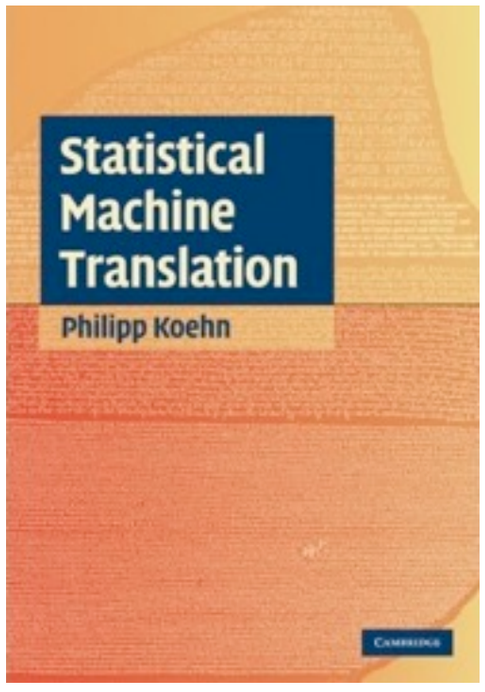
統計的機械翻訳の最先端

渡辺太郎

情報通信研究機構

taro.watanabe @ nict.go.jp

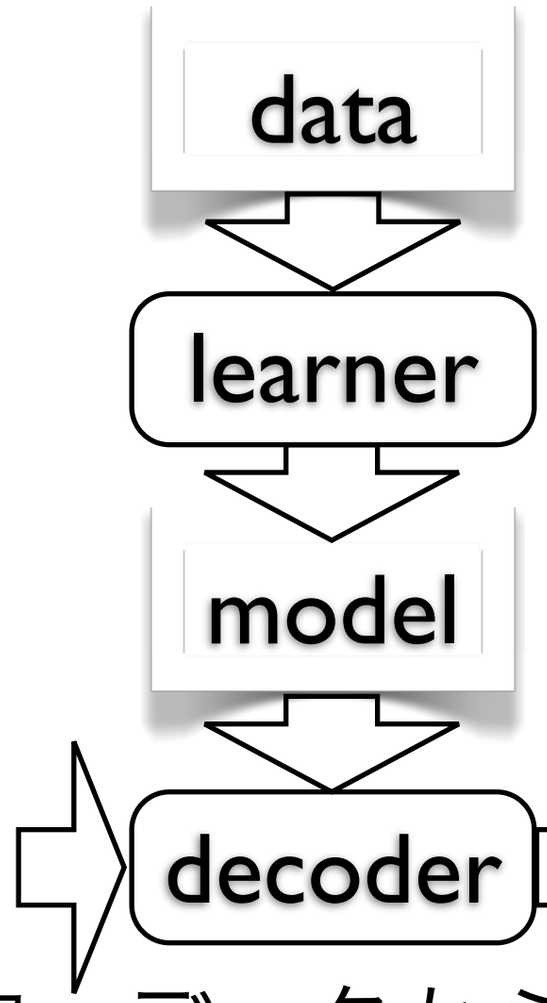
注意

- いろんな言語が混ざっています。
- 基礎的な内容は 

も読んでください。
(Koehn, 2009)
- このスライドの最新版

http://mastarpj.nict.go.jp/~t_watana

機械翻訳



黑山头口岸联检部门将原来要二至三天办完的出入境手续改为一天办完。

The United Inspection Department of Heishantou Port has shortened the procedures for leaving and entering the territory from originally 2 - 3 days to 1 day.

- モデルを仮定、データからパラメータを学習
- 学習されたモデルでデコード
- ルール翻訳、用例翻訳などの区別は無意味

主な問題

- 翻訳をどのような過程でモデル化するか?
- (データ、モデルがあったとして)パラメータの学習法 ?
- (モデル、パラメータがあったとして)デコードの手法 ?
- 翻訳結果の評価法?
  - どのようにデータを集めるか? (対象外)

最先端

- より複雑な構造: 単語、句、木、...
  - 効率のよい探索、学習
- 構文解析、機械学習からの応用

内容

- 統計的機械翻訳の基礎
- 最先端
  - 木構造に基づく機械翻訳
  - 最適化

統計的機械翻訳の基礎

内容

- 統計的機械翻訳の枠組み
- 単語アライメント
- 句に基づく機械翻訳
- 自動評価

通信路モデル

通信路モデル
+ noise

$$\begin{aligned}\hat{y} &= \operatorname*{argmax}_y Pr(y|x) \\ &= \operatorname*{argmax}_y \frac{Pr(x|y)Pr(y)}{Pr(x)} \\ &= \operatorname*{argmax}_y Pr(x|y)Pr(y)\end{aligned}$$

f = 原言語

e =目的言語

$$\hat{\mathbf{e}} = \operatorname*{argmax}_{\mathbf{e}} Pr(\mathbf{f}|\mathbf{e})Pr(\mathbf{e})$$

- 応用技術: 音声認識、OCR、機械翻訳...

翻訳モデル

$$\hat{\mathbf{e}} = \underset{\mathbf{e}}{\operatorname{argmax}} \boxed{Pr(\mathbf{f}|\mathbf{e})} \boxed{Pr(\mathbf{e})}$$

翻訳モデル 言語モデル

(Brown et al., 1990)

- 翻訳モデル: 翻訳としての正しさ(adequecy)
  - 本チュートリアルの中心
- 言語モデル: 文法エラーの修正、「スタイル」の統一、流暢さ(fuency)

言語モデル

$$Pr(\text{I do not know}) = ?$$
$$Pr(\text{I not do know}) = ?$$

- 目的言語の文の尤度
- ngramで表現

$$W = w_1, w_2, w_3, \cdots w_N$$

$$\begin{aligned} p(W) &= p(w_1, w_2, w_3, \cdots, w_N) \\ &= p(w_1)p(w_2|w_1)p(w_3|w_1, w_2)\cdots \\ &\quad p(w_N|w_1, w_2, w_3, \cdots, w_{N-1}) \end{aligned}$$

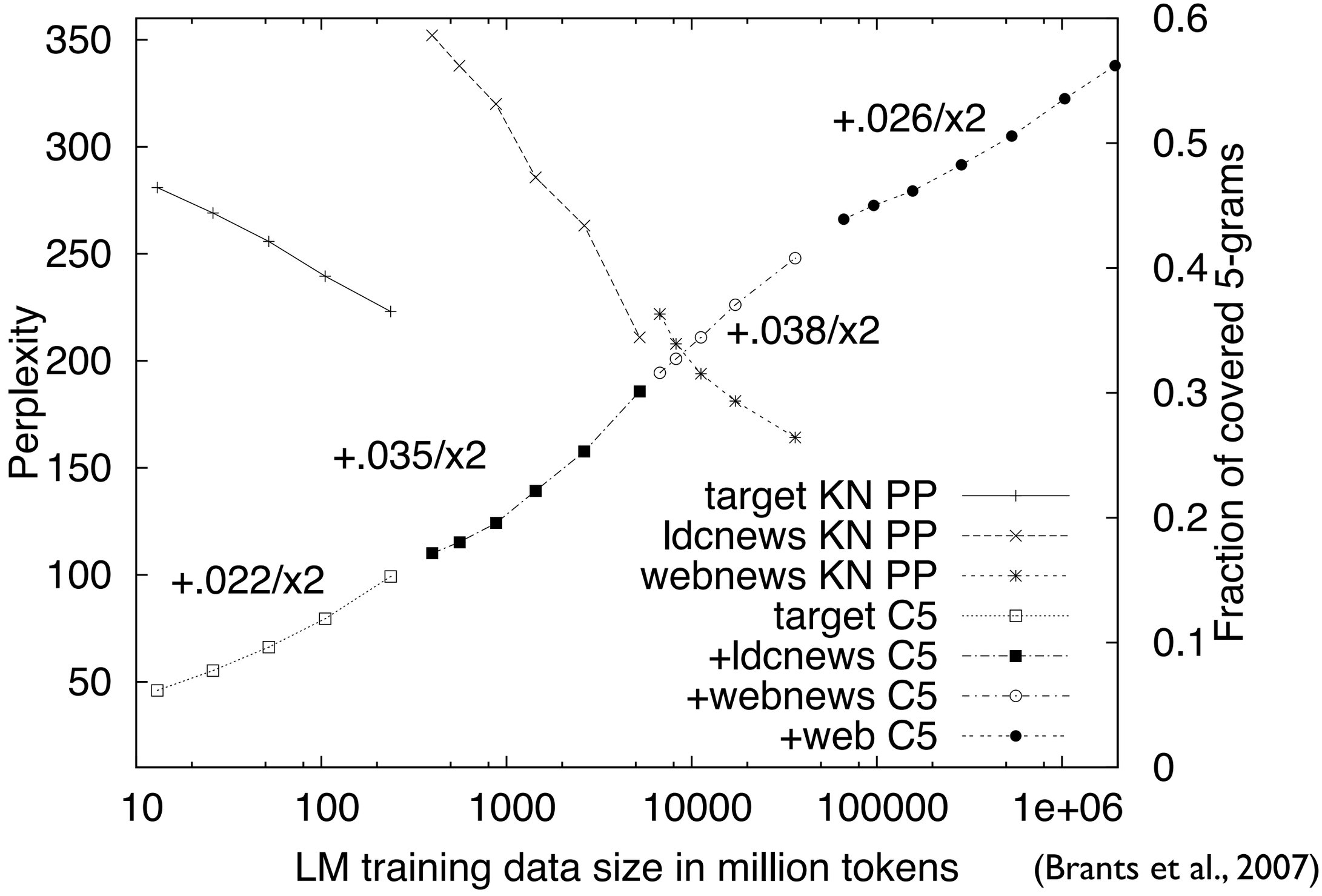
ngram 言語モデル

- マルコフな仮定:n単語だけ覚えましょう
- Bigram:

$$p(\text{I do not know}) \quad = \quad p(\text{I})p(\text{do}|\text{I})p(\text{not}|\text{do})p(\text{know}|\text{not})$$

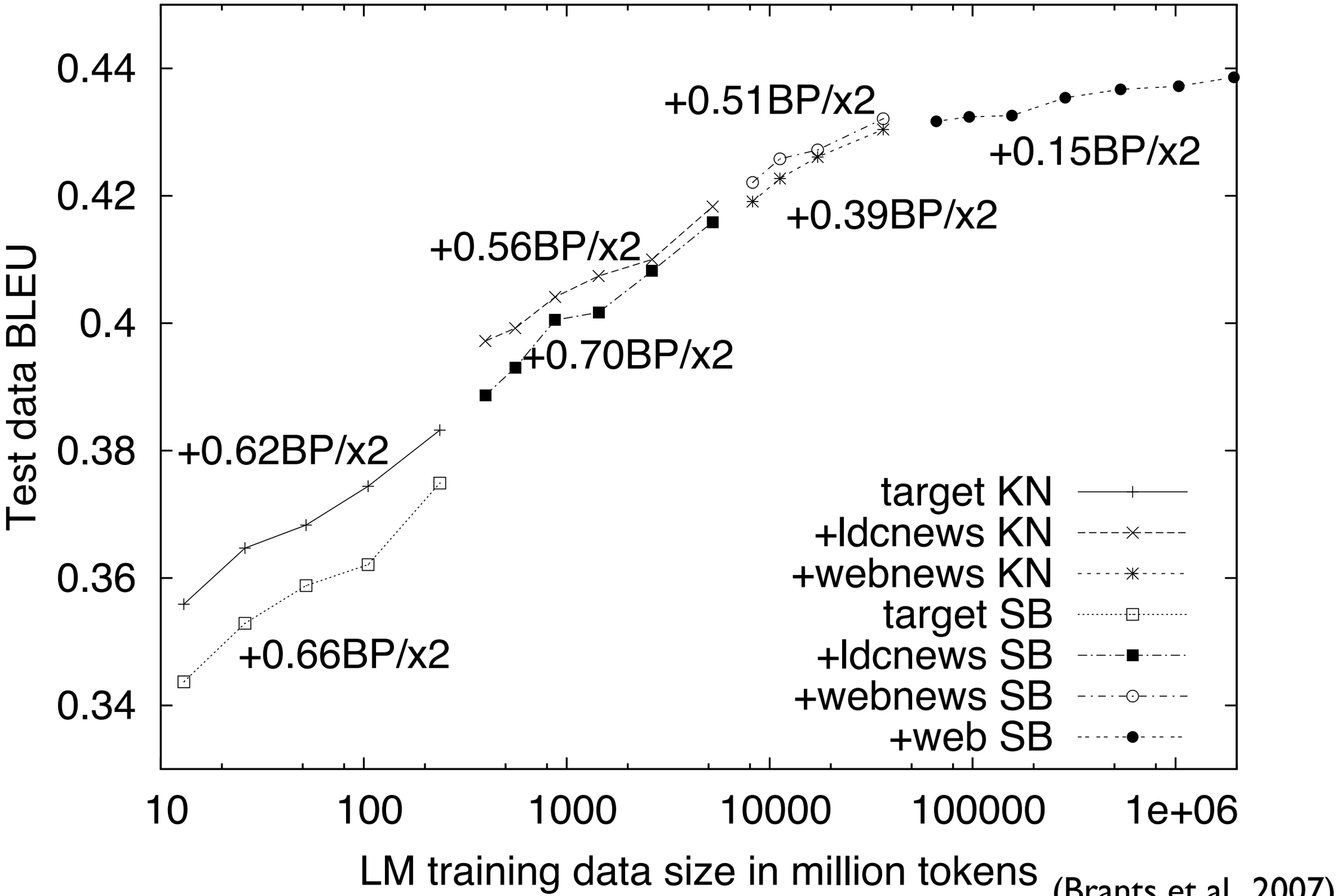
- 学習: 最尤推定 + smoothing (Good-Turing, Witten-Bell, Kneser-Ney etc.)

Larger Data, Better LM



(Brants et al., 2007)

Better LM, Better MT



(Brants et al., 2007)

内容

- 統計的機械翻訳の枠組み
- 単語アライメント
- 句に基づく機械翻訳
- 自動評価

翻訳モデル

$$\mathbf{f} = \text{je ne sais pas}$$

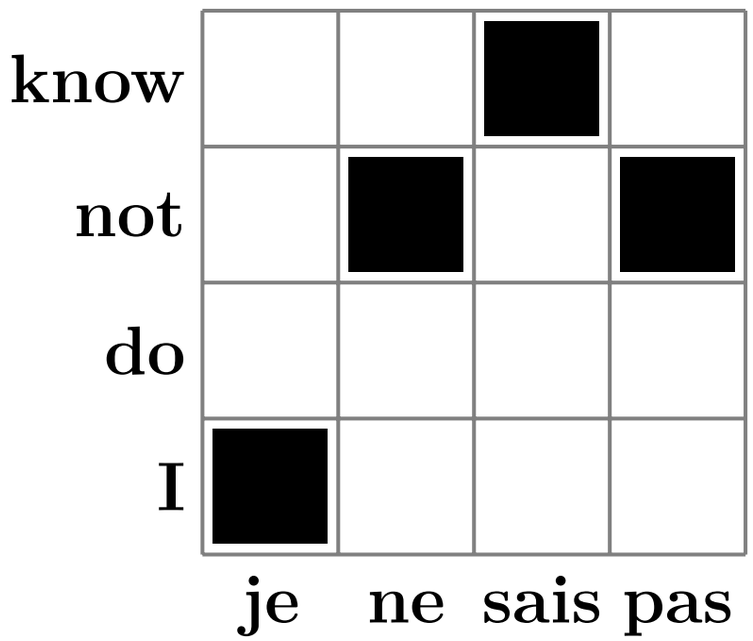
$$\mathbf{e} = \text{I do not know}$$

$$Pr(\mathbf{f}|\mathbf{e}) = ??$$

- 「単語アライメント」に基づく翻訳モデル
- Model 1 (Brown et al., 1993):
  - どのようにP(f|e)を表現するか
  - どのようにP(f|e)を推定するか

アライメントの表現

$$Pr(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} Pr(\mathbf{f}, \mathbf{a}|\mathbf{e})$$

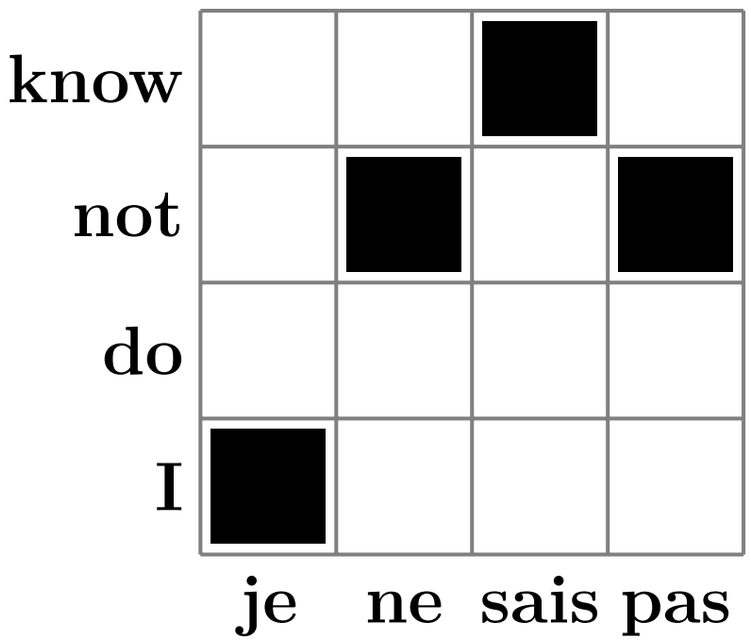


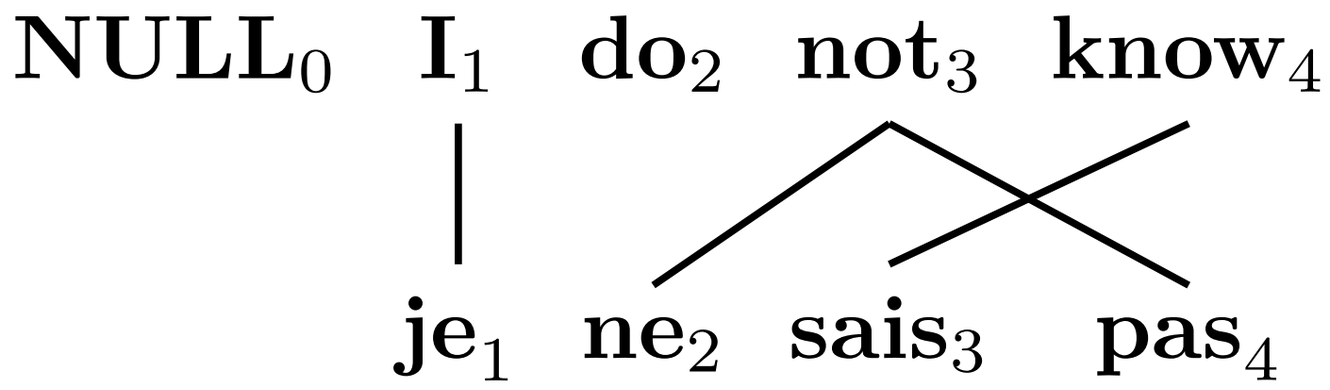
$$\mathbf{a} = \{(1 \to 1), (2 \to 3), (3 \to 4), (4 \to 3)\}$$

- P(f|e) を分解: P(f,a|e)
- “a”: 原言語と目的言語との単語単位のマッピング
- “a”の数?

$$2^{|\mathbf{e}| \times |\mathbf{f}|}$$

一対多の近似





$$\mathbf{a} = \{1, 3, 4, 3\}$$

$$\begin{aligned} \mathbf{f} &= f_1^m &&= f_1, f_2, f_3, \cdots \\ \mathbf{e} &= e_0^l &&= e_0, e_1, e_2, e_3, \cdots \\ \mathbf{a} &= a_1^m &&= a_1, a_2, a_3, \cdots \end{aligned}$$

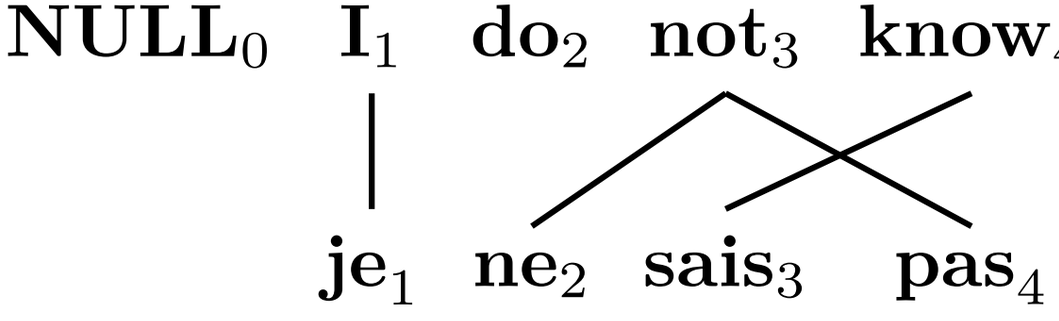
- fの各単語がeの一単語へと対応
- 特殊なNULLがeにあると仮定
- “a”の数?

$$(|\mathbf{e}| + 1)^{|\mathbf{f}|}$$

さらに分解: Model 1

$$
\begin{aligned}
Pr(\mathbf{f}|\mathbf{e}) &= \sum_{\mathbf{a}} Pr(\mathbf{f},\mathbf{a}|\mathbf{e}) \\
&= \sum_{\mathbf{a}} Pr(\mathbf{f}|\mathbf{a},\mathbf{e}) Pr(\mathbf{a}|\mathbf{e}) \\
&= Pr(m|\mathbf{e}) \sum_{\mathbf{a}} Pr(\mathbf{f}|\mathbf{a},m,\mathbf{e}) Pr(\mathbf{a}|m,\mathbf{e}) \\
&\approx \epsilon \sum_{\mathbf{a}} \prod_{j=1}^{m} t(f_j|e_{a_j}) \frac{1}{(l+1)^m} \\
&\qquad \text{s.t.} \forall e : \sum_f t(f|e) = 1
\end{aligned}
$$

- “a”の一例:



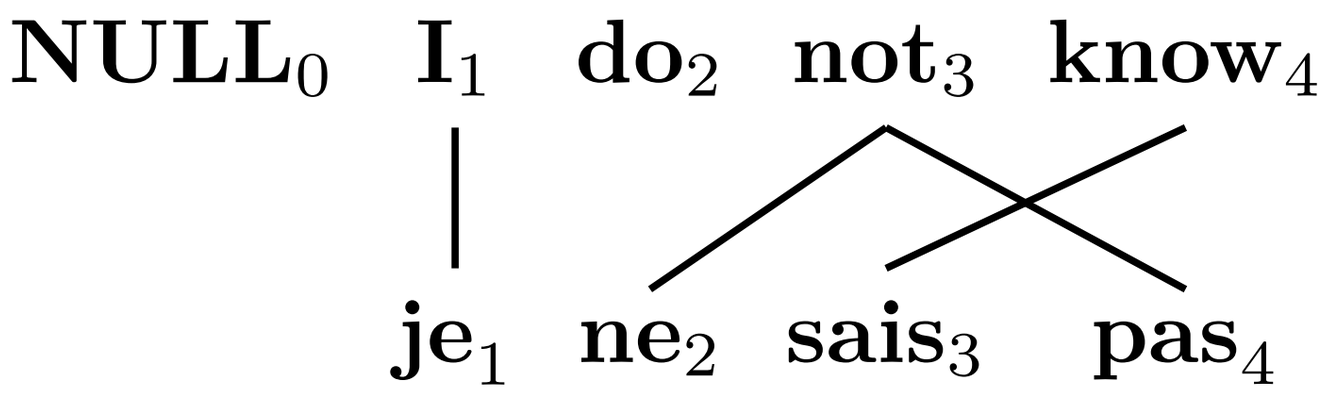
$$
\begin{aligned}
&\epsilon \times t(\mathbf{je}_1|\mathbf{I}_1) \times t(\mathbf{ne}_2|\mathbf{not}_3) \\
&\quad \times t(\mathbf{sais}_3|\mathbf{know}_4) \times t(\mathbf{pas}_4|\mathbf{not}_3) \\
&\quad \times \frac{1}{5^4}
\end{aligned}
$$

推定: Model 1

- (f, e)から成る対訳データ: $\mathcal{D} = \langle \mathcal{F}, \mathcal{E} \rangle$
- データの尤度: $\prod_{\langle \mathbf{f}, \mathbf{e} \rangle \in \mathcal{D}} Pr(\mathbf{f}|\mathbf{e})$
- データの対数尤度を最大化するパラメータΘを学習: $\hat{\theta} = \operatorname{argmax}_{\theta} \sum_{\langle \mathbf{f}, \mathbf{e} \rangle \in \mathcal{D}} \log P_{\theta}(\mathbf{f}|\mathbf{e})$
  - Model 1では、Θ = t(f | e)のテーブル

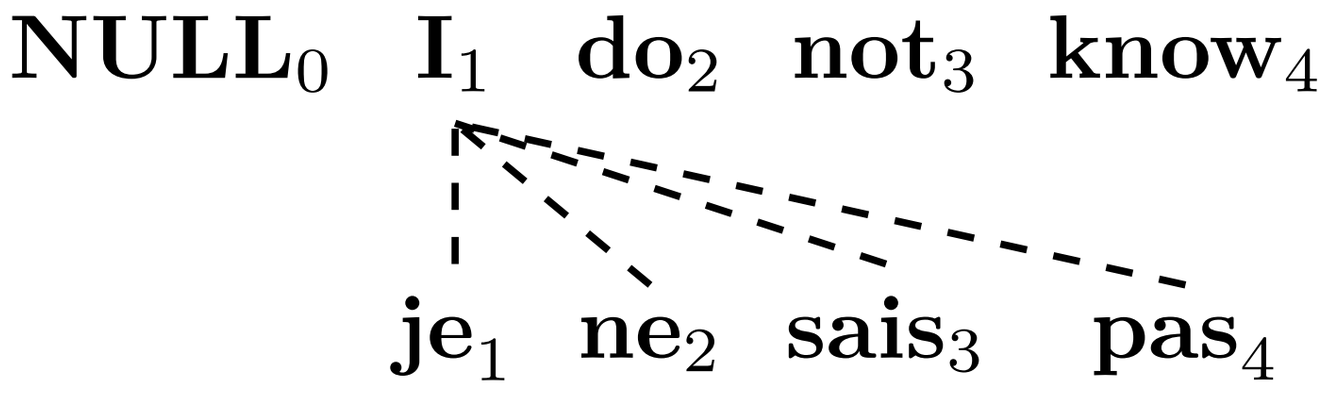
EMアルゴリズム: Model 1

- ある“a”に対して、回数を列挙



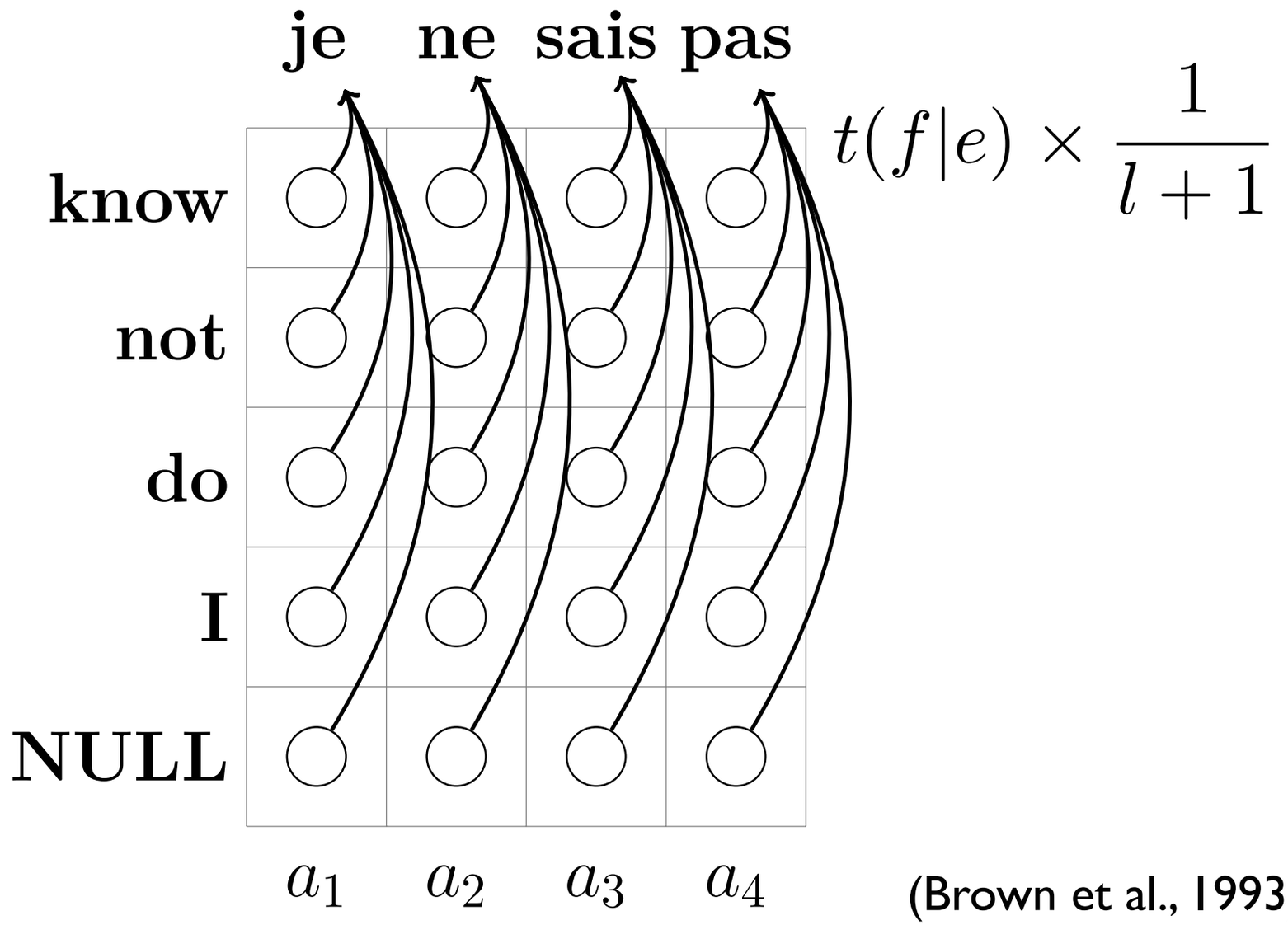
$$t(\mathbf{je}|\mathbf{I}) = \frac{\mathrm{count}(\mathbf{je}, \mathbf{I})}{\sum_f \mathrm{count}(f, \mathbf{I})}$$

- EMアルゴリズム: t(f|e) による“fractional counts”



$$t(\mathbf{je}|\mathbf{I}) = \frac{\mathrm{count}(\mathbf{je}, \mathbf{I}; \theta)}{\sum_f \mathrm{count}(f, \mathbf{I}; \theta)}$$

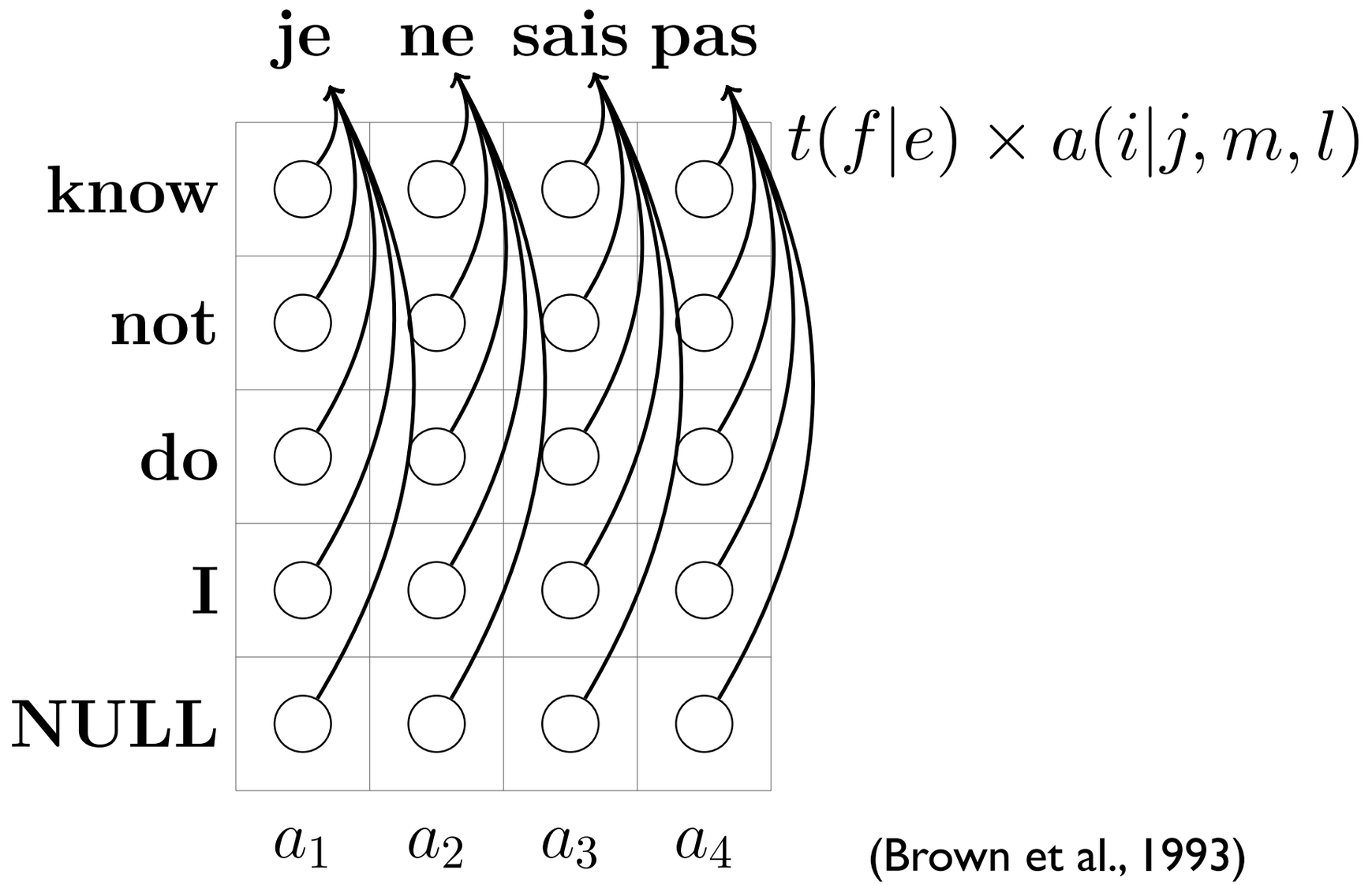
Model 1



(Brown et al., 1993)

- Generative story: Model 1
  - fの各単語は、eから生成

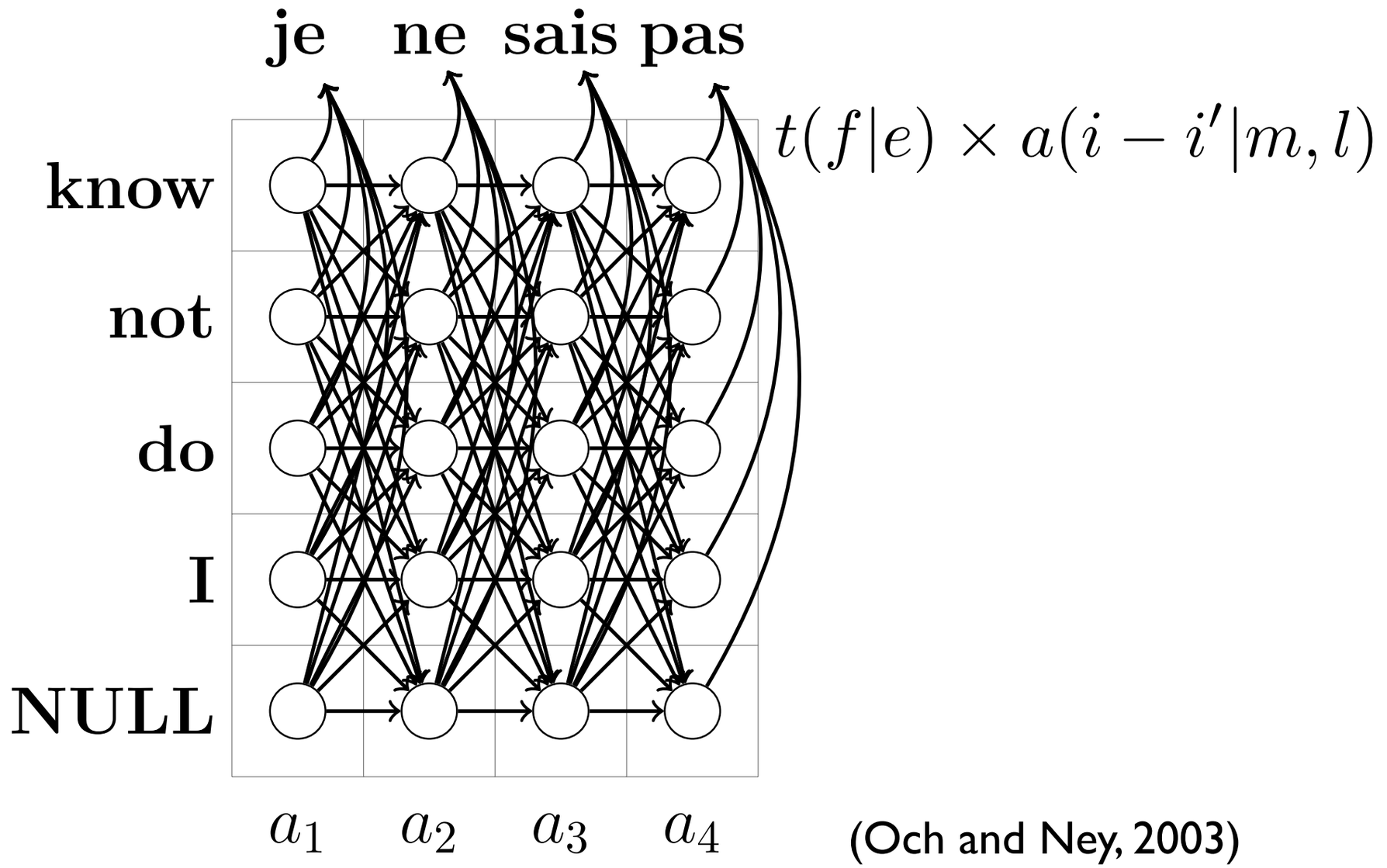
Model 2



(Brown et al., 1993)

- Model 1と同様に生成+アライメント確率

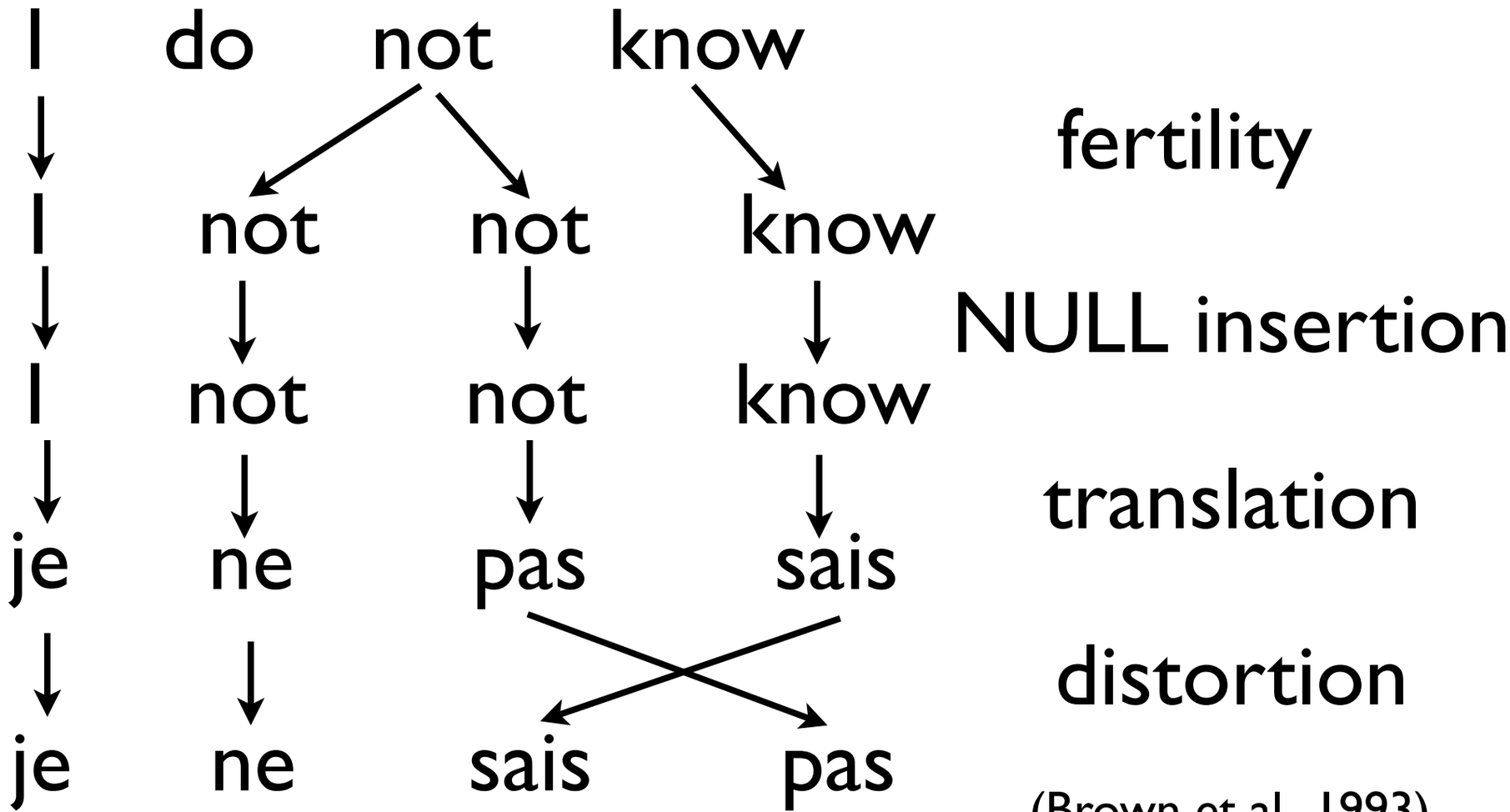
HMM Model



(Och and Ney, 2003)

- アライメント確率は、一つ前の生成に依存

Model 3-5



(Brown et al., 1993)

- Model 1やModel 2、HMMと全く異なる
- fertilityにより、明示的に一対多の関係を表現
- 動的計画法(Dynamic Programming)が使えない

他にも...(教師なし学習)

- 一対多の制約を無くしたい
  - ヒューリスティック(Och and Ney, 2003; Koehn et al., 2003)
  - 学習中に制約(Liang et al., 2006; Ganchev et al., 2008)
- Fertilityのモデル化(Zhao and Gildea, 2010; Lin and Bilmes, 2011)
- 統語論的な制約(DeNero and Klein, 2007; Pauls et al., 2010)
- 大量の素性(Berg-Kirkpatrick et al., 2010; Dyer et al., 2011)

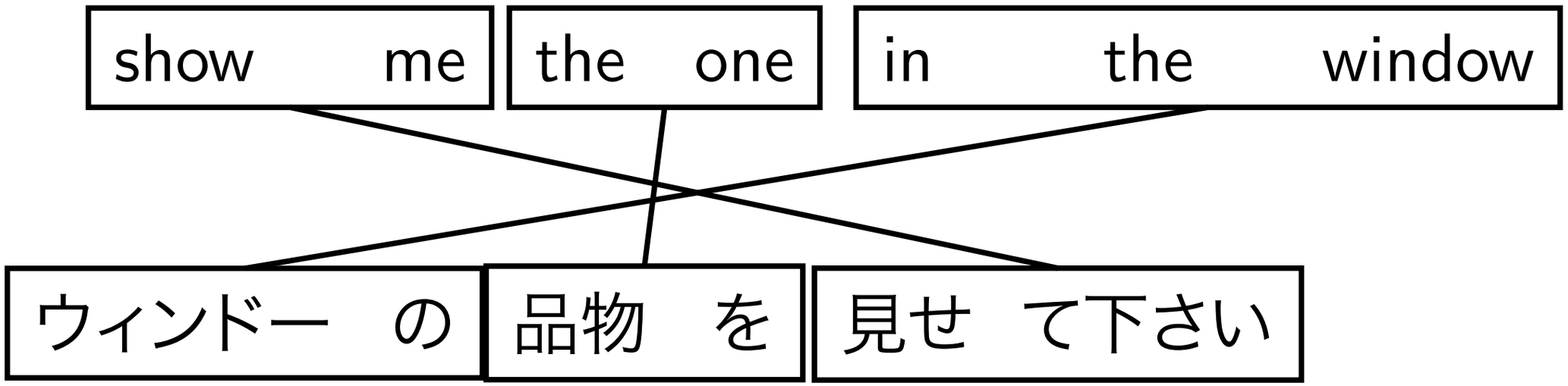
内容

- 統計的機械翻訳の枠組み
- 単語アライメント
- 句に基づく機械翻訳
- 自動評価

なぜ、句?

