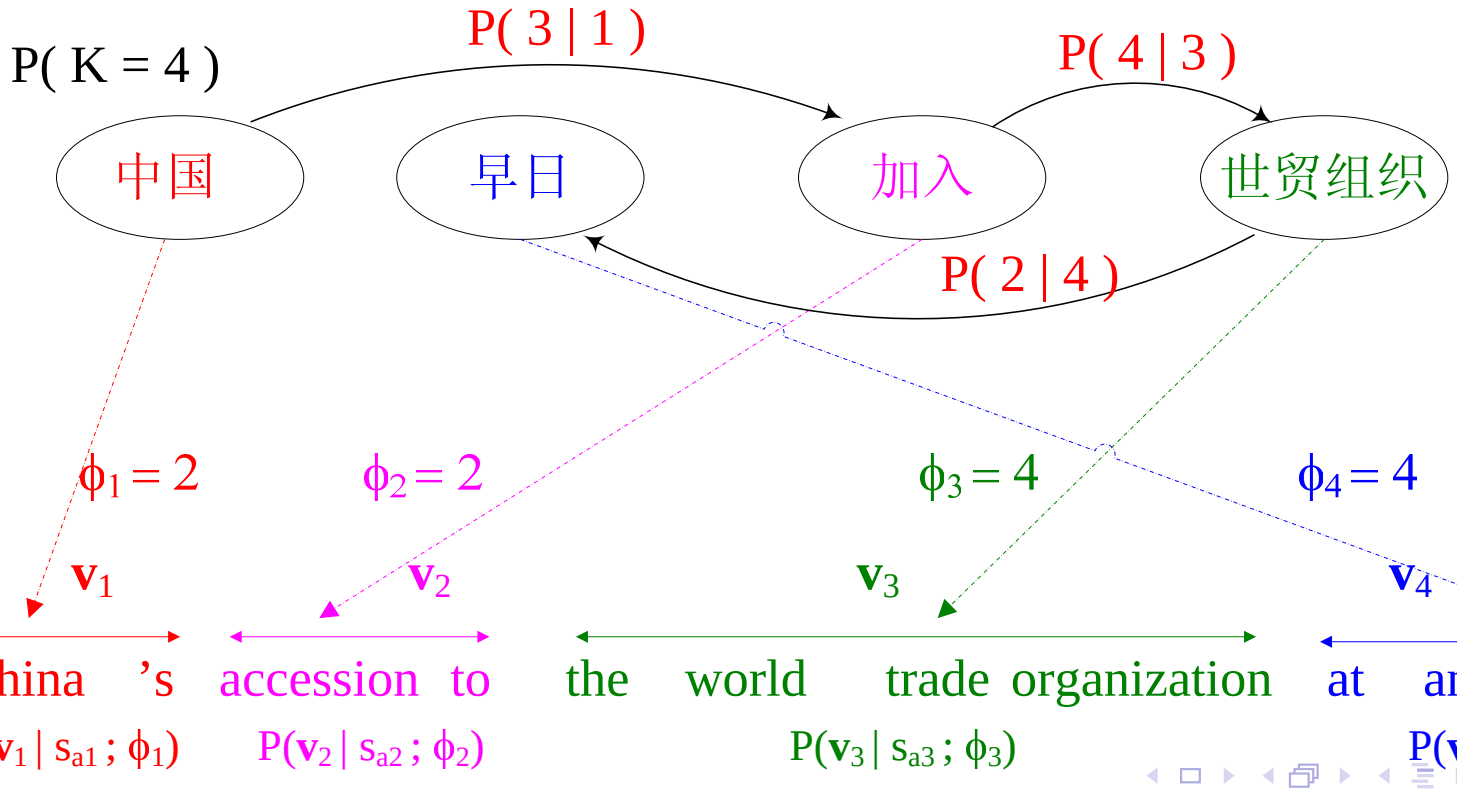
Make HMM More Powerful in Generating Observations



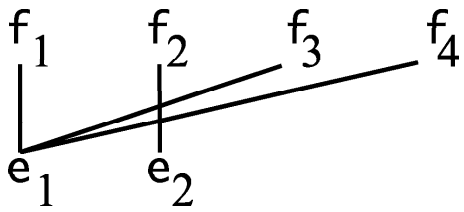
- Target phrases rather than words are emitted after jumping into a state
- Building links from source words to target phrases explicitly
- Can model dependencies between words within phrases

Word-to-Phrase HMM Alignment Models

- Word-to-phrase alignment **a** : $s_{a_j} \rightarrow \mathbf{v}_j$
- Phrase Count Model $P(K)$, depends on words
- Markov Transition $P(i|i')$
- Phrase Length Model $n(\phi;$ source word$)$, similar to fertility
- Word-to-Phrase Translation Probabilities $P(\mathbf{v}_j|s_{a_j};\phi)$

