Bitext Alignment for Statistical Machine Translation

Yonggang Deng
Advisor: Prof. William Byrne
Thesis Committee: Prof. William Byrne, Prof. Trac Tran
Prof. Jerry Prince and Prof. Gerard Meyer

Center for Language and Speech Processing
The Johns Hopkins University
Baltimore, MD 21218

20th December 2005

Bitext and Bitext Alignment

- Bitext: a collection of text in two languages
- Bitext Alignment: finding translation equivalence within bitext

Chinese	English
要 做 好 河 湖 清 障 工作 ，对 各种 河 湖 障碍 ，坚决 予以 清除 .	It is necessary to resolutely remove obstacles in rivers and lakes .
四 、 加强 监测 预报 ，科学 调度 .	4 . It is necessary to strengthen monitoring and forecast work and scientifically dispatch people and materials .
要 采取 有效 措施 ，千方百计 提高 预报 精度 .	It is necessary to take effective measures and try by every possible means to provide precision forecast .
汛 前 要 抓紧 修订 洪水 预报 方案 ，有 针对性 地 开展工作 .	Before the flood season comes , it is necessary to seize the time to formulate plans for forecasting floods and to carry out work with clear

Bitext and Bitext Alignment

- Bitext: a collection of text in two languages
- Bitext Alignment: finding translation equivalence within bitext

要 做 好 河 湖 清 障 工作 ，对 各种 河 湖 障碍 ，坚决 予以 清除 .	It is necessary to resolutely remove obstacles in rivers and lakes .
四 、 加强 监测 预报 ，科学 调度 .	4 . It is necessary to strengthen monitoring and forecast work and scientifically dispatch people and materials .
要 采取 有效 措施 ，千方百计 提高 预 报 精度 .	It is necessary to take effective measures and try by every possible means to provide precision forecast .
汛 前 要 抓紧 修订 洪水 预报 方案 ，有 针对性 地 开展工作 .	Before the flood season comes , it is necessary to seize the time to formulate plans for forecasting floods and to carry out work with clear
Chinese	English

- Bitext: a collection of text in two languages
- Bitext Alignment: finding translation equivalence within bitext

Chinese	English
要 做 好 河 湖 清 障 工作 , 对 各种 河 湖 障碍 , 坚决 予以 清除 .	It is necessary to resolutely remove obstacles in rivers and lakes .
四 、 加强 监测 预报 , 科学 调度	4 . It is necessary to strengthen monitoring and forecast work and scientifically dispatch people and materials .
要 采取 有效 措施 , 千方百计 提高 预 报 精度 .	It is necessary to take effective measures and try by every possible means to provide precision forecast .
汛 前 要 抓紧 修订 洪水 预报 方案 , 有 针对性 地 开展工作 .	Before the flood season comes , it is necessary to seize the time to formulate plans for forecasting floods and to carry out work with clear

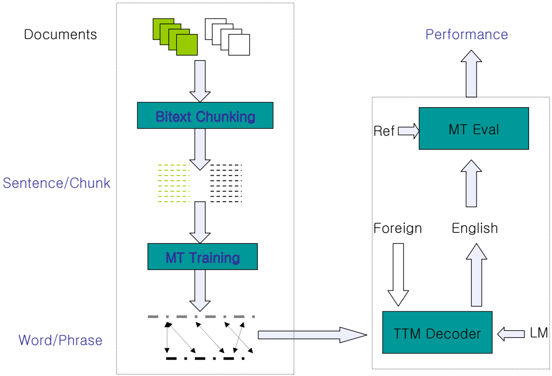
Why automatic bitext alignment?

- Critical and beneficial in many multilingual NLP tasks
  - provides basic ingredients in building a Machine Translation system
- Hand alignment is expensive for large corpora
- Desired properties
  - language independent: Chinese, Arabic, Spanish, French ...
  - no linguistic knowledge: from scratch, unsupervised, statistical
  - huge amount of data: effectiveness and efficiency

Statistical Machine Translation (SMT)

Source	$\longrightarrow$ Channel $\longrightarrow$	Target	Source Decoding
E	$P(F\|E)$	F	$\hat{E} = \mathrm{argmax}_E P(E)P(F\|E)$

Translation Model $P(F|E)$ needs BITEXTs



Outline

1. Bitext Chunk Alignment
   - Chunk Alignment
   - Chunking Algorithms
   - Sentence Alignment Results
2. Bitext Word Alignment
   - Introduction/Motivation
   - Word-to-Phrase HMM Model
   - Parameter Estimation
   - Word Alignment Results
3. Bitext Phrase Alignment
   - Inducing from Word Alignments
   - Model-based Phrase Pair Posterior
   - Translation Results
4. Conclusions

Outline

1. Bitext Chunk Alignment
   - Chunk Alignment
   - Chunking Algorithms
   - Sentence Alignment Results
2. Bitext Word Alignment
   - Introduction/Motivation
   - Word-to-Phrase HMM Model
   - Parameter Estimation
   - Word Alignment Results
3. Bitext Phrase Alignment
   - Inducing from Word Alignments
   - Model-based Phrase Pair Posterior
   - Translation Results
4. Conclusions

Chunk Alignment

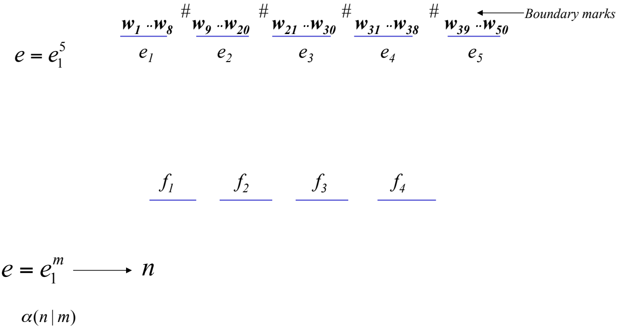
- Problem: sentences are not translated 1-to-1 in sequence
  - 1-to-n, n-to-1, m-to-n, order changes, real data challenge
- A Statistical Generative Chunk Alignment Model (Deng et al, ’04)
  - introduce a hidden chunk alignment variable
  - document generating: fill in the blank
  - two alignment algorithms are derived in a straightforward manor

		#		#		#		#		
	$w_1 \cdot\cdot w_8$		$w_9 \cdot\cdot w_{20}$		$w_{21} \cdot\cdot w_{30}$		$w_{31} \cdot\cdot w_{38}$		$w_{39} \cdot\cdot w_{50}$	← *Boundary marks*
$e = e_1^5$	e_1		e_2		e_3		e_4		e_5	

$$e = e_1^m$$

Chunk Alignment

- Problem: sentences are not translated 1-to-1 in sequence
  - 1-to-n, n-to-1, m-to-n, order changes, real data challenge
- A Statistical Generative Chunk Alignment Model (Deng et al, '04)
  - introduce a hidden chunk alignment variable
  - document generating: fill in the blank
  - two alignment algorithms are derived in a straightforward manor



Chunk Alignment

- Problem: sentences are not translated 1-to-1 in sequence
  - 1-to-n, n-to-1, m-to-n, order changes, real data challenge
- A Statistical Generative Chunk Alignment Model (Deng et al, ’04)
  - introduce a hidden chunk alignment variable
  - document generating: fill in the blank
  - two alignment algorithms are derived in a straightforward manor