- フレーズ機械翻訳、句に基づく機械翻訳(Koehn et al., 2003)
- 「句」を翻訳の単位に使うと、
  - 多対多の単語アライメント + 句内部の局所的な並び替え
  - 局所的なコンテキスト + 統語的に分解不可能な句

句に基づくモデル



- Generative story:
  - fを句へと分解 + 各句を翻訳 + 並び替え

句に基づくモデル

$$\begin{aligned}\hat{\mathbf{e}} &= \underset{\mathbf{e}}{\operatorname{argmax}} \frac{\exp\left(\mathbf{w}^\top \cdot \mathbf{h}(\mathbf{e}, \phi, \mathbf{f})\right)}{\sum_{\mathbf{e}', \phi'} \exp\left(\mathbf{w}^\top \cdot \mathbf{h}(\mathbf{e}', \phi', \mathbf{f})\right)} \\ &= \underset{\mathbf{e}}{\operatorname{argmax}}\, \mathbf{w}^\top \cdot \mathbf{h}(\mathbf{e}, \phi, \mathbf{f})\end{aligned}$$

- 複数の素性h(e, Φ, f)をlog-linearに組み合わせ、最大化
- Φ: (f, e)の句単位の分割
- w: 各素性の重み付け

Questions

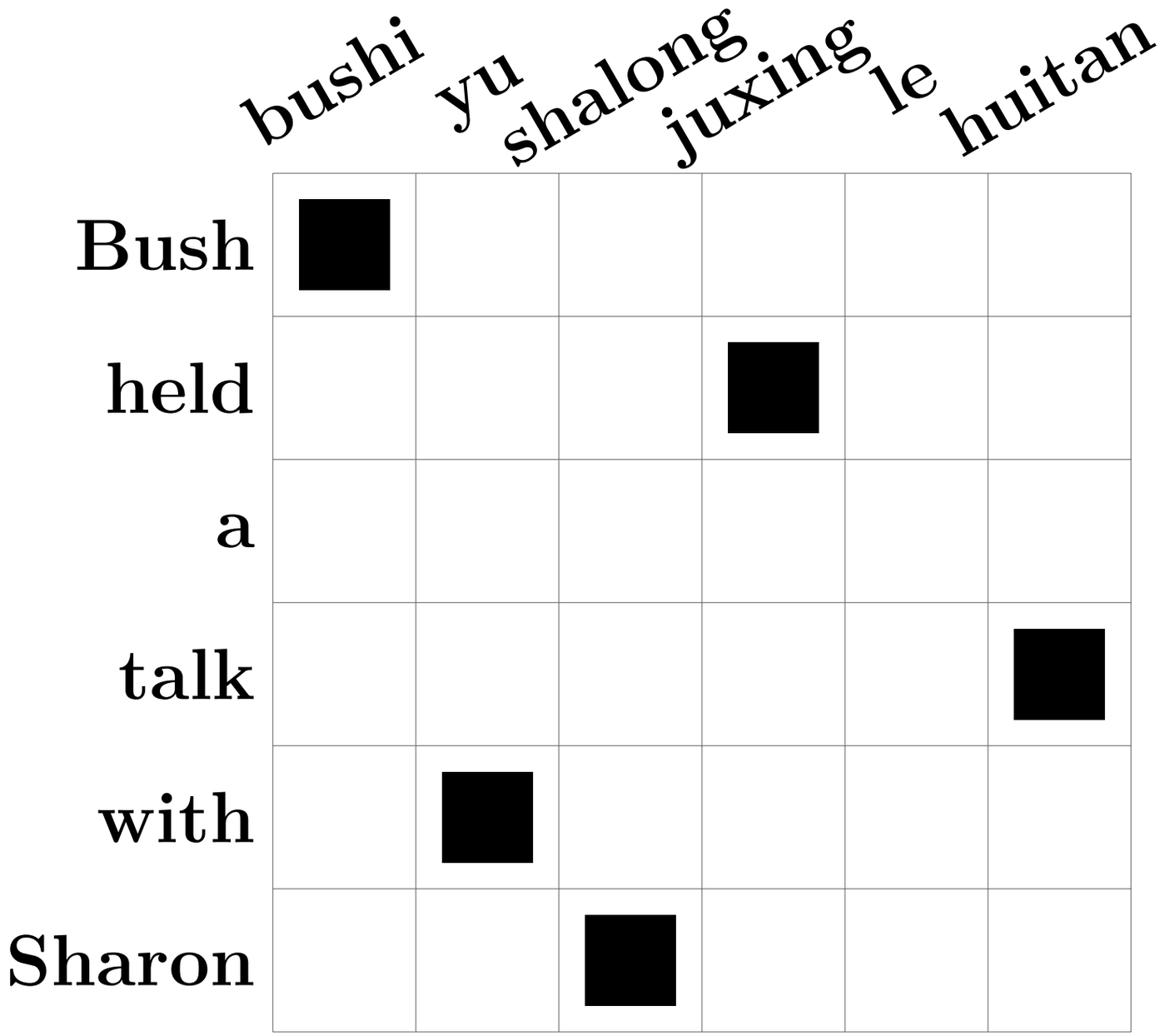
$$\hat{\mathbf{e}} = \underset{\mathbf{e}}{\operatorname{argmax}}\, \mathbf{w}^\top \cdot \mathbf{h}(\mathbf{e}, \phi, \mathbf{f})$$

- 学習: 句とパラメータをどのように学習するか (Φ and h)?
- デコード(探索): どのようにして最適な翻訳をみつけるか(argmax)?
- チューニング (最適化): どのようにして重み付けをするか(w)?

学習

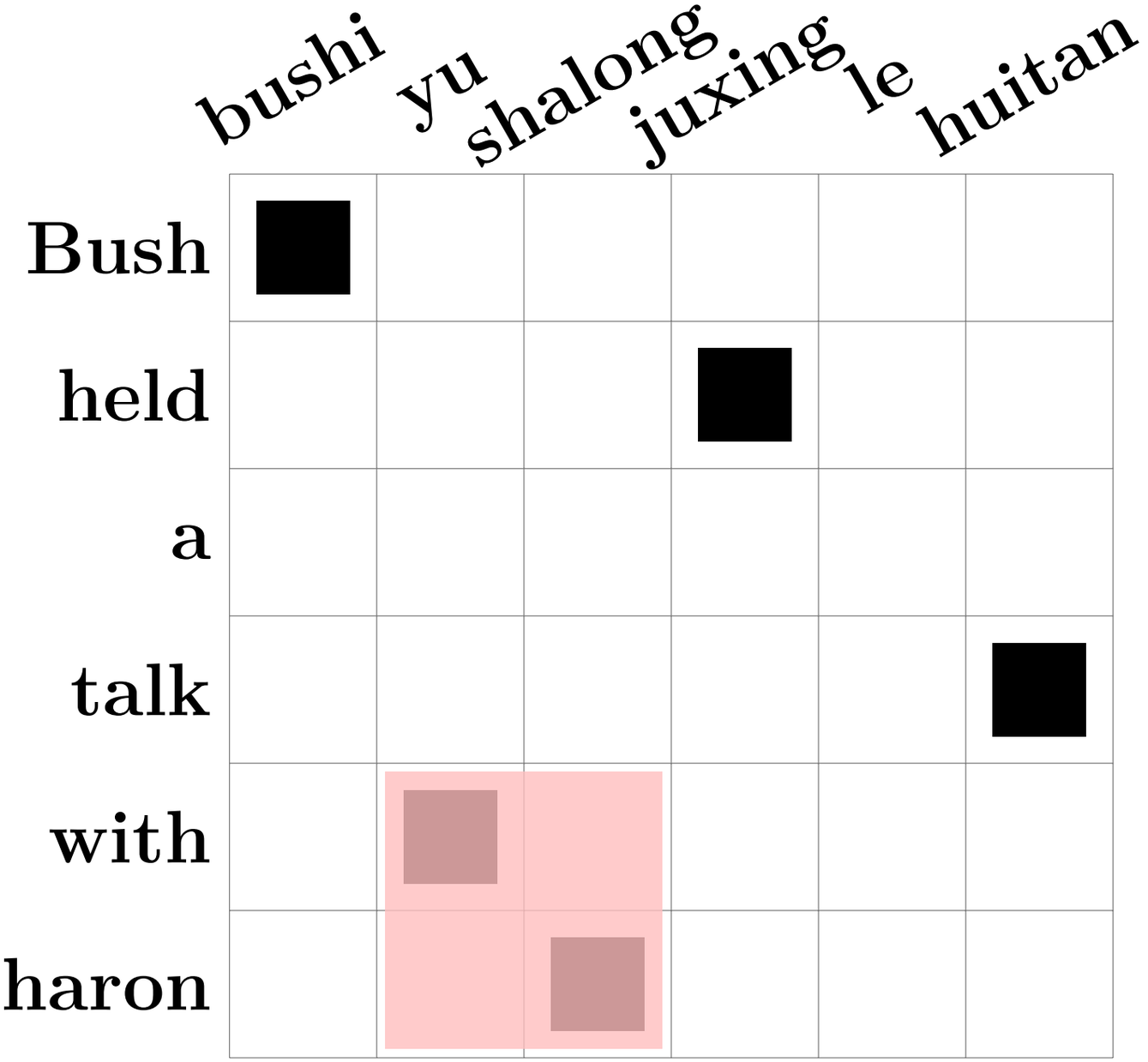
- $\mathcal{D} = \langle \mathcal{F}, \mathcal{E} \rangle$ からフーレーズペア Φ を学習
- 標準的なヒューリスティックな手法
(Koehn et al., 2003)
  - 単語アライメントの計算
  - フレーズペアの抽出
  - フレーズペアのスコアリング

単語アライメント



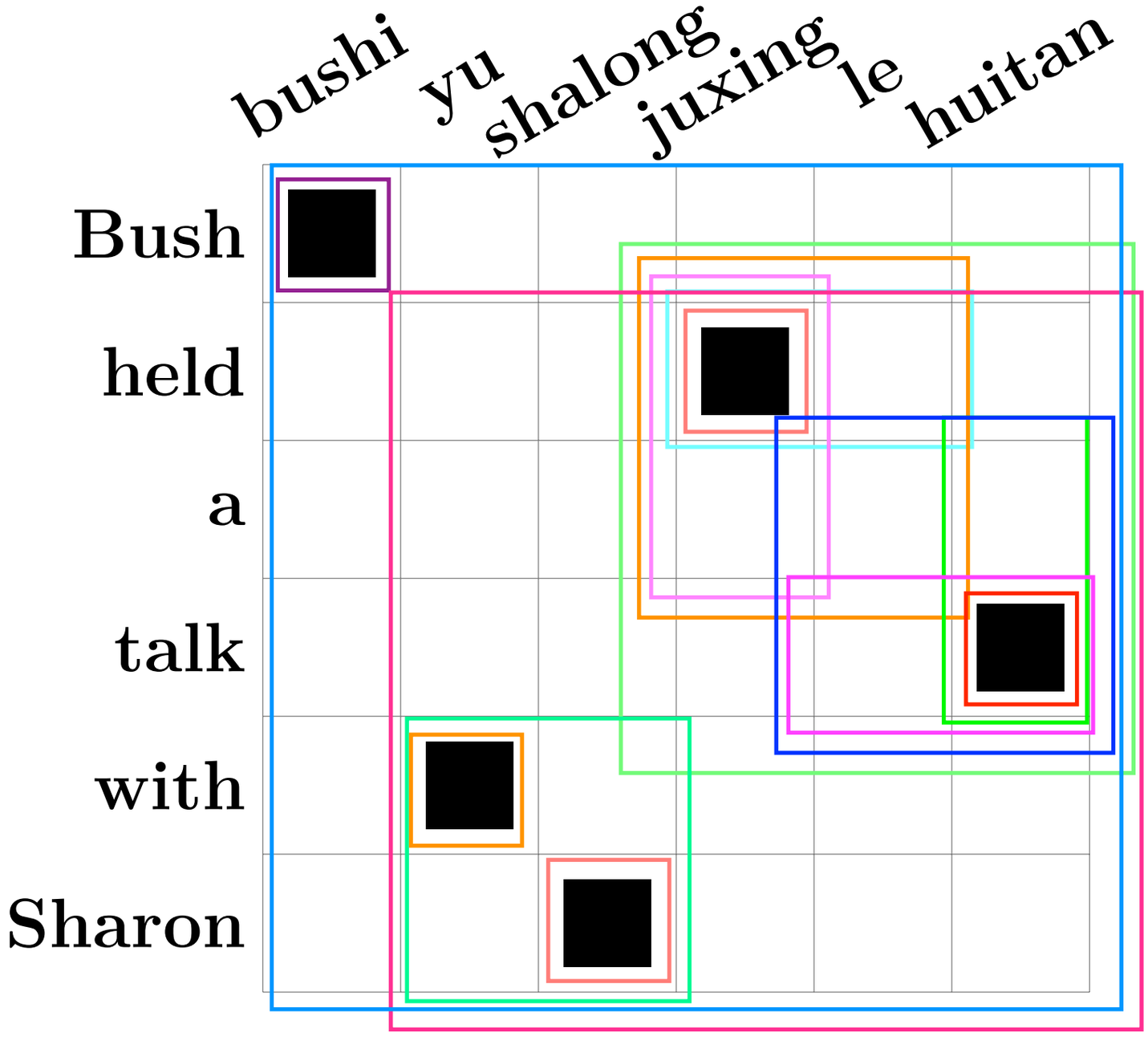
(Example from Huang and Chiang, 2007)

フレーズペアの抽出



- 一貫した句(単語アライメントが閉じている句)を抽出

網羅的に抽出


bushi
yu
shalong
juxing
le
huitan
Bush
held
a
talk
with
Sharon

句に対応した素性

$$\log p_\phi(\bar{\mathbf{f}}|\bar{\mathbf{e}}) = \log \frac{\mathrm{count}(\bar{\mathbf{e}}, \bar{\mathbf{f}})}{\sum_{\bar{\mathbf{f}}'} \mathrm{count}(\bar{\mathbf{e}}, \bar{\mathbf{f}}')}$$

$$\log p_\phi(\bar{\mathbf{e}}|\bar{\mathbf{f}}) = \log \frac{\mathrm{count}(\bar{\mathbf{e}}, \bar{\mathbf{f}})}{\sum_{\bar{\mathbf{e}}'} \mathrm{count}(\bar{\mathbf{e}}', \bar{\mathbf{f}})}$$

- データから全ての句を抽出
- 頻度に基づく、最尤推定
- 二方向の素性を使用

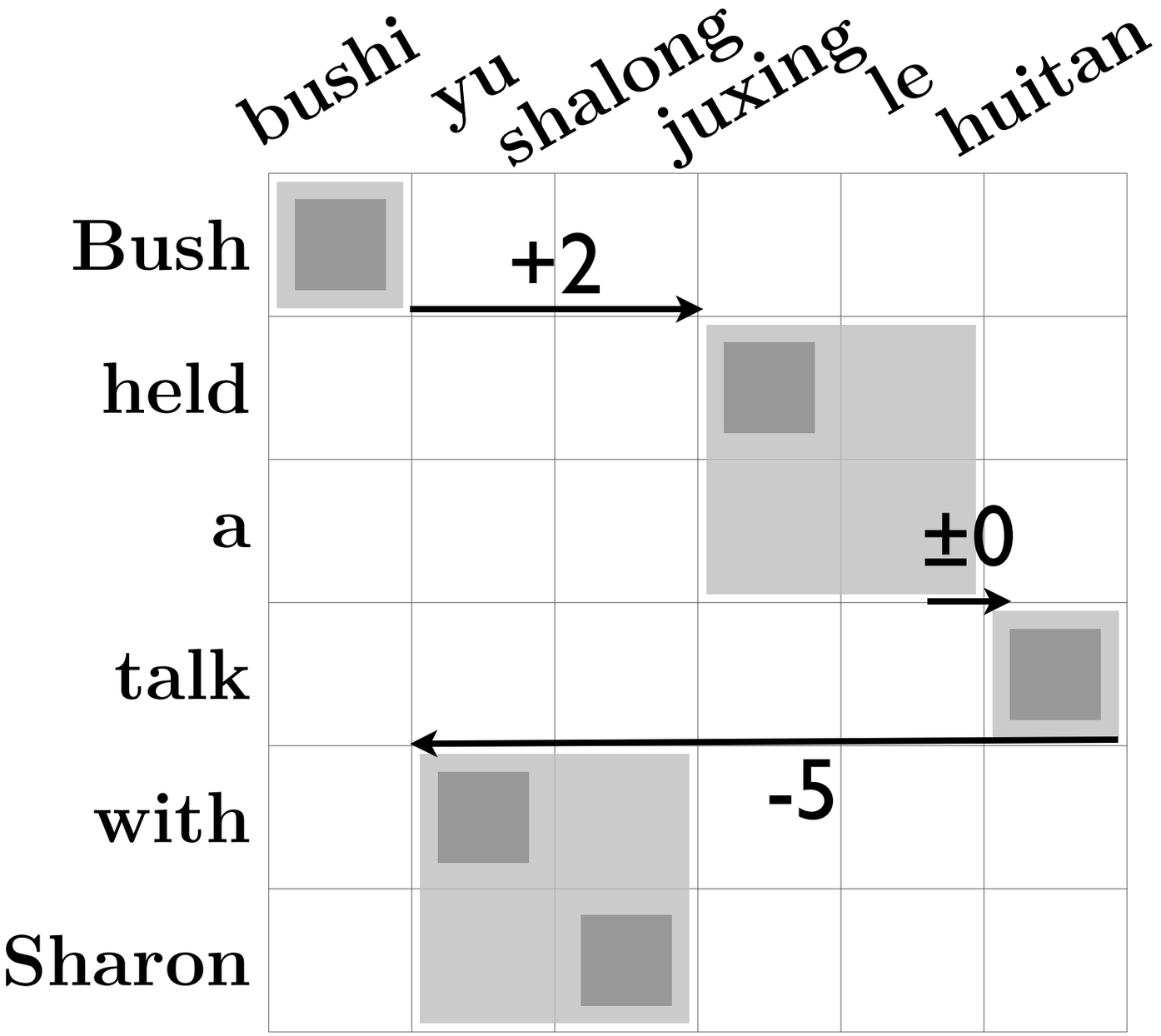
アライメントに基づく素性

$$\log p_{lex}(\bar{\mathbf{f}}|\bar{\mathbf{e}},\bar{\mathbf{a}}) = \log \prod_{i}^{|\bar{\mathbf{e}}|} \frac{1}{|\{j|(i,j)\in\bar{\mathbf{a}}\}|} \sum_{\forall(i,j)\in\bar{\mathbf{a}}} t(e_i|f_j)$$

$$\log p_{lex}(\bar{\mathbf{e}}|\bar{\mathbf{f}},\bar{\mathbf{a}}) = \log \prod_{j}^{|\bar{\mathbf{f}}|} \frac{1}{|\{i|(j,i)\in\bar{\mathbf{a}}\}|} \sum_{\forall(j,i)\in\bar{\mathbf{a}}} t(f_j|e_i)$$

- 単語アライメントモデルに基づくスコア
- 低頻度な句に対してもスコアを割り当てる

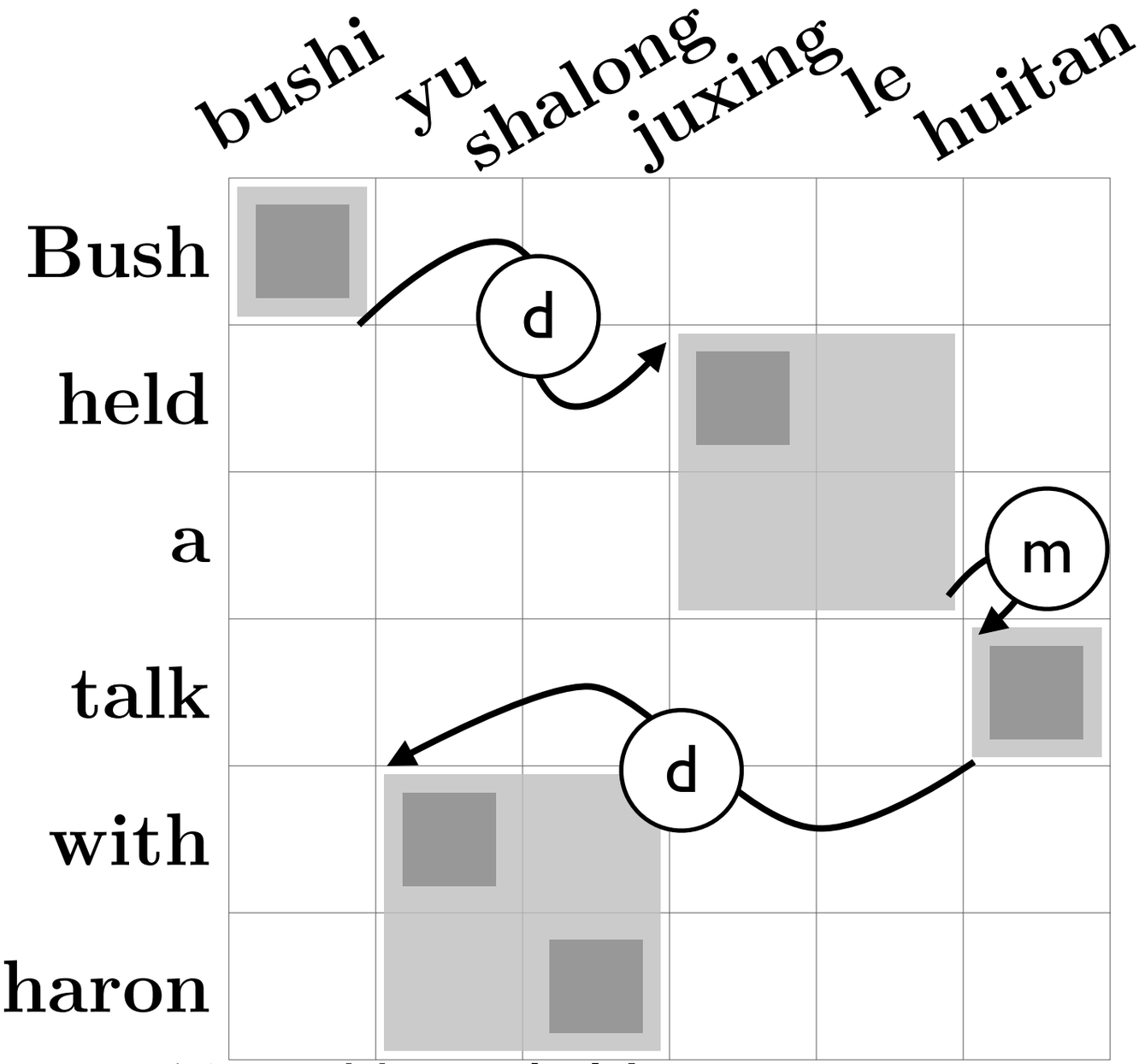
並び替え素性



- 距離に基づく素性

$$d(\mathbf{f}, \phi, \mathbf{e}) = |+2| + |0| + |-5| = 7$$

並び替え素性



- 各句ごとの並び替え素性: $\log p_o(o \in \{m, s, d\} \,|\bar{\mathbf{f}}, \bar{\mathbf{e}})$
- monotone, swap, discontinuous

他の素性

- (複数の) 言語モデル
- 単語数: 言語モデルに対するバイアス
- 句の数: 「長い」あるは「短い」句を使用

Questions

$$\hat{\mathbf{e}} = \underset{\mathbf{e}}{\operatorname{argmax}}\, \mathbf{w}^\top \cdot \mathbf{h}(\mathbf{e}, \phi, \mathbf{f})$$

- 学習: 句とパラメータをどのように学習するか (Φ and h)?
- デコード(探索): どのようにして最適な翻訳をみつけるか(argmax)?
- チューニング (最適化): どのようにして重み付けをするか(w)?

フレーズペアの列挙

Bush and

with Sharon

with

held a talk

talked

held

meeting

Bush

and

Sharon

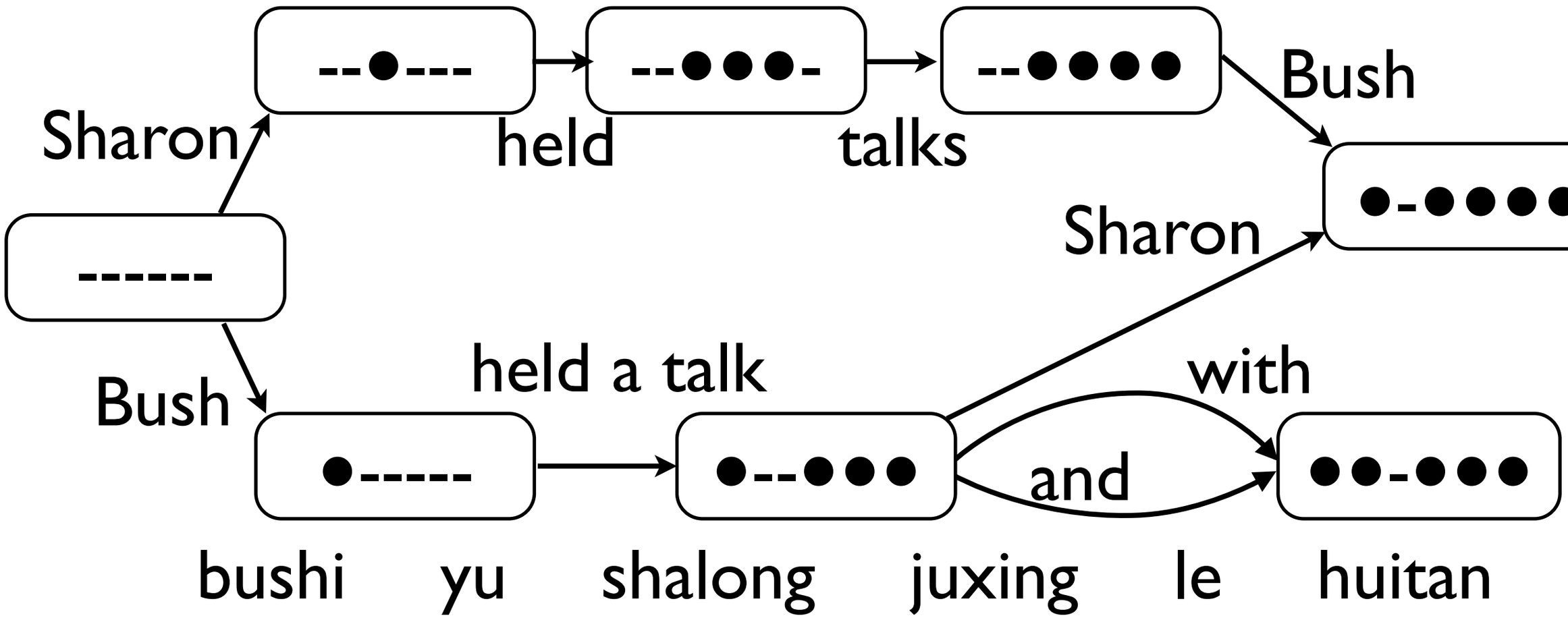
hold

talks

bushi yu shalong juxing le huitan

- 入力文fに対し、原言語側がマッチする句を列挙
- 最もよい、フレーズペアの選択 + 並び替え

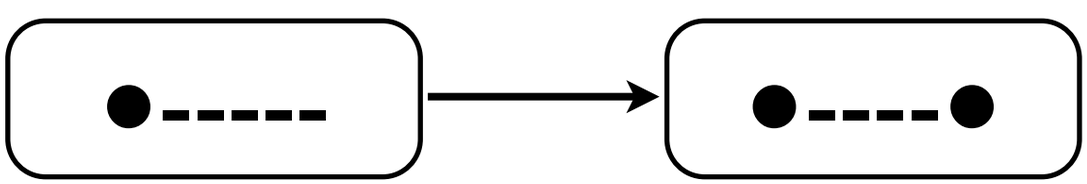
フレーズベースな探索空間



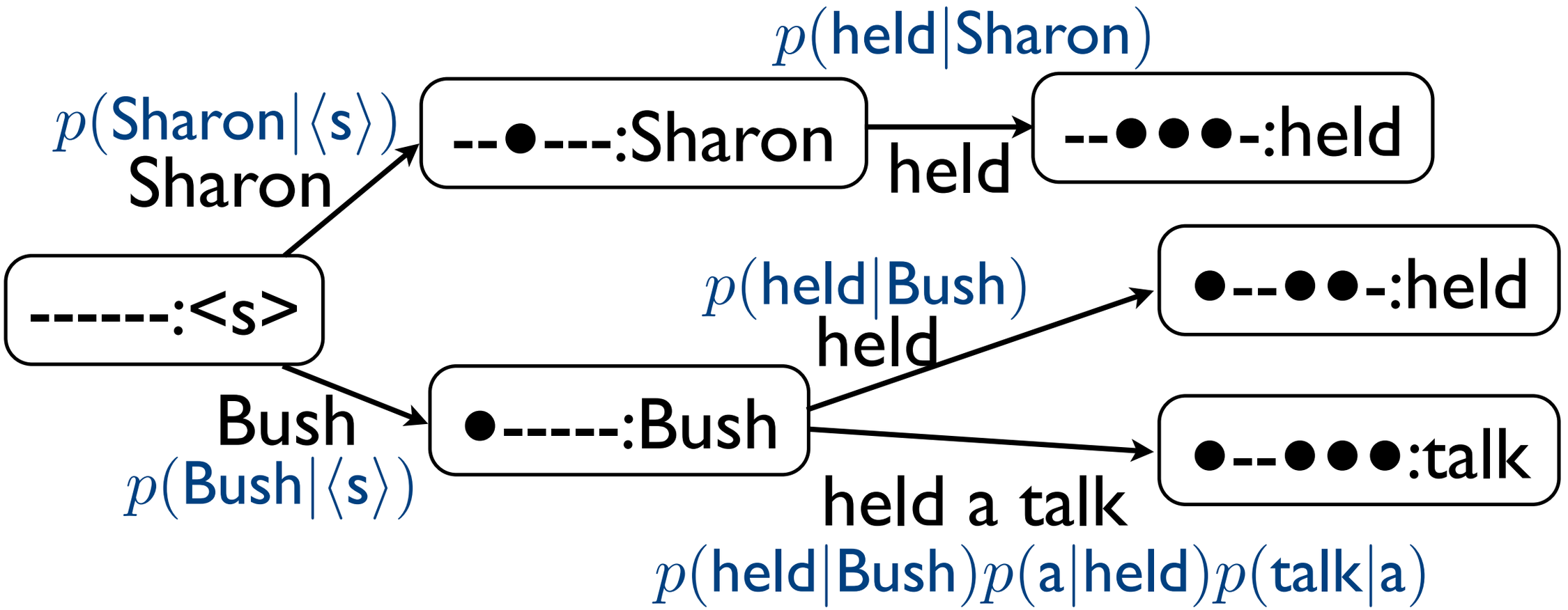
- ノード: 翻訳された原言語の単語位置を表すbit-vector
- エッジ: left-to-rightに組み合わされる目的言語側の句
- 探索空間: $O(2^n)$、時間: $O(2^n n^2)$ (Why?)

巡回セールスマン問題

- NP-hard problem:各都市を一度だけ訪れる
- 巡回セールスマン問題としてのMT(Knight, 1999)
  - 原言語の各単語 = 都市
  - 動的計画法(DP)による解:
    - State: 訪れた都市 (bit-vector)
    - 探索空間: $O(n^2)$
  - 探索空間を小さくするため、並び替えに制約

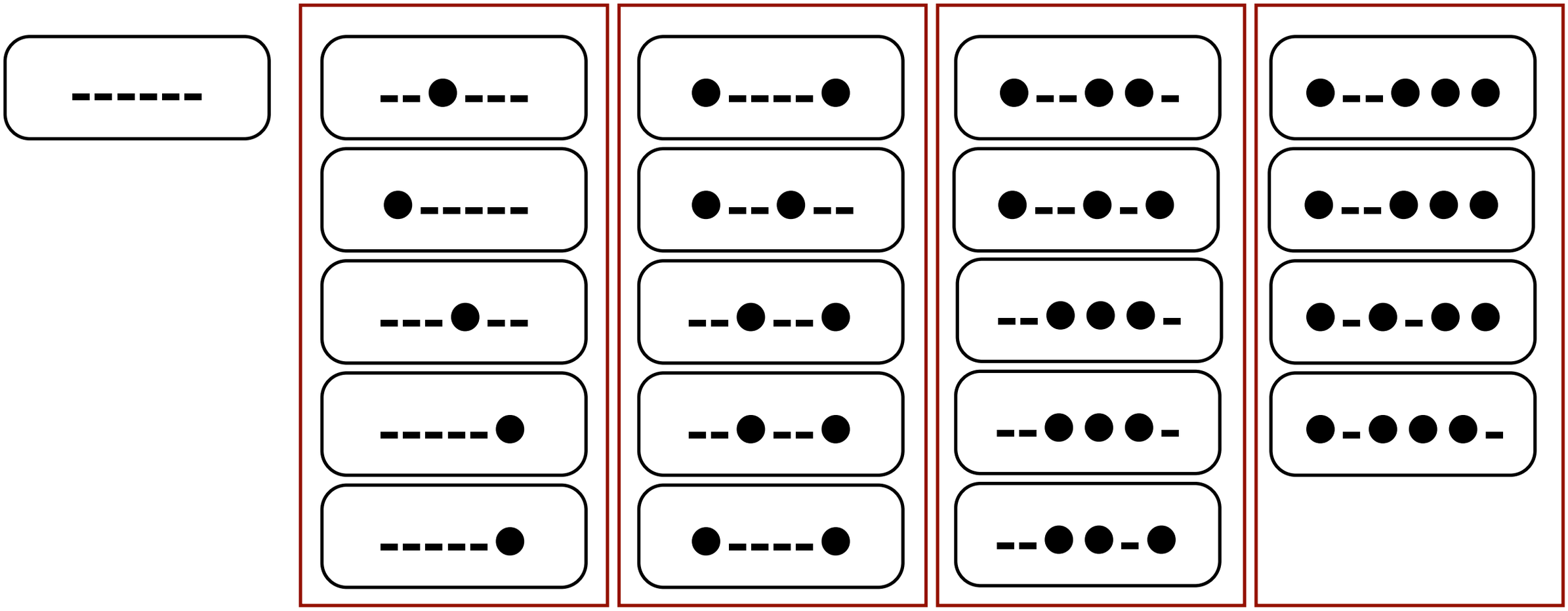
    i.e. long distortion: 

局所的でない素性



- フレーズに閉じていない素性: bigram言語モデル
- 「将来のスコアの計算」のために、一単語保持
- m-gram LM: 探索空間: $O(2^n V^{m-1})$, 時間: $O(2^n V^{m-1} n^2)$

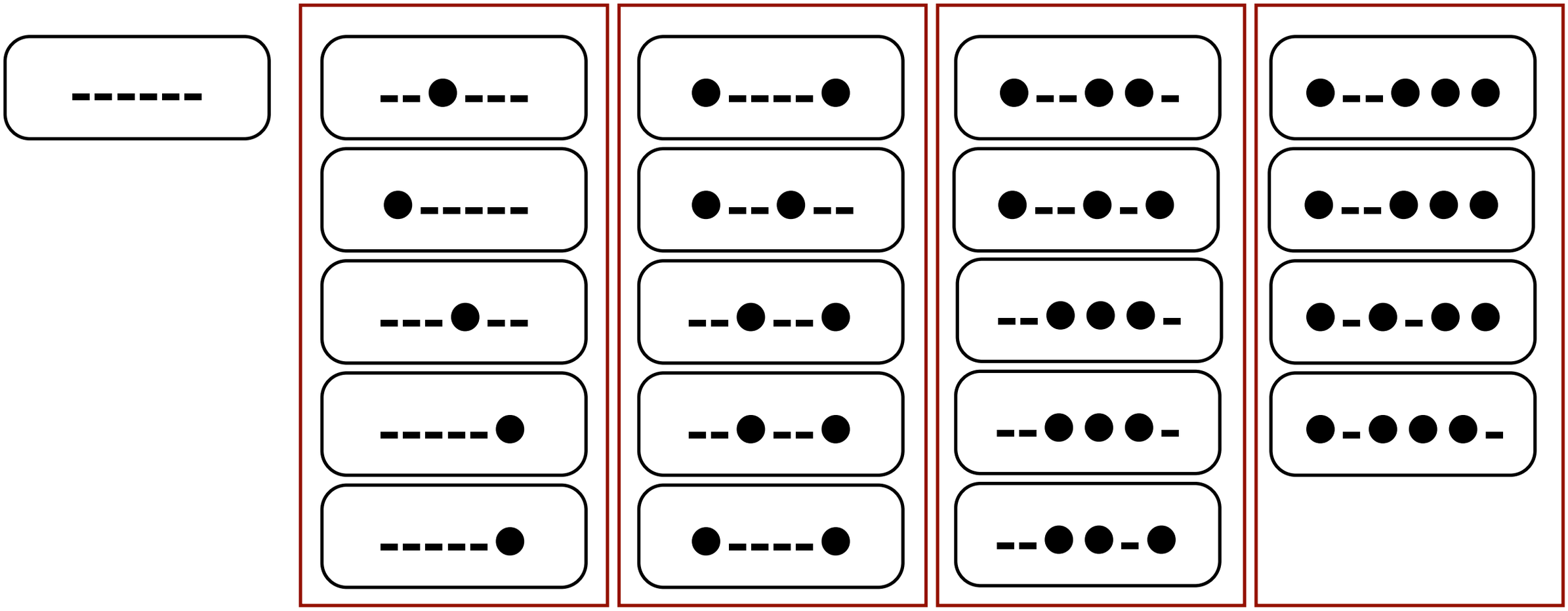
フレーズベースなデコーディング



- 探索空間を「翻訳された単語数 = cardinality」でグループ化

- 小さいcardinalityを持つ仮説から展開

プルーニング



- 同じグループの仮説内部でプルーニング
- 数あるいはスコアによるプルーニング
- O(2ⁿ)の項をO(nb)へ縮小

Questions

$$\hat{\mathbf{e}} = \underset{\mathbf{e}}{\operatorname{argmax}}\, \mathbf{w}^\top \cdot \mathbf{h}(\mathbf{e}, \phi, \mathbf{f})$$

- 学習: 句とパラメータをどのように学習するか (Φ and h)?
- デコード(探索): どのようにして最適な翻訳をみつけるか(argmax)?
- チューニング (最適化): どのようにして重み付けをするか(w)?