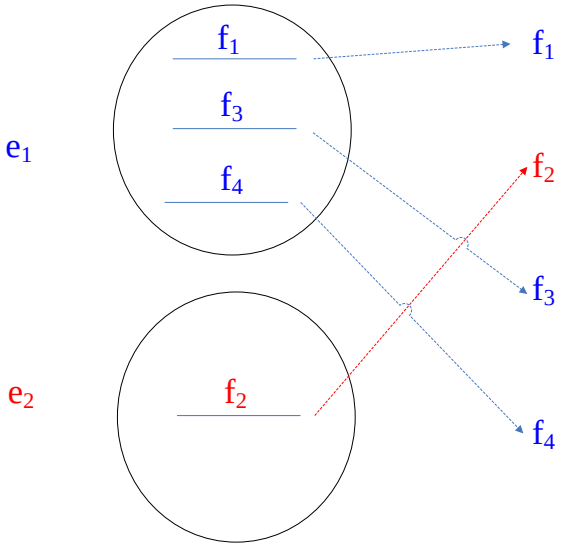


Word-to-Phrase HMM vs. IBM Model-4



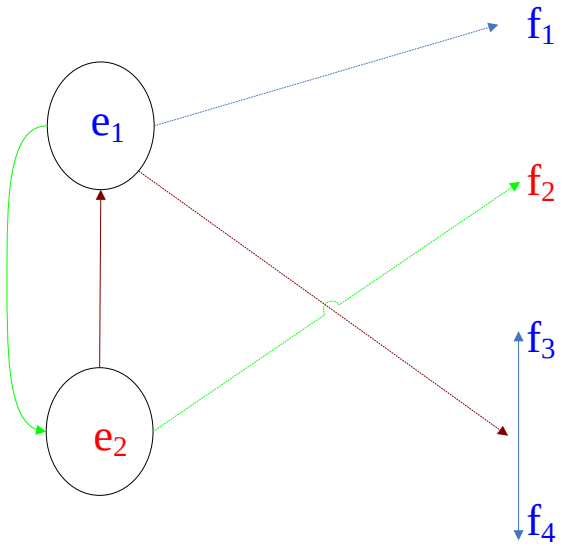
IBM Model-4

word fertility + distortion



WtoP HMM

phrase length model



Not exactly equivalent, but similar descriptive power

Inspired by features of Model-4

but incorporated within HMM

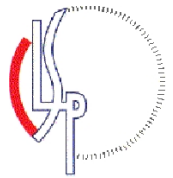
to allow efficient estimation and alignment

Word-to-Phrase Translation Probabilities

- Basic model components can be refined in various ways
- Replace weak i.i.d. word-by-word translation
- $P(\text{world trade organization}|\text{世贸组织}; 3) = ?$
  - $= t(\text{world}|\text{世贸组织}) \cdot t(\text{trade}|\text{世贸组织}) \cdot t(\text{organization}|\text{世贸组织})$
    $\longleftarrow$ i.i.d. t-table
  - $=$
    $t(\text{world}|\text{世贸组织}) \cdot t_2(\text{trade}|\text{world}, \text{世贸组织}) \cdot t_2(\text{organization}|\text{trade}, \text{世贸组织})$
    $\longleftarrow$ bigram t-table

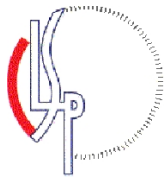
Model	i.i.d.	bigram
$P(\text{world}\|\text{世贸组织})$	0.06	0.06
$P(\text{trade}\|\text{world}, \text{世贸组织})$	0.06	0.99
$P(\text{organization}\|\text{trade}, \text{世贸组织})$	0.06	0.99
$P(\text{world trade organization}\|\text{世贸组织}, 3)$	0.0002	0.0588

  - Assigns higher probability to correct translation than i.i.d
  - Incorporates context without losing algorithmic efficiency
    - DP still possible
  - Use same estimation techniques as used for bigram LMs
  - Data sparseness, Witten-Bell smoothing



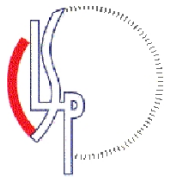
Embedded Estimation of Word-to-Phrase HMM

- Forward-Backward Procedure
  - Incremental build (experience from ASR)
  - Pruning can be applied
- Unsupervised training from scratch
  - Model-1, 10 its (initial t-table)
  - Model-2, 5 its (better t-table)
  - WtoW HMM, 5 its (initial Markov model)
  - WtoP HMM N=2, 3, .., each 5 its (Markov model, phrase length)
  - WtoP HMM with bigram t-table, 5 its (bigram t-table)
- Parallel Implementation
  - Partitioning training bitext
  - E-step: Collect counts from each partition parallel
  - M-step: Merge counts to update model parameters
  - Memory efficient, virtually no limitation on training bitext size