$\#$ $\#$ $\#$ $\#$ ← *Boundary marks*

$$e = e_1^5 \quad \underbrace{w_1 \cdots w_8}_{e_1} \; \# \; \underbrace{w_9 \cdots w_{20}}_{e_2} \; \# \; \underbrace{w_{21} \cdots w_{30}}_{e_3} \; \# \; \underbrace{w_{31} \cdots w_{38}}_{e_4} \; \# \; \underbrace{w_{39} \cdots w_{50}}_{e_5}$$

$K=3$

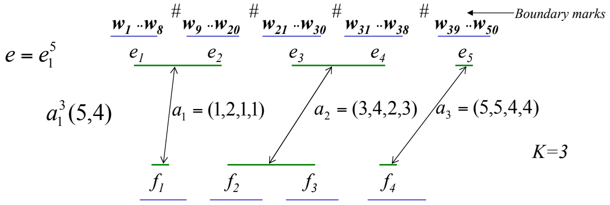
f_1 f_2 f_3 f_4

$$e = e_1^m \longrightarrow n \longrightarrow K$$

$$\alpha(n \mid m)\beta(K \mid m, n)$$

Chunk Alignment

- Problem: sentences are not translated 1-to-1 in sequence
  - 1-to-n, n-to-1, m-to-n, order changes, real data challenge
- A Statistical Generative Chunk Alignment Model (Deng et al, ’04)
  - introduce a hidden chunk alignment variable
  - document generating: fill in the blank
  - two alignment algorithms are derived in a straightforward manor

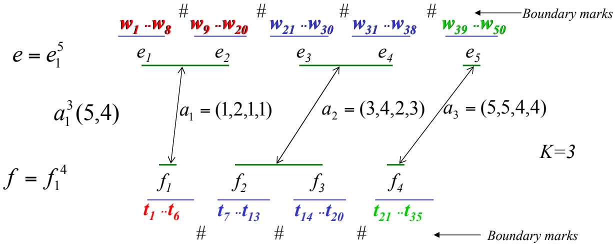


$$e = e_1^m \longrightarrow n \longrightarrow K \longrightarrow a_1^K(m,n)$$

$$\alpha(n \mid m)\beta(K \mid m,n)\; P(a_1^K \mid m,n,K)$$

Chunk Alignment

- Problem: sentences are not translated 1-to-1 in sequence
  - 1-to-n, n-to-1, m-to-n, order changes, real data challenge
- A Statistical Generative Chunk Alignment Model (Deng et al, ’04)
  - introduce a hidden chunk alignment variable
  - document generating: fill in the blank
  - two alignment algorithms are derived in a straightforward manor



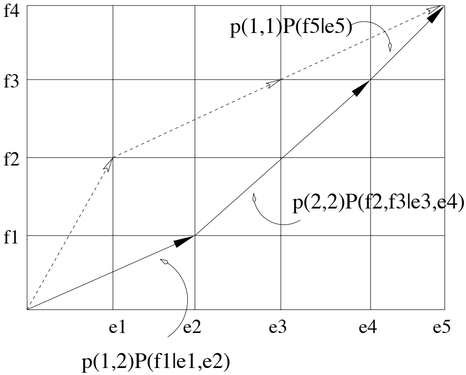
$$e = e_1^m \longrightarrow n \longrightarrow K \longrightarrow a_1^K(m,n) \longrightarrow f_1^n = f$$

$$\alpha(n\,|\,m)\beta(K\,|\,m,n)\,P(a_1^K\,|\,m,n,K)\prod_k P^{(w)}(f(a_k)\,|\,e(a_k)) = P(f_1^n, K, a_1^K\,|\,e_1^m)$$

Outline

1. Bitext Chunk Alignment
   - Chunk Alignment
   - Chunking Algorithms
   - Sentence Alignment Results
2. Bitext Word Alignment
   - Introduction/Motivation
   - Word-to-Phrase HMM Model
   - Parameter Estimation
   - Word Alignment Results
3. Bitext Phrase Alignment
   - Inducing from Word Alignments
   - Model-based Phrase Pair Posterior
   - Translation Results
4. Conclusions

Dynamic Programming (DP)



- Monotone chunk alignment
- Global optimum

Divisive Clustering (DC)

divide and conquer, iterative binary parallel splitting, reorder

自从 朝鲜半岛 被 分裂 成 两个 国家 以来 ， 韩国 在 背 靠 美国 这 棵 大 树 以求 自 安 的 同时 ， 还 小心翼翼 但 却 坚持不懈 地 向 美国 寻求 先进武器 ， 以 抗衡 朝鲜 。

据 汉城 的 消息灵通人士 向 《 华盛顿邮报 》 透露 ， 今年 早些时候 ， 美国 已 秘 而 不 宣 地 同意 韩国 “ 可以 扩展 它 现有 导弹 的 射程 ” ， 使 之 能够 直捣 朝鲜 首都 平壤 。

这 本 应 是 韩国 感到 欣喜 的 事儿 ， 可 眼下 半岛 局势 有 了 重大 变化 ， 朝 韩 首脑 面对 面地 会 了 晤 ， 并 签署 了 联合声明 。 韩国 怎么办 ？ 只好 把 到 嘴的 “ 肥肉 ” 先 吐 出来 ， 搁置 自己的 “ 导弹 射程 扩展 计划 ” 。

一 名 韩国 知情 人士 道 出 了 实情 ：

“ 因为 有 了 首脑 会谈 ， 所以 我们 已 搁置 了 自己的 导弹 计划 ， 如果 我们 再 那么 干 ， 就 会 弄糟 首脑 峰 会 开创 的 良好 局面 。 ”

Since the Korean Peninsula was split into two countries , the Republic of Korea has , while leaning its back on the " big tree " of the United States for security , carefully and consistently sought advanced weapons from the United States in a bid to confront the Democratic People 's Republic of Korea .

An informed source in Seoul revealed to the Washington Post that the United States had secretly agreed to the request of South Korea earlier this year to " extend its existing missile range " to strike Pyongyang direct .

This should have elated South Korea . But since the situation surrounding the peninsula has changed dramatically and the two heads of state of the two Koreas have met with each other and signed a joint statement , what should South Korea do now ? It has no choice but spit back the " greasy meat " from its mouth and put the " missile expansion plan " on the back burner .

A knowledgeable South Korean speaks the truth :

" Because of the summit meeting , we have shelved our own missile plan . If we go ahead with it , it will spoil the excellent situation opened up by the summit meeting . "