チューニング

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{s=1}^{S} \ell(\underset{\mathbf{e}}{\operatorname{argmax}}\, \mathbf{w}^\top \cdot \mathbf{h}(\mathbf{e}, \mathbf{f}_s), \mathbf{e}_s)$$

- MERT (Minimum Error Rate Training) (Och, 2003)
- 統計的機械翻訳では標準(でも他のNLPなタスクでは使われない)
  - l(.)に対して、様々なエラー関数を使用可能(BLEU)
  - ∑に対して、エラー関数に特有な操作が可能(BLEU)
  - 10+程度の整数値の素性

MERT

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{s=1}^{S} \ell(\underset{\mathbf{e}}{\operatorname{argmax}}\, \mathbf{w}^\top \cdot \mathbf{h}(\mathbf{e}, \mathbf{f}_s), \mathbf{e}_s)$$

- 制約なし最小化: Powell法 、 Downhill-Simplex法
- wを更新するたびに、argmaxを計算し直さないといけない
- n-bestにより、eの空間を近似 (Och and Ney, 2002)

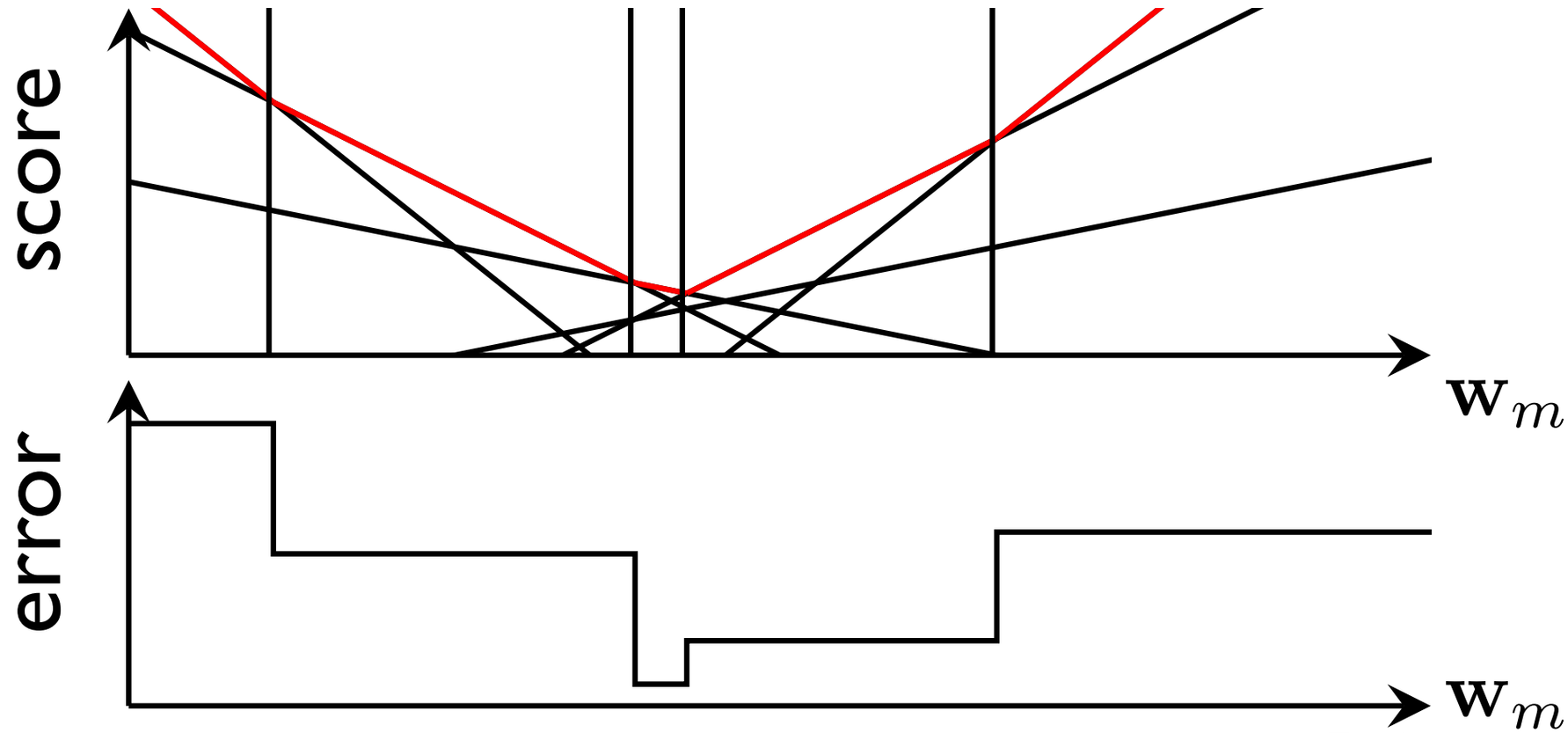
n-best 結合による近似

1: **procedure** MERT($\{(\mathbf{e}_s, \mathbf{f}_s)\}_{s=1}^{S}$)
2: **for** $n = 1...N$ **do**
3: Decode and generate nbest list using $\mathbf{w}$
4: Merge nbest list
5: **for** $k = 1...K$ **do**
6: **for** each parameter $m = 1...M$ **do**
7: Solve one dimensional optimization
8: **end for**
9: update $\mathbf{w}$
10: **end for**
11: **end for**
12: **end procedure**

- 現在のwでn-bestを生成、結合(N回)
  - M次元(M = 素性の数)の各次元に対して、最適化、wを更新(K回)

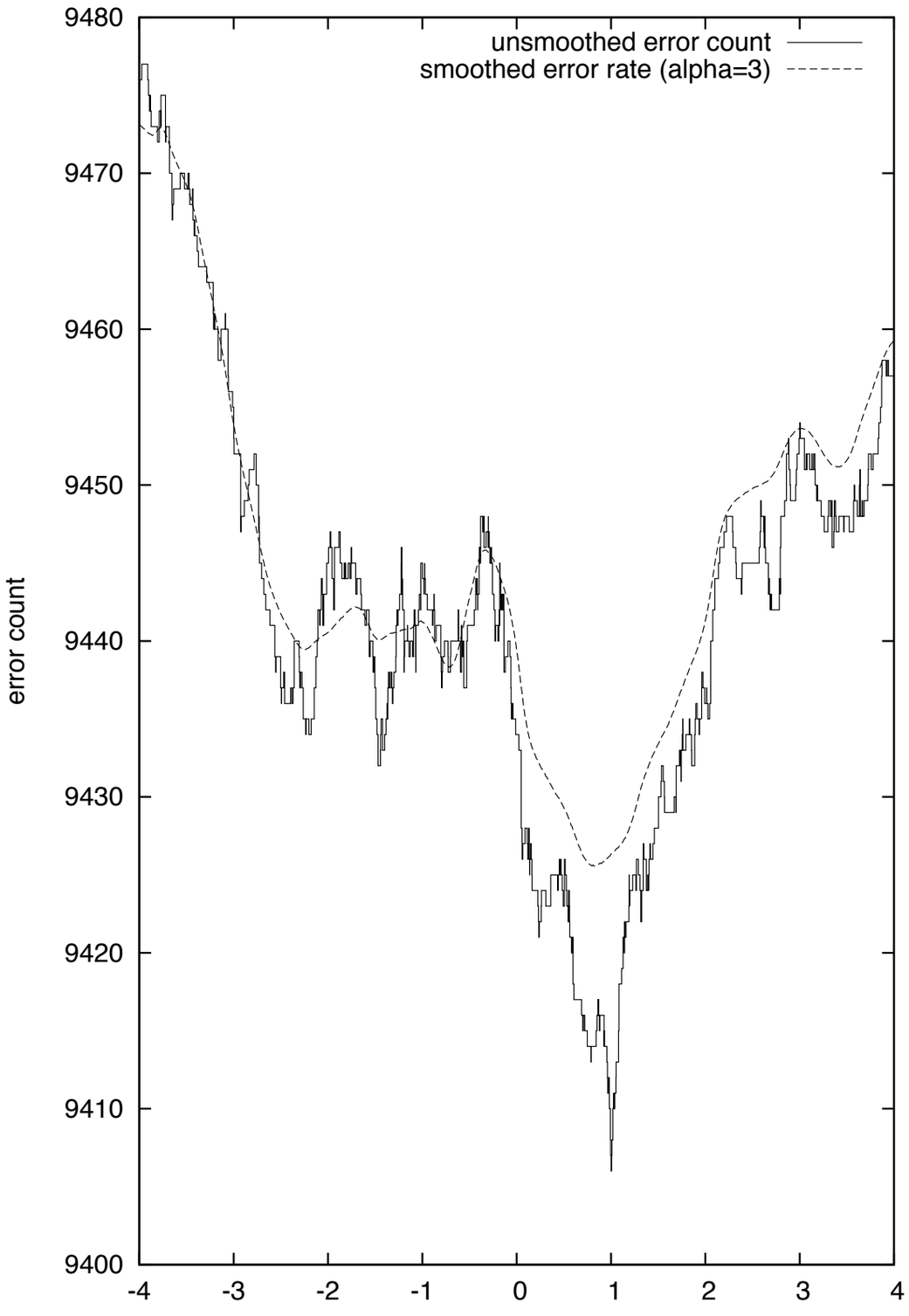
Line Searchによる効率化

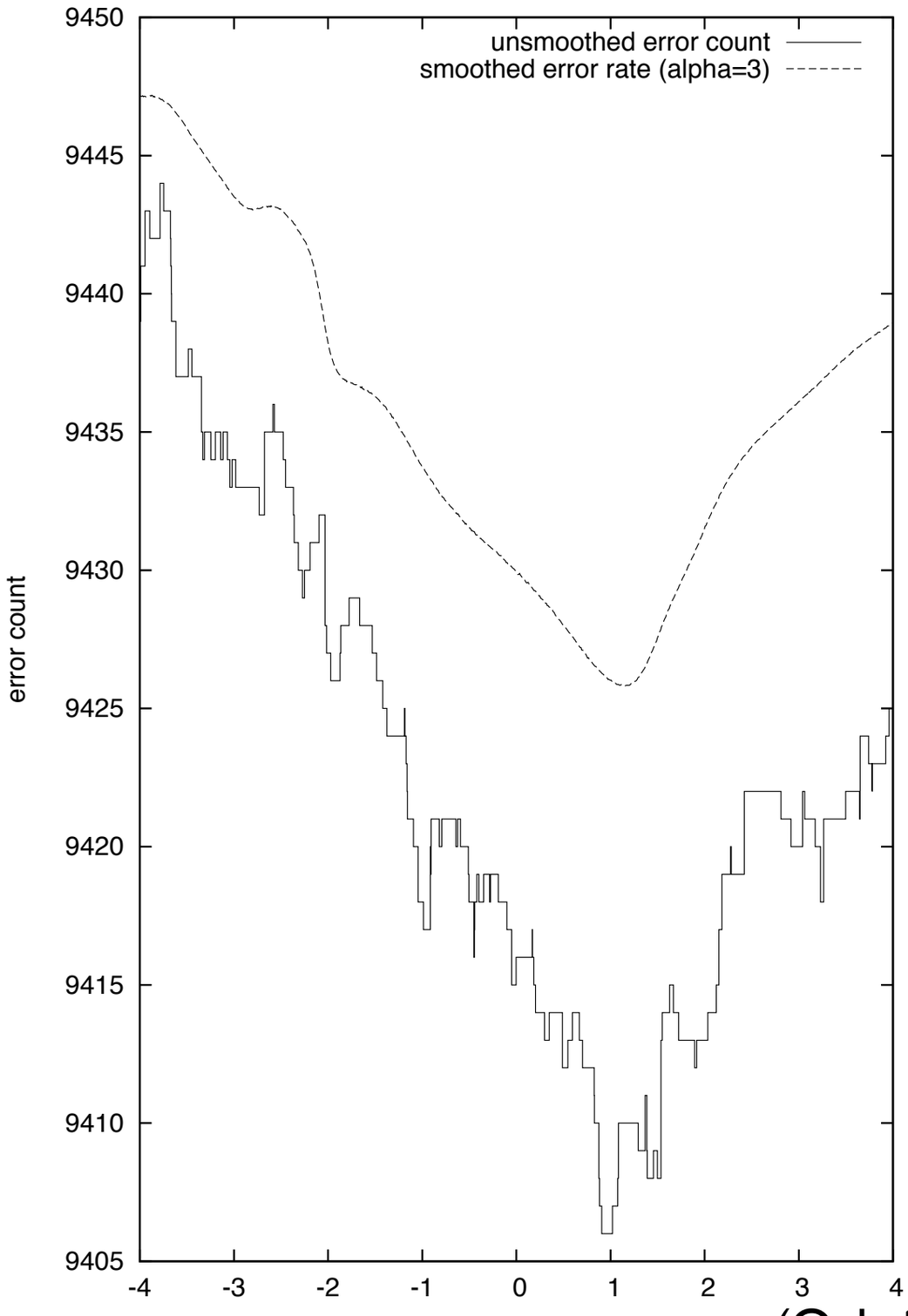
$$\hat{\mathbf{e}} = \underset{\mathbf{e}}{\operatorname{argmax}}\, \mathbf{w}_m^\top \cdot \underbrace{\mathbf{h}_m(\mathbf{e}, \mathbf{f}_s)}_{\text{slope}} + \underbrace{\mathbf{w}_{m\_}^\top \cdot \mathbf{h}_{m\_}(\mathbf{e}, \mathbf{f}_s)}_{\text{constant}}$$



- 一つの次元を選択した場合、その仮説を「線」として見なせる
- 「線」の集合から、凸包(convex hull)を計算

エラー曲線





(Och, 2003)

MERTの現実

- ランダムな初期値 (Macherey et al., 2008; Moore and Quirk, 2008)
- ランダムな方向 (Macherey et al., 2008)
- エラーの統計量のスムージング (Cer et al., 2008)
- Regularization(Hayashi et al., 2009)
- Forest/LatticeからのMERT(Macherey et al., 2008; Kumar et al., 2009)
- 凸包を計算、その後最適化 (Galley and Quirk, 2011)
- 最低3回MERT、平均BLEUを報告しなさい(Clark et al., 2011)(そんなアホな)

Answered?

- 文法のないモデル(でも結構頑健)
- 高速なデコーディング
- なぜMERT? (整数値を使った素性に結構強い)

内容

- 統計的機械翻訳の枠組み
- 単語アライメント
- 句に基づく機械翻訳
- 自動評価

評価: ngramの適合率

Well , I 'd like to stay five nights beginning October twenty-fifth to thirty .

- I 'd like to stay there for five nights , from October twenty fifth to the thirtieth .
- I want to stay for five nights , from October twenty fifth to the thirtieth .
- I 'd like to stay for five nights , from October twenty fifth to the thirtieth .
- I would like to reserve a room for five nights , from October twenty fifth to the thirtieth .

評価: ngramの適合率

Well , I 'd like to stay five nights beginning October twenty-fifth to thirty .

$$p_1 = \frac{11}{15}$$

- I 'd like to stay there for five nights , from October twenty fifth to the thirtieth .
- I want to stay for five nights , from October twenty fifth to the thirtieth .
- I 'd like to stay for five nights , from October twenty fifth to the thirtieth .
- I would like to reserve a room for five nights , from October twenty fifth to the thirtieth .

評価: ngramの適合率

Well , I 'd like to stay five nights beginning October twenty-fifth to thirty .

$$p_1 = \frac{11}{15} \qquad p_2 = \frac{5}{14}$$

- I 'd like to stay there for five nights , from October twenty fifth to the thirtieth .
- I want to stay for five nights , from October twenty fifth to the thirtieth .
- I 'd like to stay for five nights , from October twenty fifth to the thirtieth .
- I would like to reserve a room for five nights , from October twenty fifth to the thirtieth .

評価: ngramの適合率

Well , I 'd like to stay five nights beginning October twenty-fifth to thirty .

$$p_1 = \frac{11}{15} \qquad p_2 = \frac{5}{14} \qquad p_3 = \frac{3}{13}$$

- I 'd like to stay there for five nights , from October twenty fifth to the thirtieth .
- I want to stay for five nights , from October twenty fifth to the thirtieth .
- I 'd like to stay for five nights , from October twenty fifth to the thirtieth .
- I would like to reserve a room for five nights , from October twenty fifth to the thirtieth .

評価: ngramの適合率

Well , I 'd like to stay five nights beginning October twenty-fifth to thirty .

$$p_1 = \frac{11}{15} \qquad p_2 = \frac{5}{14} \qquad p_3 = \frac{3}{13} \qquad p_4 = \frac{2}{12}$$

- I 'd like to stay there for five nights , from October twenty fifth to the thirtieth .
- I want to stay for five nights , from October twenty fifth to the thirtieth .
- I 'd like to stay for five nights , from October twenty fifth to the thirtieth .
- I would like to reserve a room for five nights , from October twenty fifth to the thirtieth .

評価: BLEU

$$\exp\left(\sum_{n=1}^{N} w_n \log p_n + \min(1 - \frac{r}{c}, 0)\right)$$

- 重み付け適合率(Papineni et al., 2002)

- brevity penalty:短すぎる文に対するペナルティー
  - r = 参照訳の長さ, c = 翻訳の長さ
  - 複数の参照役の場合、cに「近い、短い」長さ
  - ドキュメント全体に対するスコア

なぜBLEU?

- 標準的な評価尺度として10年以上: BLEUと共にSMTは発展
  - ngramなので扱いやすい
  - 文に対して非線形な分解(必ずコーパス単位にスコアを計算、最適化困難)
  - BP問題(Chiang et al., 2009):ある文で長い翻訳を生成しても、他の文で短い翻訳を生成しても同じペナルティー
- 他にも: NIST(Doddington, 2002), METEOR(Banerjee and Lavie, 2005), TER(Snover et al., 2006), RIBES(Isozaki et al., 2010) etc.

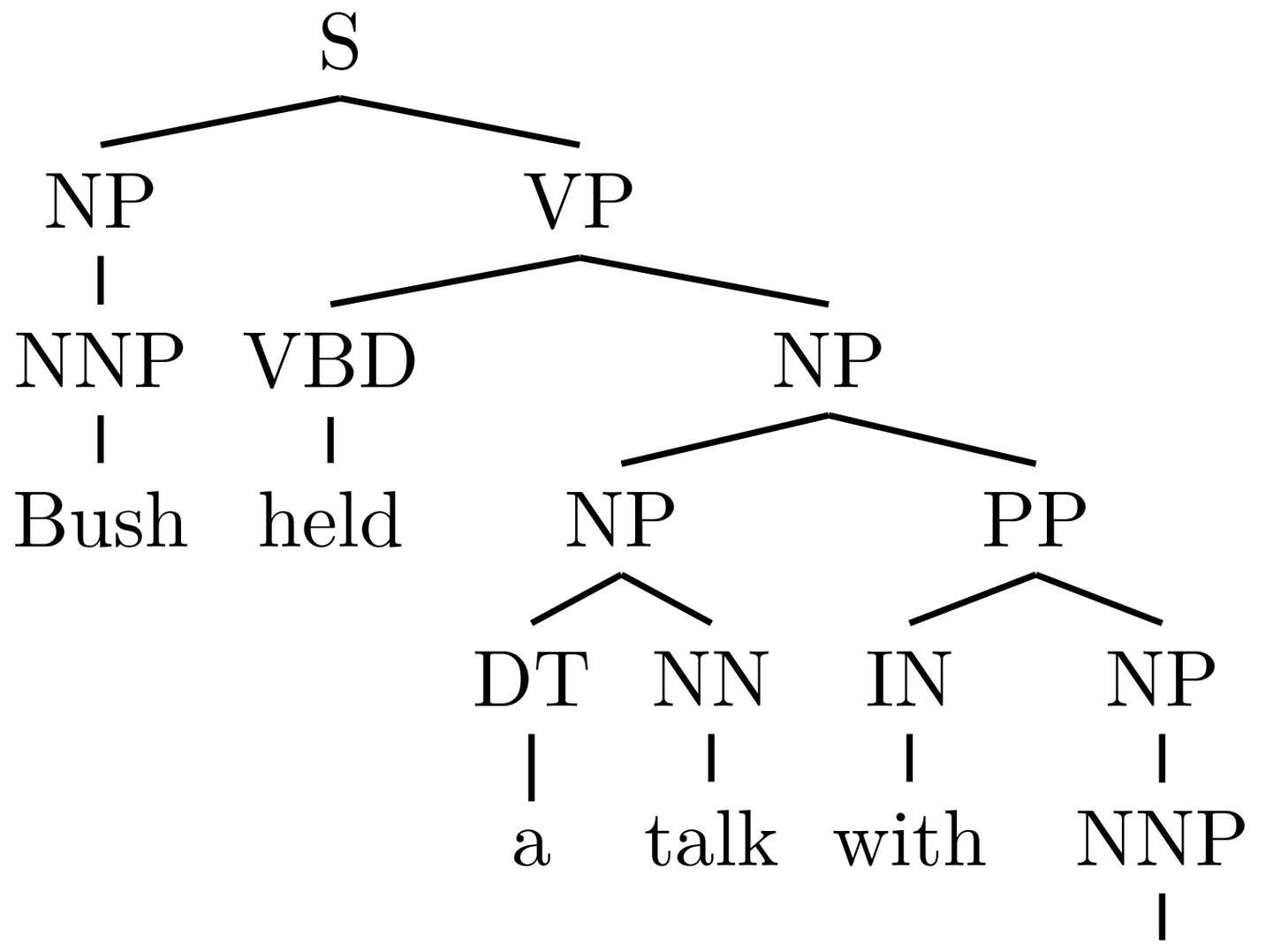
統計的機械翻訳の最先端

内容

- 木構造に基づく機械翻訳
  - 背景: CFG, hypergraph, deductive system
  - 同期文脈自由文法 (synchronous-CFG)
    - 同期文法: {string,tree}-to-{string,tree}
    - 二言語の構文解析(biparsing)
    - 同期から非同期
- 最適化

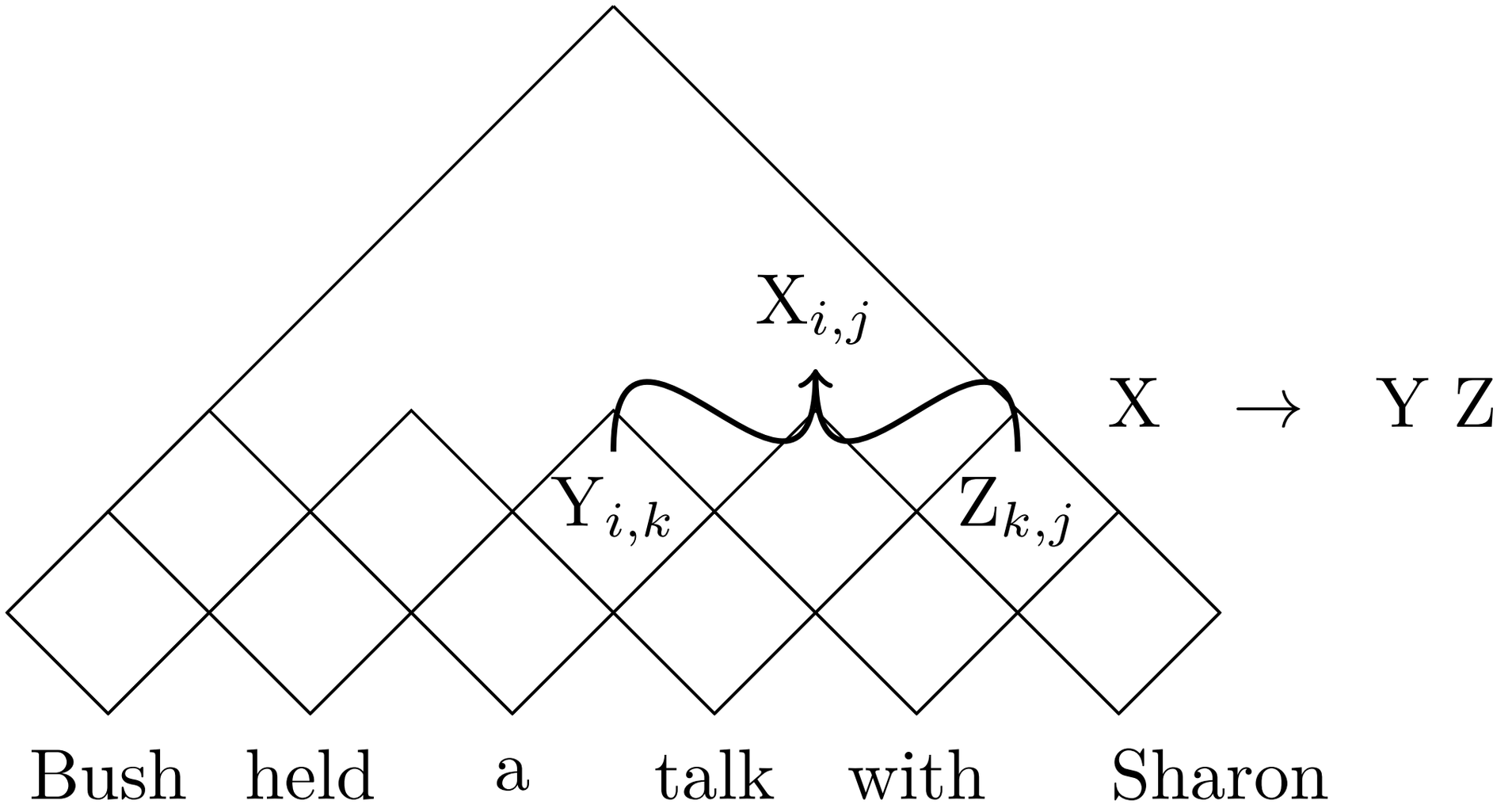
背景: CFG

$$
\begin{array}{rcl}
\mathrm{S} & \rightarrow & \mathrm{NP\ VP} \\
\mathrm{NP} & \rightarrow & \mathrm{NNP} \\
\mathrm{NP} & \rightarrow & \mathrm{NP\ PP} \\
\mathrm{NP} & \rightarrow & \mathrm{DP\ NN} \\
\mathrm{NP} & \rightarrow & \mathrm{DT\ NN} \\
\mathrm{VP} & \rightarrow & \mathrm{VBD\ NP} \\
\mathrm{NNP} & \rightarrow & \mathrm{Bush} \\
\mathrm{VBD} & \rightarrow & \mathrm{held} \\
 & \vdots &
\end{array}
$$



- 構文解析 = CFGと文字列(正規文法)との交差(intersection)

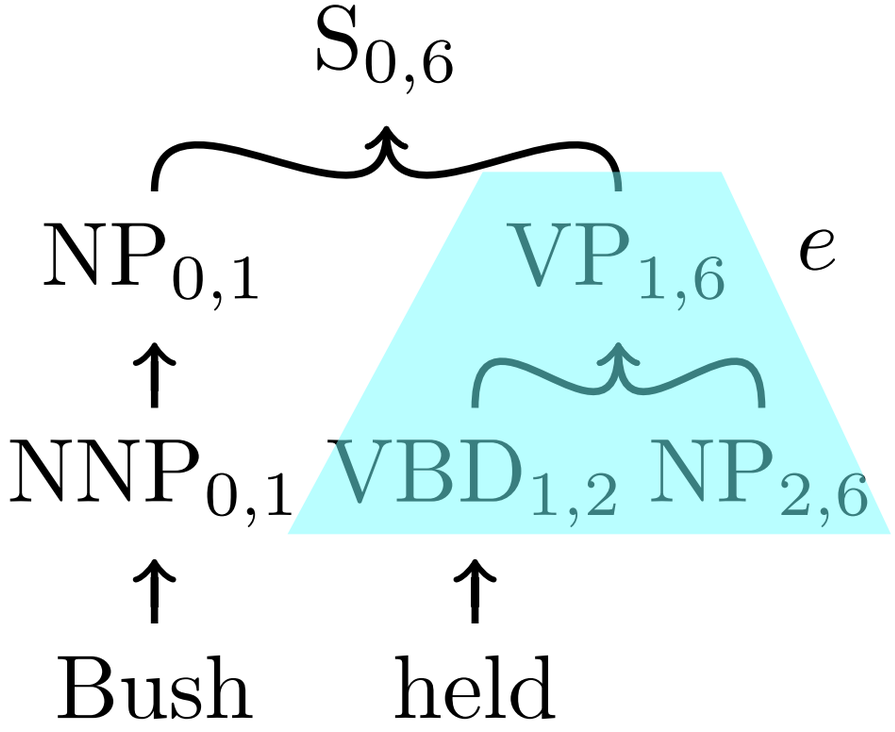
構文解析: CKY



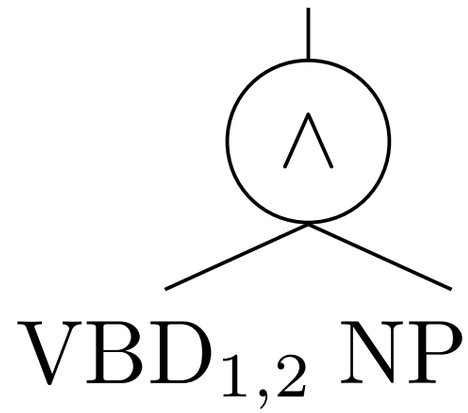
- $O(n^3)$: 各長さn、各位置i、各ルール X → Y Z、各分岐点k

- (Bottom-up) topological order

Hypergraph



$$e = \langle \underbrace{\mathrm{VP}_{1,6}}_{h(e)}, \underbrace{\{\mathrm{VBD}_{1,2}, \mathrm{NP}_{2,6}\}}_{T(e)} \rangle$$



(Klein and Manning, 2001)

- グラフの一般化:
  - h(e): 超辺 (hyperedge) eのheadノード、T(e): 超辺eのtailノード、arity = |T(e)|
  - 超辺 =インスタンス化されたルール
- and-or グラフとしても表記可能

Deductive System

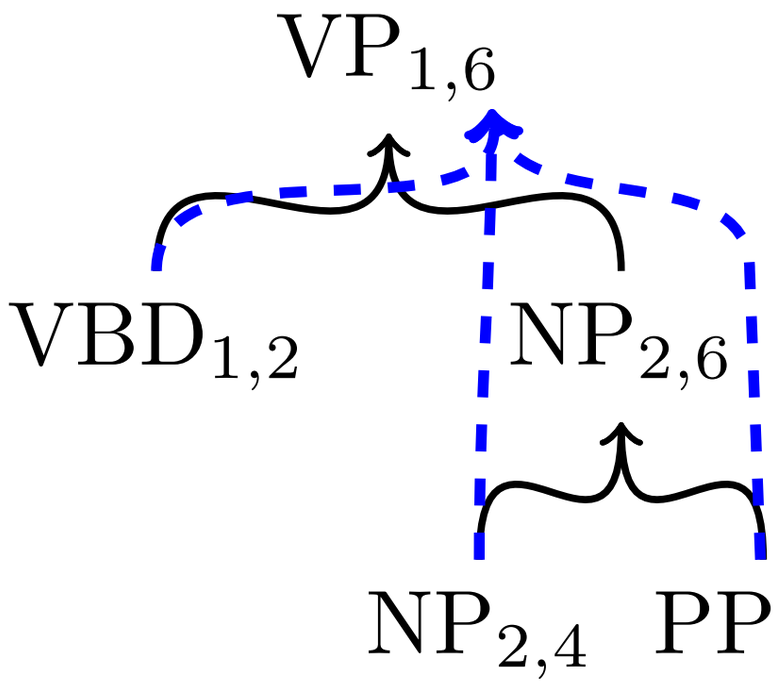
$$\underbrace{\mathrm{VBD}_{1,2} \quad \mathrm{NP}_{2,6}}_{\mathrm{VP}_{1,6}}$$

$$\frac{\overbrace{\mathrm{VBD}_{1,2}\ \mathrm{NP}_{2,6}}^{\text{antecedents}}}{\underbrace{\mathrm{VP}_{1,6}}_{\text{consequent}}}\ \mathrm{VP}_{[i,j]} \rightarrow \mathrm{VBZ}_{[j,k]}\ \mathrm{NP}_{[i,k]}$$

(Shieber et al., 1995)

- 構文解析アルゴリズムは、演繹法(deduction system)で記述可能
- 公理(axiom)から始め、goalへたどり着くまで推論規則を適用
- 前件(antecedent)が証明されたら、その後件(consequent)が証明される
- 推論規則の導出 = 超辺

Packed Forest



$$\frac{\mathrm{VBD}_{1,2} \quad \frac{\mathrm{NP}_{2,4} \ \mathrm{PP}_{4,6}}{\mathrm{NP}_{2,6}}}{\mathrm{VP}_{1,6}}$$

$$\frac{\mathrm{VBD}_{1,2} \ \mathrm{NP}_{2,4} \ \mathrm{PP}_{4,6}}{\mathrm{VP}_{1,6}}$$

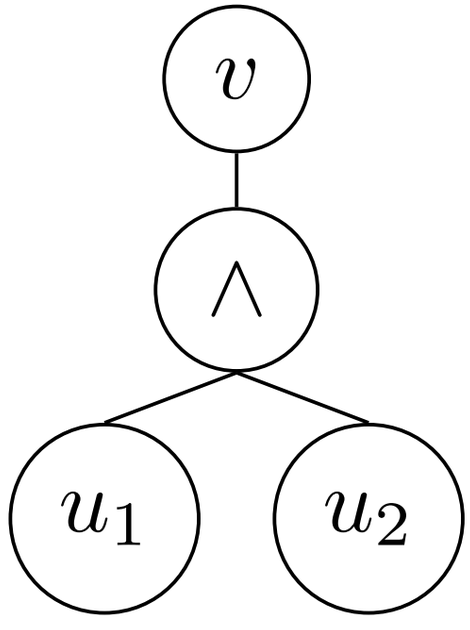
(Klein and Manning, 2001; Huang and Chiang, 2005)

- ノードを共有することにより、複数の導出をコンパクトに表現
- 一つの導出 = 木

Summary of Formalisms

hypergraph	AND/OR graph	CFG	deductive system
vertex	OR-node	symbol	item
source-vertex	leaf OR-node	terminal	axiom
target-vertex	root OR-node	start symbol	goal item
hyperedge	AND-node	production	instantiated deduction

$\langle v, \{u_1, u_2\} \rangle$



$v \rightarrow u_1 \ u_2$

$$\frac{u_1 \ u_2}{v}$$

Weight and Semiring

$$\mathrm{VP} \xrightarrow{w_1} \mathrm{VBD\ NP}$$

$$\mathrm{NP} \xrightarrow{w_2} \mathrm{NP\ PP}$$

$$\mathrm{VP}_{1,6} : w_1 \otimes c \otimes d$$

$$\mathrm{VBD}_{1,2} : c \quad \mathrm{NP}_{2,6} : d$$

$$\frac{\mathrm{VBD}_{1,2} : c\ \mathrm{NP}_{2,6} : d}{\mathrm{VP}_{1,6} : w_1 \otimes c \otimes d} : w_1$$

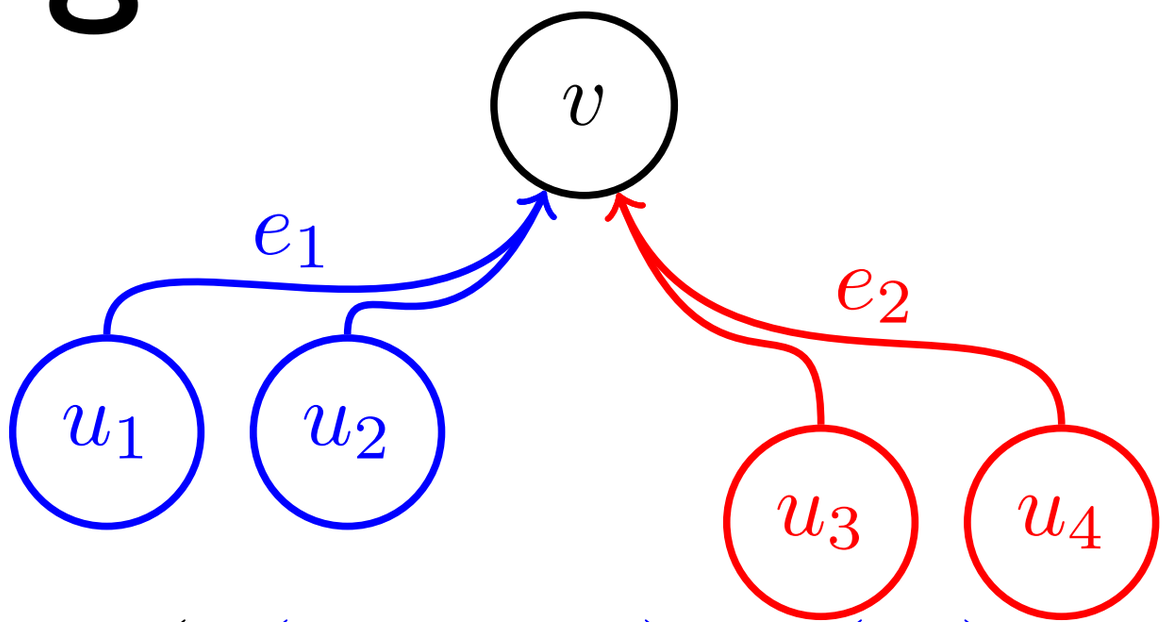
$$\mathrm{NP}_{2,6} : w_2 \otimes a \otimes b$$

$$\mathrm{NP}_{2,4} : a \quad \mathrm{PP}_{4,6} : b$$

$$\frac{\mathrm{NP}_{2,4} : a\ \mathrm{PP}_{4,6} : b}{\mathrm{NP}_{2,6} : w_2 \otimes a \otimes b} : w_2$$

- WFSTのように、各超辺にweightを関連付ける
- $\otimes$: extension (multiplicative), $\oplus$: summary (additive)

Weight and Semiring



$$d(v) = (w(e_1, u_1, u_2) \otimes d(u_1) \otimes d(u_2)) \oplus (w(e_2, u_3, u_4) \otimes d(u_3) \otimes d(u_4))$$

- 超辺の各weightは、その前件のノードに依存(non-monotonic)
- 一つの導出のweight = 超辺の各weightの積
- あるノードのweightは、それを含む導出のweightの和

Semirings

$$\mathbf{K} = \langle K, \oplus, \otimes, \mathbf{0}, \mathbf{1} \rangle$$

semiring	K	⊕	⊗	0	I
Viterbi	[0,1]	max	×	0	I
Real	R	+	x	0	I
Log	R	logsumexp	+	+∞	0
Tropical	R	min	+	+∞	0
Expectation	<P,R>	$\langle p_1 \oplus p_2,$ $r_1 \oplus r_2 \rangle$	$\langle p_1 \otimes p_2,$ $p_1 \otimes r_2 \oplus p_2 \otimes r_1 \rangle$	<0,0>	<1,0>

まとめ

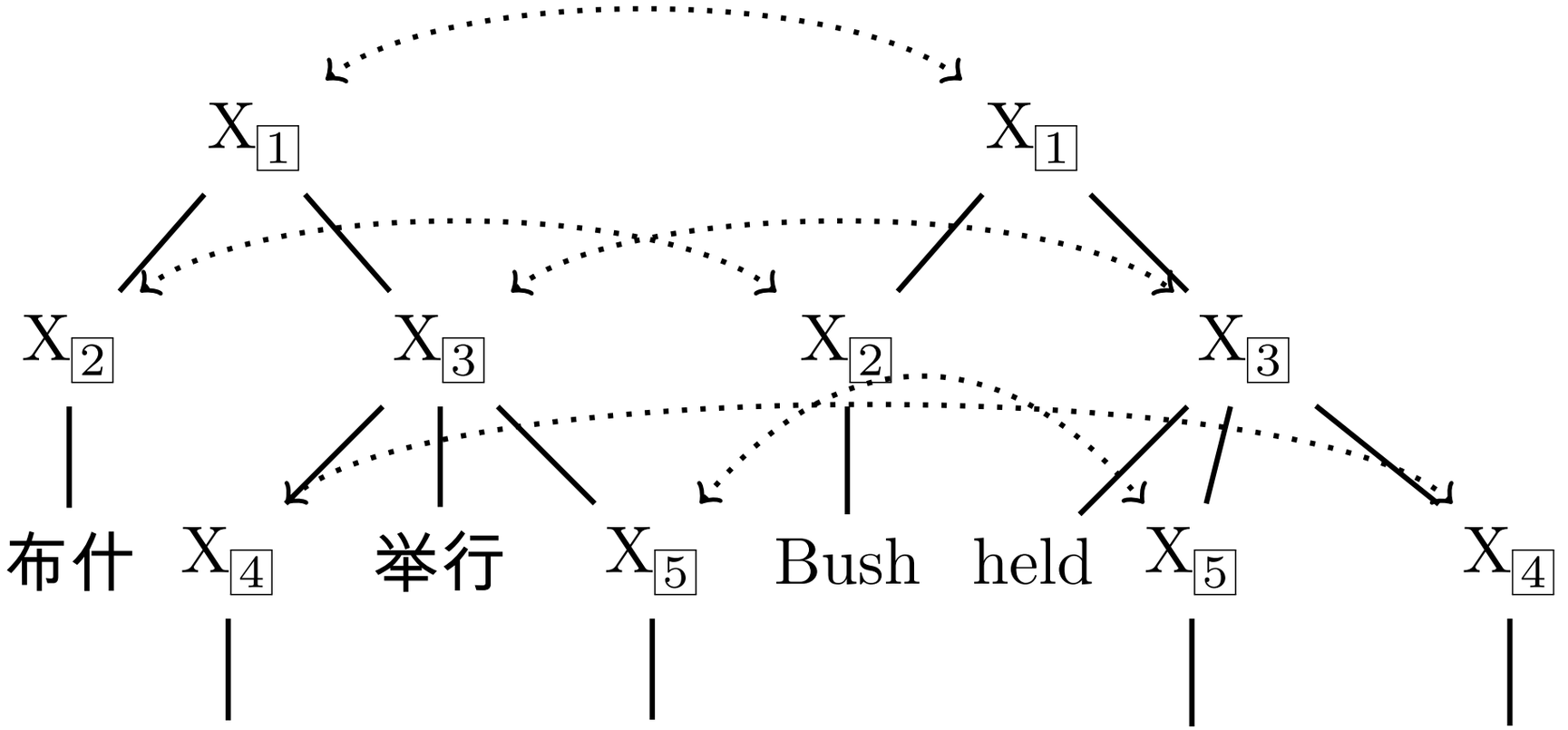
- 「構文解析」に関する復習(注意: あくまでも機械翻訳のチュートリアルです)

- CFG, parsing, hypergraph, deductive system, weight, semiring

内容

- 木構造に基づく機械翻訳
  - 背景: CFG, hypergraph, deductive system
  - 同期文脈自由文法 (synchronous-CFG)
    - 同期文法: {string,tree}-to-{string,tree}
    - 二言語の構文解析(biparsing)
    - 同期から非同期
- 最適化