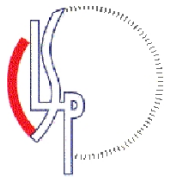
Embedded Estimation of Word-to-Phrase HMM

- Forward-Backward Procedure
  - Incremental build (experience from ASR)
  - Pruning can be applied
- Unsupervised training from scratch
  - Model-1, 10 its (initial t-table)
  - Model-2, 5 its (better t-table)
  - WtoW HMM, 5 its (initial Markov model)
  - WtoP HMM N=2, 3, .., each 5 its (Markov model, phrase length)
  - WtoP HMM with bigram t-table, 5 its (bigram t-table)
- Parallel Implementation
  - Partitioning training bitext
  - E-step: Collect counts from each partition parallel
  - M-step: Merge counts to update model parameters
  - Memory efficient, virtually no limitation on training bitext size



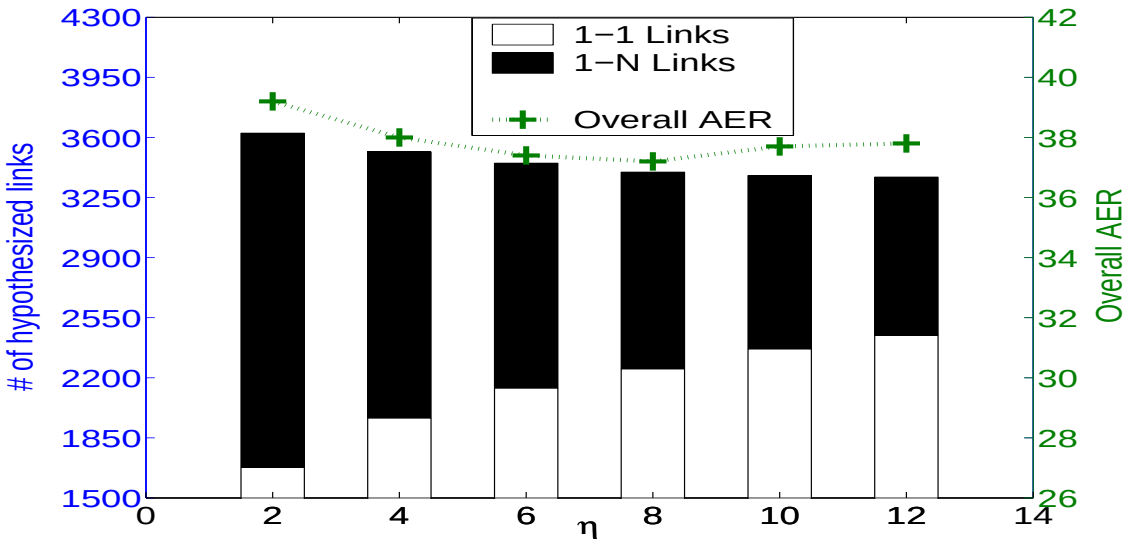
Embedded Estimation of Word-to-Phrase HMM

- Forward-Backward Procedure
  - Incremental build (experience from ASR)
  - Pruning can be applied
- Unsupervised training from scratch
  - Model-1, 10 its (initial t-table)
  - Model-2, 5 its (better t-table)
  - WtoW HMM, 5 its (initial Markov model)
  - WtoP HMM N=2, 3, .., each 5 its (Markov model, phrase length)
  - WtoP HMM with bigram t-table, 5 its (bigram t-table)
- Parallel Implementation
  - Partitioning training bitext
  - E-step: Collect counts from each partition parallel
  - M-step: Merge counts to update model parameters
  - Memory efficient, virtually no limitation on training bitext size