Divisive Clustering (DC)

divide and conquer, iterative binary parallel splitting, reorder

自从 朝鲜半岛 被 分裂 成 两个 国家 以来 ， 韩国 在 背 靠 美国 这 棵 大 树 以求 自 安 的 同时 ， 还 小心翼翼 但 却 坚持不懈 地 向 美 国 寻求 先进武器 ， 以 抗衡 朝鲜 。 据 汉城 的 消息灵通人士 向 《 华盛顿邮报 》 透露 ， 今年 早些时候 ， 美国 已 秘 而 不 宣 地 同意 韩国 " 可以 扩展 它 现有 导弹 的 射 程 " ， 使 之 能够 直捣 朝鲜 首都 平壤 。 这 本 应 是 韩国 感到 欣喜 的 事儿 ， 可 眼下 半岛 局势 有 了 重大 变化 ， 朝 韩 首脑 面对 面地 会 了 晤 ， 并 签署 了 联合声明 。 韩国 怎么办 ？ 只好 把 到 嘴的 " 肥肉 " 先 吐 出 来 ， 搁置 自己的 " 导弹 射程 扩展 计划 " 。 一 名 韩国 知情 人士 道 出 了 实情 ：	Since the Korean Peninsula was split into two countries , the Republic of Korea has , while leaning its back on the " big tree " of the United States for security , carefully and consistently sought advanced weapons from the United States in a bid to confront the Democratic People 's Republic of Korea . An informed source in Seoul revealed to the Washington Post that the United States had secretly agreed to the request of South Korea earlier this year to " extend its existing missile range " to strike Pyongyang direct . This should have elated South Korea . But since the situation surrounding the peninsula has changed dramatically and the two heads of state of the two Koreas have met with each other and signed a joint statement , what should South Korea do now ? It has no choice but spit back the " greasy meat " from its mouth and put the " missile expansion plan " on the back burner . A knowledgeable South Korean speaks the truth :
" 因为 有 了 首脑 会谈 ， 所以 我们 已 搁置 了 自己的 导弹 计划 ， 如果 我们 再 那么 干 ， 就 会 弄糟 首脑 峰 会 开创 的 良好 局面 。 "	" Because of the summit meeting , we have shelved our own missile plan . If we go ahead with it , it will spoil the excellent situation opened up by the summit meeting . "

1

Divisive Clustering (DC)

divide and conquer, iterative binary parallel splitting, reorder

自从 朝鲜半岛 被 分裂 成 两个 国家 以来， 韩国 在 背 靠 美国 这 棵 大 树 以求 自 安 的 同时， 还 小心翼翼 但 却 坚持不懈 地 向 美 国 寻求 先进武器， 以 抗衡 朝鲜 。 据 汉城 的 消息灵通人士 向 《 华盛顿邮报 》 透露， 今年 早些时候， 美国 已 秘 而 不 宣 地 同意 韩国 " 可以 扩展 它 现有 导弹 的 射 程 "， 使 之 能够 直捣 朝鲜 首都 平壤 。 这 本 应 是 韩国 感到 欣喜 的 事儿， 可 眼下 半岛 局势 有 了 重大 变化， 朝 韩 首脑 面对 面地 会 了 晤， 并 签署 了 联合声明 。 韩国 怎么办？ 只好 把 到 嘴的 " 肥肉 " 先 吐 出 来， 搁置 自己的 " 导弹 射程 扩展 计划 " 。	Since the Korean Peninsula was split into two countries , the Republic of Korea has , while leaning its back on the " big tree " of the United States for security , carefully and consistently sought advanced weapons from the United States in a bid to confront the Democratic People 's Republic of Korea . An informed source in Seoul revealed to the Washington Post that the United States had secretly agreed to the request of South Korea earlier this year to " extend its existing missile range " to strike Pyongyang direct . This should have elated South Korea . But since the situation surrounding the peninsula has changed dramatically and the two heads of state of the two Koreas have met with each other and signed a joint statement , what should South Korea do now ? It has no choice but spit back the " greasy meat " from its mouth and put the " missile expansion plan " on the back burner .
2	
一 名 韩国 知情 人士 道 出 了 实情：	A knowledgeable South Korean speaks the truth :
1	
" 因为 有 了 首脑 会谈， 所以 我们 已 搁置 了 自己的 导弹 计划， 如果 我们 再 那么 干， 就 会 弄糟 首脑 峰 会 开创 的 良好 局面 。 "	" Because of the summit meeting , we have shelved our own missile plan . If we go ahead with it , it will spoil the excellent situation opened up by the summit meeting . "

Divisive Clustering (DC)

divide and conquer, iterative binary parallel splitting, reorder

自从 朝鲜半岛 被 分裂 成 两个 国家 以来， 韩国 在 背 靠 美国 这 棵 大 树 以求 自 安 的 同时， 还 小心翼翼 但 却 坚持不懈 地 向 美 国 寻求 先进武器， 以 抗衡 朝鲜 。	Since the Korean Peninsula was split into two countries , the Republic of Korea has , while leaning its back on the " big tree " of the United States for security , carefully and consistently sought advanced weapons from the United States in a bid to confront the Democratic People 's Republic of Korea .
3	
据 汉城 的 消息灵通人士 向 《 华盛顿邮报 》 透露， 今年 早些时候， 美国 已 秘 而 不 宣 地 同意 韩国 " 可以 扩展 它 现有 导弹 的 射 程 "， 使 之 能够 直捣 朝鲜 首都 平壤 。	An informed source in Seoul revealed to the Washington Post that the United States had secretly agreed to the request of South Korea earlier this year to " extend its existing missile range " to strike Pyongyang direct .
4	
这 本 应 是 韩国 感到 欣喜 的 事儿， 可 眼下 半岛 局势 有 了 重大 变化， 朝 韩 首脑 面对 面地 会 了 晤， 并 签署 了 联合声明 。 韩国 怎么办？	This should have elated South Korea . But since the situation surrounding the peninsula has changed dramatically and the two heads of state of the two Koreas have met with each other and signed a joint statement , what should South Korea do now ?
5	
只好 把 到 嘴的 " 肥肉 " 先 吐 出来， 搁置 自 己的 " 导弹 射程 扩展 计划 " 。	It has no choice but spit back the " greasy meat " from its mouth and put the " missile expansion plan " on the back burner .
2	
一 名 韩国 知情 人士 道 出 了 实情：	A knowledgeable South Korean speaks the truth :
1	
" 因为 有 了 首脑 会谈， 所以 我们 已 搁置 了 自己的 导弹 计划， 如果 我们 再 那么 干， 就 会 弄糟 首脑 峰 会 开创 的 良好 局面 。 "	" Because of the summit meeting , we have shelved our own missile plan . If we go ahead with it , it will spoil the excellent situation opened up by the summit meeting . "

A hierarchical chunking scheme

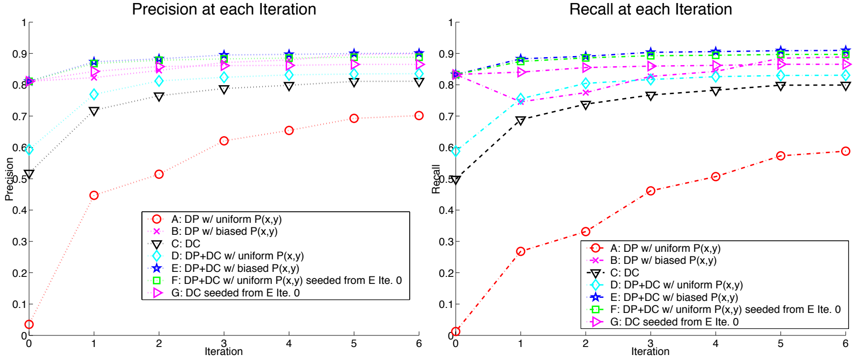
- DP+DC
  - DP at sentence level followed by DC at sub-sentence level
  - from coarse to fine, deriving short chunk pairs
- Advantage
  - significantly reduce machine training time
    - 21 hrs vs. 8 hrs
  - make most of bitext usable for machine training
    - 78% vs. 98%
    - "There is no data like more data" (Robert Mercer, 1988)
  - improve system performance by higher coverage