同期文脈自由文法



$$\hat{\mathbf{e}} = \operatorname*{argmax}_{\mathbf{e}} \frac{\exp\left(\mathbf{w}^\top \cdot \mathbf{h}(\mathbf{e}, d, \mathbf{f})\right)}{\sum_{\mathbf{e}', d'} \exp\left(\mathbf{w}^\top \cdot \mathbf{h}(\mathbf{e}', d', \mathbf{f})\right)} = \operatorname*{argmax}_{\mathbf{e}} \mathbf{w}^\top \cdot \mathbf{h}(\mathbf{e}, d, \mathbf{f})$$

(Chiang, 2007)

- d: 同期文脈自由文法(synchronous-CFG、SCFG)と入力との交差による導出

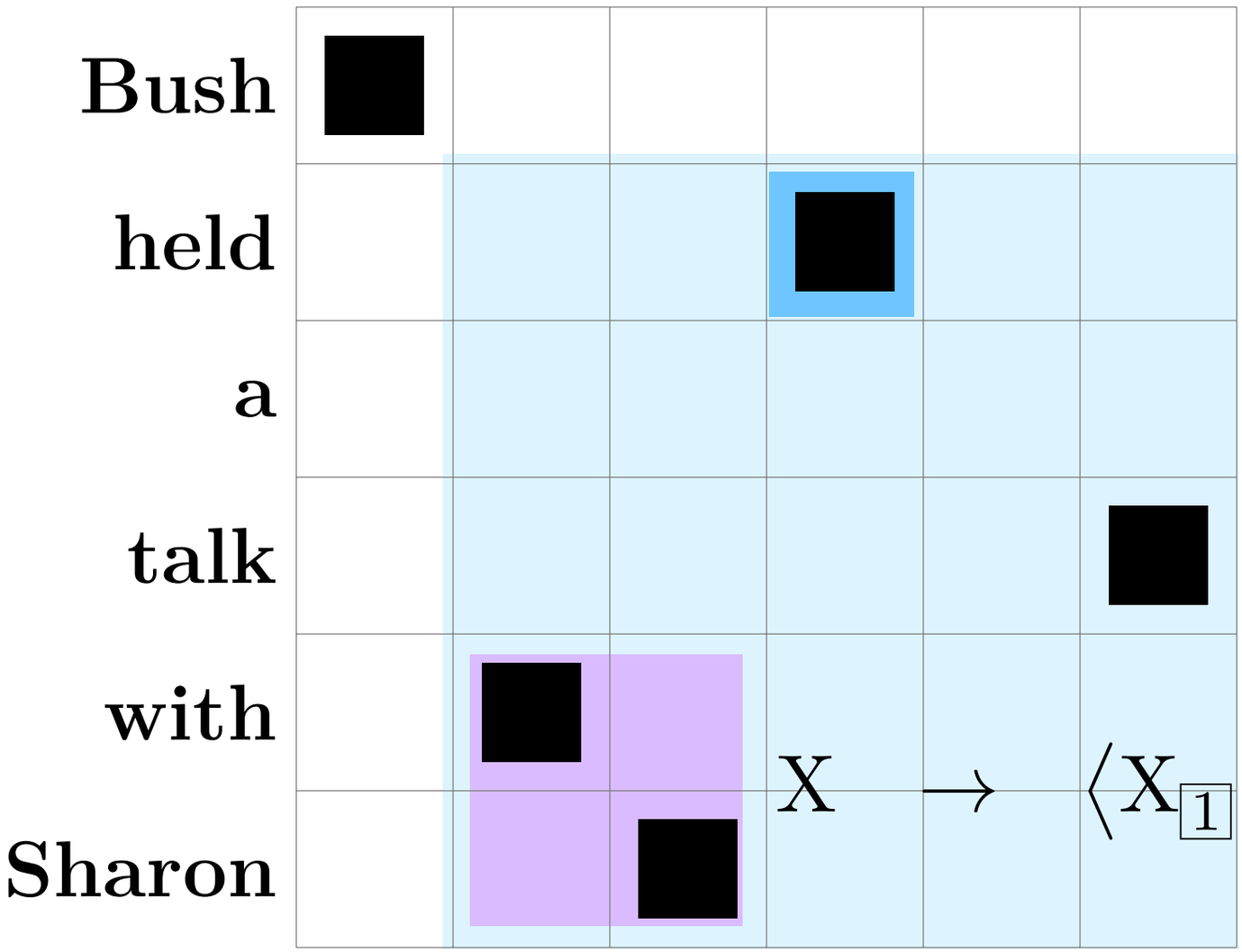
同期文脈自由文法: Model

$$
\begin{aligned}
\mathrm{S} &\rightarrow \langle \mathrm{S}_{\boxed{1}}\ \mathrm{X}_{\boxed{2}}, \mathrm{S}_{\boxed{1}}\ \mathrm{X}_{\boxed{2}} \rangle \\
\mathrm{S} &\rightarrow \langle \mathrm{X}_{\boxed{1}}, \mathrm{X}_{\boxed{1}} \rangle \\
\mathrm{X} &\rightarrow \langle \mathrm{X}_{\boxed{1}}\ \text{举行}\ \mathrm{X}_{\boxed{2}}, \text{hold}\ \mathrm{X}_{\boxed{2}}\ \mathrm{X}_{\boxed{1}} \rangle \\
\mathrm{X} &\rightarrow \langle \text{与 沙龙}, \text{with Sharon} \rangle \\
\mathrm{VP} &\rightarrow \langle \mathrm{VBD}_{\boxed{1}}\ \mathrm{NP}_{\boxed{2}},\ \mathrm{NP}_{\boxed{2}}\ \mathrm{VBD}_{\boxed{1}} \rangle \\
\mathrm{NP} &\rightarrow \langle \mathrm{NP}_{\boxed{1}}\ \mathrm{PP}_{\boxed{2}},\ \mathrm{NP}_{\boxed{1}}\ \mathrm{PP}_{\boxed{2}} \rangle \\
\mathrm{VP} &\rightarrow \langle \mathrm{VBD}_{\boxed{1}}\ \mathrm{NP}_{\boxed{2}}\ \mathrm{PP}_{\boxed{3}},\ \mathrm{NP}_{\boxed{2}}\ \mathrm{PP}_{\boxed{3}}\ \mathrm{VBD}_{\boxed{1}} \rangle
\end{aligned}
$$

- SとXという２つのカテゴリーのみ (Chiang, 2007)
- あるいは統語解析のカテゴリーを使用(Zollman and Venugopal, 2006)

ルールの抽出

布什 与 沙龙举行 了 会谈



⟨held a talk with Sharon, 与 沙龙 举行 了 会谈⟩

⟨with Sharon, 与 沙龙⟩

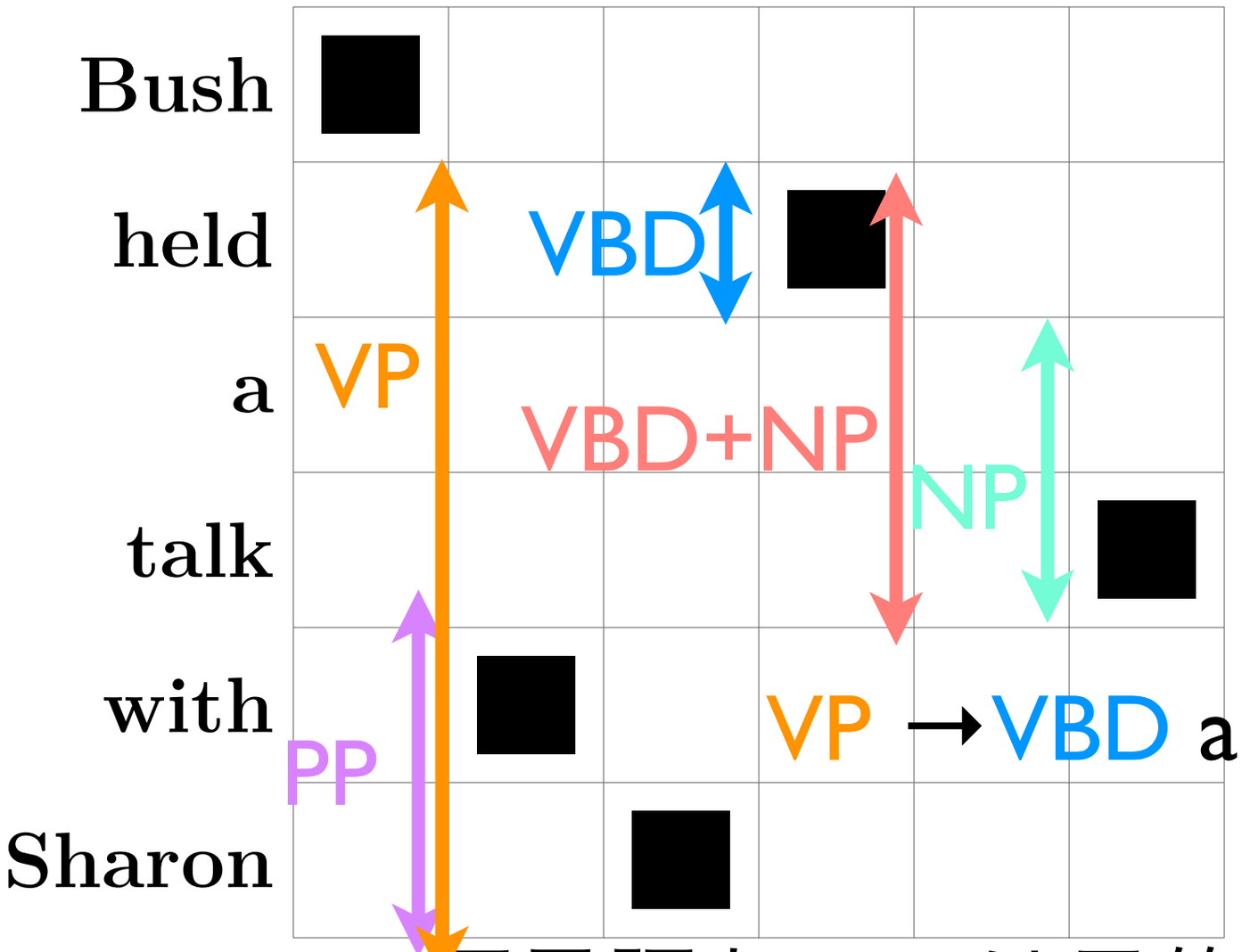
⟨held, 举行⟩

(Example from Huang and Chiang, 2007)

- Hiero文法: 句の抽出 + 小さい句で「穴」(Chiang, 2007)

統語論的なカテゴリー

布什 与 沙龙举行 了 会谈



⟨held a talk with Sharon, 与 沙龙 举行 了 会谈⟩

⟨with Sharon, 与 沙龙⟩

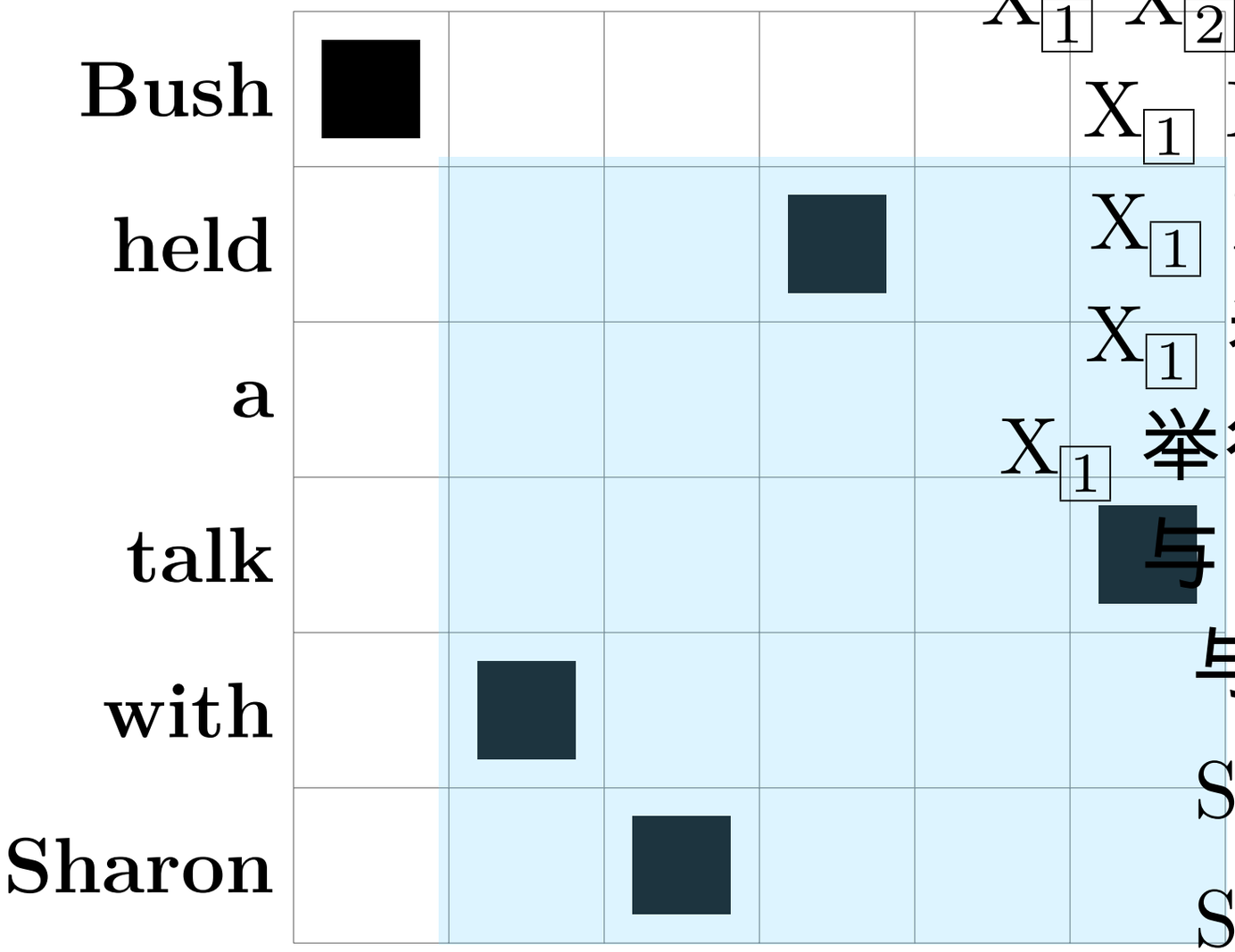
⟨held, 举行⟩

VP → VBD a talk PP, PP VBD 了 会谈

- SAMT: 原言語あるいは目的言語のカテゴリーをコピー(Zollman and Venugopal, 2006)

ルールの列挙

布什 与 沙龙举行 了 会谈



$X_{[1]}$ $X_{[2]}$ 了 会谈　$X_{[2]}$ a talk $X_{[1]}$

$X_{[1]}$ $X_{[2]}$ 会谈　$X_{[2]}$ a talk $X_{[1]}$

$X_{[1]}$ $X_{[2]}$ 会谈　$X_{[2]}$ talk $X_{[1]}$

$X_{[1]}$ 举行 $X_{[2]}$　held $X_{[2]}$ $X_{[1]}$

$X_{[1]}$ 举行 了 $X_{[2]}$　held a $X_{[2]}$ $X_{[1]}$

与 沙龙 $X_{[1]}$　$X_{[1]}$ with Sharon

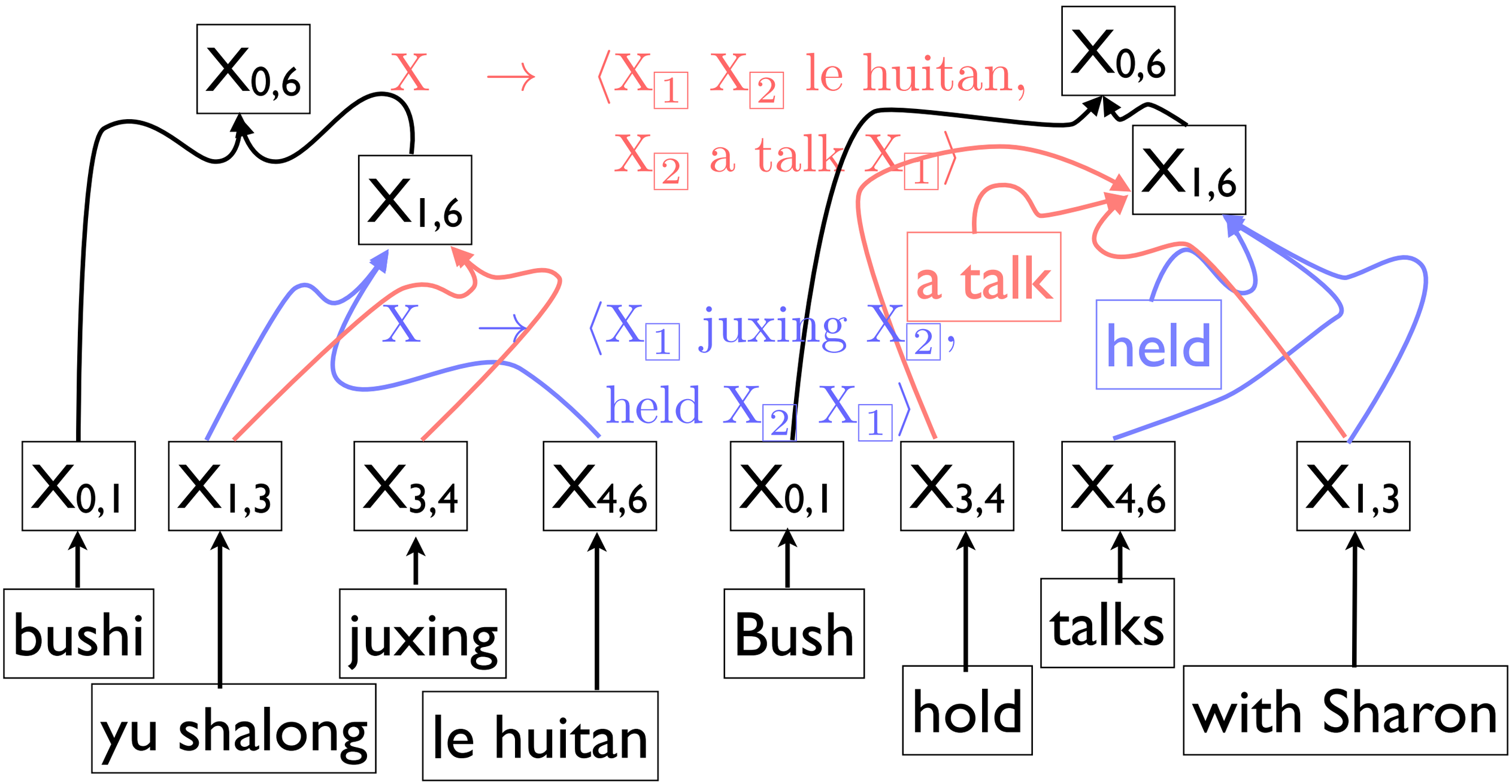
与 $X_{[1]}$ $X_{[2]}$　$X_{[2]}$ with $X_{[1]}$

$S \rightarrow \langle S_{[1]}\ X_{[2]}, S_{[1]}\ X_{[2]} \rangle$

$S \rightarrow \langle X_{[1]}, X_{[1]} \rangle$

- 句に基づく機械翻訳同様、可能なルールを列挙
- + glue rules

同期文脈自由文法:構文解析



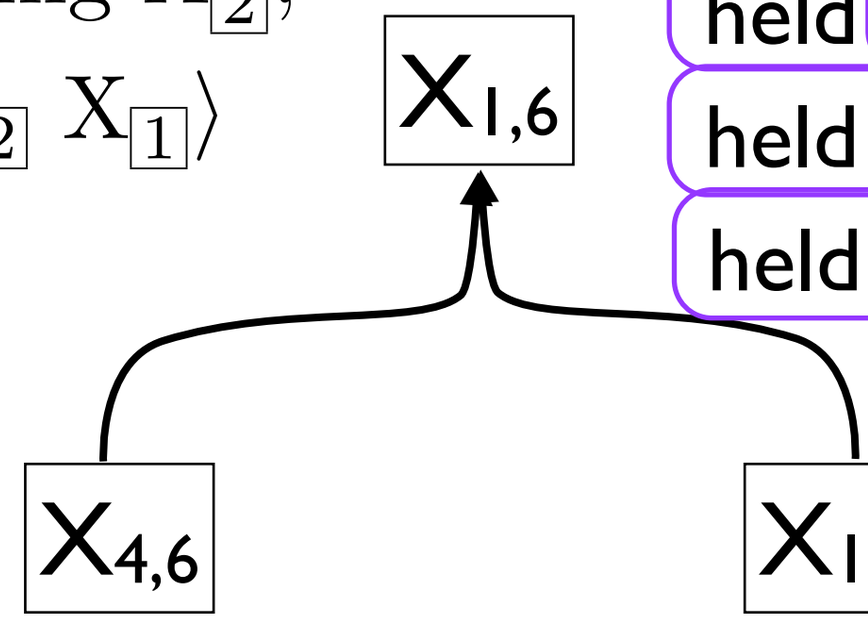
- 原言語側で構文解析、目的言語側で翻訳森を生成

同期文脈自由文法:構文解析

- SCFGによるデコーディング(Chiang, 2007)
  - 原言語側で単言語構文解析
  - 交差したルールの目的言語側で翻訳森を生成
  - 翻訳森から最適な導出を求める(Huang and Chiang, 2005)
- 計算量: $O(n^3)$ = 単言語CKY

局所的でない素性

$$X \rightarrow \langle X_{[1]} \text{ juxing } X_{[2]}, \text{ held } X_{[2]} \ X_{[1]} \rangle$$



held a talk with Sharon
held talks with Sharon
held a talk and Sharon
held meeting Sharon with

境界にある単語に

関して更新

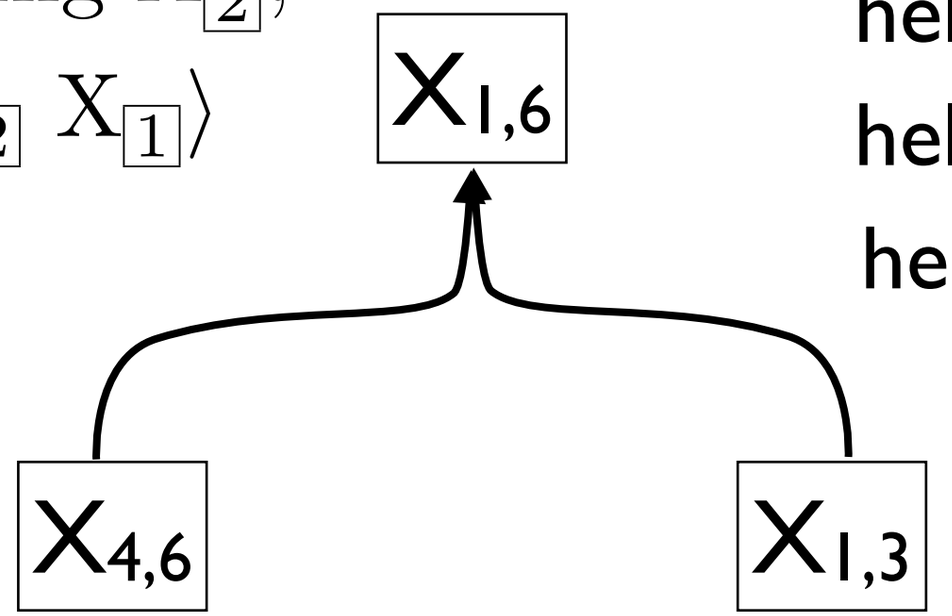
p(talk | a)　a talk
talks
meeting
meetings

with Sharon　p(Sharon | with)
and Sharon　p(Sharon | and)
Sharon with　p(with | Sharon)
Sharon and　p(and | Sharon)

- スパンの外側の情報が必要(例、bigram LM)

Bigram素性

$$X \rightarrow \langle X_{\boxed{1}} \text{ juxing } X_{\boxed{2}}, \text{ held } X_{\boxed{2}} \ X_{\boxed{1}} \rangle$$



held * Sharon
held * Sharon
held * Sharon
held * with

a * talk
talks
meeting
meetings

with * Sharon
and * Sharon
Sharon * with
Sharon * and

- bigramに必要な情報を保持:2単語(why?)

Language Model Scoring

- 各仮説に二つのコンテキストを保持:
  - Prefix:将来計算されるngram
  - Suffix: 将来のngramの計算のためのコンテキスト (i.e. フレーズベースMT)
- 計算量: $O(n^3V^{2(m-1)})$
- 非常に非効率:T(e)からたどれる、前件の全ての組み合わせを考慮

Forest Rescoring

- SCFGによるデコーディング(Chiang, 2007)
  - 原言語側で単言語構文解析
  - 交差したルールの目的言語側で翻訳森を生成 + 非局所的な素性によるリスコア
  - 翻訳森から最適な導出を求める(Huang and Chiang, 2005)
- ~~計算量: $O(n^3)$ = 単言語CKY~~

Cube Pruning

$$\mathrm{X} \rightarrow \langle \mathrm{X}_{\boxed{1}} \text{ juxing } \mathrm{X}_{\boxed{2}}, \text{ held } \mathrm{X}_{\boxed{2}}\ \mathrm{X}_{\boxed{1}} \rangle$$

		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.5	1.7	2.6	3.2
a * talk	1.0	2.5	2.7	3.6	4.2
talks	1.3	2.8	3.0	3.9	4.5
meeting	2.2	3.7	3.9	4.8	5.4
meetings	2.6	4.1	4.3	5.2	5.8

- 各超辺に対して、前件の組み合わせを表した“cube” を作成(Huang and Chiang, 2007)

Cube Pruning

$$X \rightarrow \langle X_{\boxed{1}} \text{ juxing } X_{\boxed{2}}, \text{ held } X_{\boxed{2}} \; X_{\boxed{1}} \rangle$$

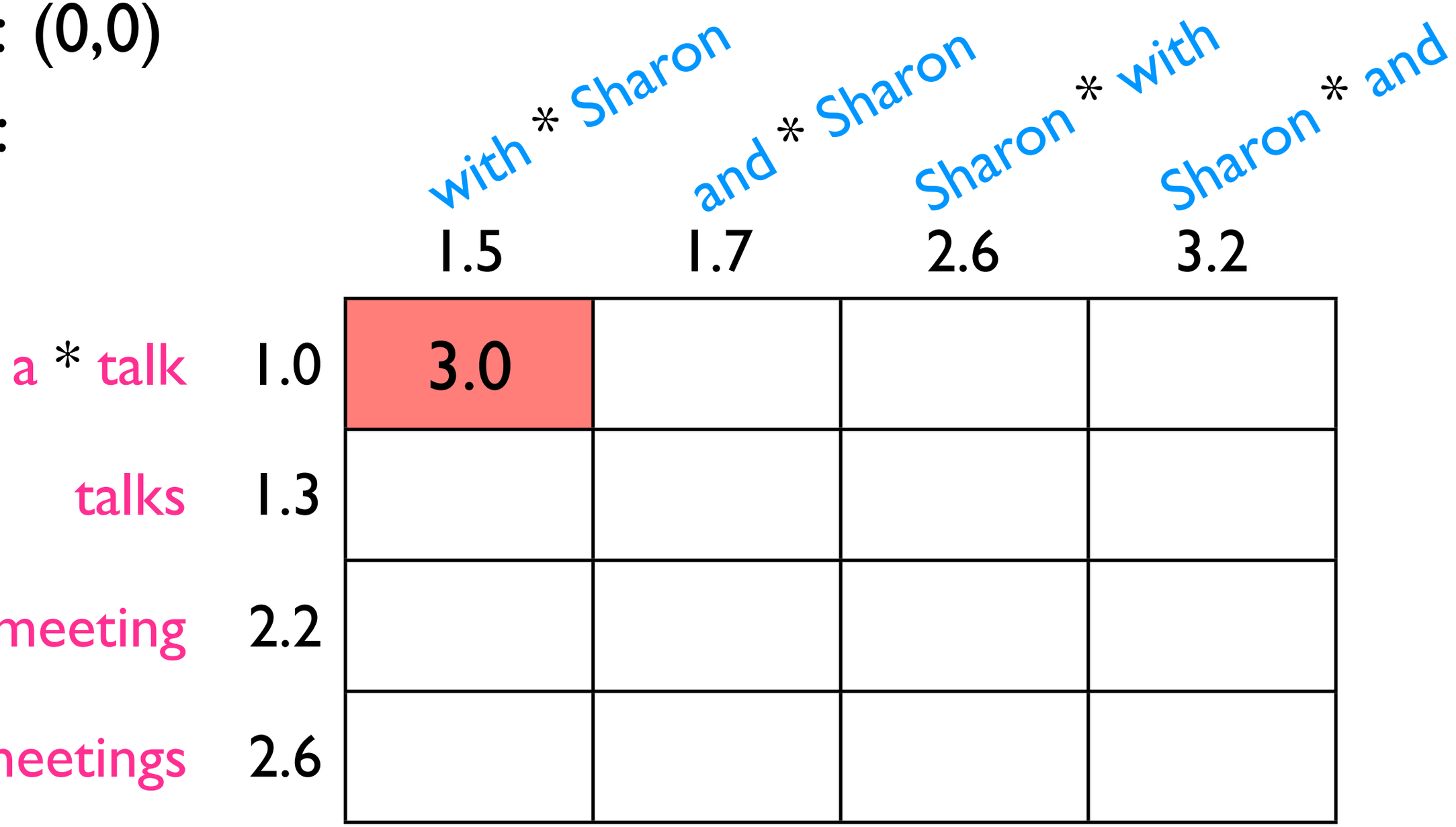
		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.5	1.7	2.6	3.2
a * talk	1.0	2.5 +0.5	2.7 +1.0	3.6 +1.5	4.2 +1.5
talks	1.3	2.8 +0.3	3.0 +1.5	3.9 +2.0	4.5 +2.0
meeting	2.2	3.7 +0.5	3.9 +1.0	4.8 +1.5	5.4 +1.5
meetings	2.6	4.1 +0.3	4.3 +1.5	5.2 +2.0	5.8 +2.0

- Bigramは前件のコンテキストが必要(非局所的な素性)

Cube Pruning

queue: (0,0)

k-best:



- 左上の隅から組み合わせを列挙(min-costを仮定)

Cube Pruning

queue:

k-best: (0,0)

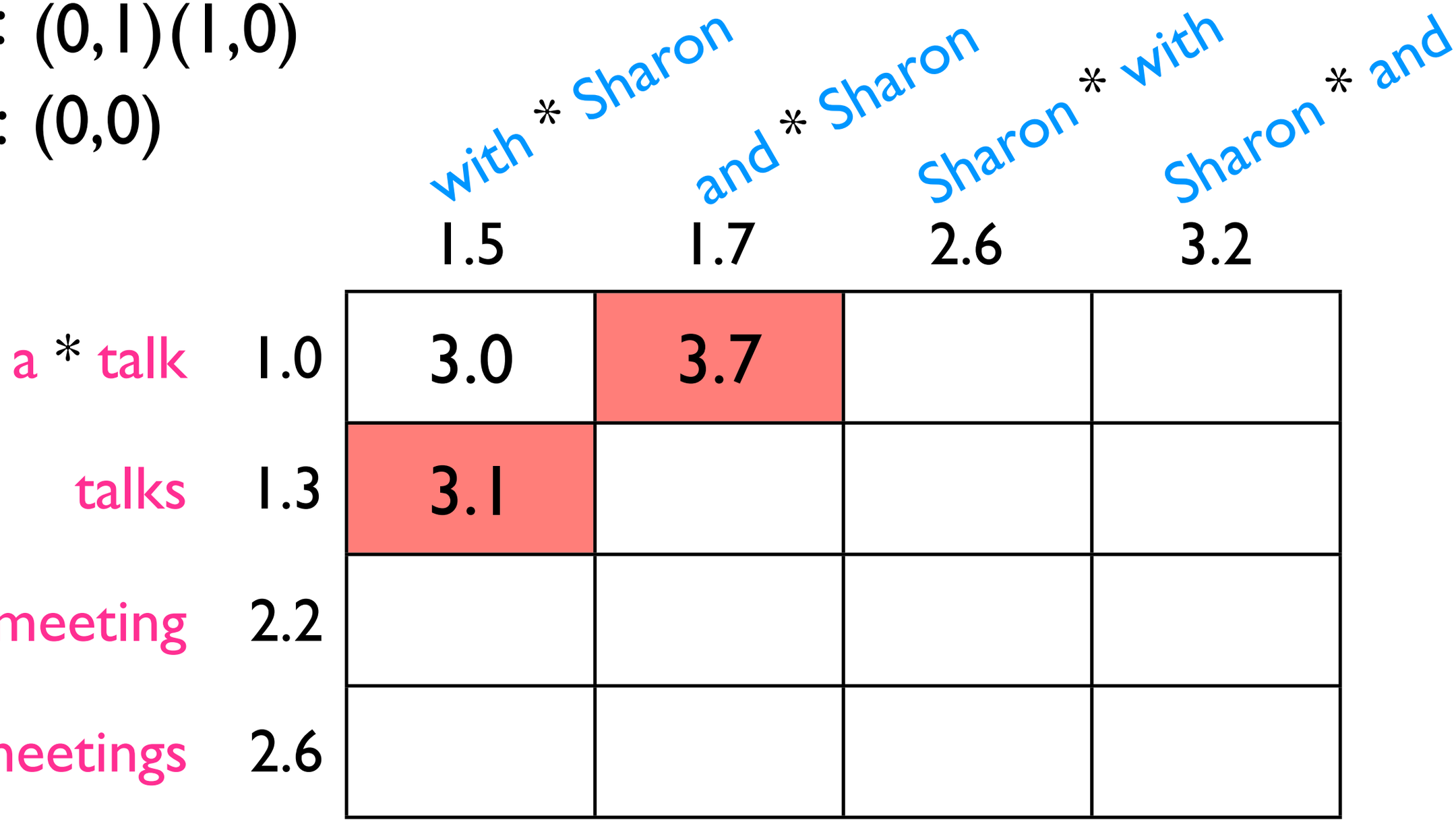
		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.5	1.7	2.6	3.2
a * talk	1.0	3.0			
talks	1.3				
meeting	2.2				
meetings	2.6				

- 左上の隅から組み合わせを列挙(min-costを仮定)

Cube Pruning

queue: (0,1)(1,0)

k-best: (0,0)

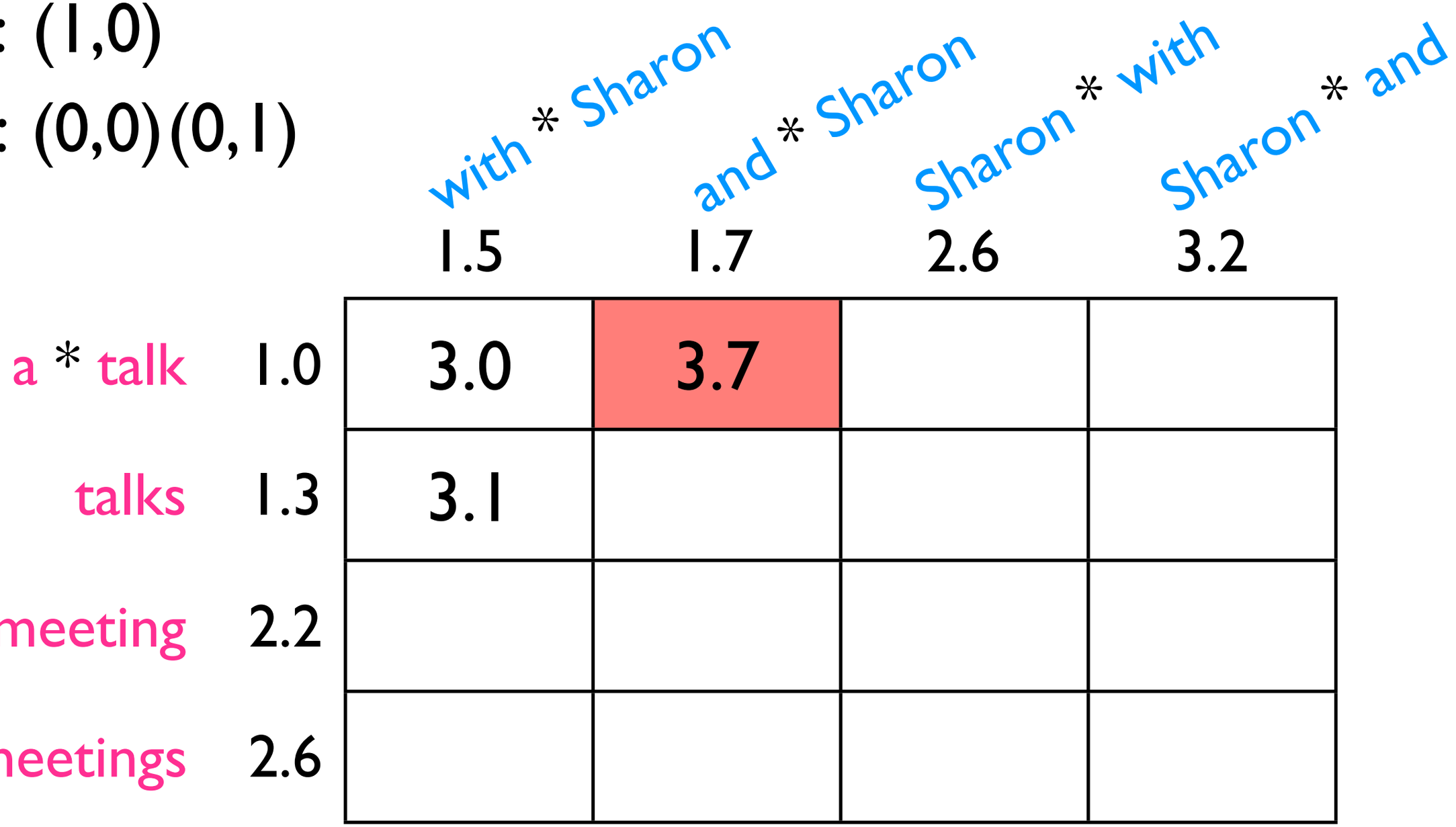


- 左上の隅から組み合わせを列挙(min-costを仮定)

Cube Pruning

queue: (1,0)

k-best: (0,0)(0,1)

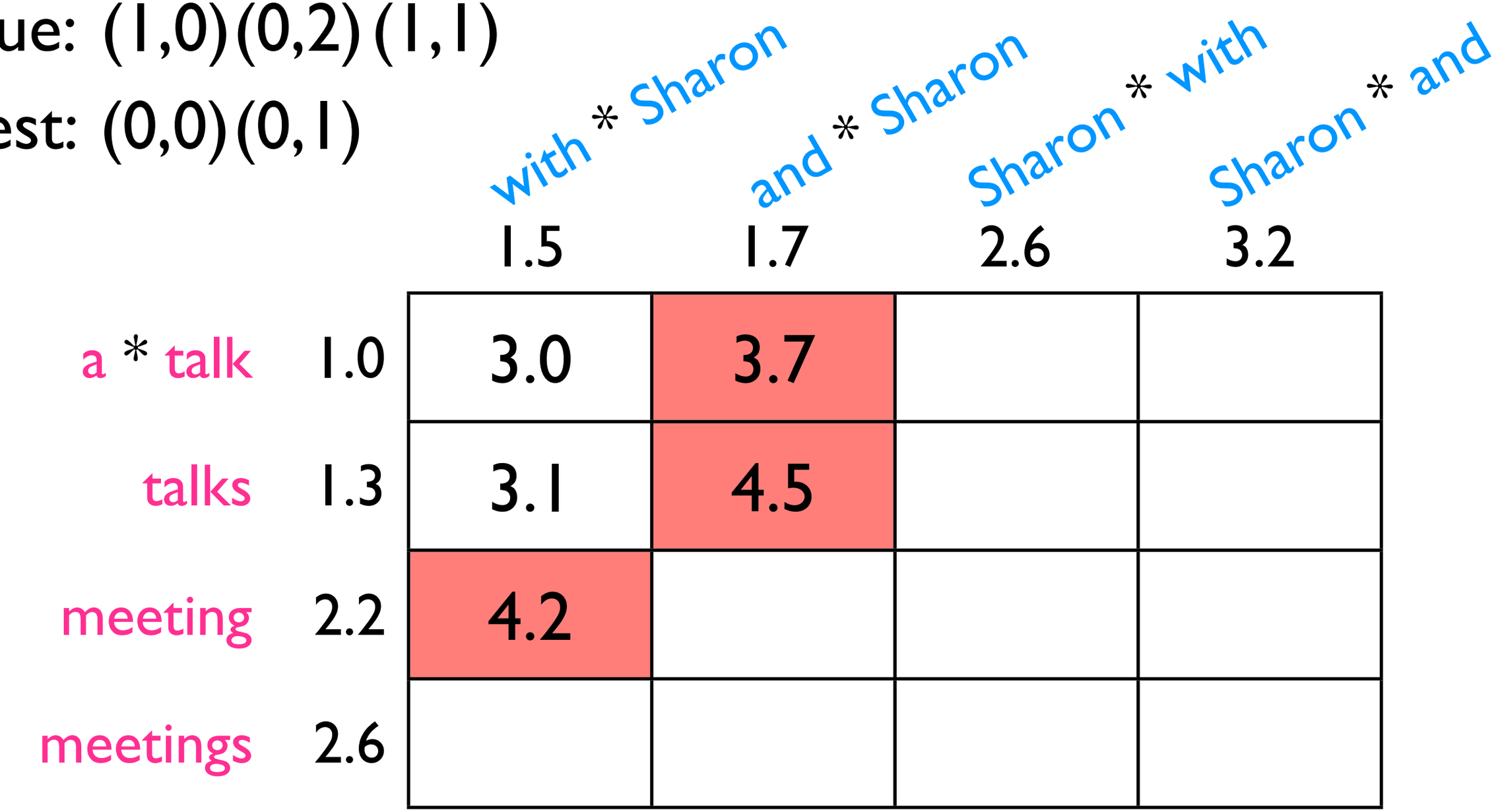


- 左上の隅から組み合わせを列挙(min-costを仮定)

Cube Pruning

queue: (1,0)(0,2)(1,1)

k-best: (0,0)(0,1)



- 左上の隅から組み合わせを列挙(min-costを仮定)