Bitext Alignment Results

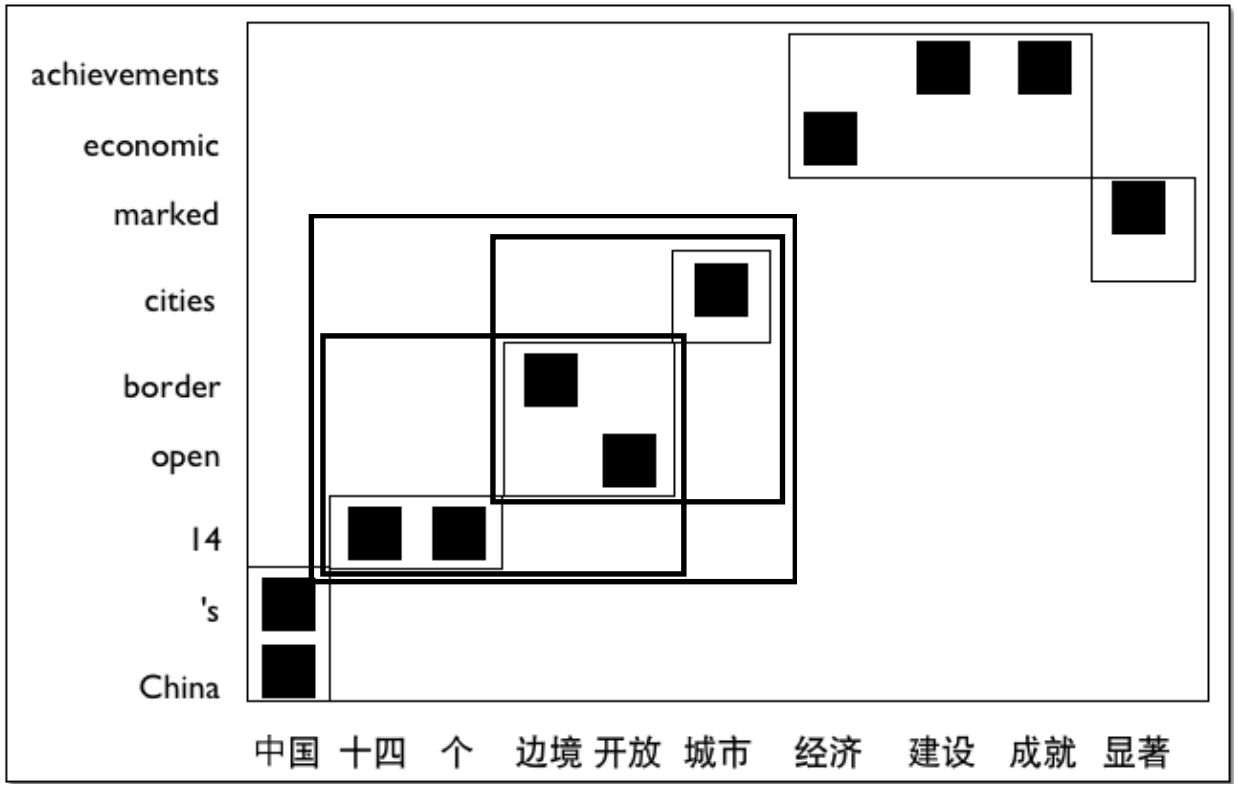
- Test: NIST 2001 MT-eval set, 124 sentence pairs w/ manual word alignments
- Comparable performance to Model-4 on FBIS training bitext
- Increasing max phrase length N improves quality in $C \rightarrow E$ direction
- Bigram translation probability improves word-to-phrase links
- A good balance between 1-1 and 1-N distribution can be achieved



- Comparable performance when extending to large scale bitexts

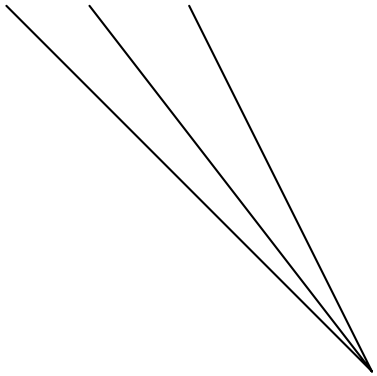
Statistical Phrase Translation Models

- Phrase-based SMT performs better than word-based SMT
- Phrases Pair Inventory (PPI) extracted from word aligned bitext (Och et al, '99)



But word alignments are imperfect ...

There is no **gang and money linked** **politics** in hong kong and there will not be such **politics** in future either

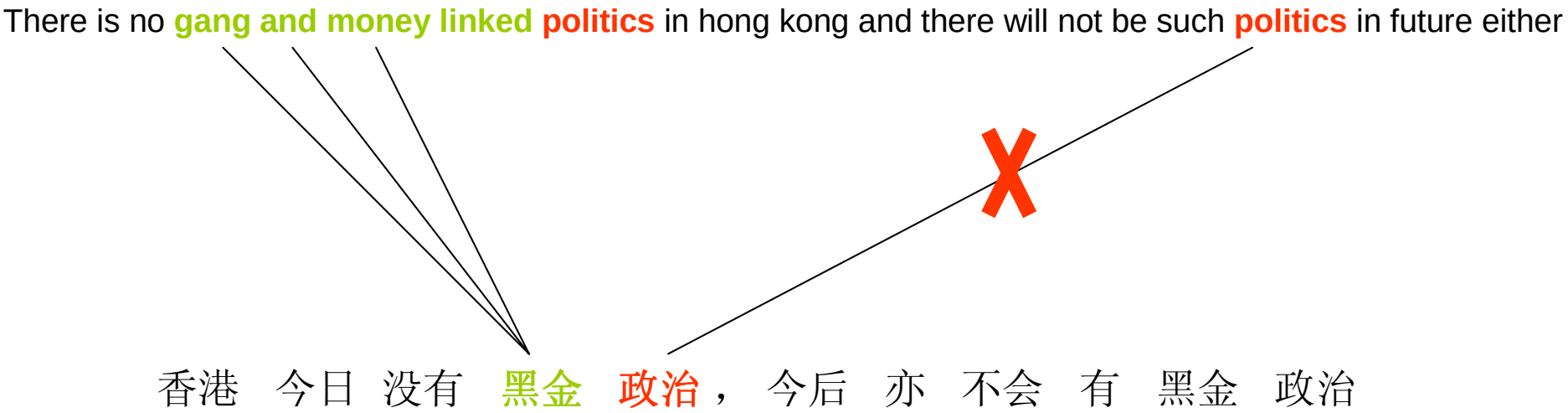


?