Outline

1. Bitext Chunk Alignment
   - Chunk Alignment
   - Chunking Algorithms
   - Sentence Alignment Results
2. Bitext Word Alignment
   - Introduction/Motivation
   - Word-to-Phrase HMM Model
   - Parameter Estimation
   - Word Alignment Results
3. Bitext Phrase Alignment
   - Inducing from Word Alignments
   - Model-based Phrase Pair Posterior
   - Translation Results
4. Conclusions

Unsupervised Sentence Alignment

- 122 Chinese/English document pairs selected from FBIS corpus
- sentence aligned by humans, $\sim$ 2,200 sentence pairs
- unsupervised from scratch, measured by Pre/Rec

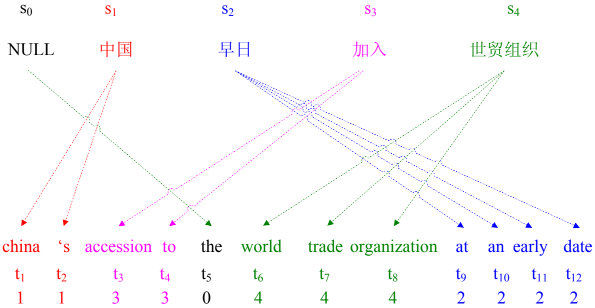


Outline

1. Bitext Chunk Alignment
   - Chunk Alignment
   - Chunking Algorithms
   - Sentence Alignment Results
2. Bitext Word Alignment
   - Introduction/Motivation
   - Word-to-Phrase HMM Model
   - Parameter Estimation
   - Word Alignment Results
3. Bitext Phrase Alignment
   - Inducing from Word Alignments
   - Model-based Phrase Pair Posterior
   - Translation Results
4. Conclusions

Word Alignment

- Fundamental problem in Machine Translation
- Basis for phrase/syntax models
- Model relations from source $\mathbf{s} = s_1^I$ to target $\mathbf{t} = t_1^J$
  - Word alignment $\mathbf{a} = a_1^J$: $s_{a_j} \rightarrow t_j, j = 1, 2, \cdots, J \Longleftarrow$ hidden r.v.
  - Conditional likelihood $P(\mathbf{t}, \mathbf{a}|\mathbf{s}) \Longleftarrow$ complete data
  - Sentence translation $P(\mathbf{t}|\mathbf{s}) = \sum_{\mathbf{a}} P(\mathbf{t}, \mathbf{a}|\mathbf{s}) \Longleftarrow$ incomplete data



State of the Art

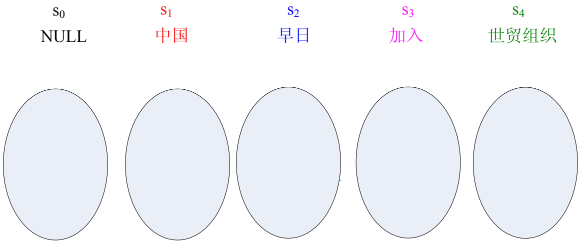
- IBM Model-4 generated by GIZA++ Toolkit (Och & Ney, ’03)
  - The state of the art word alignments especially on large bitexts
- But
  - Exact-EM is problematic, sub-optimal estimation algorithms used
  - Difficult to compute statistics under the model
  - Applications limited by word alignments only
- GOAL: improve word alignments of bitexts for better translation
  - Comparable performance to Model-4
  - Fast efficient training, with controlled memory usage
  - Use the model, not just the alignments

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

s_0	s_1	s_2	s_3	s_4
NULL	中国	早日	加入	世贸组织

- What makes the model powerful also makes computation complex
- Typical training procedure: Model-1, HMM, Model-4
- Can we do something to HMM?

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



Create a tablet for each source word

- What makes the model powerful also makes computation complex
- Typical training procedure: Model-1, HMM, Model-4
- Can we do something to HMM?

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

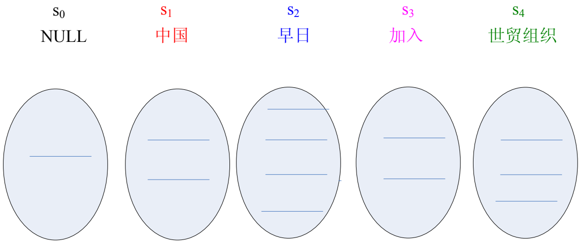
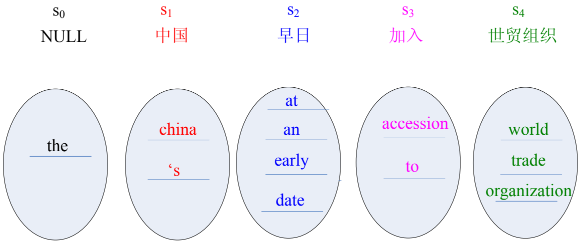


Table lookup to decide fertility: # of target words connected

- What makes the model powerful also makes computation complex
- Typical training procedure: Model-1, HMM, Model-4
- Can we do something to HMM?

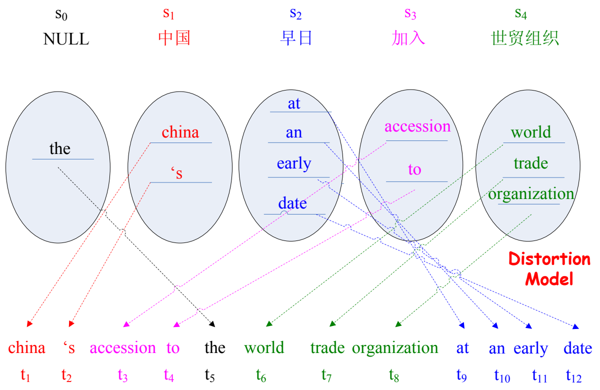
IBM Model-4 Word Alignments (Brown et al, ’93)



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

- What makes the model powerful also makes computation complex
- Typical training procedure: Model-1, HMM, Model-4
- Can we do something to HMM?

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



- What makes the model powerful also makes computation complex
- Typical training procedure: Model-1, HMM, Model-4
- Can we do something to HMM?

HMM WtoW Model (Vogel et al, '96; Och & Ney, '03)

s_1 s_2 s_3 s_4

中国 早日 加入 世贸组织

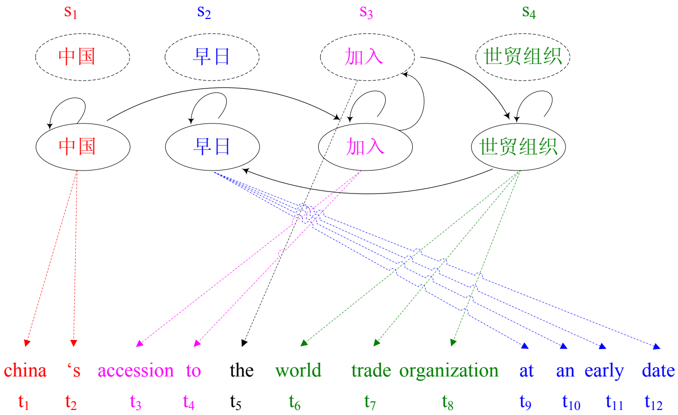
china	's	accession	to	the	world	trade	organization	at	an	early	date
t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}

- State sequences $\Longleftrightarrow$ word to word alignments
- Words are generated one by one, one transition emits one target word

HMM WtoW Model (Vogel et al, ’96; Och & Ney, ’03)