Cube Pruning

queue: (0,2) (1,1)

k-best: (0,0) (0,1) (1,0)

		with * Sharon 1.5	and * Sharon 1.7	Sharon * with 2.6	Sharon * and 3.2
a * talk	1.0	3.0	3.7		
talks	1.3	3.1	4.5		
meeting	2.2	4.2			
meetings	2.6				

- 左上の隅から組み合わせを列挙(min-costを仮定)

Cube Pruning

queue: (0,2) (1,1) (3,0)

k-best: (0,0) (0,1) (1,0)

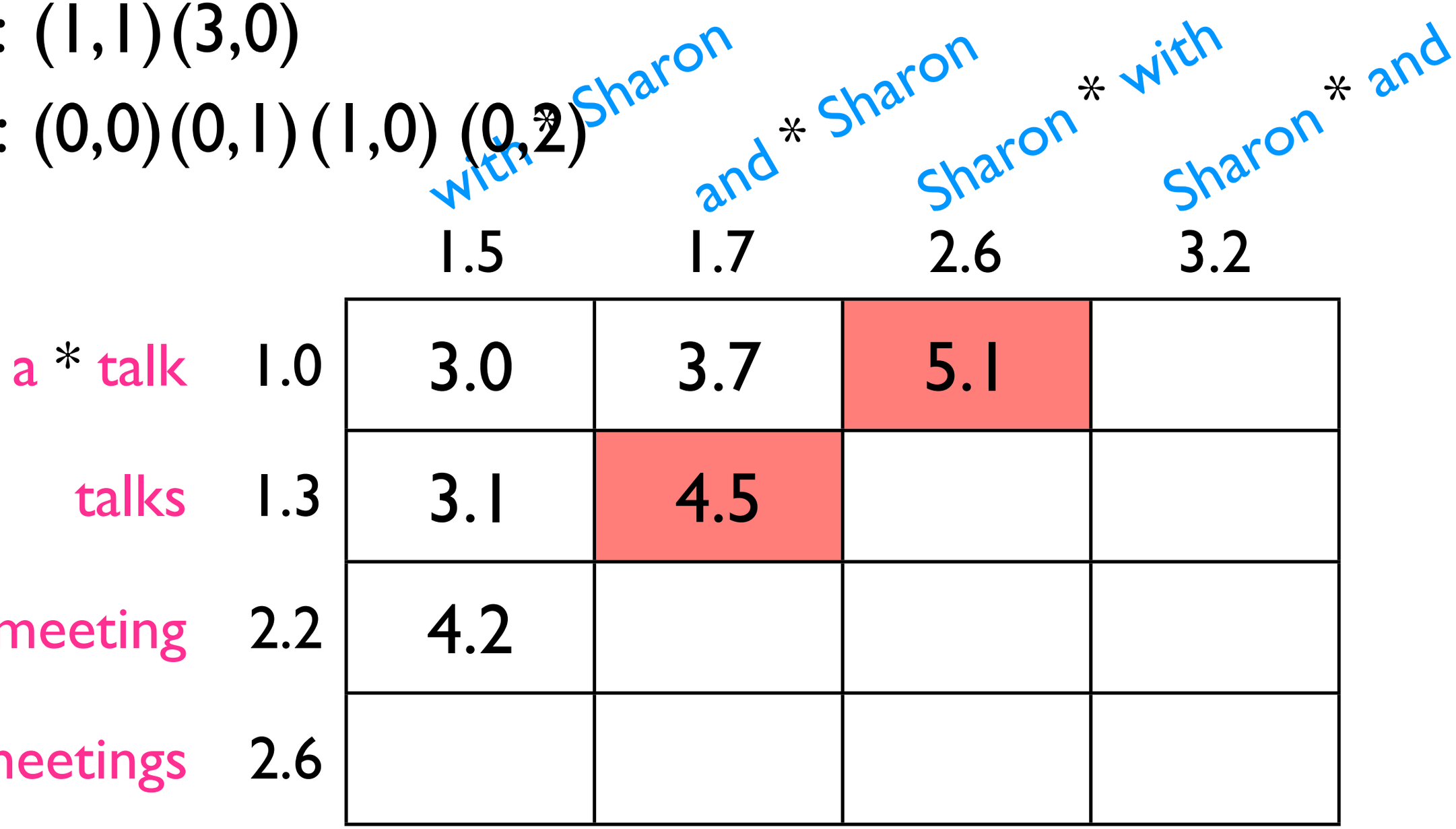
		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.5	1.7	2.6	3.2
a * talk	1.0	3.0	3.7	5.1	
talks	1.3	3.1	4.5		
meeting	2.2	4.2			
meetings	2.6				

- 左上の隅から組み合わせを列挙(min-costを仮定)

Cube Pruning

queue: (1,1)(3,0)

k-best: (0,0)(0,1)(1,0) (0,2)



- 左上の隅から組み合わせを列挙(min-costを仮定)

Cube Pruning

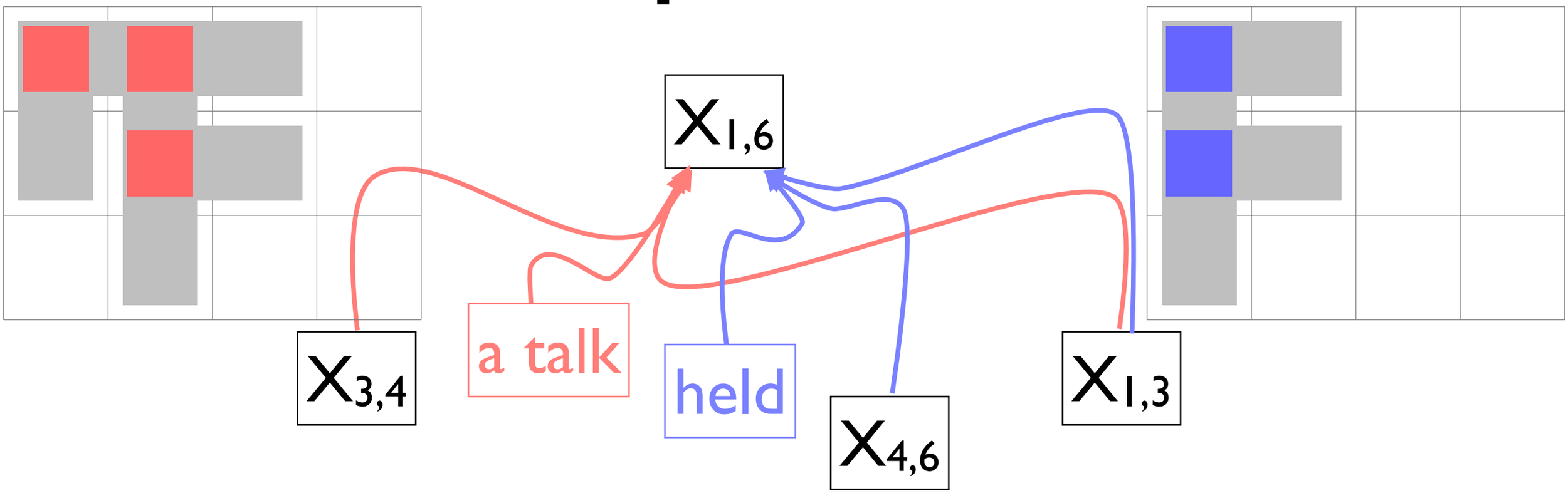
queue: (0,4) (1,1) (1,2) (3,0)

k-best: (0,0) (0,1) (1,0) (0,2)

		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.5	1.7	2.6	3.2
a * talk	1.0	3.0	3.7	5.1	
talks	1.3	3.1	4.5		
meeting	2.2	4.2	4.9		
meetings	2.6	4.4			

- 左上の隅から組み合わせを列挙(min-costを仮定)

Multiple Cubes



- 同じh(e)を持つ超辺を、同じqueueに入れる
  - 仮説(cube) = 超辺 + cubeの位置

Further Faster Decoding

- Cube Growing (Huang and Chiang, 2007)
  - bottom upにk個の仮説を列挙するのではなく、
    top downで「必要な数だけ」列挙
- Faster Cube Pruning (Gesmundo and Henderson, 2010)
  - cubeの列挙の順序を決定的にすることで余分な「メモリー」を除去(Alg. 2)
  - 全ての親が列挙された時のみに展開(Alg. 3)
- Incremental (Huang and Mi, 2010)
  - Top-downにデコーディング(Watanabe et al., 2006と似ている)

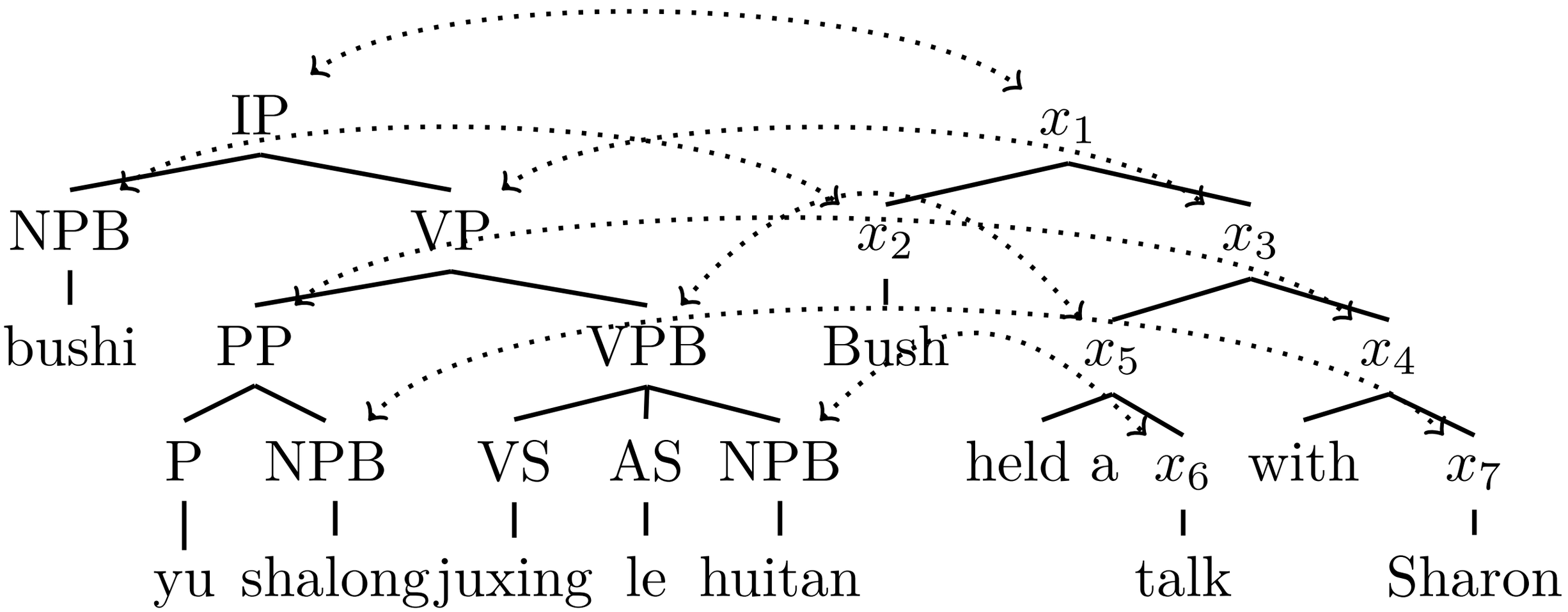
まとめ

- 同期文脈自由文法: 非終端記号を共有したルールの対
- 学習:フレーズベースSMTと同様に学習
  - 小さい句は大きい句の非終端記号
- デコード: 原言語側での構文解析
  - cube pruningによる効率的な組み合わせの列挙

内容

- 木構造に基づく機械翻訳
  - 背景: CFG, hypergraph, deductive system
  - 同期文脈自由文法 (synchronous-CFG)
  - 同期文法: {string,tree}-to-{string,tree}
  - 二言語の構文解析(biparsing)
  - 同期から非同期
- 最適化

{Tree,String}-to-{Tree,String}

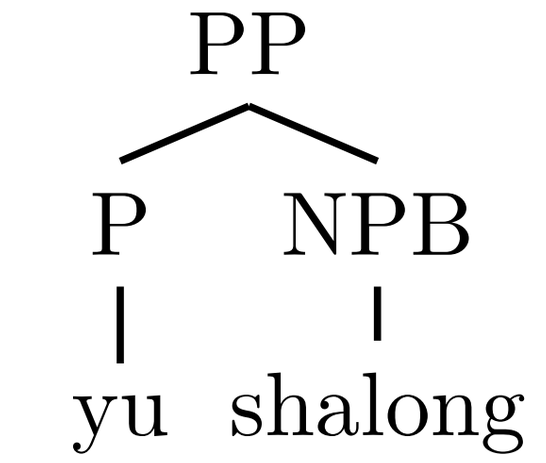


(Galley et al., 2004)

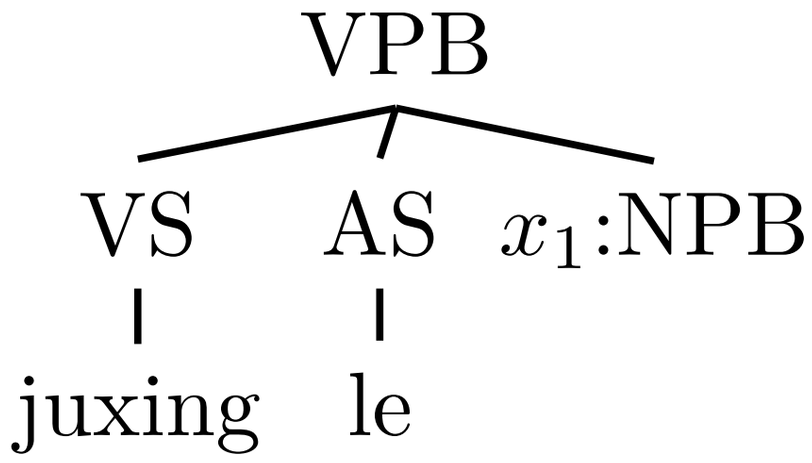
- 木構造を持ったルールのペア
- 同期木置換文法(Tree Substitution Grammars)

(Eisner, 2003)

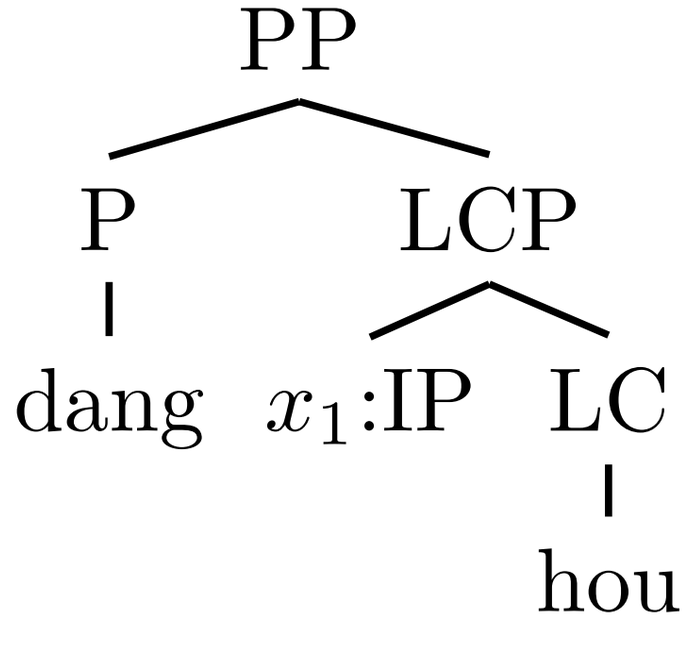
Tree-to-String Rules



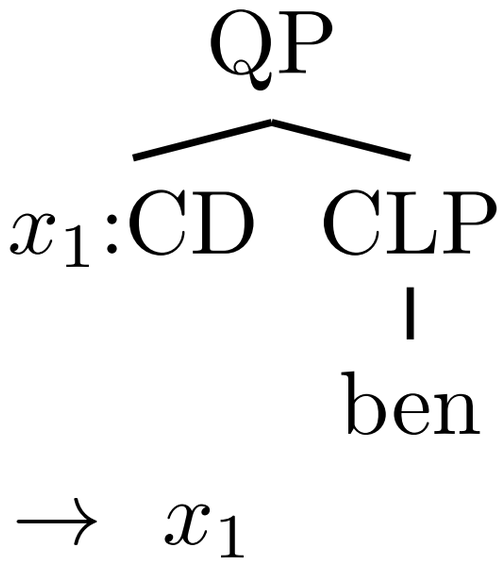
$\rightarrow$ with Sharon



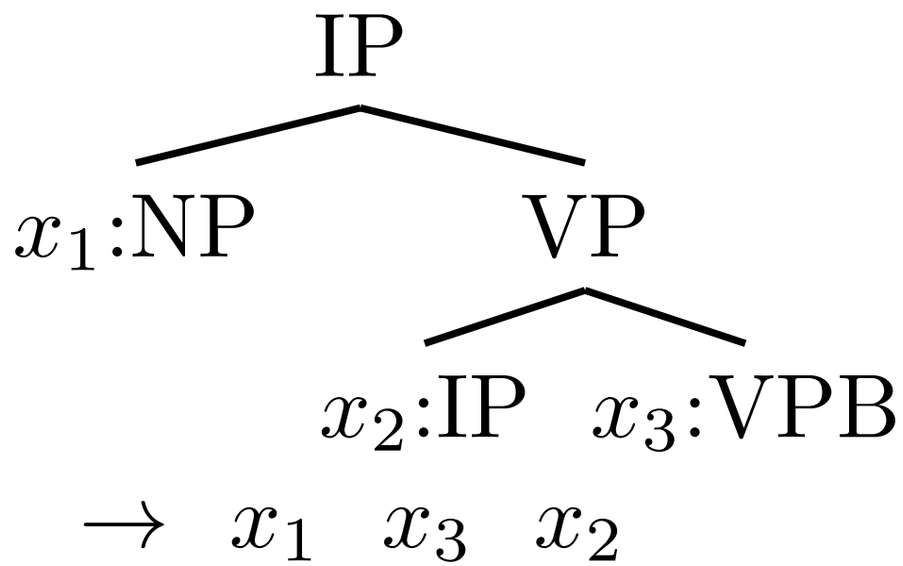
$\rightarrow$ held a x_1



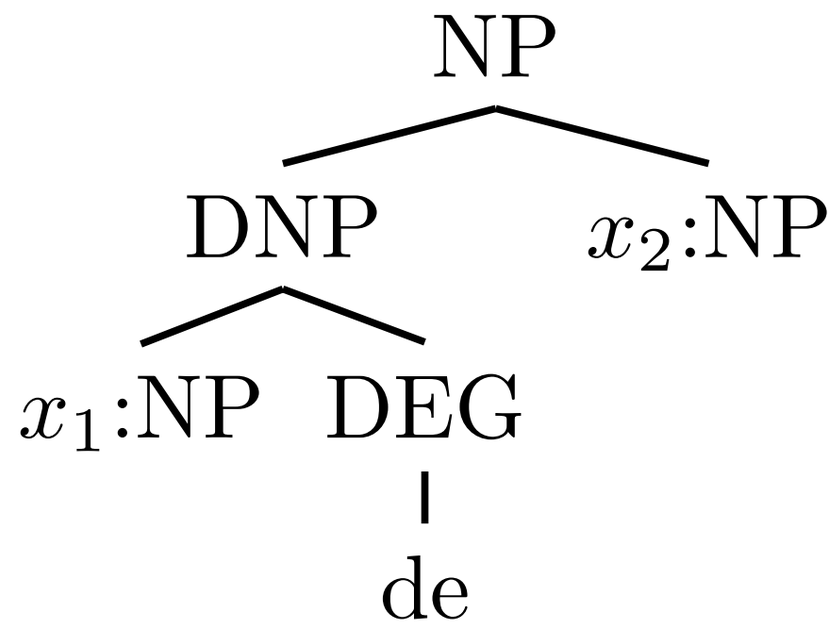
$\rightarrow$ when x_1



$\rightarrow$ x_1

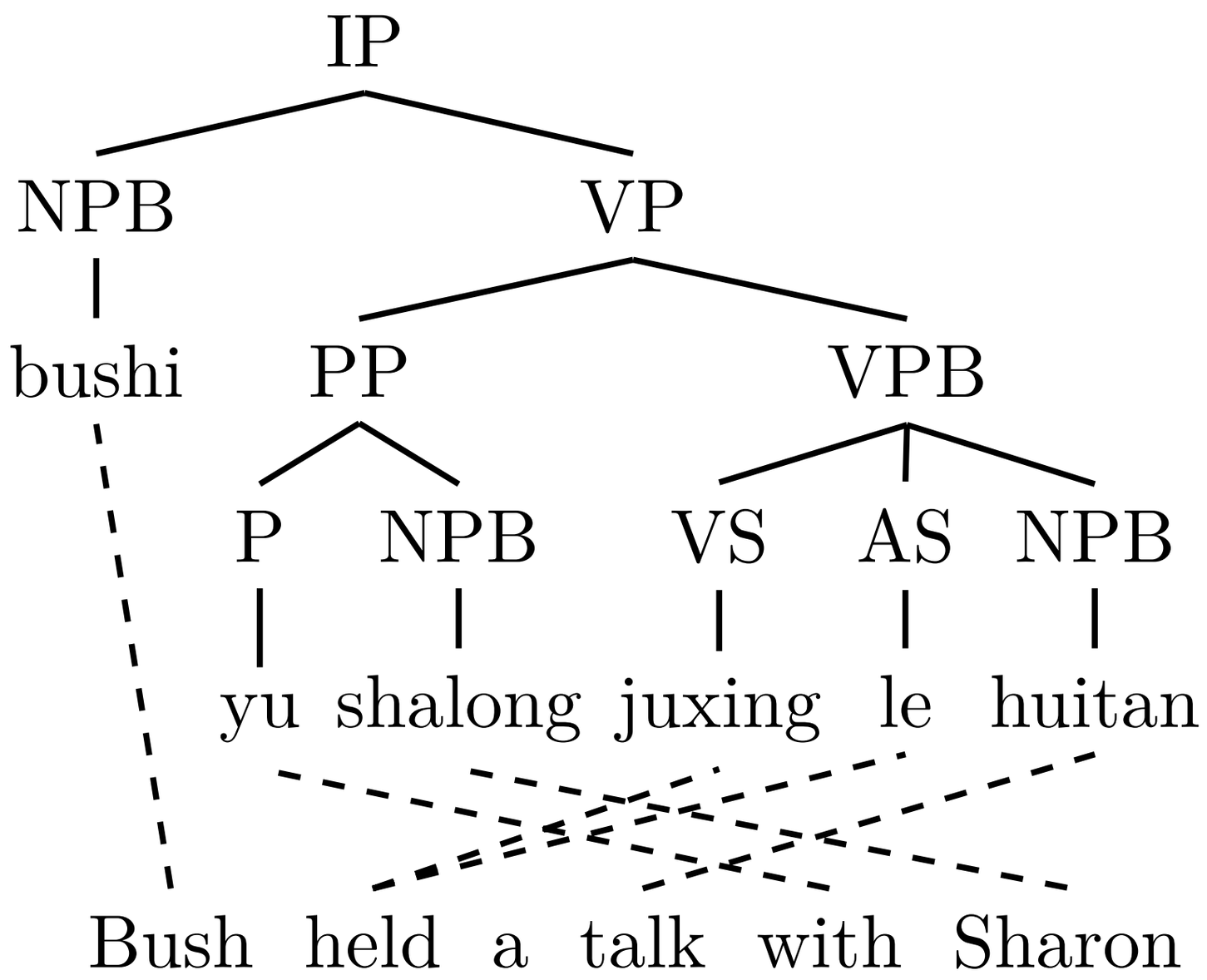


$\rightarrow$ x_1 x_3 x_2



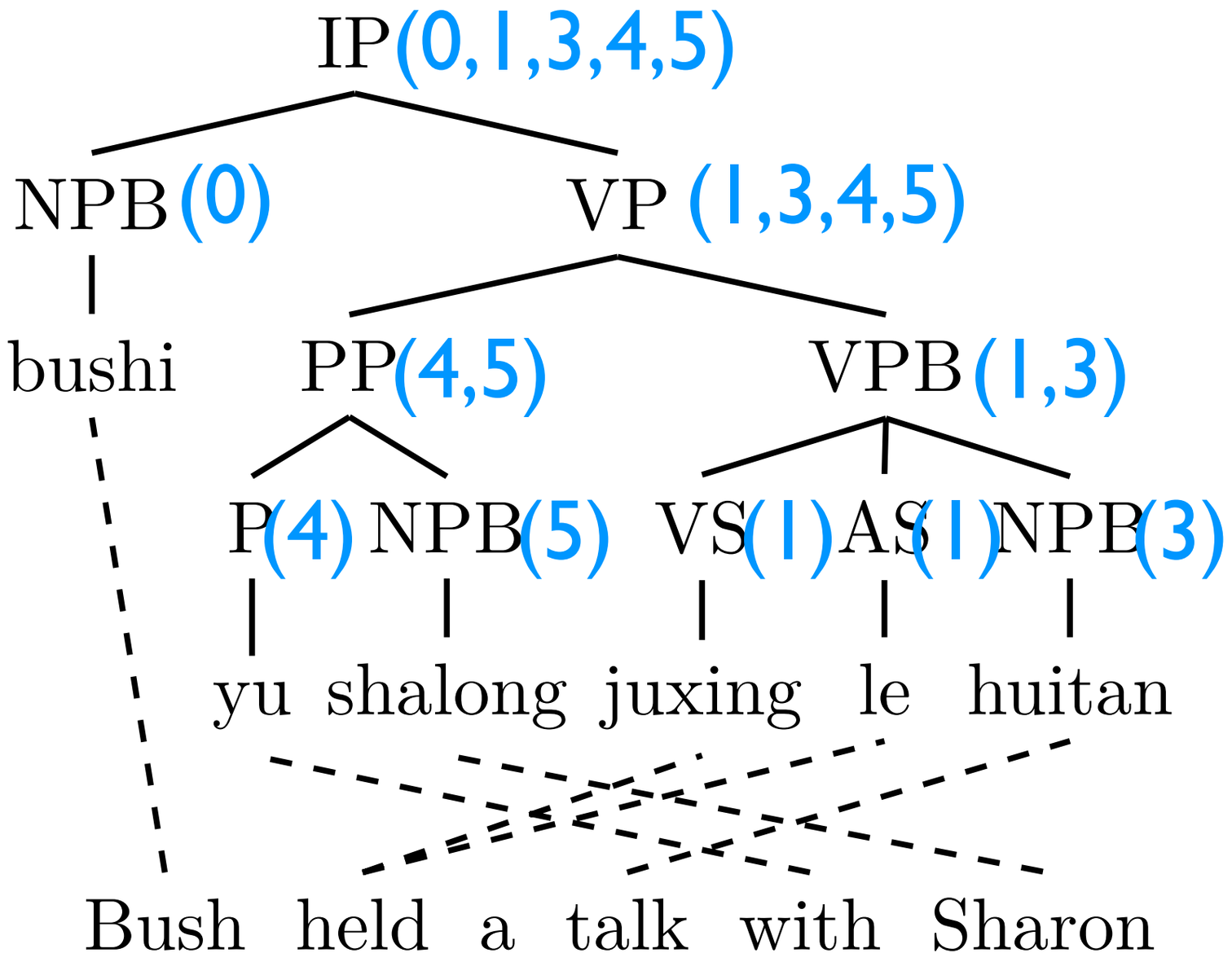
$\rightarrow$ x_2 of x_1

ルールの抽出



- フレーズのように、「最小ルールペア」を求める (Galley et al., 2004)

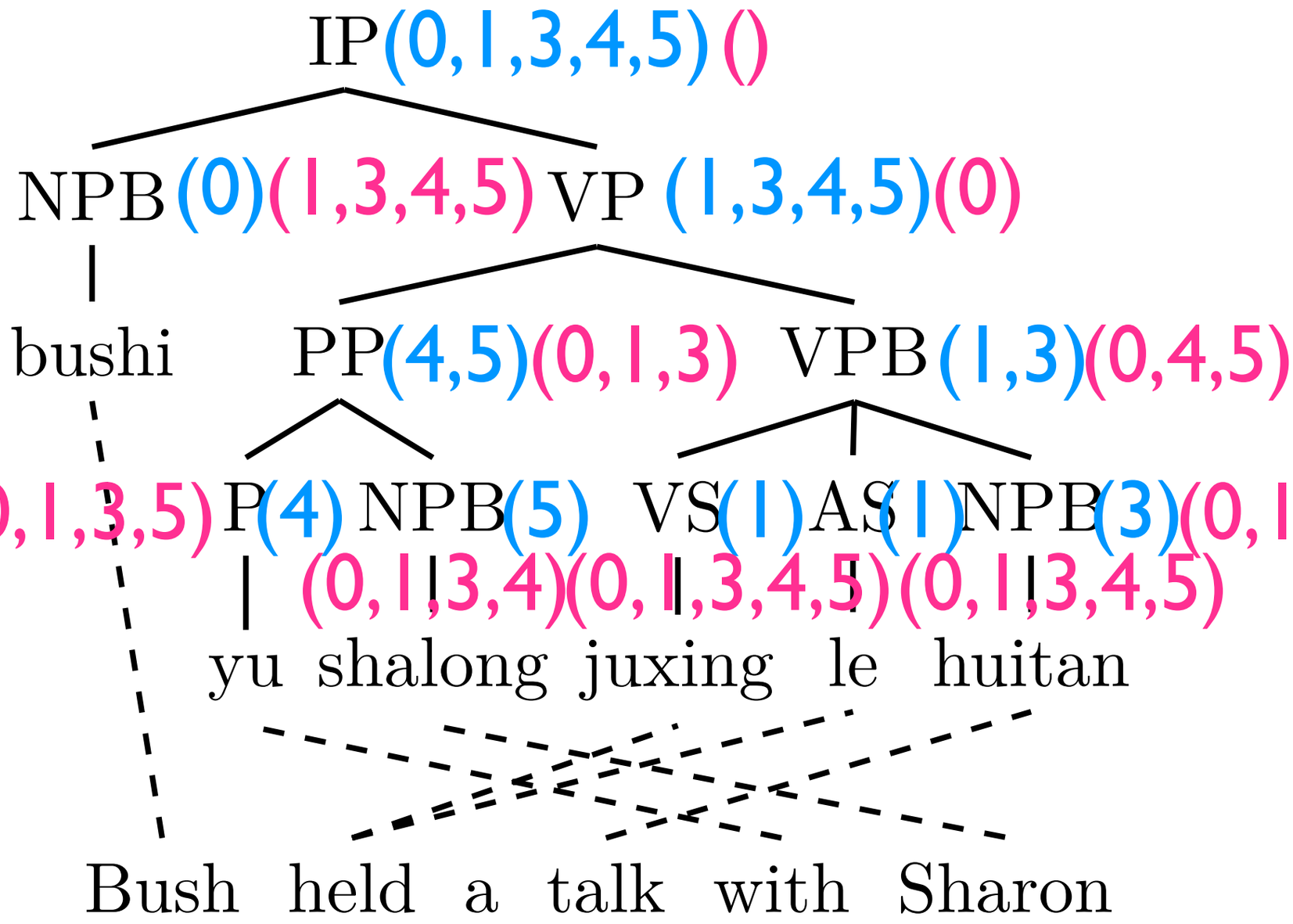
ルールの抽出



- bottom-upにアライメントを伝搬、「スパン」を計算(内側スパン)

(Galley et al., 2004)

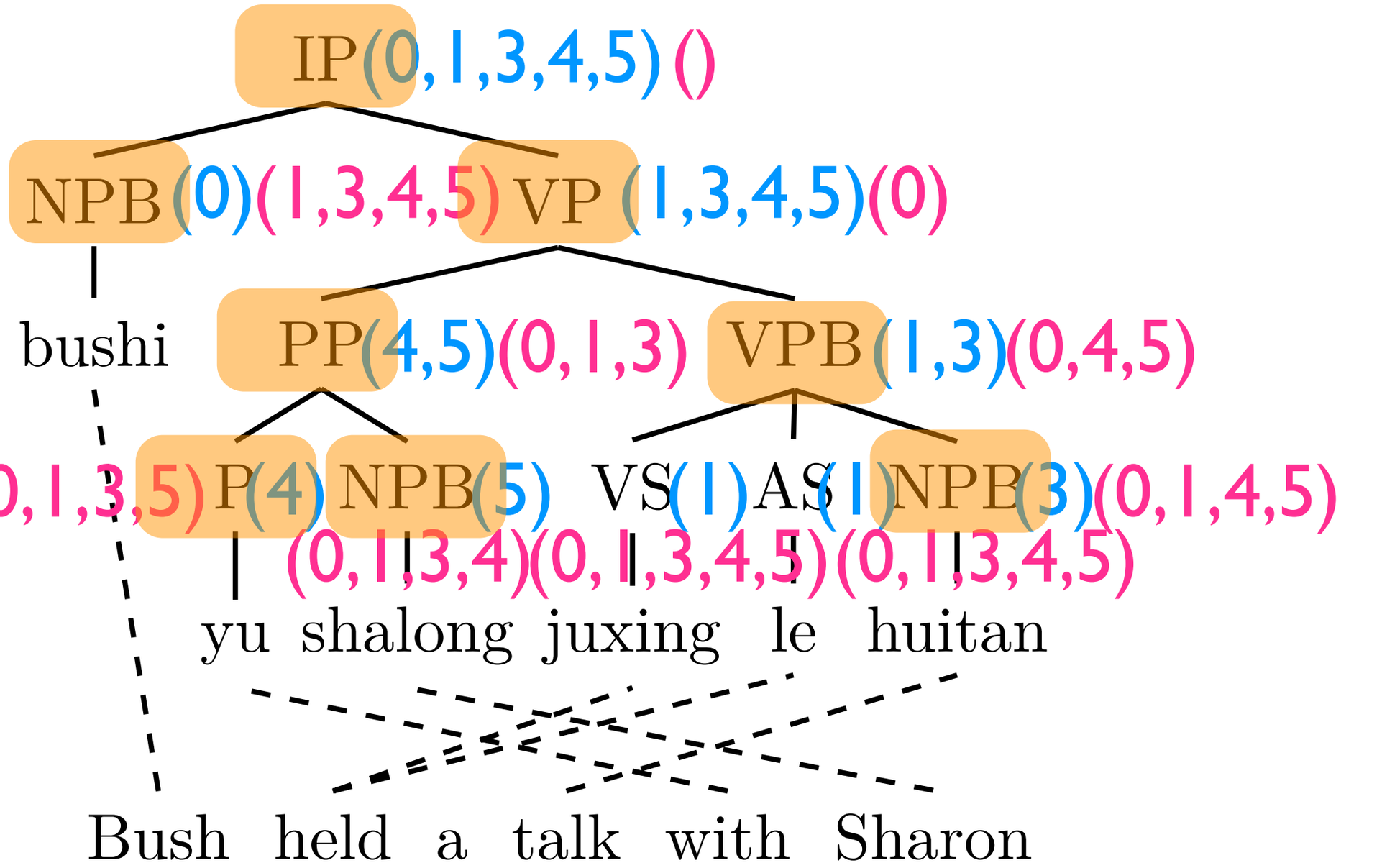
ルールの抽出



(Galley et al., 2004)

- top-downで「補完(complement)」したアライメントを計算(外側スパン)

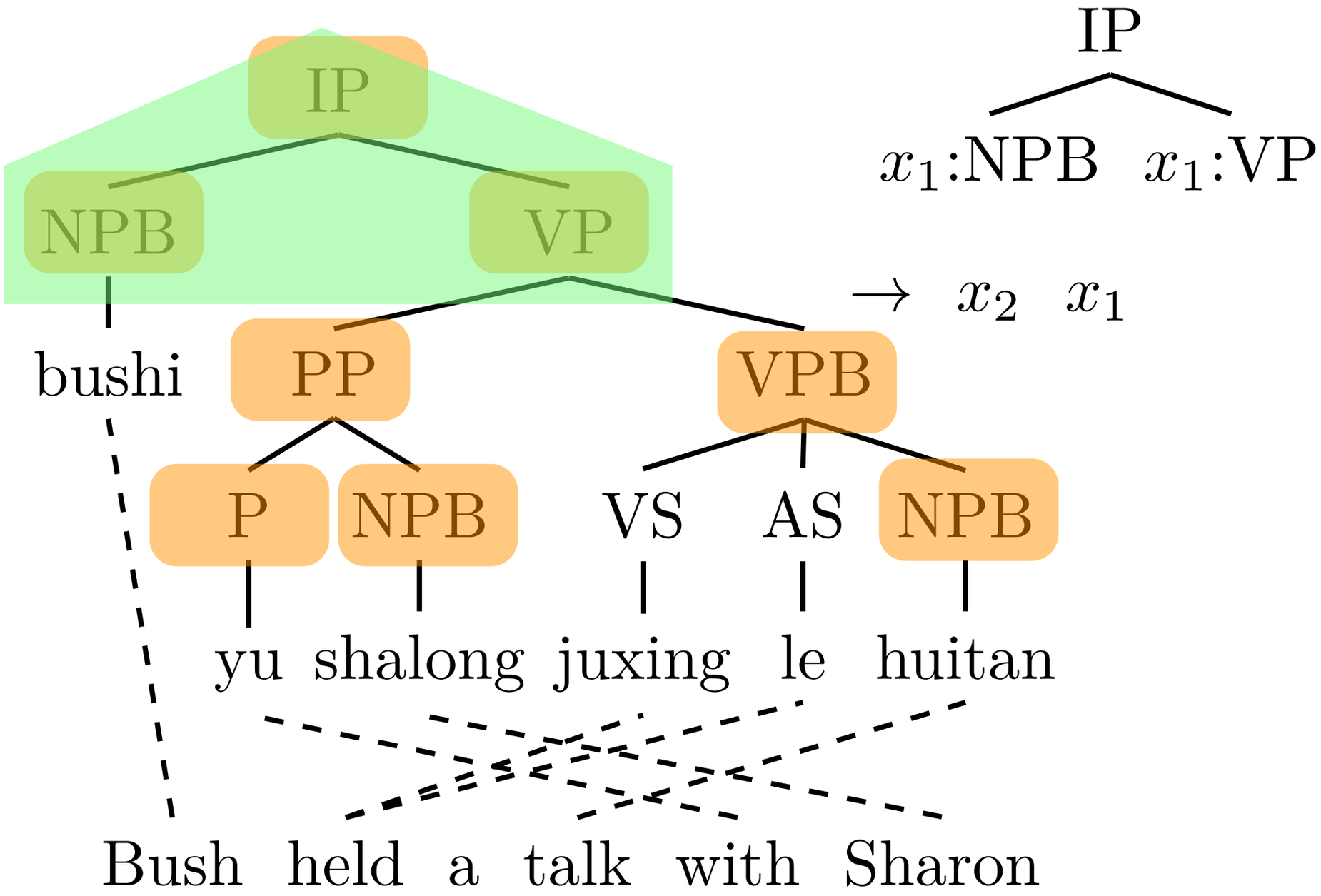
ルールの抽出



(Galley et al., 2004)

- “frontier”: spanとcomplementとの積集合が空集合

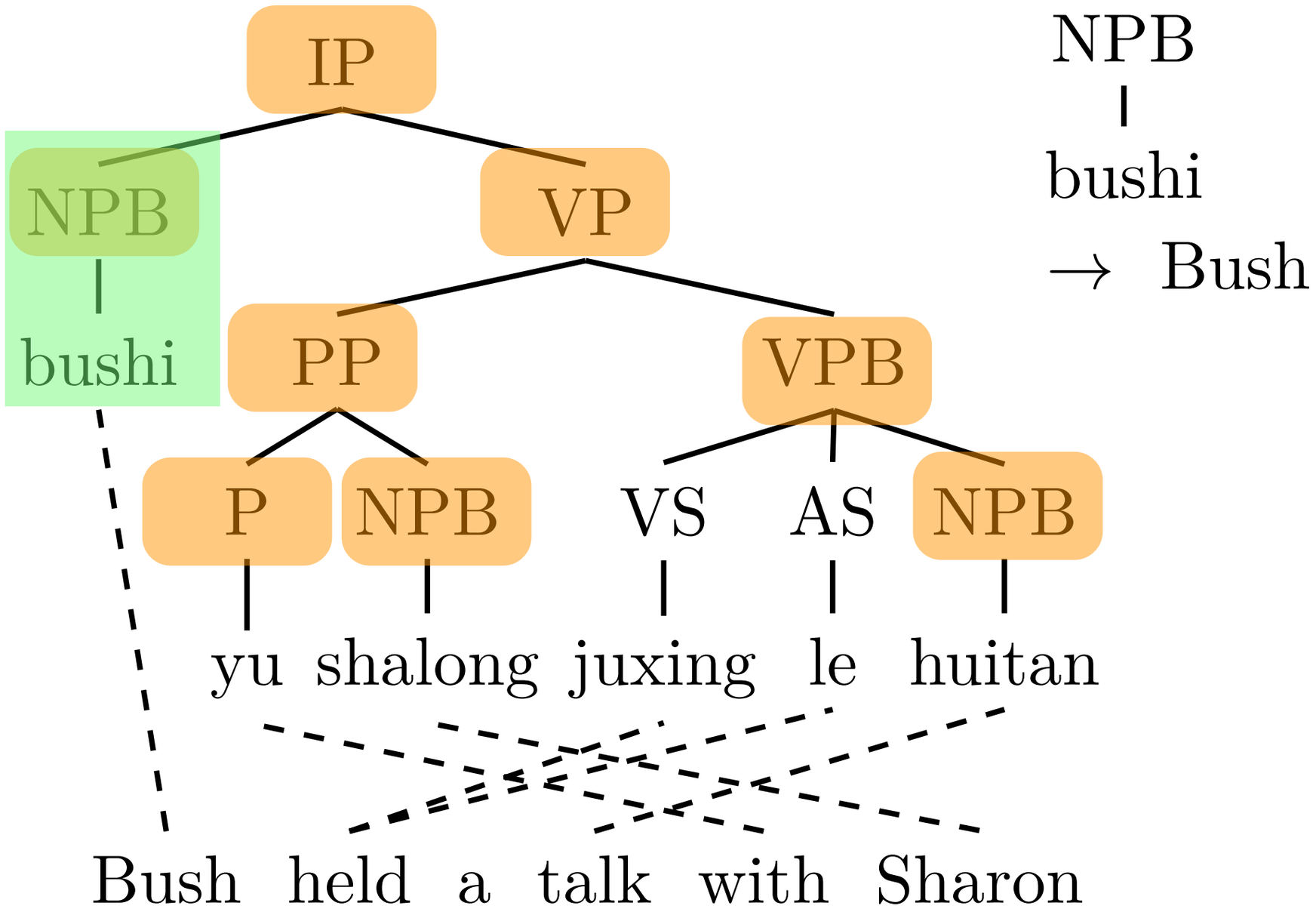
ルールの抽出



(Galley et al., 2004)