香港 今日 没有 黑金 政治， 今后 亦 不会 有 黑金 政治

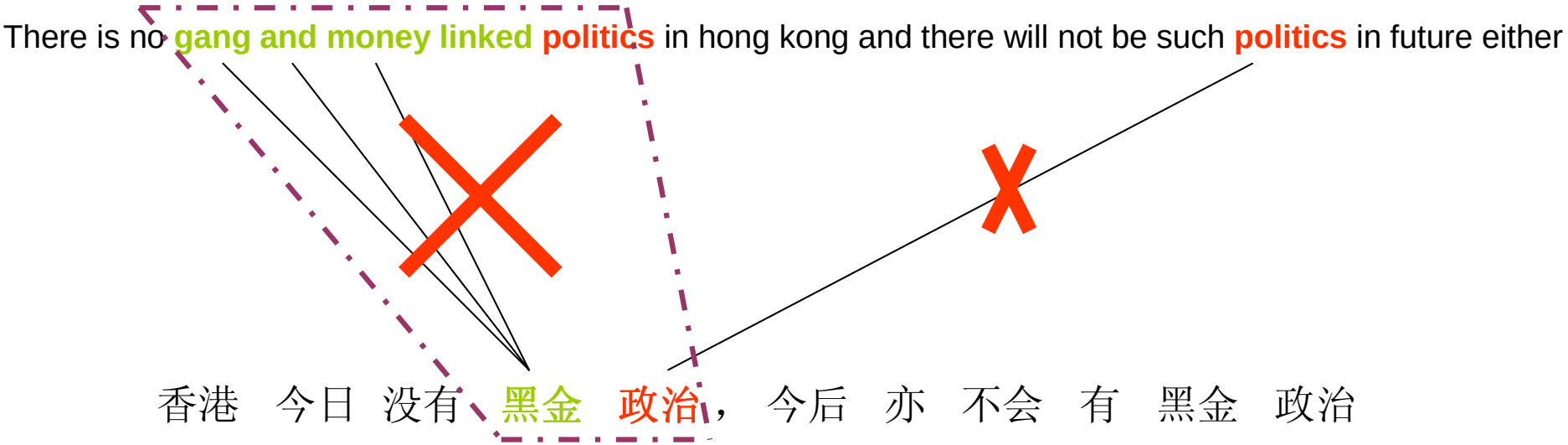
- Relying on the one-best word alignment may exclude some valid phrase pairs
- Goal is to define a probability distribution over phrase pairs
  - Allows more control over generation of phrase pairs

But word alignments are imperfect ...



- Relying on the one-best word alignment may exclude some valid phrase pairs
- Goal is to define a probability distribution over phrase pairs
  - Allows more control over generation of phrase pairs

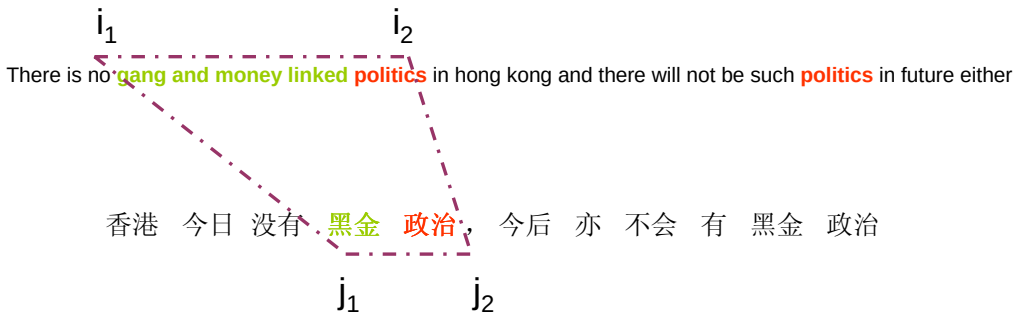
But word alignments are imperfect ...