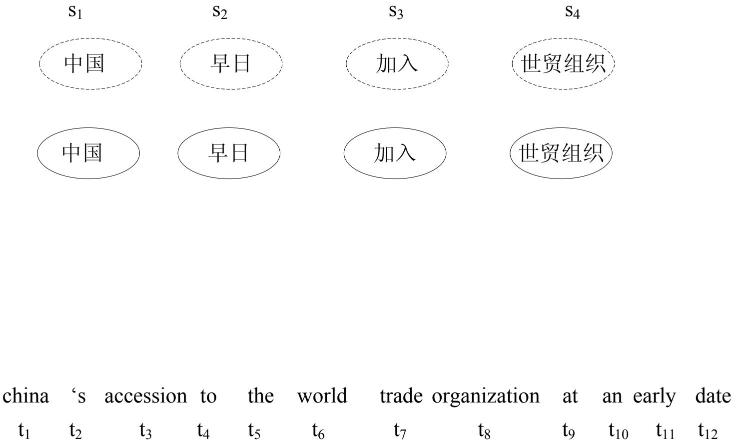
- State sequences $\Longleftrightarrow$ word to word alignments
- Words are generated one by one, one transition emits one target word

HMM WtoW Model (Vogel et al, ’96; Och & Ney, ’03)



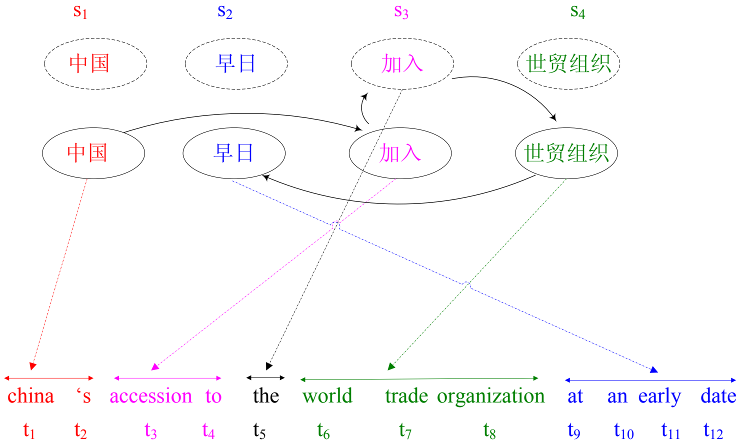
- State sequences $\Longleftrightarrow$ word to word alignments
- Words are generated one by one, one transition emits one target word

Make HMM More Powerful in Generating Observations



- Target phrases rather than words are emitted after jumping into a state
- State sequences $\Longleftrightarrow$ word to phrase alignments
- Word-to-Phrase (WtoP) HMM (Deng & Byrne, '05)

Make HMM More Powerful in Generating Observations



- Target phrases rather than words are emitted after jumping into a state
- State sequences $\Longleftrightarrow$ word to phrase alignments
- Word-to-Phrase (WtoP) HMM (Deng & Byrne, '05)

Outline

1. Bitext Chunk Alignment
   - Chunk Alignment
   - Chunking Algorithms
   - Sentence Alignment Results
2. Bitext Word Alignment
   - Introduction/Motivation
   - Word-to-Phrase HMM Model
   - Parameter Estimation
   - Word Alignment Results
3. Bitext Phrase Alignment
   - Inducing from Word Alignments
   - Model-based Phrase Pair Posterior
   - Translation Results
4. Conclusions

Word-to-Phrase HMM Alignment Models

- Target sentence segmented into K phrases
- Phrase length sequence ϕ_1^K, $\mathbf{t} = \mathbf{v}_1^K$
- Phrase alignment sequence a_1^K
- NULL: h_1^K is a Bernoulli process, $d(h_k = 1) = 1 - p_0, d(h_k = 0) = p_0$
  - $h_k = 1 \Rightarrow s_{a_k} \rightarrow \mathbf{v}_k$
  - $h_k = 0 \Rightarrow \text{NULL} \rightarrow \mathbf{v}_k$
- Hidden random variable: Word-to-phrase alignment $\mathbf{a} = (K, a_1^K, \phi_1^K, h_1^K)$

$$
\begin{aligned}
P(\mathbf{t}, \mathbf{a}|\mathbf{s}) &= P(\mathbf{v}_1^K, K, a_1^K, h_1^K, \phi_1^K | \mathbf{s}) \\
&= P(K|J, \mathbf{s}) \times P(a_1^K, \phi_1^K, h_1^K | K, J, \mathbf{s}) \times P(\mathbf{v}_1^K | a_1^K, h_1^K, \phi_1^K, K, J, \mathbf{s}) \\
&= P(K|J, I) \Longleftarrow \text{Phrase Count} \propto \eta^K \\
&\quad \times \prod_{k=1}^{K} p(a_k | a_{k-1}, h_k; I) \cdot d(h_k) \cdot n(\phi_k; h_k \cdot s_{a_k}) \Longleftarrow \text{Markov, Phrase Length} \\
&\quad \times \prod_{k=1}^{K} P(\mathbf{v}_k | h_k \cdot s_{a_k}) \Longleftarrow \text{Word-to-Phrase Translation}
\end{aligned}
$$

Word-to-Phrase HMM Alignment Models

- Target sentence segmented into K phrases
- Phrase length sequence ϕ_1^K, $\mathbf{t} = \mathbf{v}_1^K$
- Phrase alignment sequence a_1^K
- NULL: h_1^K is a Bernoulli process, $d(h_k = 1) = 1 - p_0, d(h_k = 0) = p_0$
  - $h_k = 1 \Rightarrow s_{a_k} \rightarrow \mathbf{v}_k$
  - $h_k = 0 \Rightarrow \text{NULL} \rightarrow \mathbf{v}_k$
- Hidden random variable: Word-to-phrase alignment $\mathbf{a} = (K, a_1^K, \phi_1^K, h_1^K)$

$$
\begin{aligned}
P(\mathbf{t}, \mathbf{a}|\mathbf{s}) &= P(\mathbf{v}_1^K, K, a_1^K, h_1^K, \phi_1^K|\mathbf{s}) \\
&= P(K|J, \mathbf{s}) \times P(a_1^K, \phi_1^K, h_1^K|K, J, \mathbf{s}) \times P(\mathbf{v}_1^K|a_1^K, h_1^K, \phi_1^K, K, J, \mathbf{s}) \\
&= P(K|J, I) \Longleftarrow \text{Phrase Count} \propto \eta^K \\
&\quad \times \prod_{k=1}^{K} p(a_k|a_{k-1}, h_k; I) \cdot d(h_k) \cdot n(\phi_k; h_k \cdot s_{a_k}) \Longleftarrow \text{Markov, Phrase Length} \\
&\quad \times \prod_{k=1}^{K} P(\mathbf{v}_k|h_k \cdot s_{a_k}) \Longleftarrow \text{Word-to-Phrase Translation}
\end{aligned}
$$