- “frontier”から最小ルールを抽出

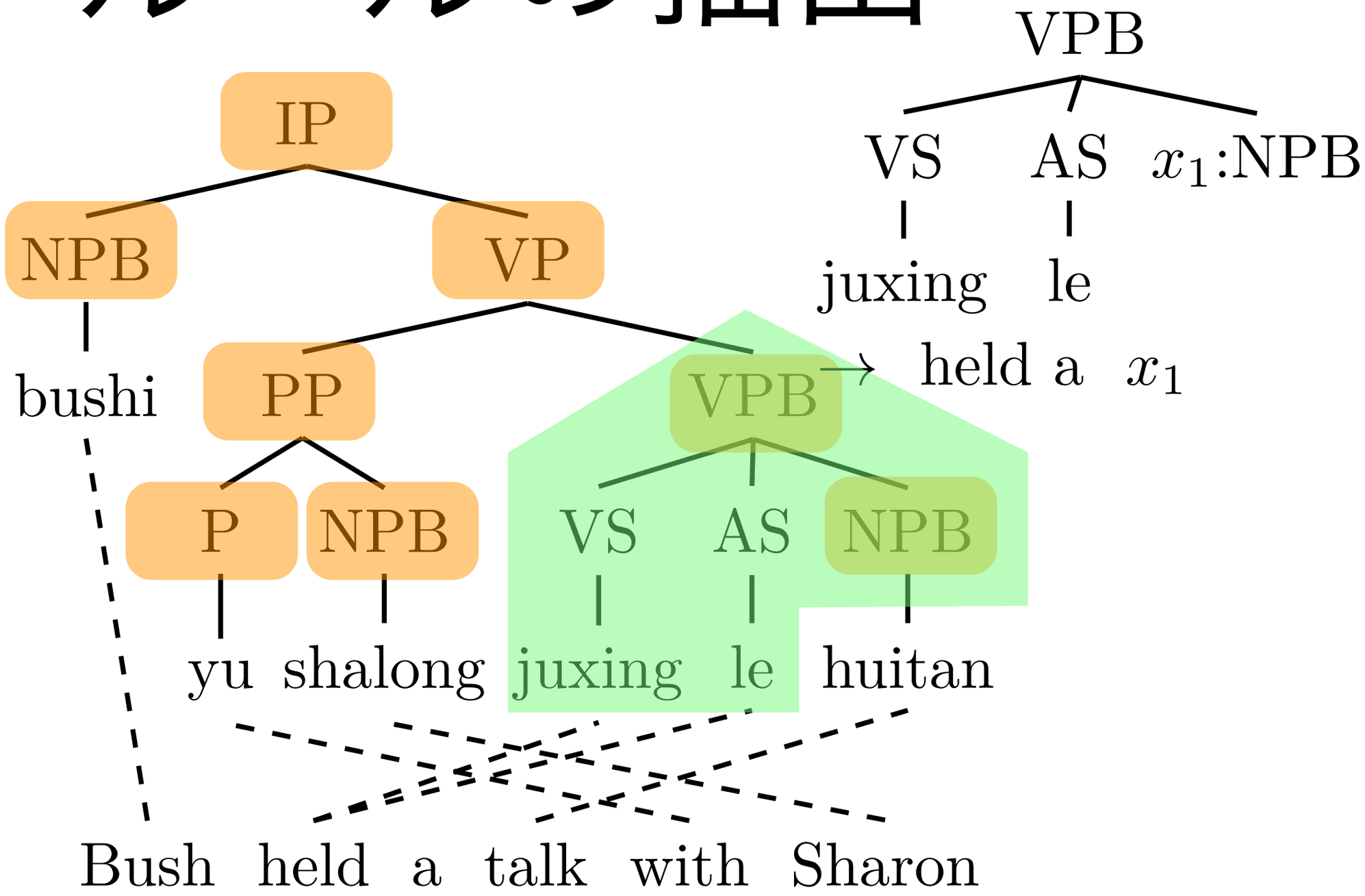
ルールの抽出



(Galley et al., 2004)

- “frontier”から最小ルールを抽出

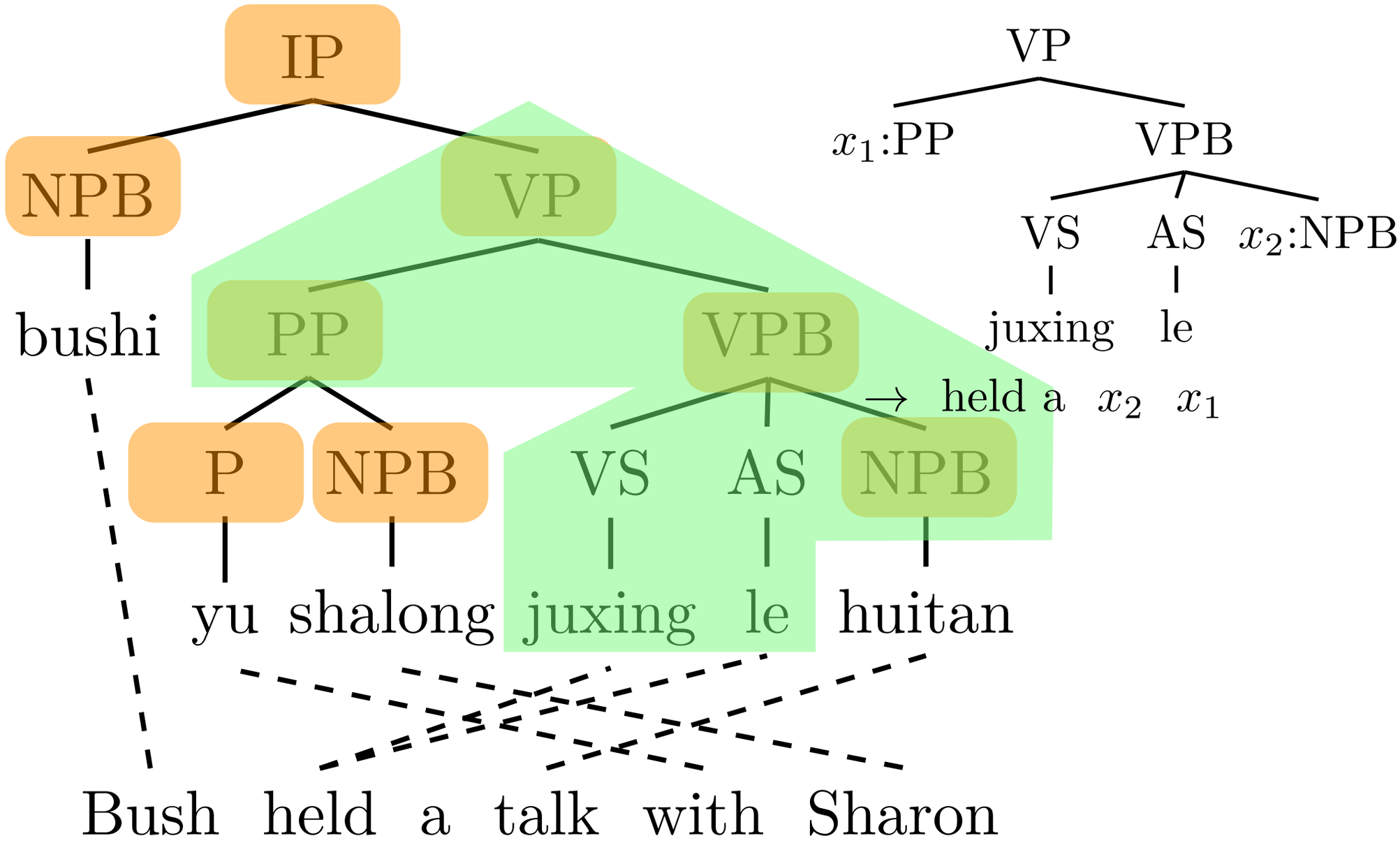
ルールの抽出



(Galley et al., 2004)

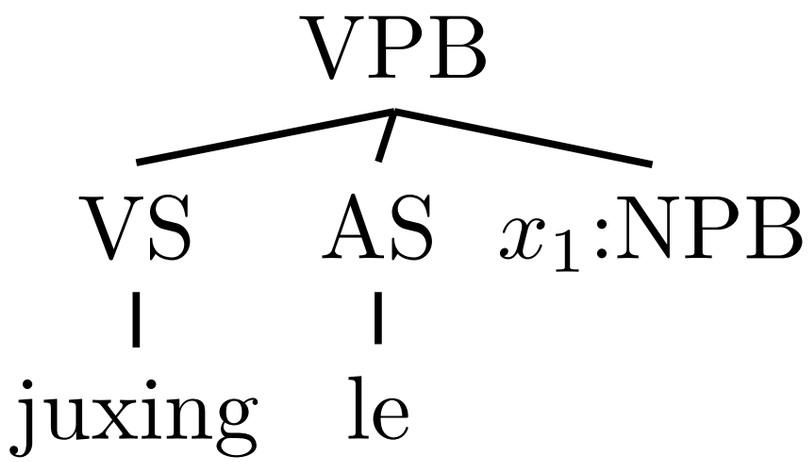
- “frontier”から最小ルールを抽出

ルールの抽出



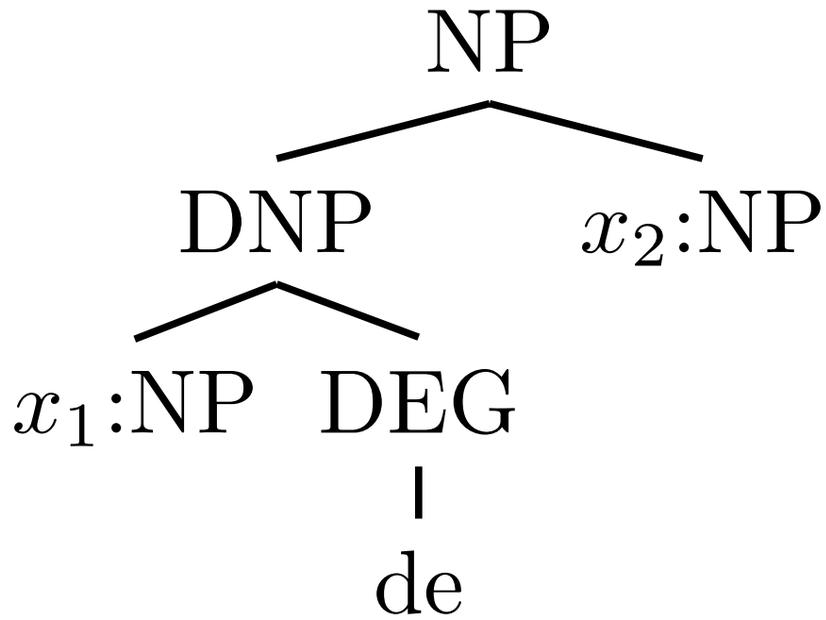
- 最小ルールを組み合わせ、“compound rules”を抽出(長いフレーズ) (Galley et al., 2006)

Decoding: String-{String, Tree}



$\rightarrow$ held a x_1

$\langle \text{VPB} \rightarrow \text{juxing le NPB}_1, \; x \rightarrow \text{hold a } x_1 \rangle$



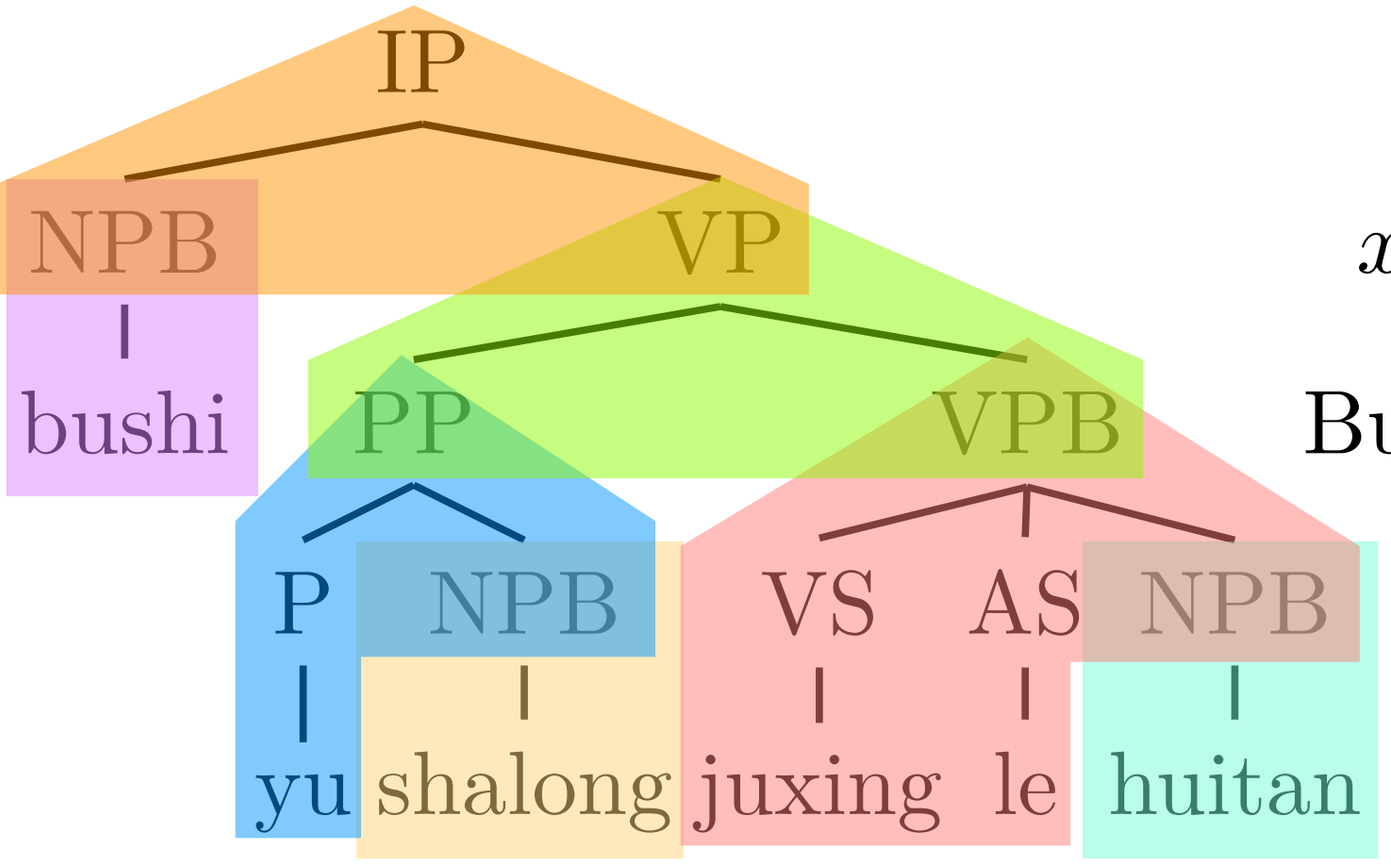
$\rightarrow$ x_2 of x_1

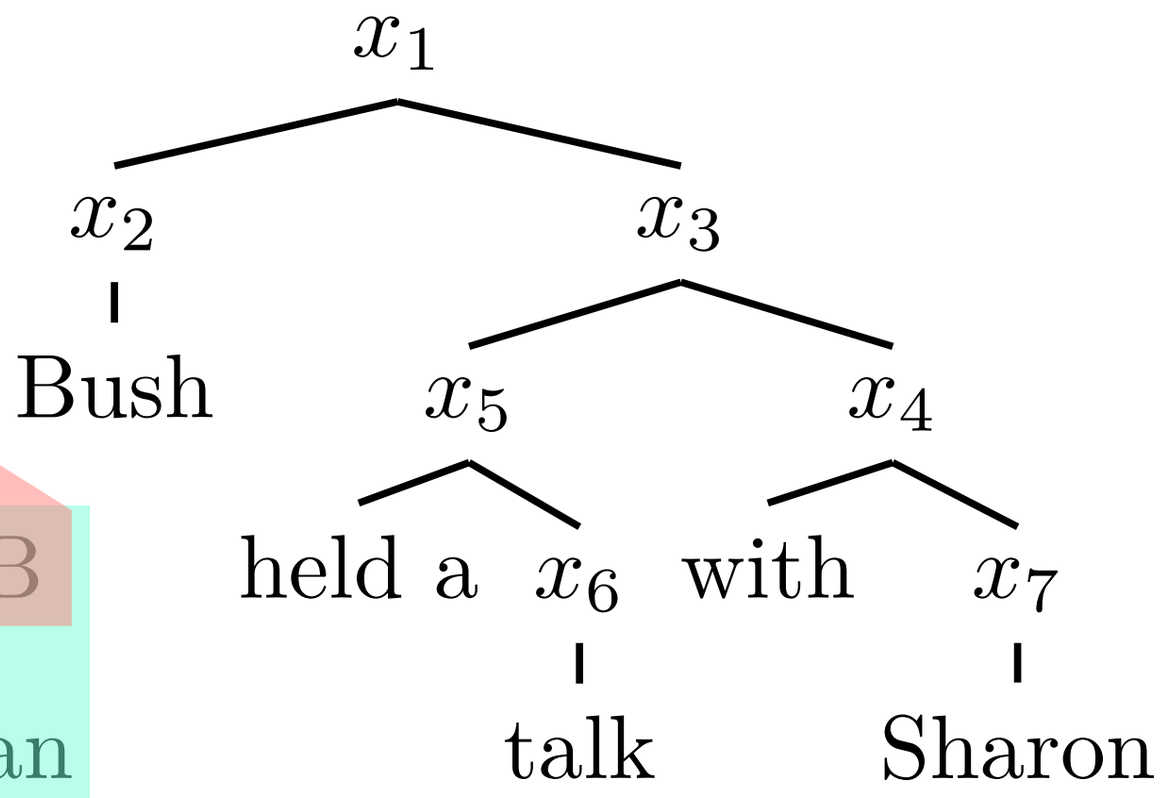
$\langle \text{NP} \rightarrow \text{NP}_1 \text{ de NP}_2, \; x \rightarrow x_2 \text{ of } x_1 \rangle$

(Galley et al., 2004)

- SCFGと同様なデコード:原言語側の内部の構造を取り除いたルールでデコード、目的言語側で翻訳森を生成

Decoding: Tree-{String, Tree}





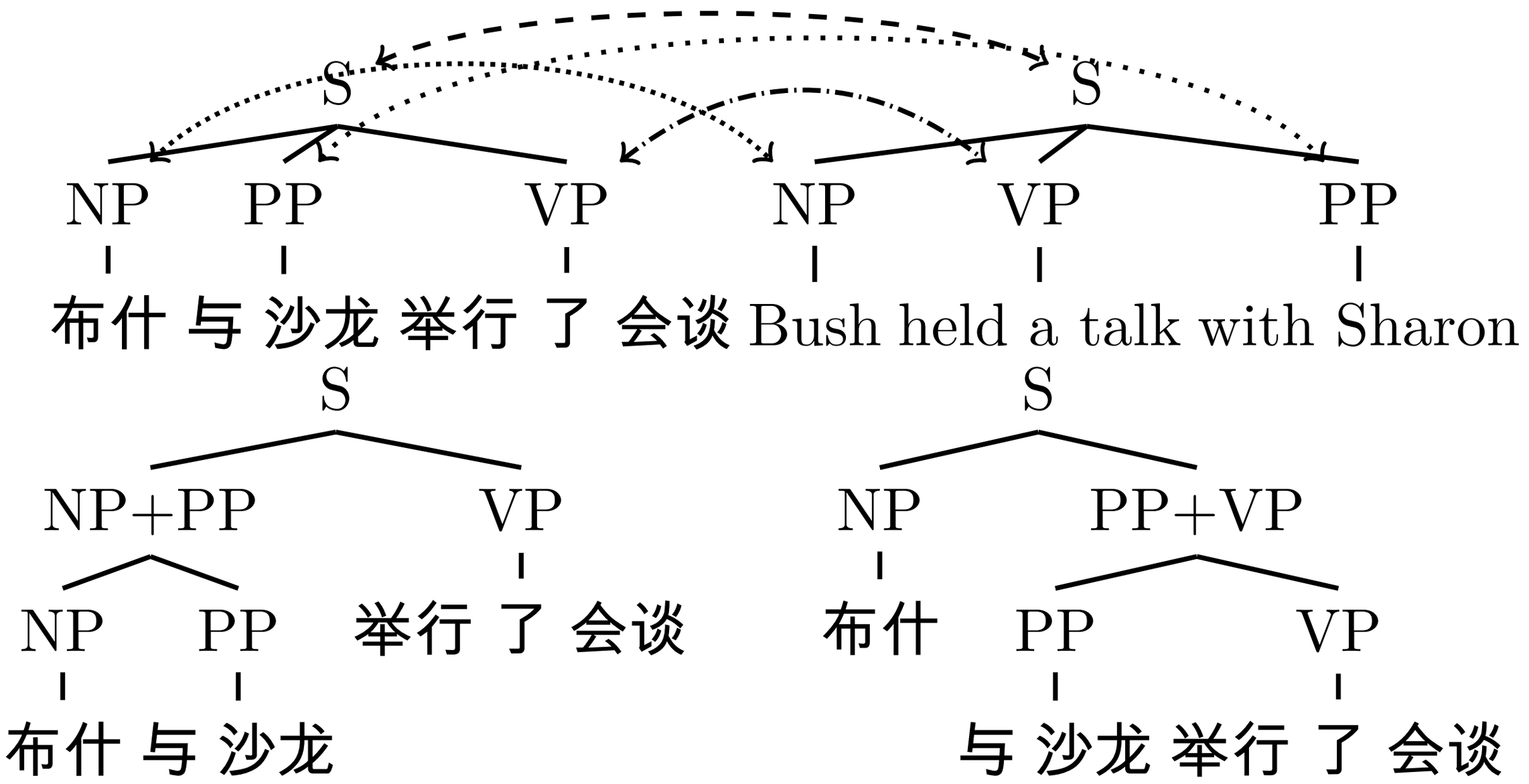
(Huang et al., 2006)

- 入力文を構文解析
- ルールの原言語側でマッチング、目的言語側で翻訳森を生成

Forest Rescoring

- {tree,string}-to-{tree,string}によるデコーディング (Huang and Chiang, 2007)
  - string-to-{tree,sting}: 原言語側で単言語構文解析
  - tree-to-{tree,string}:入力文を構文解析、原言語側で木構造のマッチング
  - 交差したルールの目的言語側で翻訳森を生成
  - 翻訳森から最適な導出を求める(Huang and Chiang, 2005)

Binarization

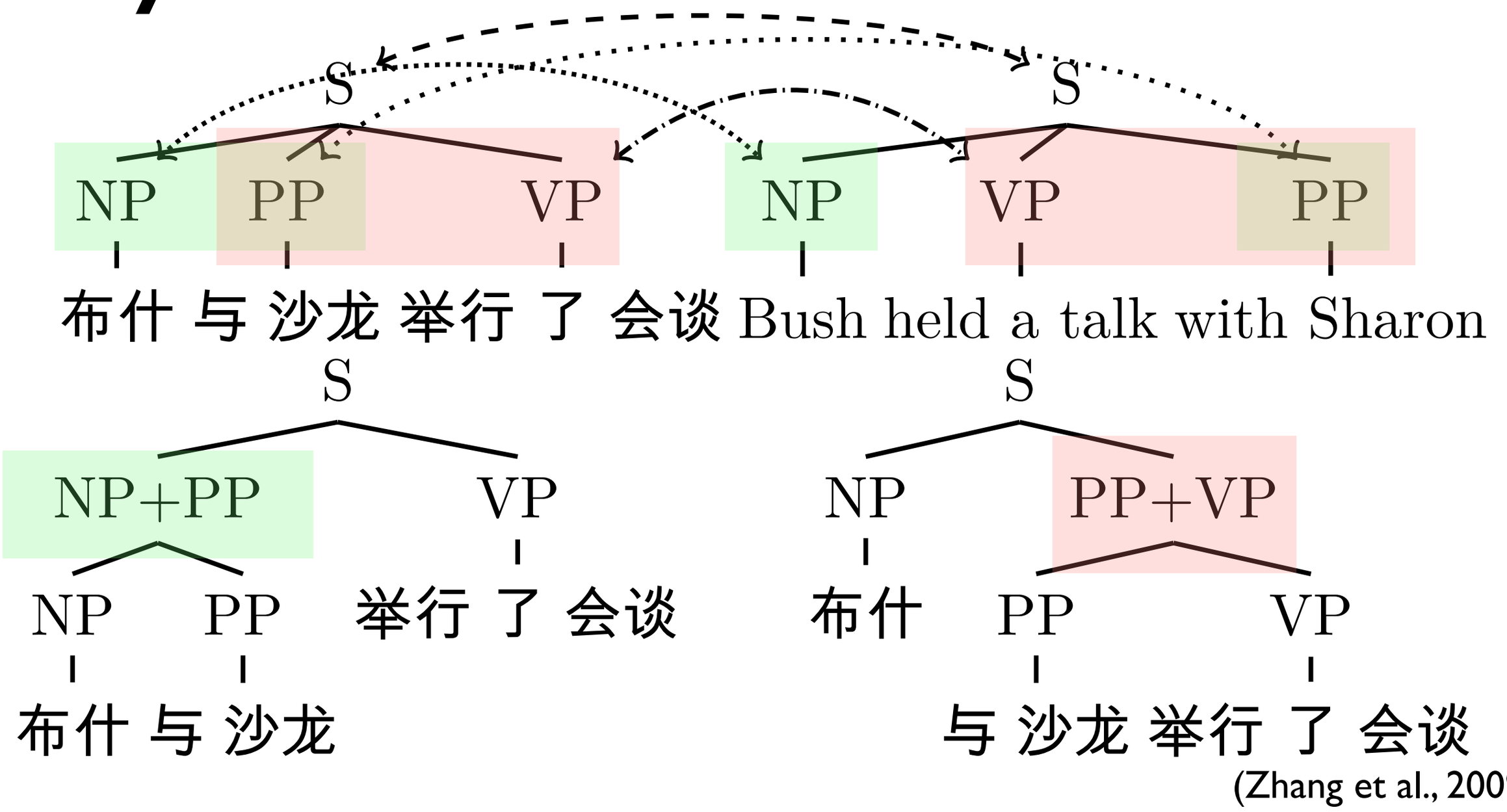


(Zhang et al., 2009)

- 単言語のCNFのように、同期文法のbinarization

- どっちがいいでしょう?

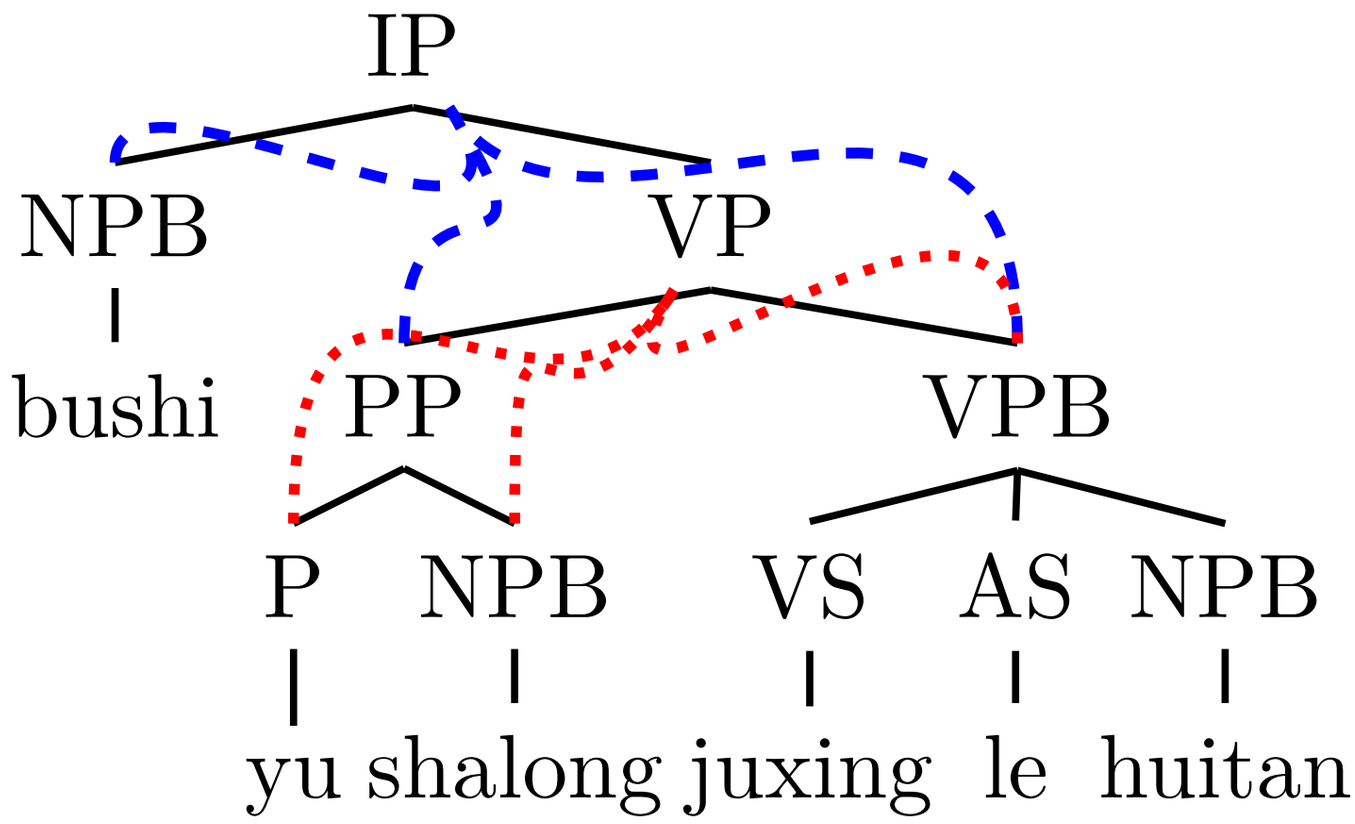
Synchronous Binarization



(Zhang et al., 2009)

- SCFGには、CNFがない(rank≥4の場合、必ずしもbinarizeできない)
- shift-reduce アルゴリズム (Zhang et al., 2009)

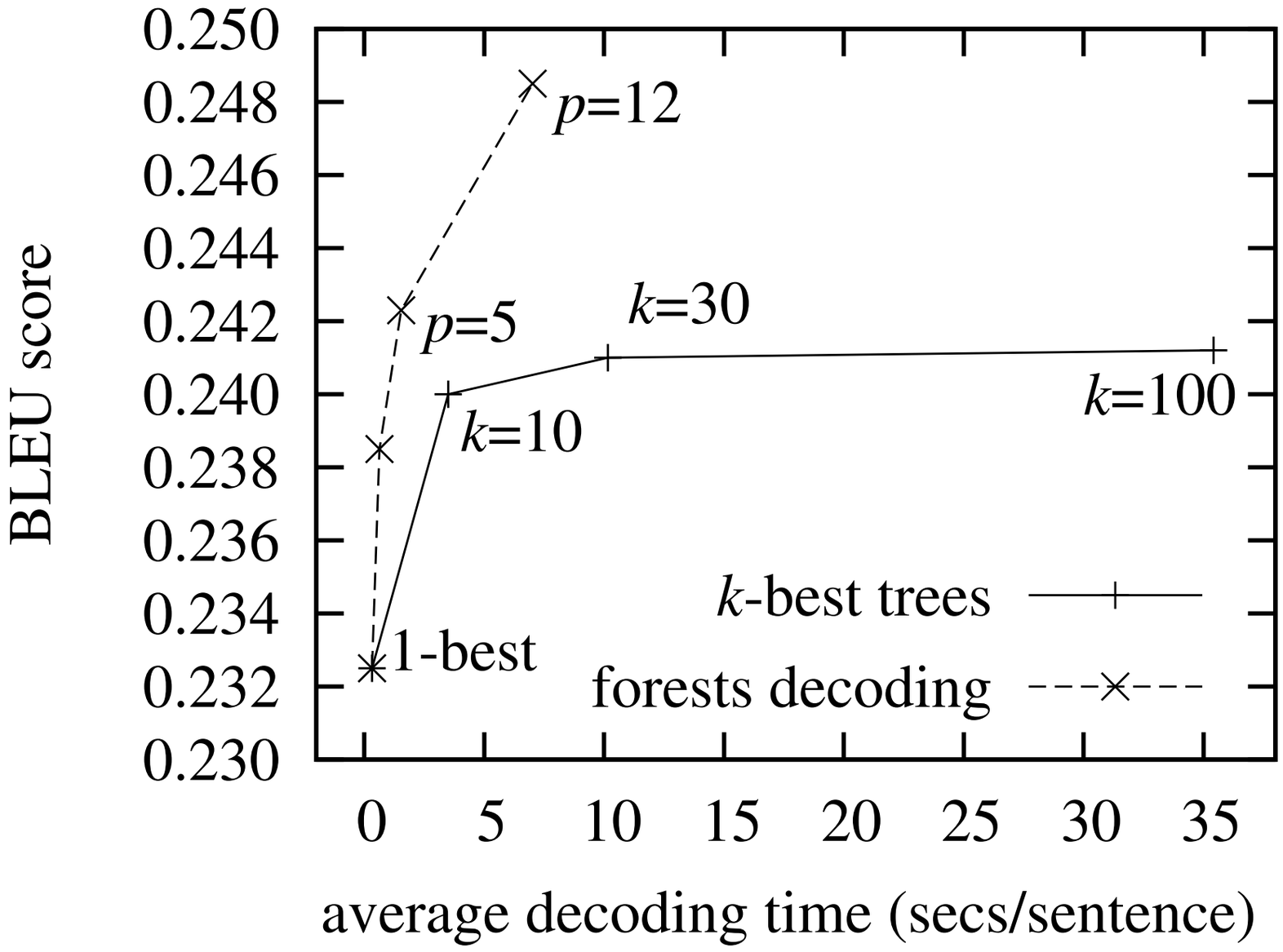
Tree or Forest



(Mi et al., 2008)

- tree-to-stringでは、入力文の構文解析誤りに弱い
- 構文解析器から(枝刈りされた)構文解析森を出力
  - 森から翻訳、あるいは、森から文法獲得

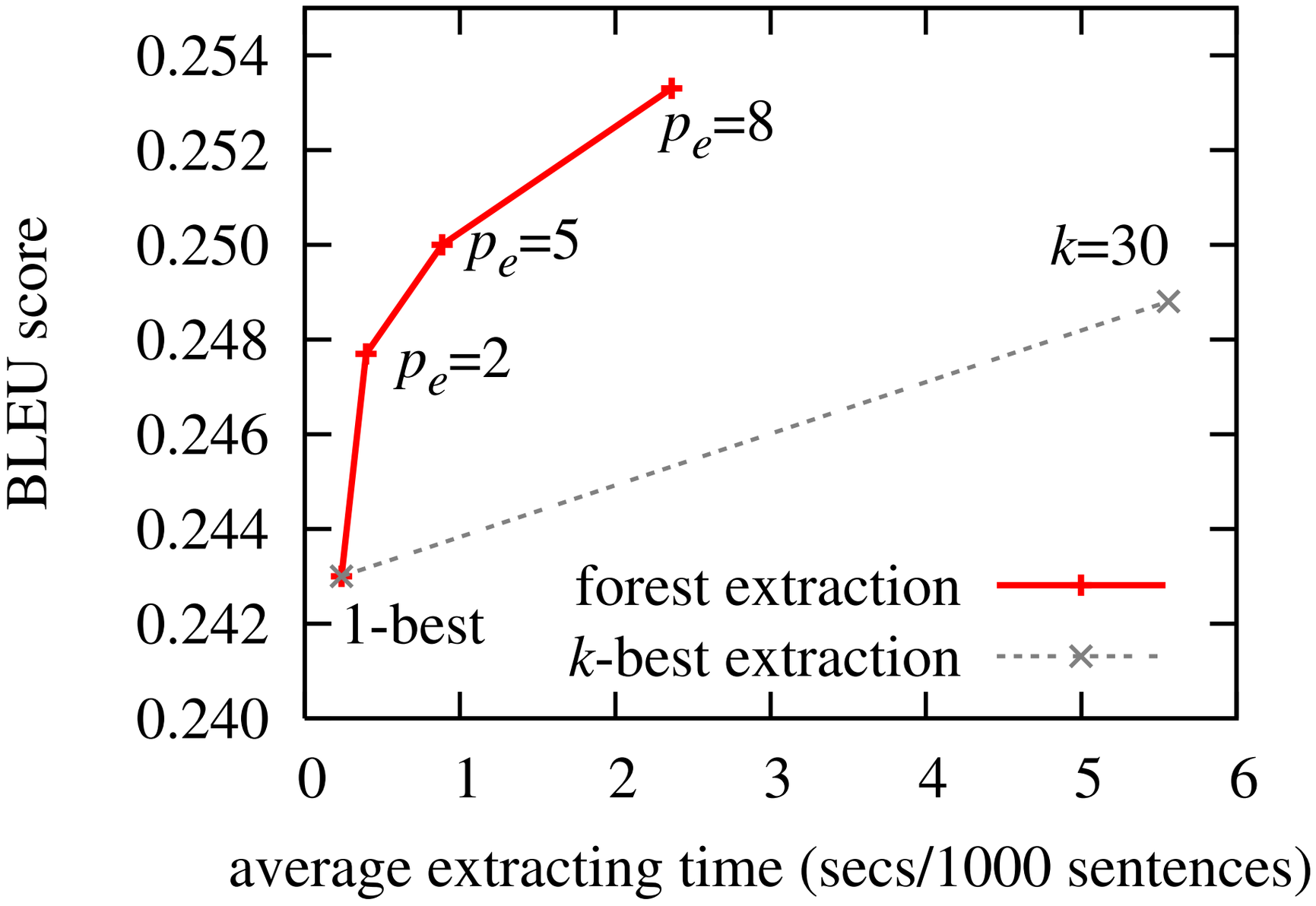
森から翻訳



(Mi et al., 2008)

- k-best構文解析木をそれぞれ翻訳するより高速
- 複数の木を効率よく森で表現することにより制度の高い翻訳

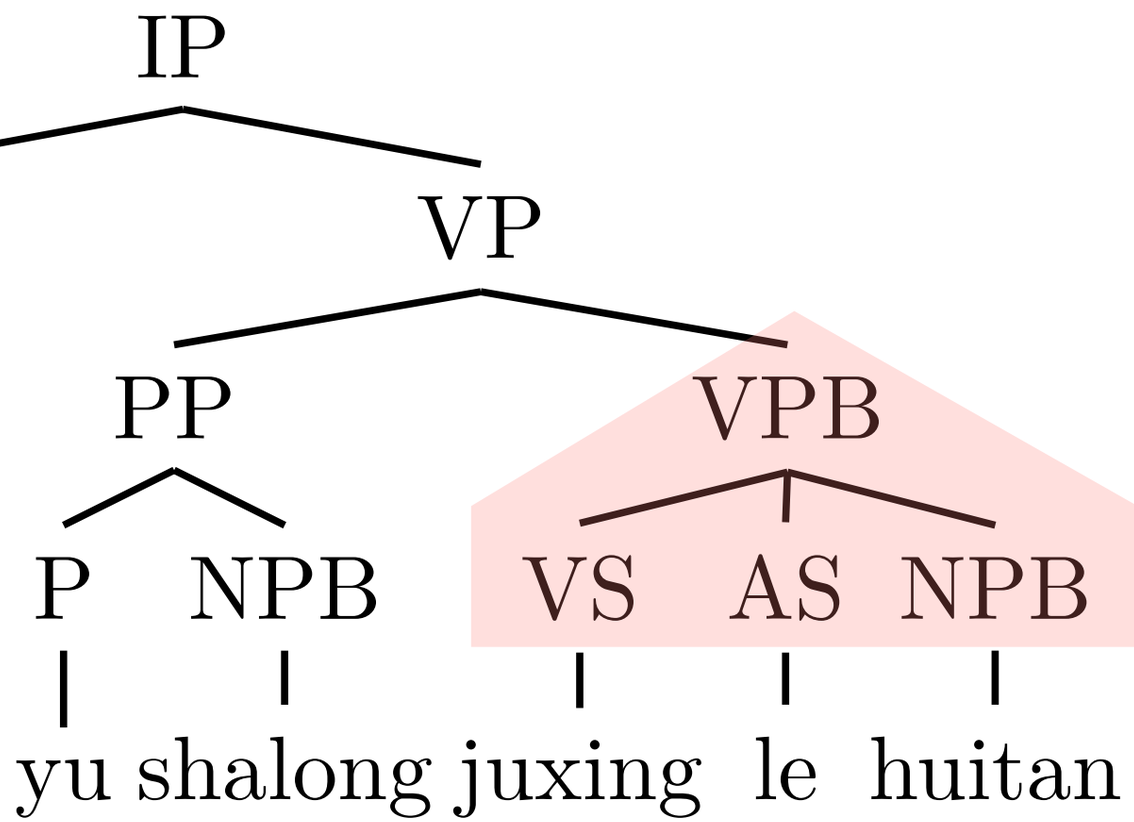
森から文法獲得



(Mi and Huang, 2008)