- Relying on the one-best word alignment may exclude some valid phrase pairs
- Goal is to define a probability distribution over phrase pairs
  - Allows more control over generation of phrase pairs

Model-based Phrase Pair Posterior

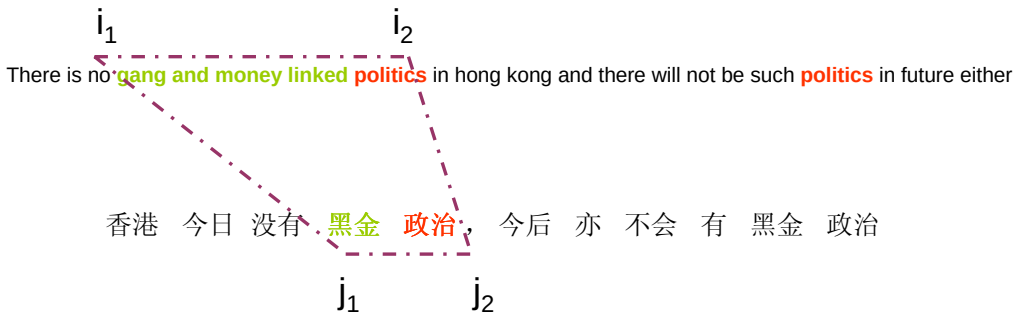
- Doesn't rely on a single alignment



- Define a set of alignments that align words to words in phrases
  $A(i_1, i_2; j_1, j_2) = \{a_1^m : a_j \in [i_1, i_2] \text{ iff } j \in [j_1, j_2]\}$
- Calculate the likelihood of the source phrase producing the target phrase
  $P(\mathbf{T}, A(i_1, i_2; j_1, j_2) \,|\, \mathbf{S}) = \sum_{\mathbf{a} \,:\, a_1^m \in A(i_1, i_2; j_1, j_2)} P(\mathbf{T}, \mathbf{a} | \mathbf{S})$
- Obtain phrase pair posterior
  $P(A(i_1, i_2; j_1, j_2) \,|\mathbf{T}, \mathbf{S}) = P(\mathbf{T}, A(i_1, i_2; j_1, j_2) \,|\, \mathbf{S}) / P(\mathbf{T} | \mathbf{S})$
- Efficient DP-based implementation for WtoP HMM, Difficult for Model-4

Model-based Phrase Pair Posterior

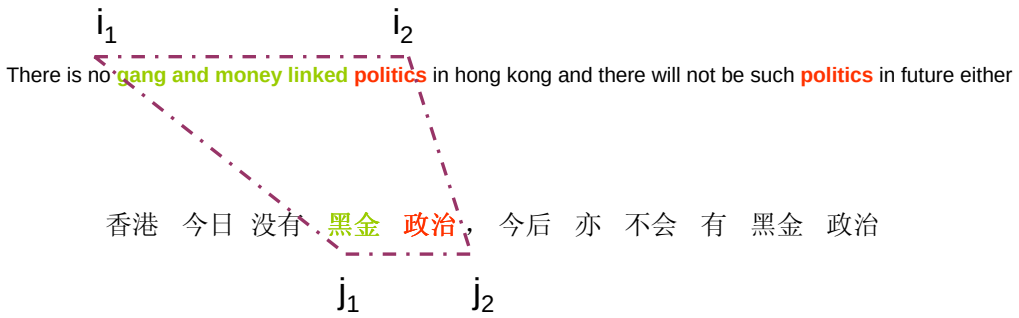
- Doesn't rely on a single alignment



- Define a set of alignments that align words to words in phrases
  $A(i_1, i_2; j_1, j_2) = \{a_1^m : a_j \in [i_1, i_2] \text{ iff } j \in [j_1, j_2]\}$
- Calculate the likelihood of the source phrase producing the target phrase
  $P(\mathbf{T}, A(i_1, i_2; j_1, j_2) \,|\, \mathbf{S}) = \sum_{\mathbf{a}\,:\, a_1^m \in A(i_1, i_2; j_1, j_2)} P(\mathbf{T}, \mathbf{a} | \mathbf{S})$
- Obtain phrase pair posterior
  $P(A(i_1, i_2; j_1, j_2) \,|\mathbf{T}, \mathbf{S}) = P(\mathbf{T}, A(i_1, i_2; j_1, j_2) \,|\, \mathbf{S}) / P(\mathbf{T} | \mathbf{S})$
- Efficient DP-based implementation for WtoP HMM, Difficult for Model-4