Word-to-Phrase HMM Alignment Models

- Target sentence segmented into K phrases
- Phrase length sequence ϕ_1^K, $\mathbf{t} = \mathbf{v}_1^K$
- Phrase alignment sequence a_1^K
- NULL: h_1^K is a Bernoulli process, $d(h_k = 1) = 1 - p_0, d(h_k = 0) = p_0$
  - $h_k = 1 \Rightarrow s_{a_k} \rightarrow \mathbf{v}_k$
  - $h_k = 0 \Rightarrow \text{NULL} \rightarrow \mathbf{v}_k$
- Hidden random variable: Word-to-phrase alignment $\mathbf{a} = (K, a_1^K, \phi_1^K, h_1^K)$

$$
\begin{aligned}
P(\mathbf{t}, \mathbf{a}|\mathbf{s}) &= P(\mathbf{v}_1^K, K, a_1^K, h_1^K, \phi_1^K|\mathbf{s}) \\
&= P(K|J, \mathbf{s}) \times P(a_1^K, \phi_1^K, h_1^K|K, J, \mathbf{s}) \times P(\mathbf{v}_1^K|a_1^K, h_1^K, \phi_1^K, K, J, \mathbf{s}) \\
&= P(K|J, I) \Longleftarrow \text{Phrase Count} \propto \eta^K \\
&\quad \times \prod_{k=1}^{K} p(a_k|a_{k-1}, h_k; I) \cdot d(h_k) \cdot n(\phi_k; h_k \cdot s_{a_k}) \Longleftarrow \text{Markov, Phrase Length} \\
&\quad \times \prod_{k=1}^{K} P(\mathbf{v}_k|h_k \cdot s_{a_k}) \Longleftarrow \text{Word-to-Phrase Translation}
\end{aligned}
$$

Word-to-Phrase Translation Probabilities

- Replace weak i.i.d. word-for-word translation
- $P(\text{world trade organization}|f = \text{世贸组织}; 3) =?$
  - $= t(\text{world}|\text{f}) \cdot t(\text{trade}|\text{f}) \cdot t(\text{organization}|\text{f}) \Longleftarrow$ i.i.d.
  - $= t(\text{world}|\text{f}) \cdot t_2(\text{trade}|\text{world}, \text{f}) \cdot t_2(\text{organization}|\text{trade}, \text{f}) \Longleftarrow$ bigram

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
- Use same estimation techniques as used for bigram LMs
- Data sparseness, Witten-Bell smoothing

Comparing Word-to-Phrase HMM to ...

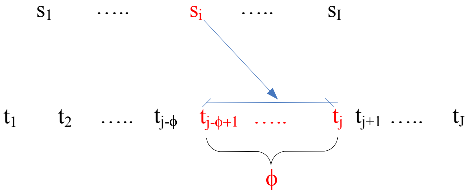
- Segmental Hidden Markov Models (Ostendorf et al, '96)
  - states emit observation sequences
- WtoW HMM (Vogel et al, '96; Och & Ney, '03)
  - $N = 1$
- Extensions to WtoW HMM (Toutanova et al, '02)
  - $P(\mathrm{stay}|s)$ vs. $P(\mathrm{stay} = \phi|s)$ in modeling state durations
- IBM Model-4
  - fertility vs. phrase length

Outline

1. Bitext Chunk Alignment
   - Chunk Alignment
   - Chunking Algorithms
   - Sentence Alignment Results
2. Bitext Word Alignment
   - Introduction/Motivation
   - Word-to-Phrase HMM Model
   - Parameter Estimation
   - Word Alignment Results
3. Bitext Phrase Alignment
   - Inducing from Word Alignments
   - Model-based Phrase Pair Posterior
   - Translation Results
4. Conclusions

Forward-backward Algorithm

State space $S = \{(i, \phi, h) : 1 \leq i \leq I, 1 \leq \phi \leq N, h = 0 \text{ or } 1\}$ Grid $2NI \times J$



$$\alpha_j(i, \phi, h) = \left\{ \sum_{i', \phi', h'} \alpha_{j-\phi}(i', \phi', h') p(i|i', h; I) \right\} \cdot \eta \cdot n(\phi; h \cdot s_i) \cdot P(t_{j-\phi+1}^{j} | h \cdot s_i, \phi)$$

$$\beta_j(i, \phi, h) = \sum_{i', \phi', h'} \beta_{j+\phi'}(i', \phi', h') p(i'|i, h'; I) \cdot \eta \cdot n(\phi'; h' \cdot s_{i'}) \cdot P(t_{j+1}^{j+\phi'} | h' \cdot s_{i'}, \phi')$$

$$\gamma_j(i, \phi, h) = P(h \cdot s_i \rightarrow v = t_{j-\phi+1}^{j} | \theta, \mathbf{s}, \mathbf{t}) = \frac{\alpha_j(i, \phi, h) \beta_j(i, \phi, h)}{\sum_{i', h', \phi'} \alpha_J(i', \phi', h')}$$

Embedded Estimation of Word-to-Phrase HMM

- 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) (experience from ASR)
  - 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

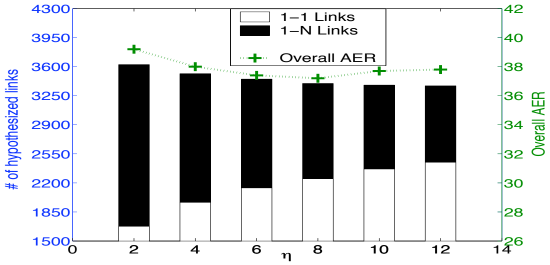
- Unsupervised training from scratch
  - Model-1, 10 its (initial t-table)
  - Model-2, 5 its (better t-table)
  - WtoW HMM, 5 its (initial Markov model)
  - WtoP HMM N=2, 3, .., each 5 its (Markov model, phrase length) (experience from ASR)
  - WtoP HMM with bigram t-table, 5 its (bigram t-table)
- 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

Outline

1. Bitext Chunk Alignment
   - Chunk Alignment
   - Chunking Algorithms
   - Sentence Alignment Results
2. Bitext Word Alignment
   - Introduction/Motivation
   - Word-to-Phrase HMM Model
   - Parameter Estimation
   - **Word Alignment Results**
3. Bitext Phrase Alignment
   - Inducing from Word Alignments
   - Model-based Phrase Pair Posterior
   - Translation Results
4. Conclusions

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



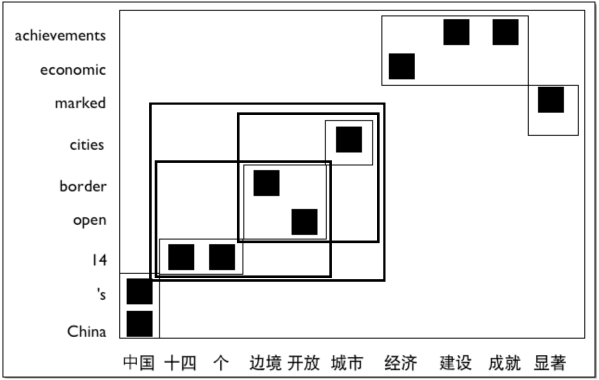
- Comparable performance when extending to large scale bitexts

Outline

1. Bitext Chunk Alignment
   - Chunk Alignment
   - Chunking Algorithms
   - Sentence Alignment Results
2. Bitext Word Alignment
   - Introduction/Motivation
   - Word-to-Phrase HMM Model
   - Parameter Estimation
   - Word Alignment Results
3. Bitext Phrase Alignment
   - Inducing from Word Alignments
   - Model-based Phrase Pair Posterior
   - Translation Results
4. Conclusions

Statistical Phrase Translation Models

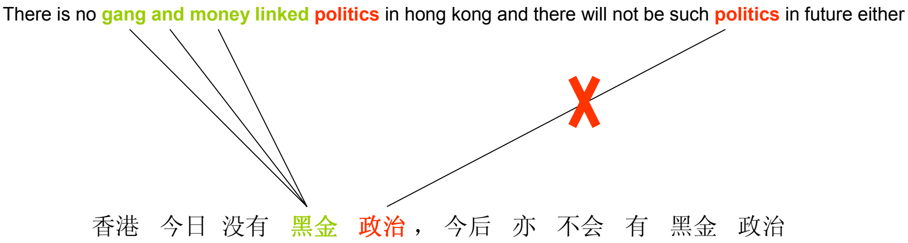
- Phrase-based SMT performs better than word-based SMT
- Phrases Pair Inventory (PPI) extracted from word aligned bitext (Och et al, '99)



But word alignments are imperfect ...