- k-best構文解析木から獲得するより高速
- 森から文法を学習することにより、よりよい翻訳

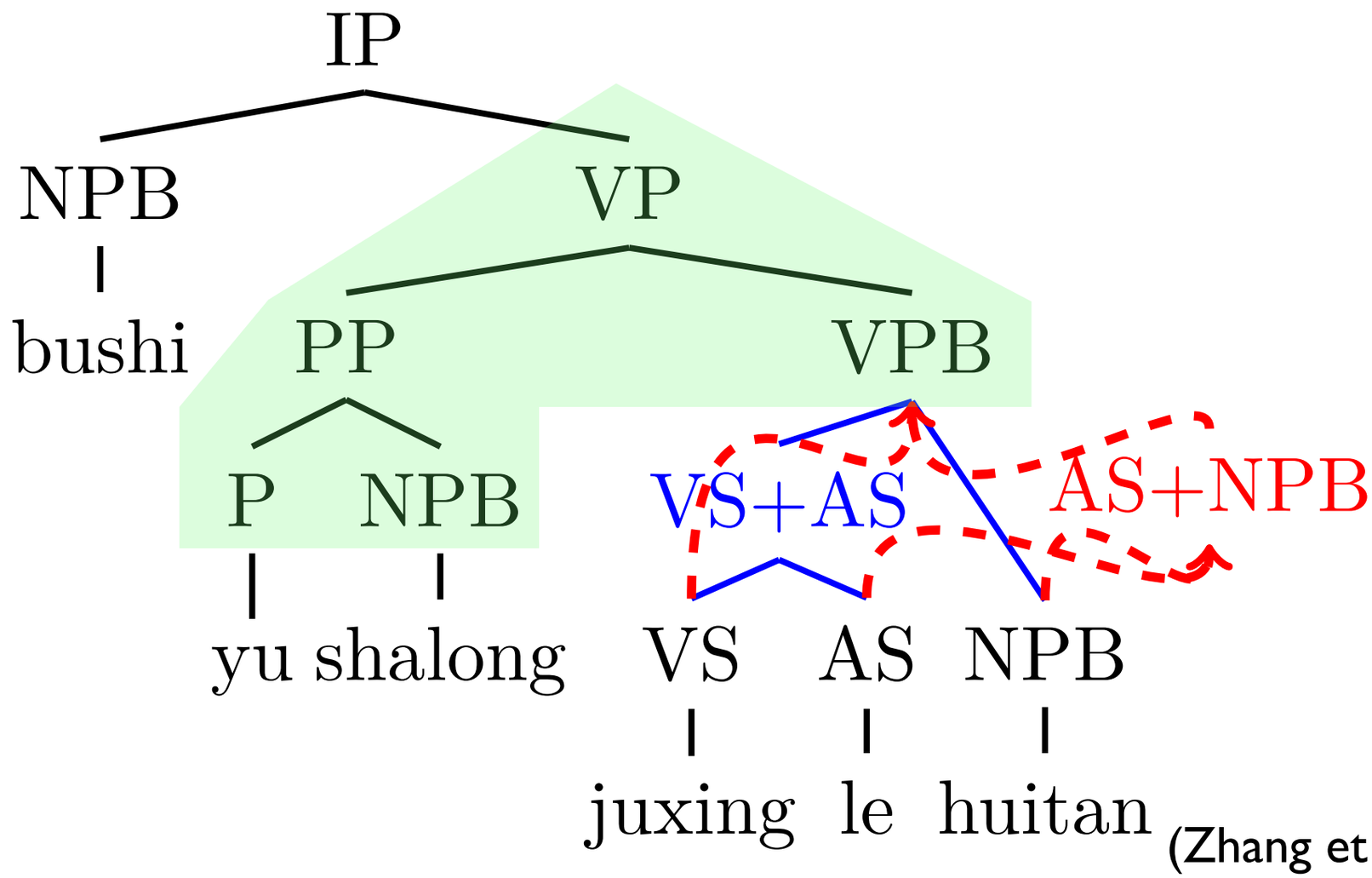
Forest or Binarized Forest



(Zhang et al., 2011)

- 構文解析の1-best出力をbinarize:全てのbinarization およびルールの境界を超えたbinarization

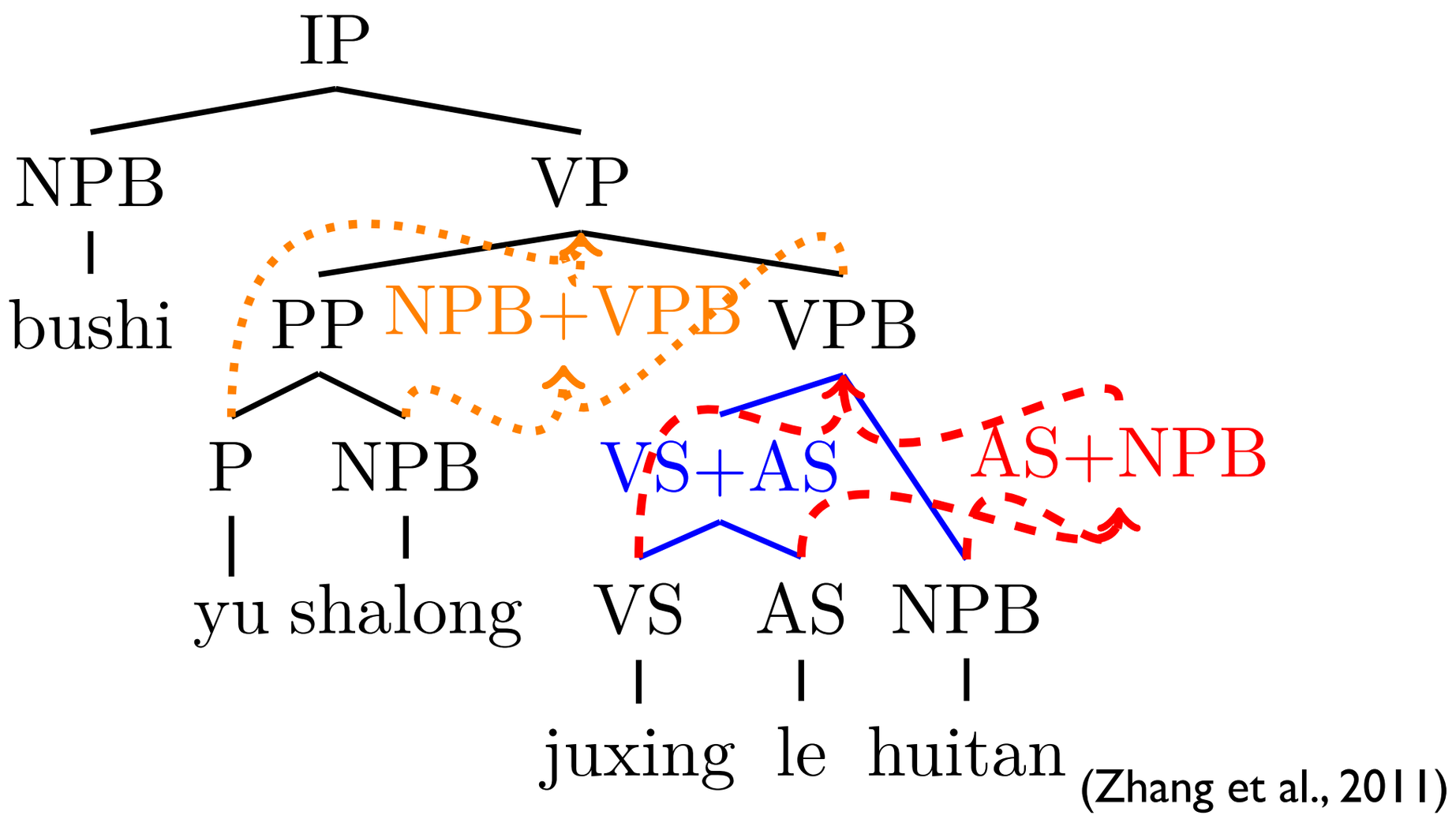
Forest or Binarized Forest



(Zhang et al., 2011)

- 構文解析の1-best出力をbinarize:全てのbinarization およびルールの境界を超えたbinarization

Forest or Binarized Forest



(Zhang et al., 2011)

- 構文解析の1-best出力をbinarize:全てのbinarization およびルールの境界を超えたbinarization

Binarized Forest

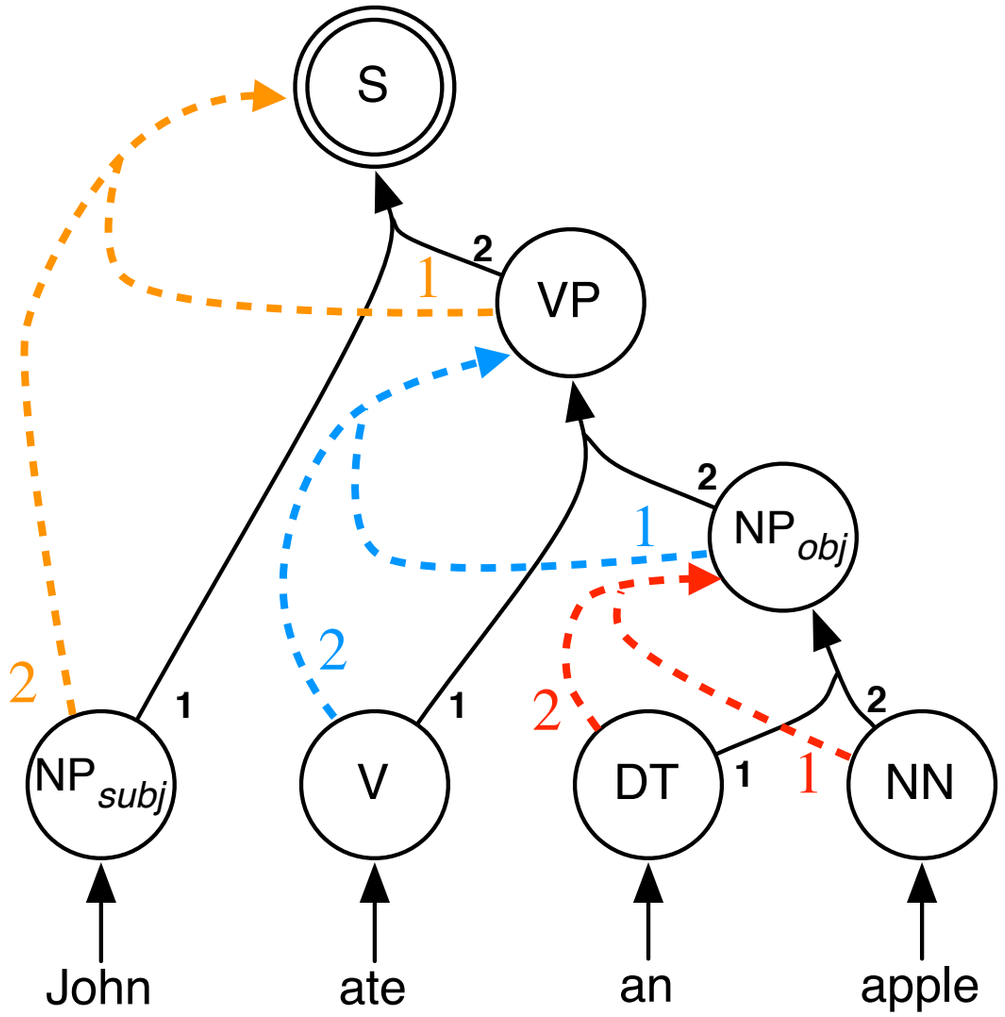
	rules	BLEU dev	BLEU test
no binarization	378M	28.0	36.3
head-out	408M	30.0	38.2
cyk-1	527M	31.6	40.5
cyk-2	803M	31.9	40.7
cyk-3	1053M	32.0	40.6
cyk-∞	1441M	32.0	40.3

	BLEU dev	BLEU test
cyk-2	14.9	16.0
parser	14.7	15.7

(Zhang et al., 2011)

- CYK binarizationの効果大
- 特に構文解析器の構文解析森よりも良い結果

CFGからの並び替え森



(Dyer and Resnik, 2010)

- CFGによる構文解析木にT(e)を並び替えた超辺を追加 (Dyer and Resnik, 2010)
- Earleyアルゴリズムによるフレーズペアとの交差 (Dyer, 2010)
  - Yamada and Knight (2001)と違い、境界は同期していない

CFGからの並び替え森

Feature	λ	note
VP → VE NP	0.995	
VP → VV VP	0.939	modal + VP
VP → VV NP	0.895	
VP → VP PP*	0.803	PP modifier of VP
VP → VV NP IP	0.763	
PP → P NP	0.753	
IP → NP VP PU	0.728	PU = punctuation
VP → VC NP	0.598	
NP → DP NP	0.538	
NP → NP CP*	0.537	rel. clauses follow

Condition	Mono	PB	Hiero	Forest
BTEC	47.4	51.8	52.4	**54.1**
Chinese-Eng.	29.0	30.9	32.1	**32.4**
Arabic-Eng.	41.2	45.8	**46.6**	44.9

(Dyer and Resnik, 2010)

- 並び替えのモデルをMaxEntにて学習
- MERT時には、一つの素性として扱う

まとめ

- {tree,string}-to-{tree,string}による翻訳
- もう完全にルール翻訳、でも
  - ルールは自動獲得
  - 統計量に基づくスコア
- 構文解析誤りに対する頑健性のため、「森」を利用

内容

- 木構造に基づく機械翻訳
  - 背景: CFG, hypergraph, deductive system
  - 同期文脈自由文法 (synchronous-CFG)
  - 同期文法: {string,tree}-to-{string,tree}
  - 二言語の構文解析(biparsing)
  - 同期から非同期
- 最適化

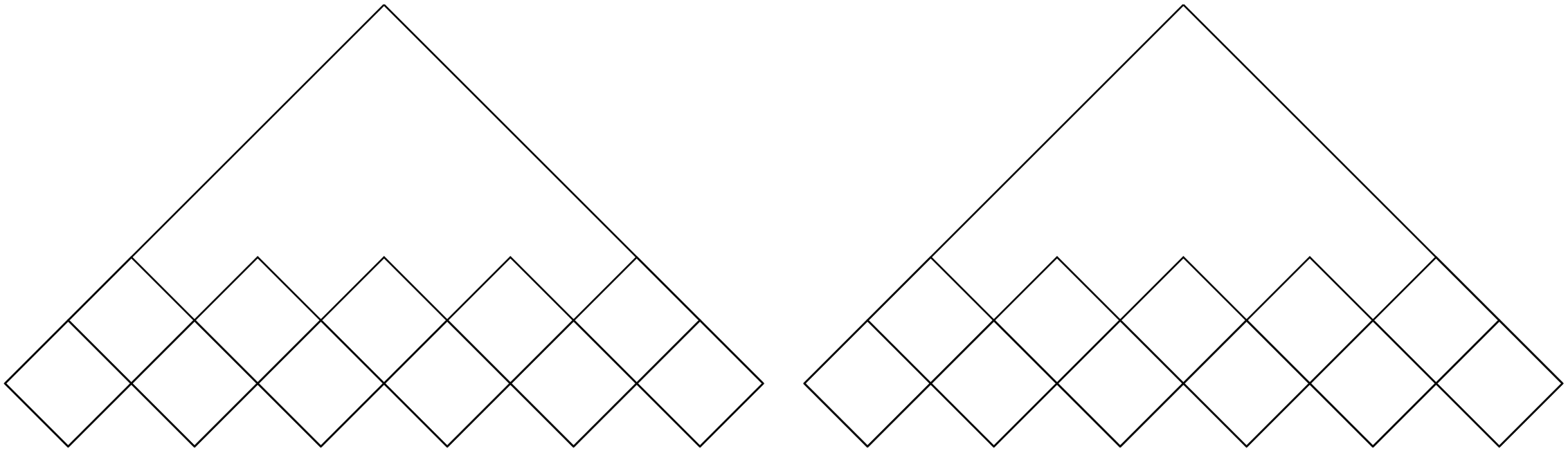
ITG

$$
\begin{aligned}
\mathrm{X} &\rightarrow \langle \mathrm{X}_{\boxed{1}}\ \mathrm{X}_{\boxed{2}}, \mathrm{X}_{\boxed{1}}\ \mathrm{X}_{\boxed{2}} \rangle \\
\mathrm{X} &\rightarrow \langle \mathrm{X}_{\boxed{1}}\ \mathrm{X}_{\boxed{2}}, \mathrm{X}_{\boxed{2}}\ \mathrm{X}_{\boxed{1}} \rangle \\
\mathrm{X} &\rightarrow \langle f, e \rangle
\end{aligned}
$$

$$\mathrm{X} \rightarrow [X\ X] \mid \langle X\ X \rangle \mid f/e$$

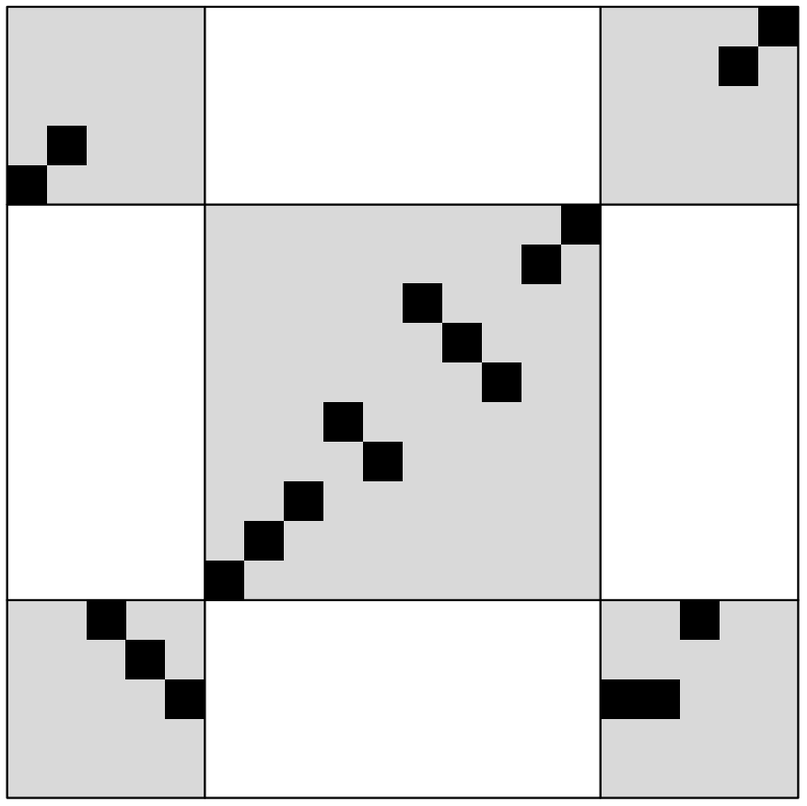
- SCFGのインスタンス、Inversion Transduction Grammar (ITG) (Wu, 1997)
- 単語アライメント(Wu, 1997; Zhang and Gildea, 2005; Haghighi et al., 2009)、フレーズアライメント(Cherry and Lin, 2007; Zhang et al., 2008)、デコード時の制約 (Zens and Ney, 2003; Zens et al., 2004)

二言語構文解析



- 二言語の文とSCFGとの交差
- ITG (Wu, 1997) では、$O(N^3 M^3)$
  - 各長さnとm、各位置iとj、各ルールX → YZ、各分岐点kとl

Span Pruning



- 予めfigure-of-meritスコアでスパンのペアを枝刈りすることで高速化
  - $O(N^4)$ for a naive algorithm (Zhang and Gildea, 2005)
  - $O(N^3)$ for a DP-based algorithm (Zhang et al., 2008)

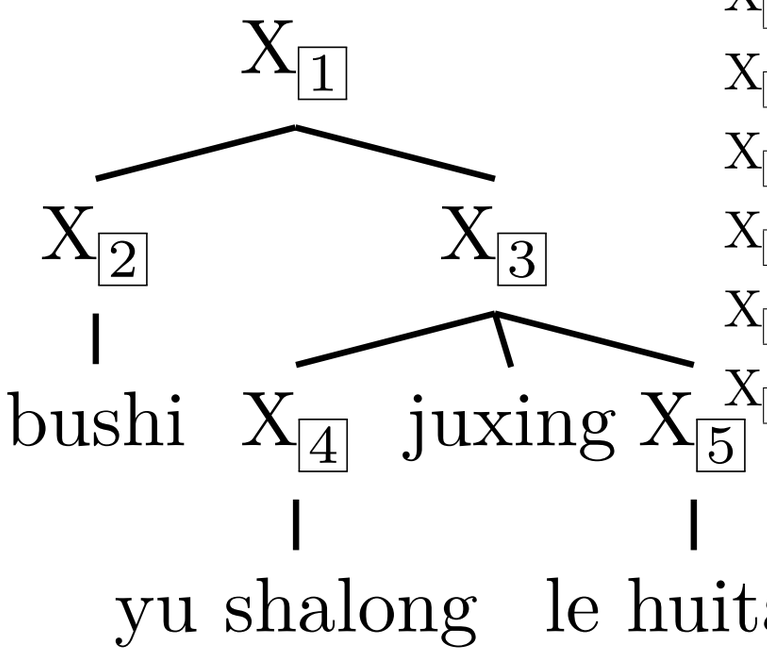
Beam Pruning

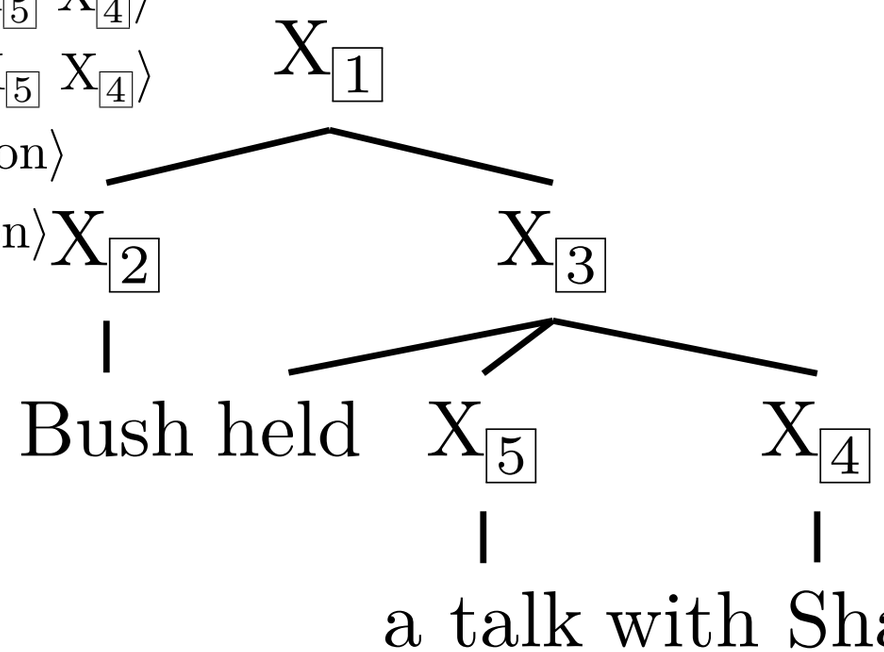
--●---,------	--●---,--●---	--●---,--●●--	--●●--,--●●--
-●----,------	-●----,--●---	-●----,--●●--	●●----,--●●--
------,----●-	---●--,----●-	---●●-,----●-	---●●-,---●●-
------,---●--	------,--●●--	---●--,--●●--	---●●-,--●●--
------,-●----	---●--,-●----	--●●--,-●----	--●●--,●●----

- cardinalityで探索空間をグループ化 (Saers et al., 2009)
  - cardinality = 構文解析された終端記号の数
- cardinality毎に枝刈り: 計算量 $O(bN^3)$

Two Parse

$$
\begin{aligned}
X_{\boxed{1}} &\rightarrow \langle X_{\boxed{2}}\ X_{\boxed{3}}, X_{\boxed{2}}\ X_{\boxed{3}} \rangle \\
X_{\boxed{1}} &\rightarrow \langle X_{\boxed{2}}\ X_{\boxed{3}}, X_{\boxed{3}}\ X_{\boxed{2}} \rangle \\
X_{\boxed{2}} &\rightarrow \langle \text{bushi}, \text{Bush} \rangle \\
X_{\boxed{3}} &\rightarrow \langle X_{\boxed{4}}\ \text{juxing}\ X_{\boxed{5}}, \text{held}\ X_{\boxed{5}}\ X_{\boxed{4}} \rangle \\
X_{\boxed{3}} &\rightarrow \langle X_{\boxed{4}}\ \text{juxing}\ X_{\boxed{5}}, \text{hold}\ X_{\boxed{5}}\ X_{\boxed{4}} \rangle \\
X_{\boxed{4}} &\rightarrow \langle \text{yu shalong}, \text{with Sharon} \rangle \\
X_{\boxed{4}} &\rightarrow \langle \text{yu shalong}, \text{and Sharon} \rangle \\
X_{\boxed{5}} &\rightarrow \langle \text{le huitan}, \text{a talk} \rangle \\
X_{\boxed{5}} &\rightarrow \langle \text{le huitan}, \text{talked} \rangle
\end{aligned}
$$



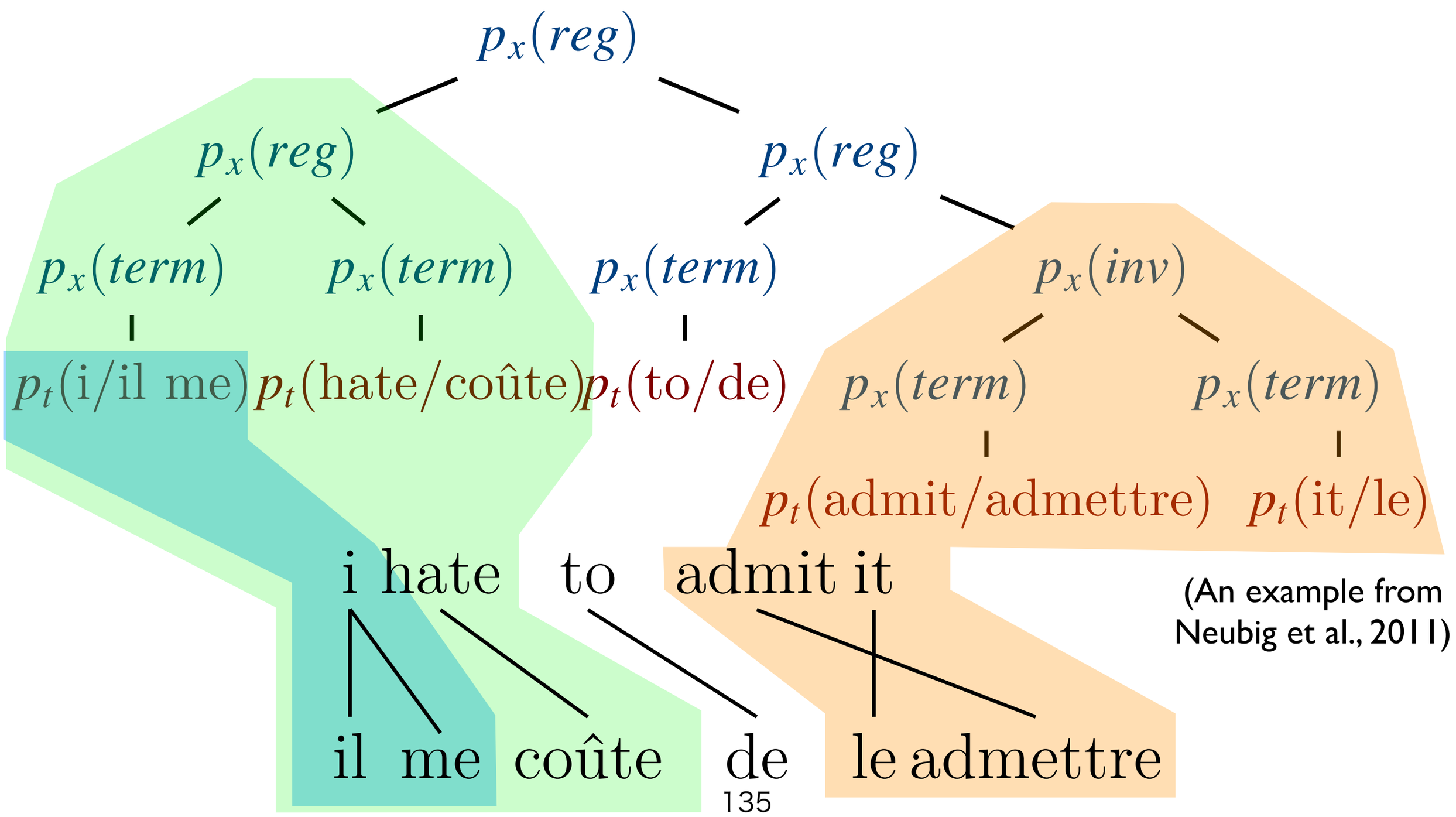


- Hiro文法では、全てのルールを列挙する必要はない

  (Dyer, 2010): Dyer and Resnik (2010) で使用

  - 原言語側で原言語を解析
  - 交差したルールの目的言語側で目的言語を解析

ITGによるアライメント



(An example from Neubig et al., 2011)

ITGアライメント

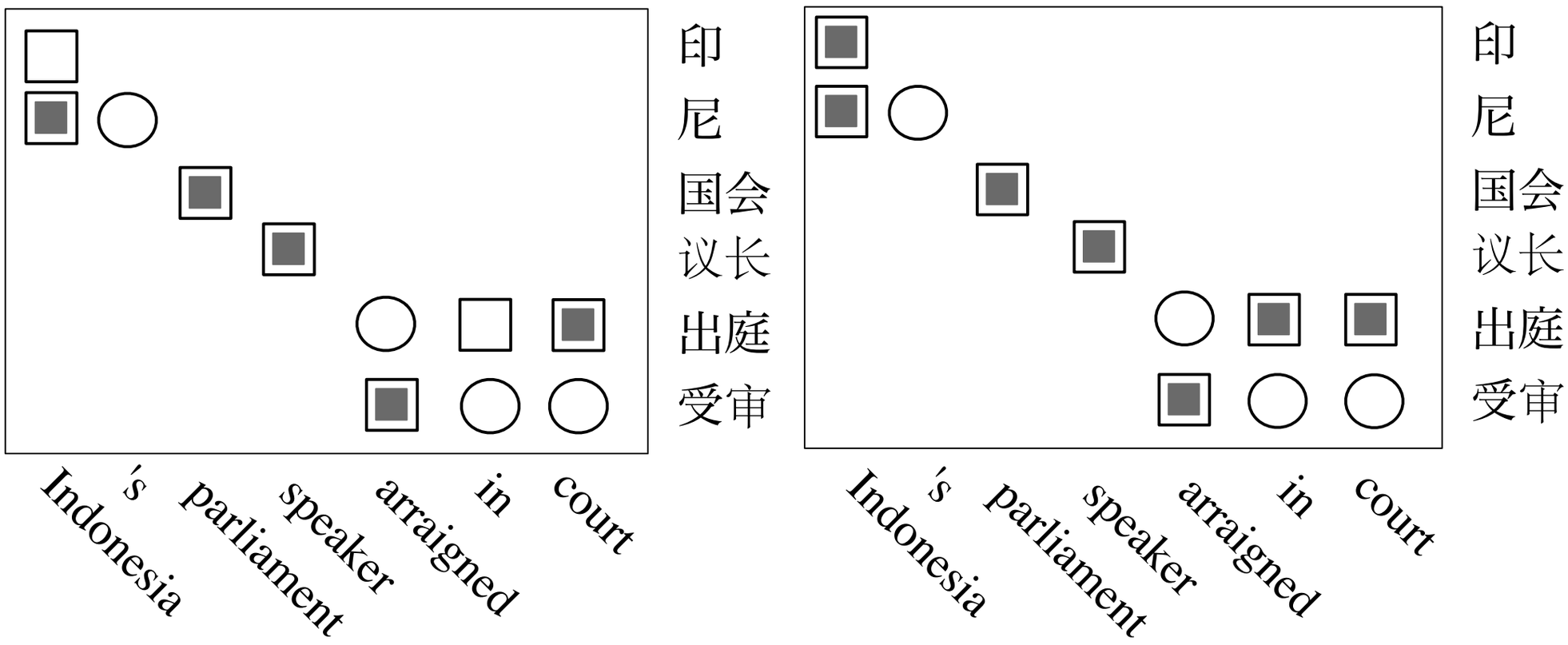
Method	Prec	Rec	AER
Matching	0.916	0.860	0.110
D-ITG	0.940	0.854	0.100
SD-ITG	0.944	0.878	0.086

(Cherry and Lin, 2006)

- マージン最大化学習による学習+依存構造の制約
- アライメント空間 (Zens and Ney, 2003): Schröder Number O(5.83^n)
- Alignment Error Rate(AER)による評価
  - A = アライメントの数、S = Sureアライメントの数、P = Possibleアライメントの数

$$AER(A,S,P) = \left(1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}\right)$$

Block-ITGアライメント



(Haghighi et al., 2009)

- フレーズ単位の制約を入れる (Haghighi et al., 2009)

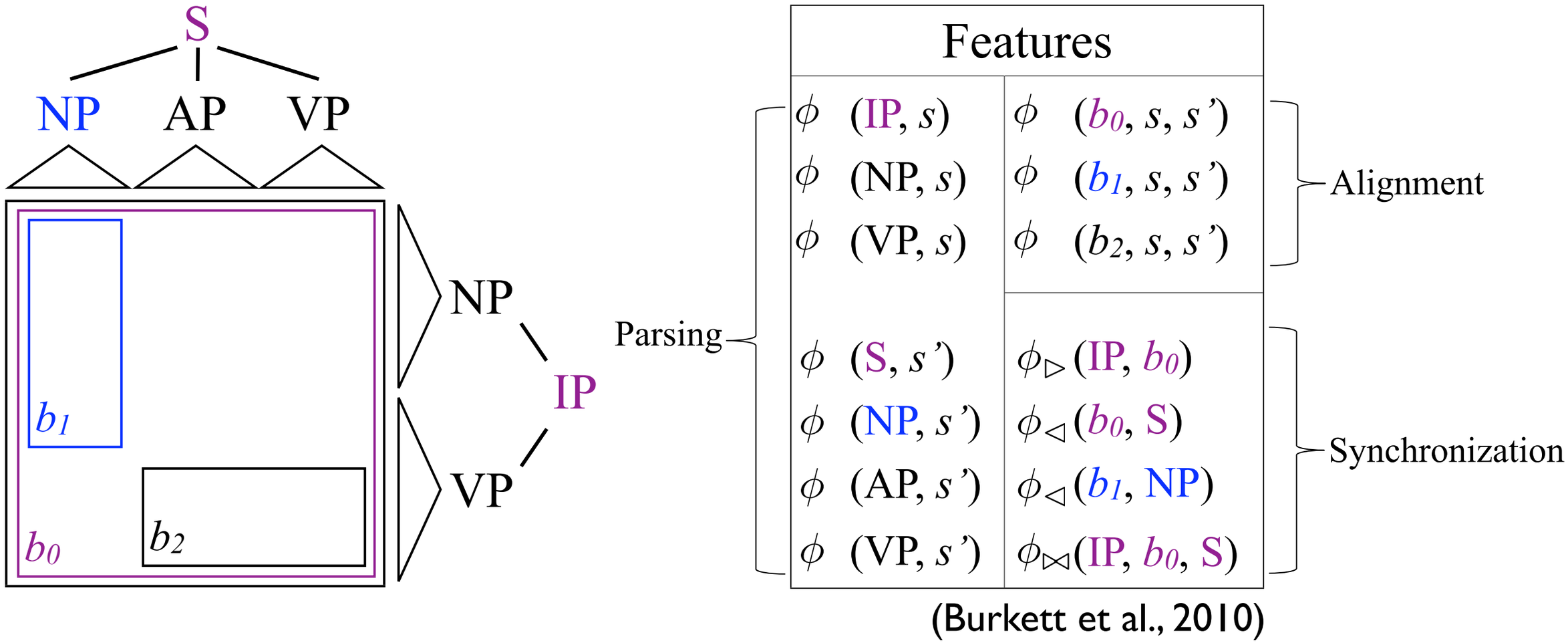
Block-ITGアライメント

Alignments			Translations	
Model	Prec	Rec	Rules	BLEU
GIZA++	62	84	1.9M	23.22
Joint HMM	79	77	4.0M	23.05
Viterbi ITG	90	80	3.8M	24.28
Posterior ITG	81	83	4.2M	**24.32**

(Haghighi et al., 2009)

- 中国語、英語アライメントタスク
- MIRAとMaxEntによる学習
- アライメントの向上によるBLEUの向上をはじめて示した結果

ITG + Bi-parsingアライメント



(Burkett et al., 2010)

- ITGアライメント+原言語、目的言語の構文木
- 非同期的な素性 + Mean Field Inferenceによる学習

ITG + Bi-parsingアライメント

	Test Results			
	Precision	Recall	AER	F_1
HMM	86.0	58.4	30.0	69.5
ITG	**86.8**	73.4	20.2	79.5
Joint	85.5	**84.6**	**14.9**	**85.0**

	Rules	Tune	Test
HMM	1.1M	29.0	29.4
ITG	1.5M	**29.9**	30.4†
Joint	1.5M	29.6	**30.6**

(Burkett et al., 2010)

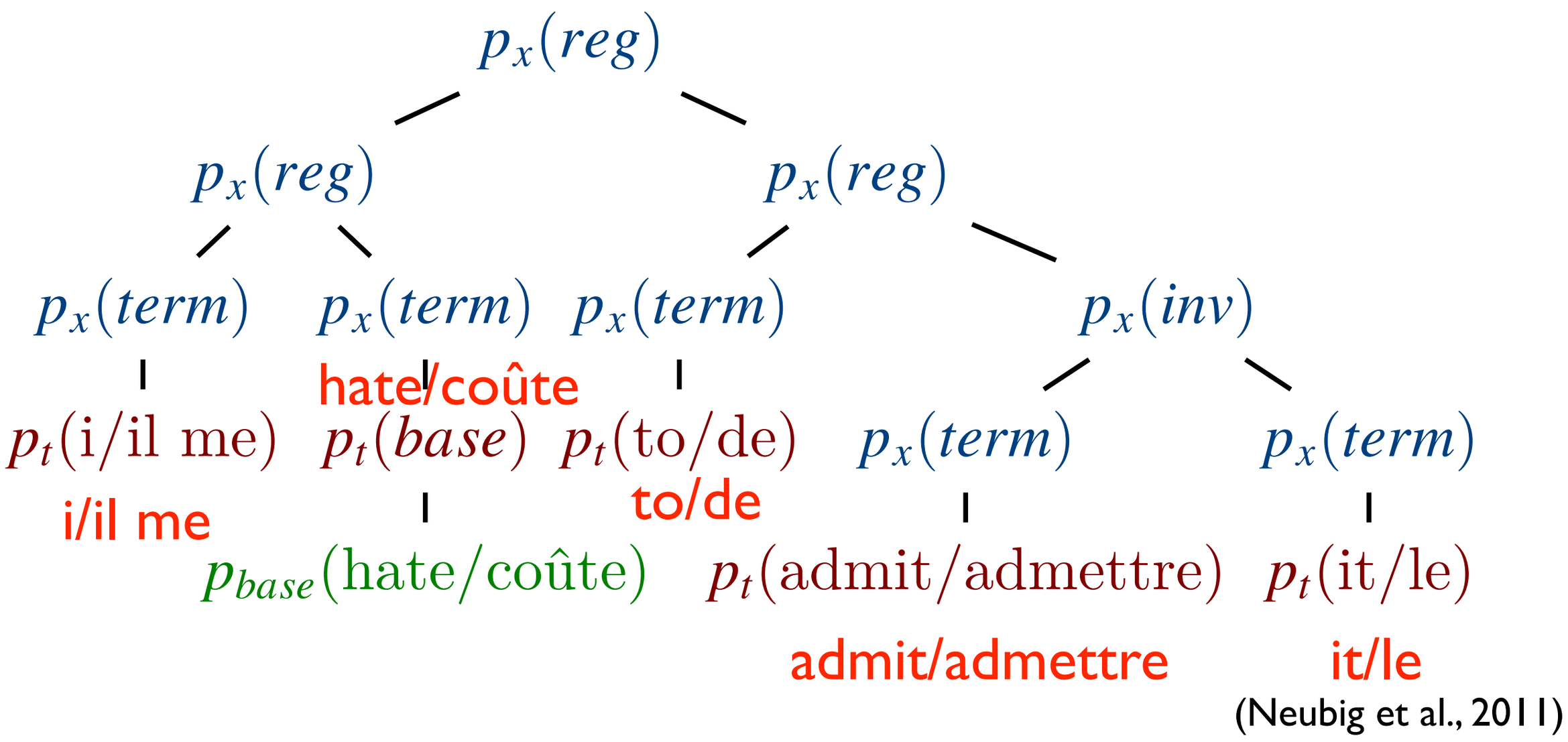
単語からフレーズへ

- 教師なし学習によりフレーズペアを直接学習

  (Marcu and Wong, 2002; DeNero et al., 2008; Arun et al., 2009)

- ITGに基づく学習(Cherry and Lin, 2007; Zhang et al., 2008; Blunsom et al., 2009)
- 実は、あまり性能向上していない
  - 結局、最後にヒューリスティックな句の抽出