Model-based Phrase Pair Posterior

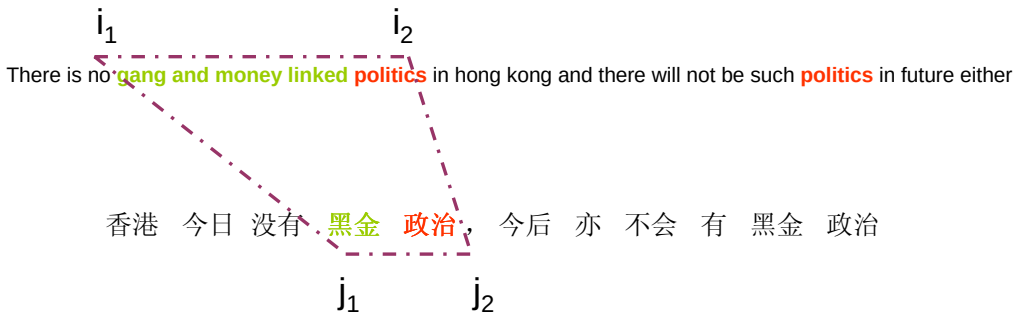
- Doesn't rely on a single alignment



- Define a set of alignments that align words to words in phrases
  $A(i_1, i_2; j_1, j_2) = \{a_1^m : a_j \in [i_1, i_2] \text{ iff } j \in [j_1, j_2]\}$
- Calculate the likelihood of the source phrase producing the target phrase
  $P(\mathbf{T}, A(i_1, i_2; j_1, j_2) \,|\, \mathbf{S}) = \sum_{\mathbf{a} \,:\, a_1^m \in A(i_1, i_2; j_1, j_2)} P(\mathbf{T}, \mathbf{a} | \mathbf{S})$
- Obtain phrase pair posterior
  $P(A(i_1, i_2; j_1, j_2) \,|\mathbf{T}, \mathbf{S}) = P(\mathbf{T}, A(i_1, i_2; j_1, j_2) \,|\, \mathbf{S}) / P(\mathbf{T}|\mathbf{S})$
- Efficient DP-based implementation for WtoP HMM, Difficult for Model-4

Model-based Phrase Pair Posterior

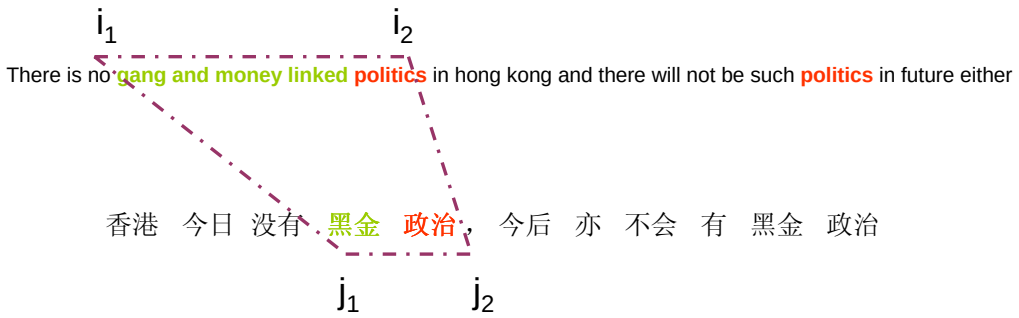
- Doesn't rely on a single alignment



- Define a set of alignments that align words to words in phrases
  $A(i_1, i_2; j_1, j_2) = \{a_1^m : a_j \in [i_1, i_2] \text{ iff } j \in [j_1, j_2]\}$
- Calculate the likelihood of the source phrase producing the target phrase
  $P(\mathbf{T}, A(i_1, i_2; j_1, j_2) \,|\, \mathbf{S}) = \sum_{\mathbf{a}\,:\, a_1^m \in A(i_1, i_2; j_1, j_2)} P(\mathbf{T}, \mathbf{a} | \mathbf{S})$
- Obtain phrase pair posterior
  $P(A(i_1, i_2; j_1, j_2) \,|\mathbf{T}, \mathbf{S}) = P(\mathbf{T}, A(i_1, i_2; j_1, j_2) \,|\, \mathbf{S}) / P(\mathbf{T} | \mathbf{S})$
- Efficient DP-based implementation for WtoP HMM, Difficult for Model-4