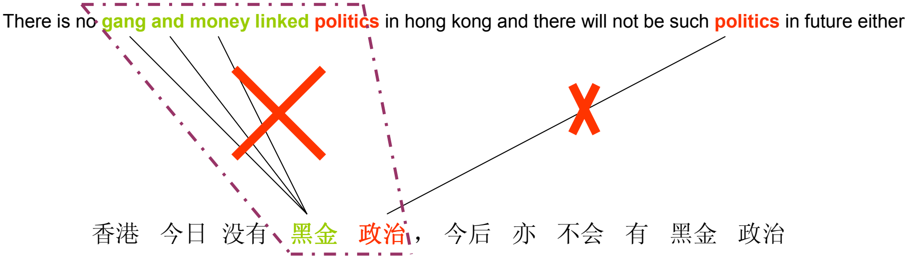
- Relying on the one-best word alignment may exclude some valid phrase pairs
- Goal is to define a probability distribution over phrase pairs
  - Allows more control over generation of phrase pairs

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

Outline

1 Bitext Chunk Alignment
- Chunk Alignment
- Chunking Algorithms
- Sentence Alignment Results

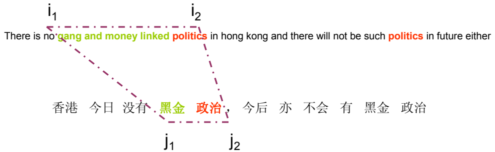
2 Bitext Word Alignment
- Introduction/Motivation
- Word-to-Phrase HMM Model
- Parameter Estimation
- Word Alignment Results

3 Bitext Phrase Alignment
- Inducing from Word Alignments
- Model-based Phrase Pair Posterior
- Translation Results

4 Conclusions

Model-based Phrase Pair Posterior

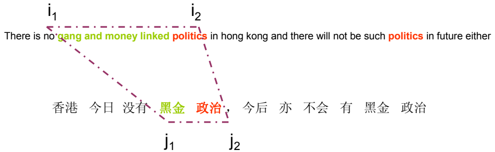
- Doesn't rely on a single alignment



- 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
  $P(\mathbf{t}, A(i_1, i_2; j_1, j_2) \mid \mathbf{s}) = \sum_{\mathbf{a} : a_1^m \in A(i_1, i_2; j_1, j_2)} P(\mathbf{t}, \mathbf{a} | \mathbf{s})$
- Obtain phrase pair posterior
  $P(A(i_1, i_2; j_1, j_2) \mid \mathbf{t}, \mathbf{s}) = P(\mathbf{t}, A(i_1, i_2; j_1, j_2) \mid \mathbf{s}) / P(\mathbf{t} | \mathbf{s})$
- Efficient DP-based implementation for WtoP HMM, Difficult for Model-4

Model-based Phrase Pair Posterior

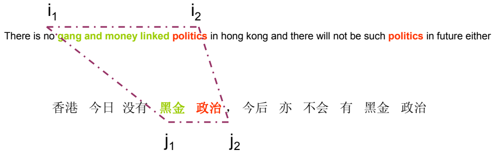
- Doesn't rely on a single alignment



- 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

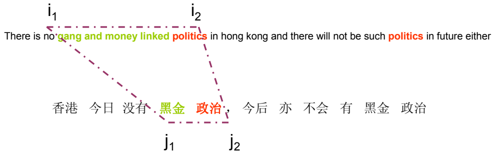
- Doesn't rely on a single alignment



- 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
  $P(\mathbf{t}, A(i_1,i_2;j_1,j_2) \mid \mathbf{s}) = \sum_{\mathbf{a}\,:\, a_1^m \in A(i_1,i_2;j_1,j_2)} P(\mathbf{t},\mathbf{a}|\mathbf{s})$
- Obtain phrase pair posterior
  $P(A(i_1,i_2;j_1,j_2) \mid \mathbf{t}, \mathbf{s}) = P(\mathbf{t}, A(i_1,i_2;j_1,j_2) \mid \mathbf{s}) / P(\mathbf{t}|\mathbf{s})$
- Efficient DP-based implementation for WtoP HMM, Difficult for Model-4

Model-based Phrase Pair Posterior

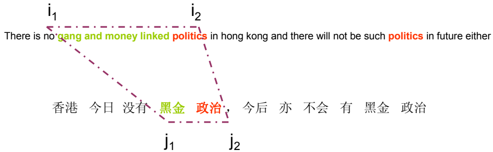
- Doesn't rely on a single alignment



- 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
  $P(\mathbf{t}, A(i_1, i_2; j_1, j_2) \mid \mathbf{s}) = \sum_{\mathbf{a}\,:\, a_1^m \in A(i_1, i_2; j_1, j_2)} P(\mathbf{t}, \mathbf{a} | \mathbf{s})$
- Obtain phrase pair posterior
  $P(A(i_1, i_2; j_1, j_2) | \mathbf{t}, \mathbf{s}) = P(\mathbf{t}, A(i_1, i_2; j_1, j_2) \mid \mathbf{s}) / P(\mathbf{t} | \mathbf{s})$
- 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
  $P(\mathbf{t}, A(i_1, i_2; j_1, j_2) \mid \mathbf{s}) = \sum_{\mathbf{a} : a_1^m \in A(i_1, i_2; j_1, j_2)} P(\mathbf{t}, \mathbf{a} | \mathbf{s})$
- Obtain phrase pair posterior
  $P(A(i_1, i_2; j_1, j_2) \mid \mathbf{t}, \mathbf{s}) = P(\mathbf{t}, A(i_1, i_2; j_1, j_2) \mid \mathbf{s}) / P(\mathbf{t} | \mathbf{s})$
- 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

$$f(i_1, i_2) = P_{F \to E}(A(i_1, i_2; j_1, j_2) \,|\, e_1^l, f_1^m)$$

$$b(i_1, i_2) = P_{E \to F}(A(i_1, i_2; j_1, j_2) \,|\, e_1^l, f_1^m)$$

$$g(i_1, i_2) = \sqrt{f(1_1, i_2)\, b(i_1, i_2)}$$

$$(\hat{i}_1, \hat{i}_2) = \operatorname*{argmax}_{1 \le i_1, i_2 \le l} g(i_1, i_2) \text{ , and set } u = e_{\hat{i}_1}^{\hat{i}_2}$$

  - add $(\mathbf{u}, \mathbf{v})$ to the baseline PPI if posterior exceeds a threshold value

Outline

1 Bitext Chunk Alignment
- Chunk Alignment
- Chunking Algorithms
- Sentence Alignment Results

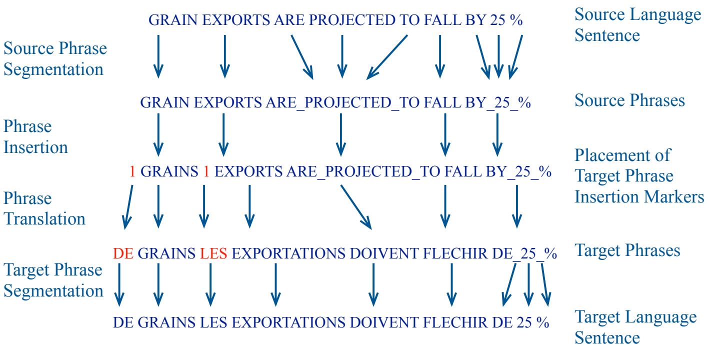
2 Bitext Word Alignment
- Introduction/Motivation
- Word-to-Phrase HMM Model
- Parameter Estimation
- Word Alignment Results

3 Bitext Phrase Alignment
- Inducing from Word Alignments
- Model-based Phrase Pair Posterior
- Translation Results

4 Conclusions

Transduce Translation Model (Kumar et al, '05)

- TTM Decoder - WFST implementation with monotone order



Automatic Machine Translation Evaluation

- hard problem !
- BLEU (Papeneni et al, '01) – an automatic MT metric
  - correlated well with human judgements
  - geomantic mean of n-gram precisions weighted by brevity penalty

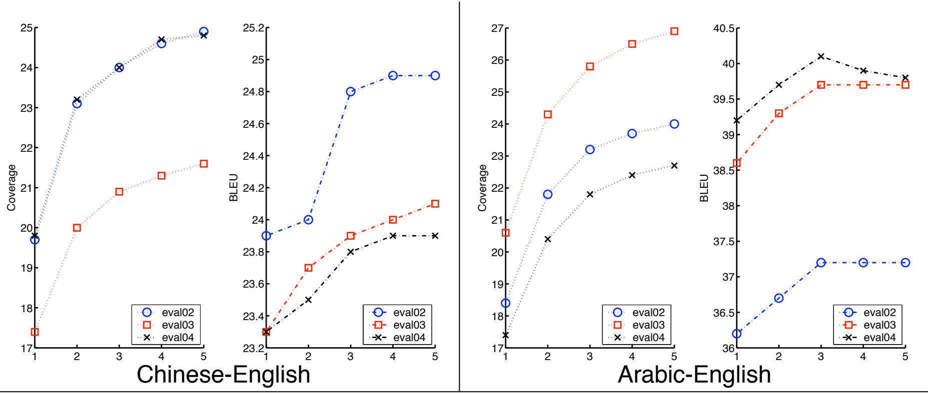
Reference : mr. speaker , in absolutely no way .

Hypothesis : in absolutely no way , mr. chairman .

BLEU computation

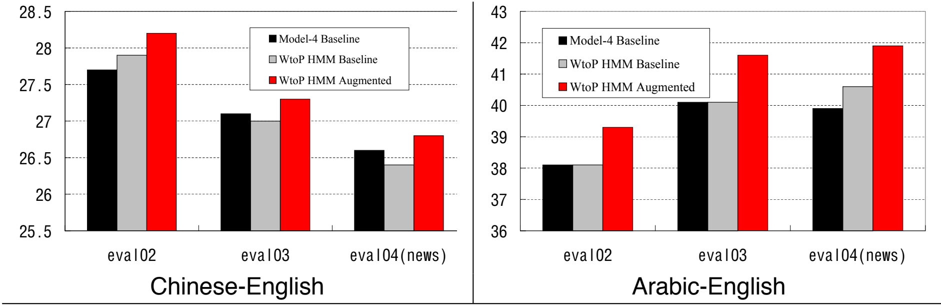
Sub-string-Matches(Truth,Hyp)				BLEU
1-word	2-word	3-word	4-word	$\left(\frac{7}{8} \times \frac{3}{7} \times \frac{2}{6} \times \frac{1}{5}\right)^{\frac{1}{4}} = 0.3976$
7/8	3/7	2/6	1/5	

Translation Results: Small Systems



- Relaxing threshold in PPI augmenting improves coverage and BLEU score
- Balance coverage against phrase translation quality
- WtoP model can even be applied to augment Model-4 PPI

Translation Results: Large Systems



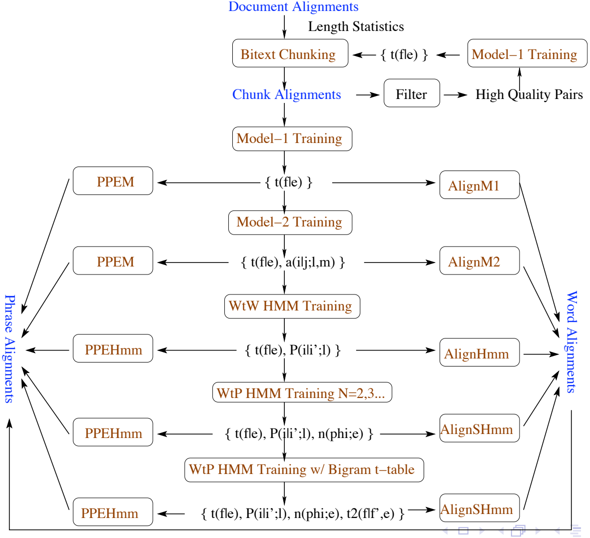
- Used all parallel corpora available from LDC
  - C-E: 200M En. words (FBIS, Xinhua, HK News, ..., all UN bitexts)
  - A-E: 130M En. words (news, all UN bitexts)

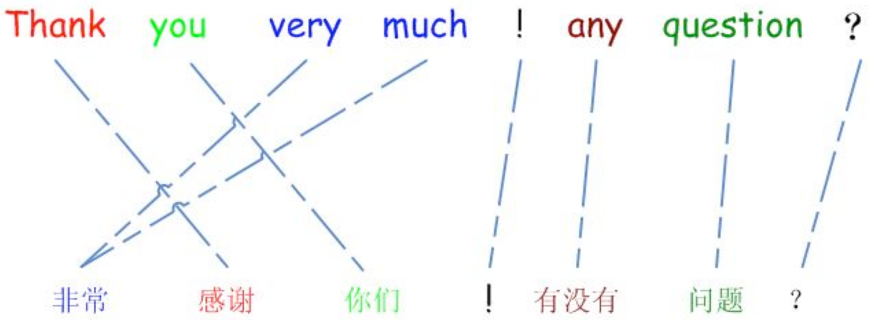
Conclusions

- A hierarchial bitext chunking approach
  - language independent, no linguistic knowledge required
  - derived short chunk pairs, retain more of the available bitext
- The word-to-phrase HMM alignment model
  - produces good quality word alignments over very large bitexts
  - has efficient training algorithm with parallel implementation
  - a powerful framework
- Model-based phrase pair distribution enables
  - an improved phrase pair extraction strategy
  - controlled balance coverage vs. quality
- WtoP HMM performs better than IBM Model-4 on large systems

Machine Translation Toolkit (MTTK)

Solutions for MT training, Used for JHU-CU 2005 NIST MT Eval Systems




Thank you very much ! any question ?
非常 感谢 你们 ! 有没有 问题 ?