Model-based Phrase Pair Posterior

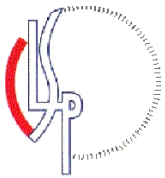
- Doesn't rely on a single alignment



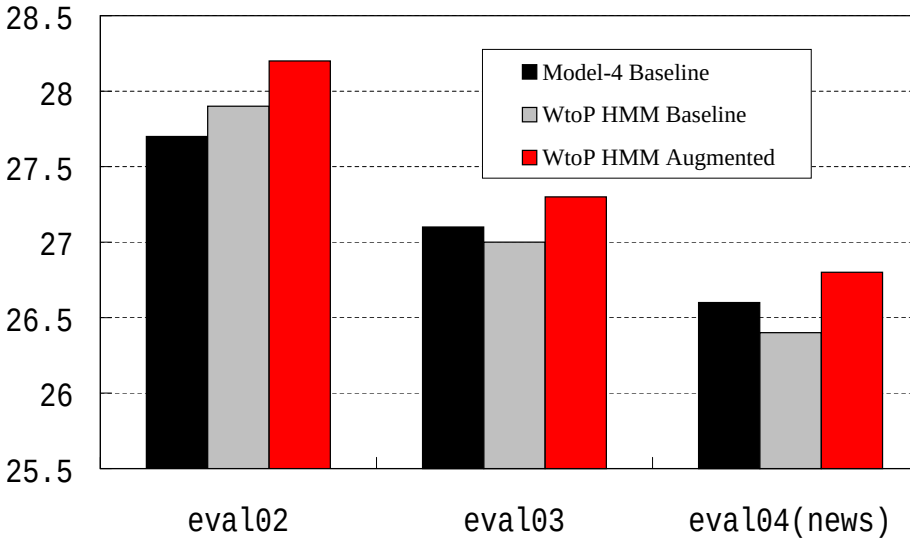
- Define a set of alignments that align words to words in phrases
  $A(i_1, i_2; j_1, j_2) = \{a_1^m : a_j \in [i_1, i_2] \text{ iff } j \in [j_1, j_2]\}$
- Calculate the likelihood of the source phrase producing the target phrase
  $P(\mathbf{T}, A(i_1, i_2; j_1, j_2) \,|\, \mathbf{S}) = \sum_{\mathbf{a}\,:\, a_1^m \in A(i_1, i_2; j_1, j_2)} P(\mathbf{T}, \mathbf{a} | \mathbf{S})$
- Obtain phrase pair posterior
  $P(A(i_1, i_2; j_1, j_2) \,|\mathbf{T}, \mathbf{S}) = P(\mathbf{T}, A(i_1, i_2; j_1, j_2) \,|\, \mathbf{S}) / P(\mathbf{T} | \mathbf{S})$
- Efficient DP-based implementation for WtoP HMM, Difficult for Model-4

Augmented PPI for a Better Coverage

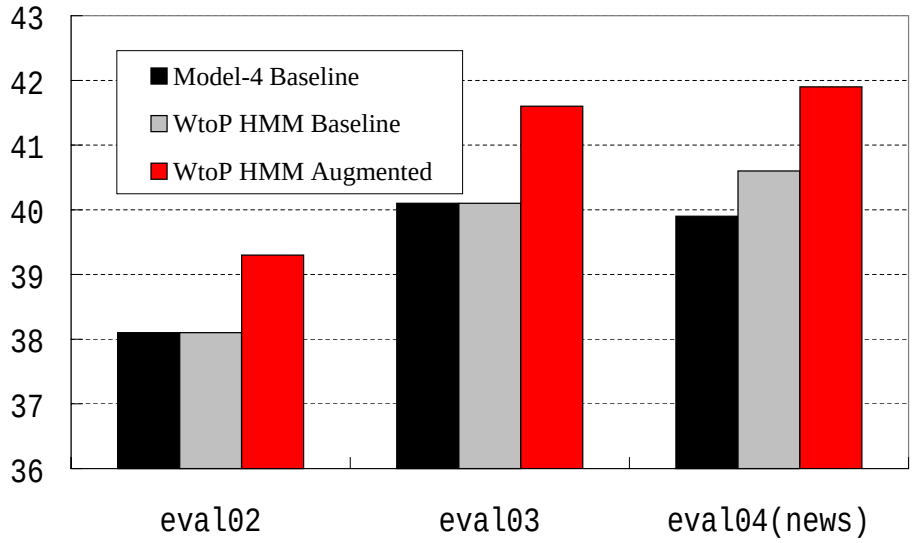
- Baseline PPI
  - extracted from 1-best alignments using establishing techniques (Och et al., '99)
- Goal: add phrase pairs to improve test set coverage
- For each foreign phrase **v** in test set not covered by the baseline
  - for each sentence pair containing **v**
  - find the English phrase **u** that maximizes the phrase pair posterior
  - add $(\mathbf{u}, \mathbf{v})$ to the baseline PPI if posterior exceeds a threshold value
- Balance coverage against phrase translation quality



Translation Results



Chinese-English



Arabic-English

- TTM Decoder - WFST implementation with monotone order (Kumar et al, '05)
- Used all parallel corpora available from LDC
  - C-E: 200M En. words (FBIS, Xinhua, HK News, ..., all UN bitexts)
  - A-E: 130M En. words (news, all UN bitexts)
- Relaxing threshold in PPI augmenting improves coverage and BLEU score
- WtoP model can even be applied to augment Model-4 PPI (see the paper)

Conclusions

- The word-to-phrase HMM alignment model
  - is capable of good quality bitext word alignment over very large training bitexts
  - has efficient training algorithm with parallel implementation
- Model-based phrase pair distribution enables
  - an improved phrase pair extraction strategy
  - controlled balance coverage vs. quality
- WtoP HMM performs
  - slightly better than Model-4 on large C-E translation system
  - significantly better than Model-4 on large A-E translation system
- WtoP HMM is a powerful framework for exploring new models of word and phrase translation