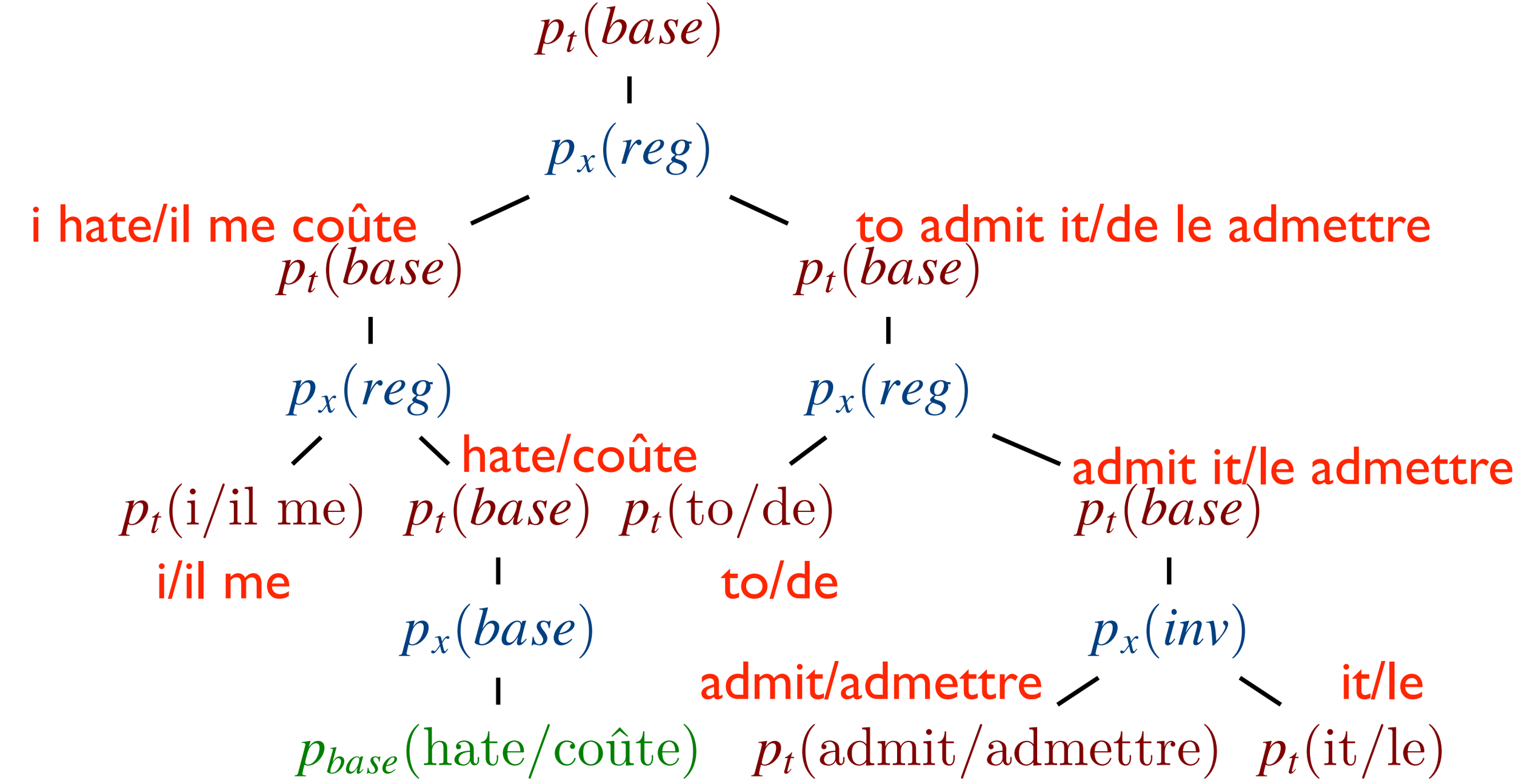
ITGによる導出



(Neubig et al., 2011)

- 最小フレーズのみモデルへ追加

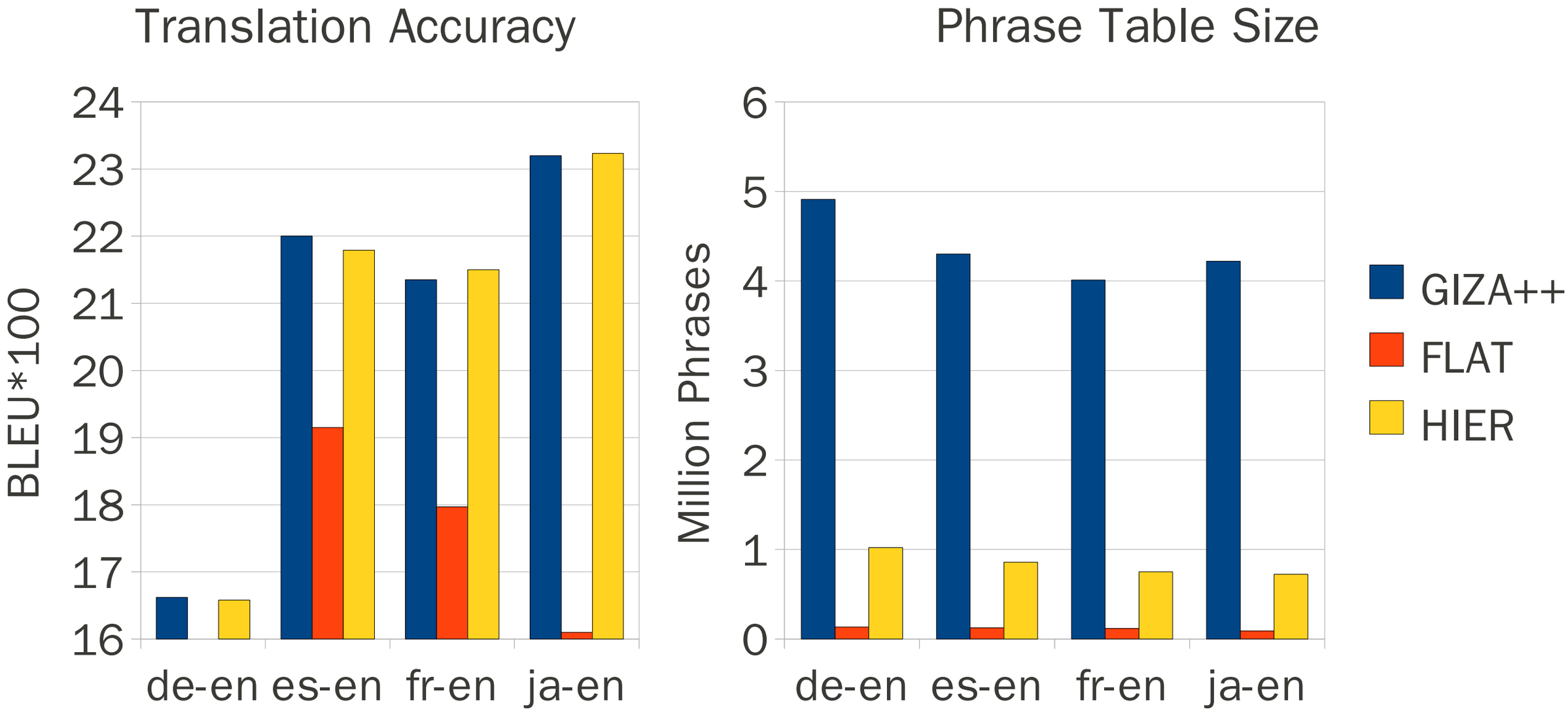
Backoff ITGの導出



- P_t で生成された全てのペアを追加

(Neubig et al., 2011)

Backoff ITGの導出



(Neubig et al., 2011)

- 小さいモデルでGIZA++と同じ精度

http://www.phontron.com/pialign

ITGの並び替え

$$\mathrm{A} \rightarrow [\mathrm{B}\ \ \mathrm{C}] \qquad \mathrm{A} \rightarrow \langle \mathrm{B}^{\mathbf{L}}\ \ \mathrm{C}^{\mathbf{R}} \rangle$$

$$\mathrm{A}^{\mathbf{L}} \rightarrow [\mathrm{B}\ \ \mathrm{C}] \qquad \mathrm{A}^{\mathbf{L}} \rightarrow \langle \mathrm{B}^{\mathbf{L}}\ \ \mathrm{C}^{\mathbf{R}} \rangle$$

$$\mathrm{A}^{\mathbf{R}} \rightarrow [\mathrm{B}\ \ \mathrm{C}] \qquad \mathrm{A}^{\mathbf{R}} \rightarrow \langle \mathrm{B}^{\mathbf{L}}\ \ \mathrm{C}^{\mathbf{R}} \rangle$$

$$\mathrm{A} \rightarrow \mathrm{A}_{\mathbf{P}} \qquad \mathrm{A}_{\mathbf{P}} \rightarrow \alpha \ / \ \beta$$

$$\mathrm{A}^{\mathbf{L}} \rightarrow \mathrm{A}^{\mathbf{L}}_{\mathbf{P}} \qquad \mathrm{A}^{\mathbf{L}}_{\mathbf{P}} \rightarrow \alpha \ / \ \beta$$

$$\mathrm{A}^{\mathbf{R}} \rightarrow \mathrm{A}^{\mathbf{R}}_{\mathbf{P}} \qquad \mathrm{A}^{\mathbf{R}}_{\mathbf{P}} \rightarrow \alpha \ / \ \beta$$

(Mylonakis and Sima'an, 2011)

- ITGは一つのレベルしか記憶していない
- 一つ前も覚えましょう (Mylonakis and Sima'an, 2011)

カテゴリーの学習

<table>
<tr><td colspan="4">X, SBAR, WHNP+VP, WHNP+VBZ+NP</td></tr>
<tr><td></td><td colspan="3">X, VBZ+NP, VP, SBAR\WHNP</td></tr>
<tr><td colspan="3">X, SBAR/NN, WHNP+VBZ+DT</td><td></td></tr>
<tr><td></td><td colspan="2">X, VBZ+DT, VP/NN</td><td></td></tr>
<tr><td colspan="2">X, WHNP+VBZ, SBAR/NP</td><td colspan="2">X, NP, VP\VBZ</td></tr>
<tr><td>X, WHNP, SBAR/VP</td><td>X, VBZ, VP/NP</td><td>X, DT, NP/NN</td><td>X, NN, NP\DT</td></tr>
<tr><td>which</td><td>is</td><td>the</td><td>problem</td></tr>
</table>

(Mylonakis and Sima'an, 2011)

- Xだけでは不十分、でも統語論的なカテゴリーを一意に決定できない (Zollman and Venugopal, 2006)
- EMアルゴリズムで学習してしまいましょう (Mylonakis and Sima'an, 2011)

カテゴリーの学習

Training set size	English to	French		German		Dutch		Chinese	
		BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST
200K	josh-base	29.20	7.2123	18.65	5.8047	21.97	6.2469	22.34	6.5540
	lts	**29.43**	**7.2611****	**19.10****	**5.8714****	**22.31***	**6.2903***	**23.67****	**6.6595****
400K	josh-base	29.58	7.3033	18.86	5.8818	22.25	6.2949	23.24	6.7402
	lts	**29.83**	**7.4000****	**19.49****	**5.9374****	**22.92****	**6.3727****	**25.16****	**6.9005****

(Mylonakis and Sima'an, 2011)

- カテゴリーの学習+前のレベルを記憶することにより大幅な向上

まとめ

- 二言語構文解析を高速化することにより、複雑なモデルを学習可能
- 単語からフレーズ
- カテゴリーの詳細化
- ITGのルールの詳細化

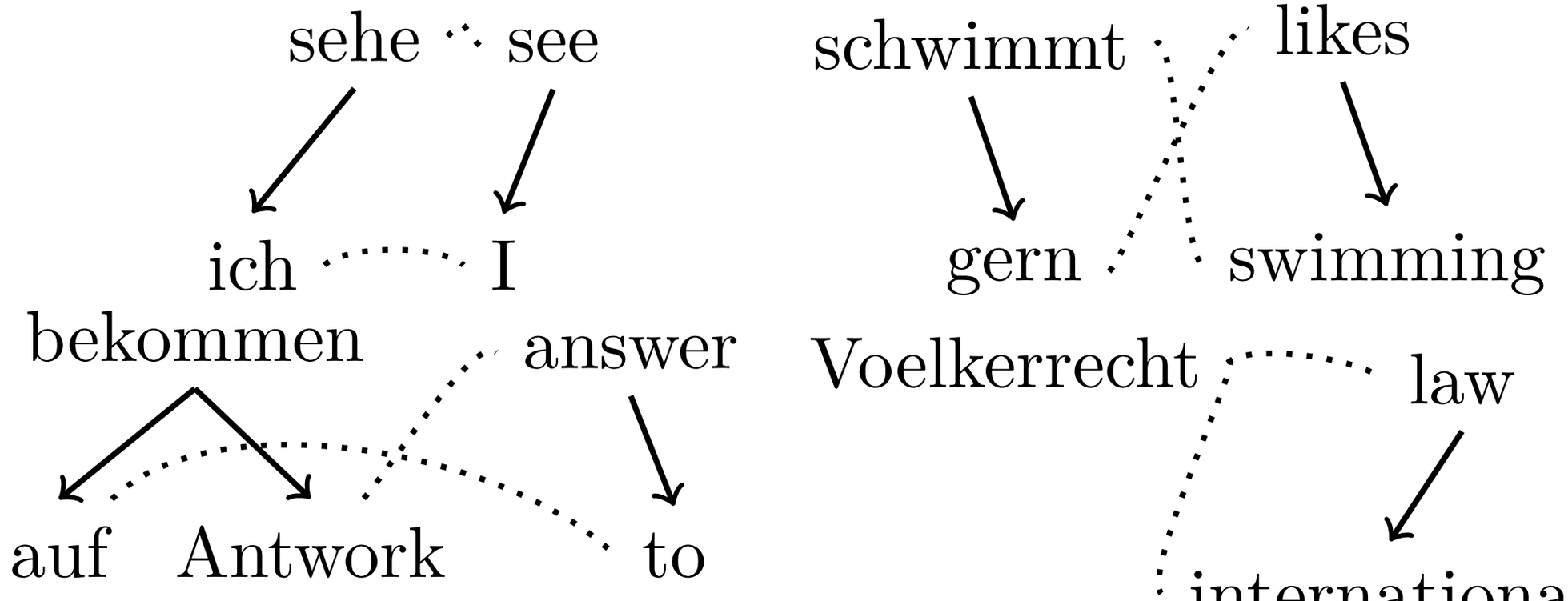
内容

- 木構造に基づく機械翻訳
  - 背景: CFG, hypergraph, deductive system
  - 同期文脈自由文法 (synchronous-CFG)
  - 同期文法: {string,tree}-to-{string,tree}
  - 二言語の構文解析(biparsing)
  - 同期から非同期
- 最適化

Asynchronous Models

- 同期的モデルでは、変数の一対一のマッピングを仮定
  - 現実の翻訳はそんなもんでない(Hwa et al., 2002; Fox 2002)
- quasi-synchronous models: 木構造の中のいずれかのノードが同期(Smith and Eisner, 2006)
- non-synchronized models: 不連続なフレーズペア (Galley and Manning, 2010)

Quasi-Synchronous Models

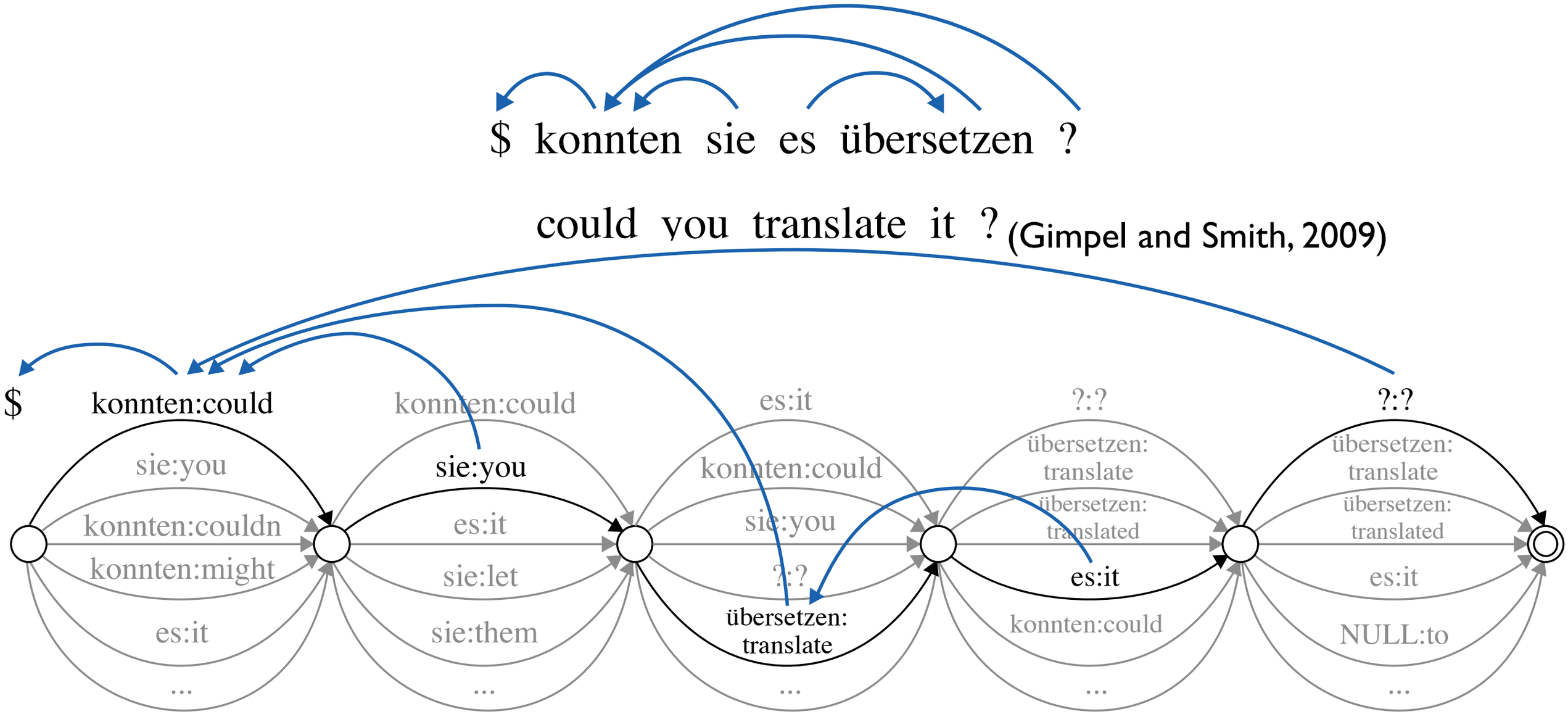


$$\langle \mathbf{e}^*, \tau_{\mathbf{e}}^*, \mathbf{a}^* \rangle = \underset{\langle \mathbf{e}, \tau_{\mathbf{e}}, \mathbf{a} \rangle}{\operatorname{argmax}} \frac{\exp\left(\mathbf{w}^\top \cdot \mathbf{h}(\mathbf{f}, \tau_{\mathbf{f}}, \mathbf{e}, \tau_{\mathbf{e}}, \mathbf{a})\right)}{\sum_{\mathbf{e}', \tau_{\mathbf{e}}', \mathbf{a}'} \exp\left(\mathbf{w}^\top \cdot \mathbf{h}(\mathbf{f}, \tau_{\mathbf{f}}, \mathbf{e}', \tau_{\mathbf{e}}', \mathbf{a}')\right)}$$

(Smith and Eisner, 2006)

- τ: 木構造(依存構造)、a:アライメント
- 原言語、目的言語の関係を無視してアライメント
  - 同期した素性を使用(同期した親子関係など)

Quasi-Synchronous Models



(Gimpel and Smith, 2009)

- 長さ固定でlatticeを作成: 単語単位で同期
- lattice上の構文解析により翻訳+依存構造を生成

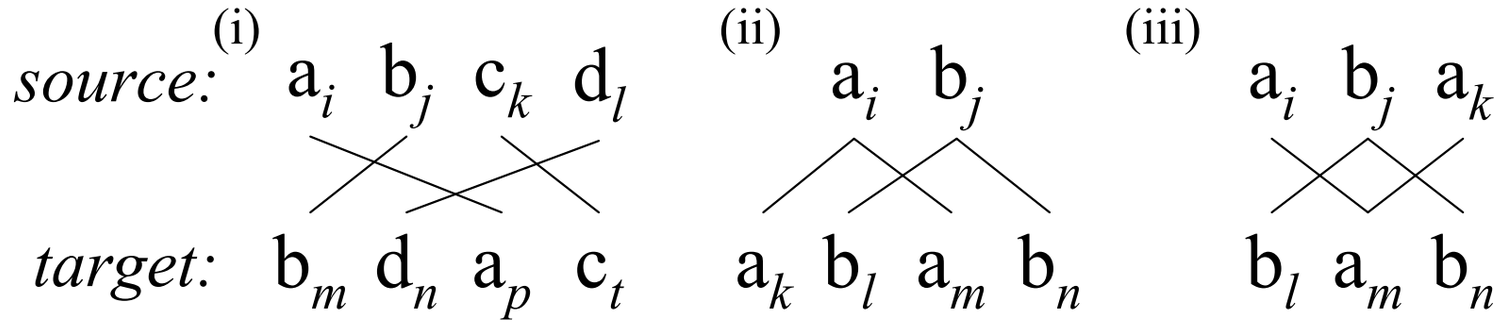
Quasi-Synchronous Models

	MT03 (tune)	MT02	MT05	MT06	Average
Moses	33.84	33.35	31.81	28.82	31.33
QPDG (TT)	34.63 (+0.79)	34.10 (+0.75)	32.15 (+0.34)	29.33 (+0.51)	31.86 (+0.53)
QPDG (TT+S2T+T2T)	34.98 (+1.14)	34.26 (+0.91)	32.34 (+0.53)	29.35 (+0.53)	31.98 (+0.65)

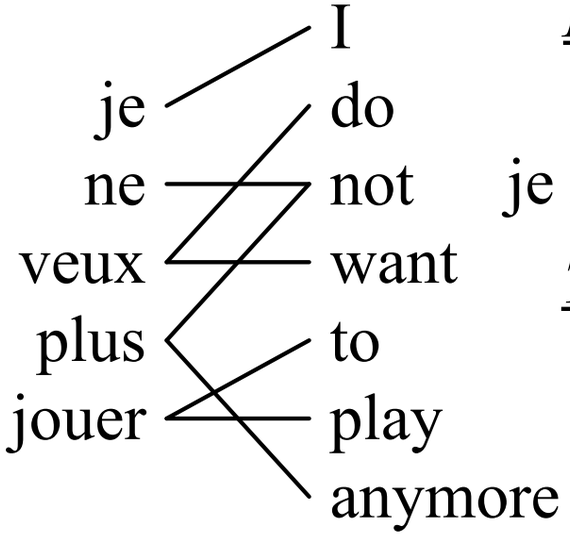
(Gimpel and Smith, 2011)

- フレーズでlatticeを作成(Gimpel and Smith, 2011)

非同期ルール



(Galley and Manning, 2010)



Hiero:

ne veux plus X ⟷ do not want X anymore

je ne veux plus X ⟷ I do not want X anymore

This work:

veux ⟷ do ... want

ne ... plus ⟷ not ... anymore

je ne ... plus ⟷ I ... not ... anymore

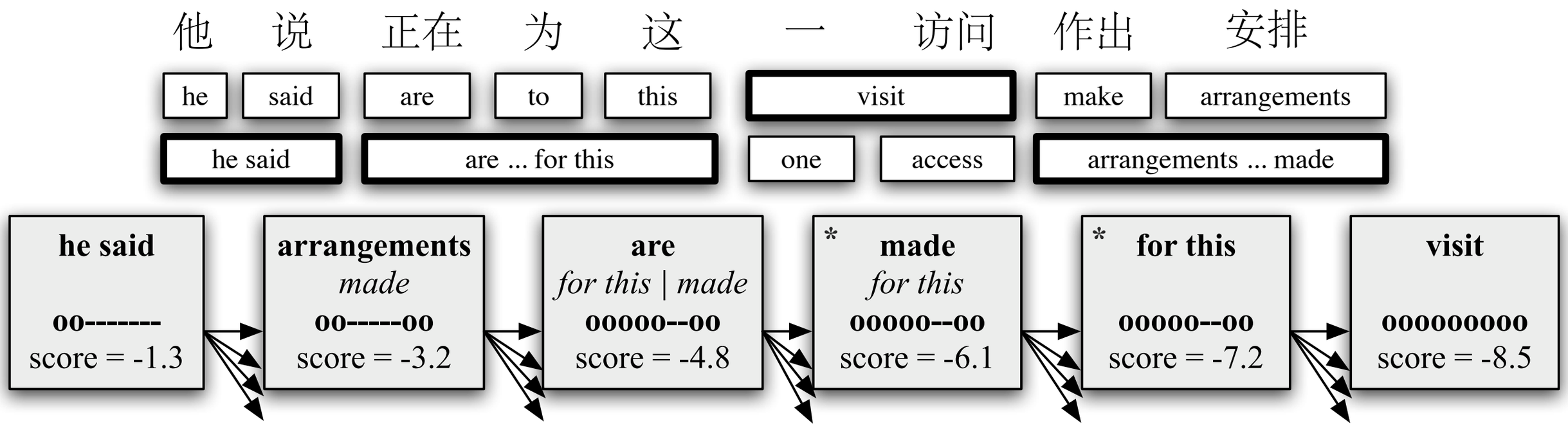
veux ... jouer ⟷ do ... want to play

ne veux plus ⟷ do not want ... anymore

je ne veux plus ⟷ I do not want ... anymore

- SCFGでは直接表現できないアライメント
- 同期していなくてもとりあえず抽出

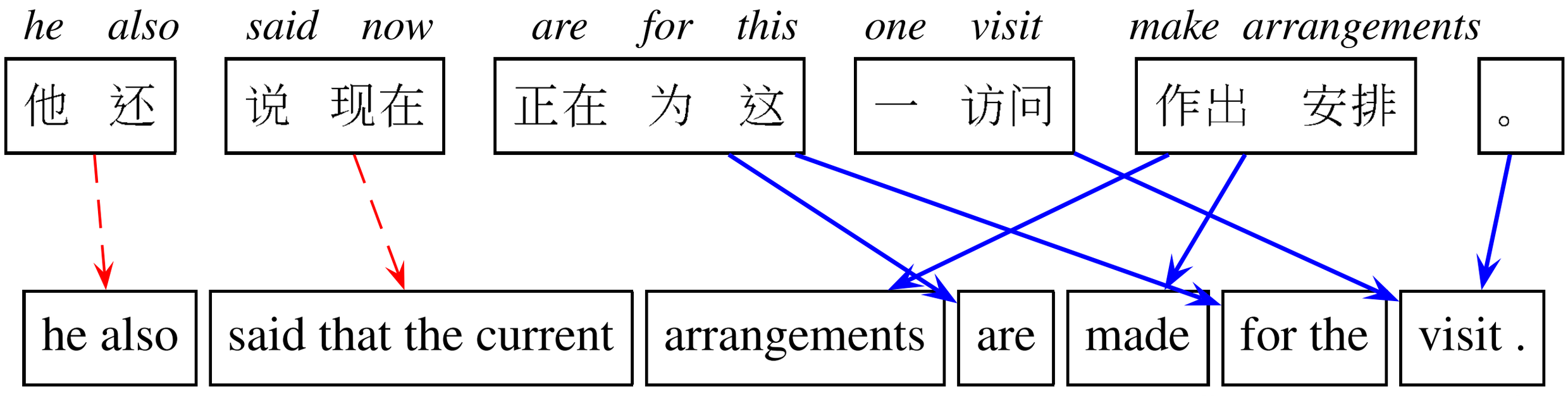
非階層的なデコーディング



(Galley and Manning, 2010)

- 基本的に句に基づくモデルのデコードと同じ
- gapがあったときに記憶:新しい句を導入するか、gapで取り残された句を結合

非同期的なモデル



(Galley and Manning, 2010)

System	Features	MT06 (tune)	MT03	MT05
Moses	Moses	34.23	33.72	32.51
Phrasal	Moses	34.25	33.72	32.49
Phrasal	Default	35.02	34.98	33.21

(Cer et al., 2010)

- Mosesと同等、さらに階層的な素性を導入することで向上

まとめ

- 非同期なモデルに表現力の向上
  - quasi-synchronous models
  - non-synchronized models
- 今後の発展に期待

内容

- 木構造に基づく機械翻訳
  - 背景: CFG, hypergraph, deductive system
  - 同期文脈自由文法 (synchronous-CFG)
    - 同期文法: {string,tree}-to-{string,tree}
    - 二言語の構文解析(biparsing)
    - 同期から非同期
- 最適化

他にも...

- 同期木接合文法(Tree Adjoining Grammars)
  - 置換以外に、挿入を許す文法(DeNeefe and Knight, 2009; Liu et al., 2011)
- 依存構造解析に基づく機械翻訳(Alshawi et al., 2000; Ding and Palmer, 2005; Quirk et al., 2005)
  - 基本的に、STAG(あるいは、STSG)と同じ
- (重み付き)有限状態木トランスデューサ(Finite State Tree Transducer) (Knight and Graehl, 2005; Graehl et al., 2008)
  - 正規文法におけるFSTのように、正規木文法におけるFSTT(文字列へと投影した場合、文脈自由文法)

内容

- 木構造に基づく機械翻訳
  - 背景: CFG, hypergraph, deductive system
  - 同期文脈自由文法 (synchronous-CFG)
    - 同期文法: {string,tree}-to-{string,tree}
    - 二言語の構文解析(biparsing)
    - 同期から非同期
- 最適化

Tuning

$$\begin{aligned}\hat{\mathbf{e}} &= \underset{\mathbf{e}}{\operatorname{argmax}} \frac{\exp\left(\mathbf{w}^\top \cdot \mathbf{h}(\mathbf{e}, \phi, \mathbf{f})\right)}{\sum_{\mathbf{e}', \phi'} \exp\left(\mathbf{w}^\top \cdot \mathbf{h}(\mathbf{e}', \phi', \mathbf{f})\right)} \\ &= \underset{\mathbf{e}}{\operatorname{argmax}}\, \mathbf{w}^\top \cdot \mathbf{h}(\mathbf{e}, \phi, \mathbf{f})\end{aligned}$$

- エラー最小化 (Och, 2003)
- MaxEnt (Och and Ney, 2002)
- マージン最大化 (Watanabe et al., 2007; Chiang et al., 2008; Hopkins and May, 2011)
- リスク最小化(Smith and Eisner, 2006; Li and Eisner 2009)
- 期待BLEU最大化(Pauls et al., 2009; Rosti et al., 2010; Rosti et al., 2011)

Maximum Entropy

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{\lambda}{2}\|\mathbf{w}\|^2 - \sum_{s=1}^{S} \log \frac{\sum_{\mathbf{e}^* \in \mathsf{ORACLE}(\mathbf{f}_s)} \exp\left(\mathbf{w}^\top \cdot \mathbf{h}(\mathbf{e}^*, \mathbf{f}_s)\right)}{\sum_{\mathbf{e}' \in \mathsf{GEN}(\mathbf{f}_s)} \exp\left(\mathbf{w}^\top \cdot \mathbf{h}(\mathbf{e}', \mathbf{f}_s)\right)}$$

- negative conditional log-likelihoodを最小化(Och and Ney, 2002)
- GENから口ス最小なORACLEの集合を求める
- 標準的な最適化アルゴリズム: LBFGS、SGD
- 様々な素性を導入可能

Why Not MaxEnt?

error criterion used in training	mWER [%]	mPER [%]	BLEU [%]	NIST	# words
confidence intervals	+/- 2.7	+/- 1.9	+/- 0.8	+/- 0.12	-
MMI	**68.0**	*51.0*	11.3	5.76	21933
mWER	*68.3*	*50.2*	13.5	6.28	22914
smoothed-mWER	*68.2*	*50.2*	13.2	6.27	22902
mPER	*70.2*	*49.8*	15.2	*6.71*	24399
smoothed-mPER	*70.0*	**49.7**	15.2	*6.69*	24198
BLEU	76.1	53.2	**17.2**	6.66	28002
NIST	73.3	*51.5*	*16.4*	**6.80**	26602

(Och, 2003)

- BLEUによるORACLEを選択していない(Och and Ney, 2002): 逆にこれが難しい(コーパス単位+BP問題)
- summation問題: n-best結合による近似をしているが、本当に正しい和集合ではない(と思う)

全ての導出

System	Test (BLEU)
Discriminative max-derivation	25.78
Hiero (p_d, gr, rc, wc)	26.48
Discriminative max-translation	27.72
Hiero ($p_d, p_r, p_d^{lex}, p_r^{lex}, gr, rc, wc$)	28.14
Hiero ($p_d, p_r, p_d^{lex}, p_r^{lex}, gr, rc, wc, lm$)	32.00

(Blunsom et al., 2008)

- Blunsom et al. (2008):森から計算、正解の導出を一つにせず、複数の導出に対して最適化
- ただし、正解は翻訳が参照訳とマッチしたもののみ

Large Margin

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_{s=1}^{S} \sum_{\mathbf{e}_s^*} \sum_{\mathbf{e}'_s} \xi_{s,\mathbf{e}_s^*,\mathbf{e}'_s}$$

$$\mathbf{w}^\top \cdot \mathbf{h}(\mathbf{e}_s^*, \mathbf{f}_s) - \mathbf{w}^\top \cdot \mathbf{h}(\mathbf{e}'_s, \mathbf{f}_s) \geq \ell(\mathbf{e}'_s, \mathbf{e}_s^*) - \xi_{s,\mathbf{e}_s^*,\mathbf{e}'_s}$$

$$\mathbf{e}_s^* \in \mathsf{ORACLE}(\mathbf{f}_s)$$

$$\mathbf{e}'_s \in \mathsf{GEN}(\mathbf{f}_s)$$

- 構造を出力とする学習(Structured Output learning)
- 元々、e' やe*を列挙することはほぼ不可能
- n-best 結合による近似、あるいは、オンライン学習による近似

Online Learning

Require: $\{(\mathbf{f}_s, \mathbf{e}_s)\}_{s=1}^{S}$

$\mathbf{w}^1 = \{0\}$
$t = 1$
for $1...N$ **do**
    $s \sim \text{random}(1, S)$
    $\hat{\mathbf{e}} \in \text{GEN}(\mathbf{f}_s, \mathbf{w}^{t-1})$
    if $l(\hat{\mathbf{e}}, \mathbf{e}_s) \geq 0$ **then**
        $\mathbf{w}^{t+1} = \mathbf{w}^t + \mathbf{h}(\mathbf{e}_s, \mathbf{f}_s) - \mathbf{h}(\hat{\mathbf{e}}, \mathbf{f}_s)$
        $t = t + 1$
    end if
end for
return $\mathbf{w}^t$ or $\frac{1}{N}\sum_{i=1}^{N} \mathbf{w}^j$

- Averaged perceptron (Liang et al., 2006)
- オンラインで学習:毎回デコード+更新

Online Large Margin

$$\hat{\mathbf{w}} = \underset{\mathbf{w}'}{\operatorname{argmin}} \frac{\lambda}{2}\|\mathbf{w}' - \mathbf{w}\|^2 + \max\left(\ell_s - \mathbf{w}'^{\top} \cdot \Delta\mathbf{h}_s\right)$$

$$\begin{aligned} \hat{\mathbf{e}}_s &= \underset{e}{\operatorname{argmax}}\, \mathbf{w}^{\top} \cdot \mathbf{h}(\mathbf{e}, \mathbf{f}_s) \\ \ell_s &= \ell(\hat{\mathbf{e}}_s) - \ell(\mathbf{e}^*_s) \\ \Delta\mathbf{h}_s &= \mathbf{h}(\hat{\mathbf{e}}_s, \mathbf{f}_s) - \mathbf{h}(\mathbf{e}^*, \mathbf{f}_s) \end{aligned}$$

- MIRA(Crammer et al., 2006)による更新
  (Watanabe et al., 2007; Chiang et al., 2008)
- では、どうやってBLEUを計算するか?

BLEUの近似

$$\mathsf{GEN}(\mathbf{f}_s, \mathbf{w})$$

$$\mathbf{e}_1^*, \cdots, \begin{pmatrix} \mathbf{e}_s^1 \\ \vdots \\ \mathbf{e}_s^i \\ \vdots \\ \mathbf{e}_s^n \end{pmatrix}, \cdots, \mathbf{e}_S^*$$

- 今までの各文に対するBLEUの統計量を保存(1-bestあるいはoracle)
- 新しいn-bestによる更新 (Watanabe et al., 2007)

減衰によるBLEUの近似

$$\begin{aligned}\mathbf{b} &\leftarrow 0.9 \times (\mathbf{b} + \mathbf{c}(\mathbf{e})) \\ l &\leftarrow 0.9 \times (l + |\mathbf{f}|)\end{aligned}$$

$$B(\mathbf{e}) = (l + |\mathbf{f}|) \times \mathsf{Bleu}(\mathbf{b} + \mathbf{c}(\mathbf{e}))$$

$$\begin{aligned}\hat{\mathbf{e}}_s &= \operatorname*{argmax}_{\mathbf{e}} -B(\mathbf{e}) + \hat{\mathbf{w}} \cdot \mathbf{h}(\mathbf{e}, \mathbf{f}_s) \\ \mathbf{e}^*_s &= \operatorname*{argmax}_{\mathbf{e}} +B(\mathbf{e}) + \hat{\mathbf{w}} \cdot \mathbf{h}(\mathbf{e}, \mathbf{f}_s)\end{aligned}$$

- sentence-BLEUに対して、今までのBLUEの履歴(×0.9)を加える(Chiang et al., 2008)
- エラーを含めたargmax

Results

System	Training	Features	#	Tune	Test
Hiero	MERT	baseline	11	35.4	36.1
	MIRA	syntax, distortion	56	35.9	36.9*
		syntax, distortion, discount	61	36.6	37.3**
		all source-side, discount	10990	38.4	37.6**
Syntax	MERT	baseline	25	38.6	39.5
	MIRA	baseline	25	38.5	39.8*
		overlap	132	38.7	39.9*
		node count	136	38.7	40.0**
		all target-side, discount	283	39.6	40.6**

(Chiang et al., 2009)

- + feature engineeringにより、MERTのベースラインより統計的に優位な向上

Ranking Approach

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_{s=1}^{S}\sum_{\mathbf{e}''_s}\sum_{\mathbf{e}'_s} \xi_{s,\mathbf{e}''_s,\mathbf{e}'_s}$$

$$-\log\left(1+\exp(-\mathbf{w}^\top \cdot \Delta\mathbf{h}_{\mathbf{e}''_s,\mathbf{e}'_s})\right) \geq -\xi_{s,\mathbf{e}''_s,\mathbf{e}'_s}$$

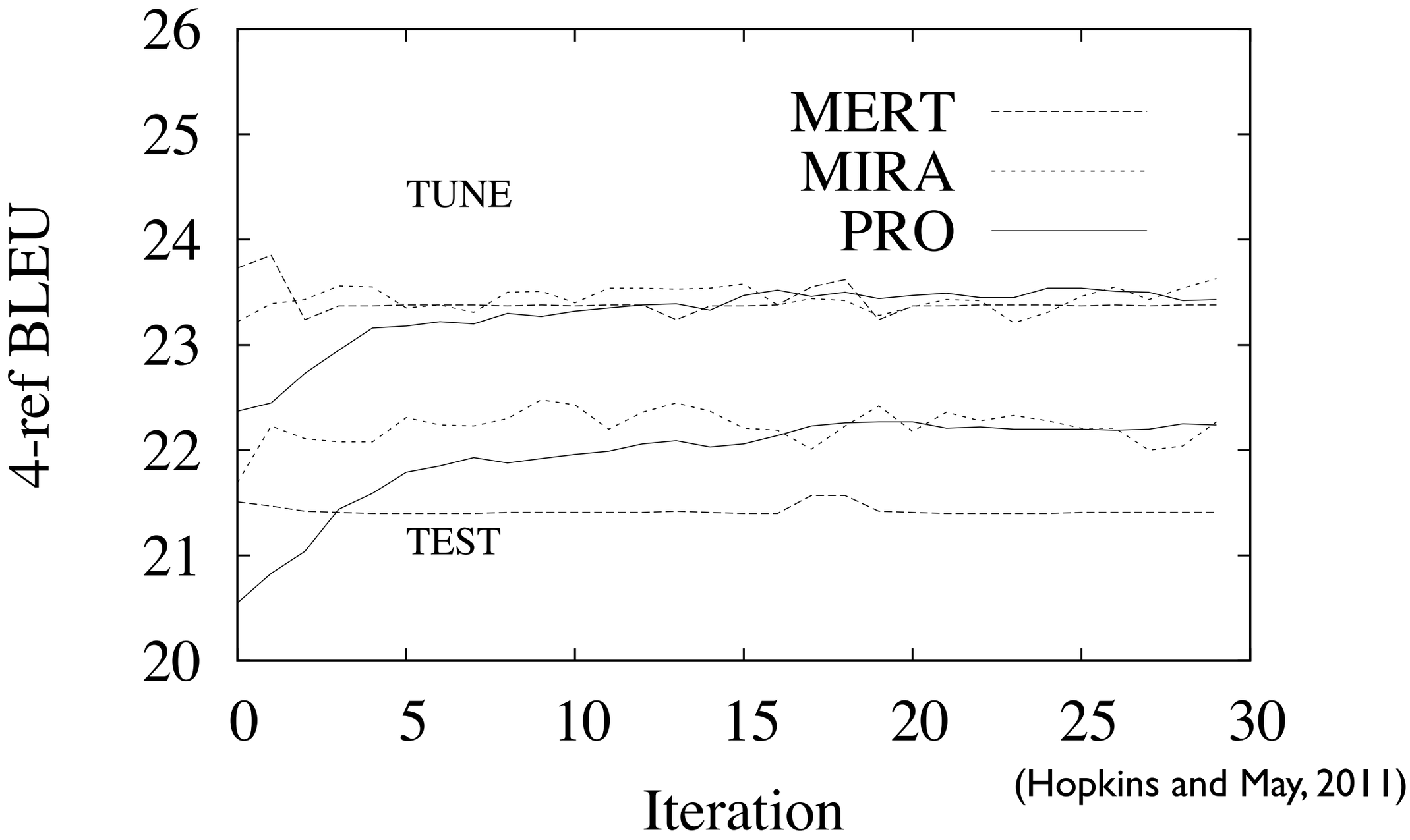
$$\mathbf{e}''_s, \mathbf{e}'_s \in \mathsf{GEN}(\mathbf{f}_s)$$

$$\ell(\mathbf{e}'_s, \mathbf{e}''_s) > 0$$

$$\Delta\mathbf{h}_{\mathbf{e}''_s,\mathbf{e}'_s} = \mathbf{h}(\mathbf{e}''_s, \mathbf{f}_s) - \mathbf{h}(\mathbf{e}'_s, \mathbf{f}_s)$$

- n-best結合による近似(Hopkins and May, 2011)
- ペア単位の比較(sentence-BLEU)+サンプリング
- 普通の二値分類器(ここでは、logistic-loss) + 前のパラメータとの線形結合

Results



(Hopkins and May, 2011)

- MERTやMIRAとほぼ同様の結果 (でもMosesの実装では...)

リスク最小化

$$\min_{\gamma,\mathbf{w}} \mathbb{E}_{p_{\gamma,\mathbf{w}}}[\ell(\mathbf{e}_s)] - T \cdot H(p_{\gamma,\mathbf{w}})$$

$$\mathbb{E}_{p_{\gamma,\mathbf{w}}}[\ell(\mathbf{e}_s)] = \sum_s \sum_i \ell(\mathbf{e}_s^i) p_{\gamma,\mathbf{w}}(\mathbf{e}_s^i|\mathbf{f}_s)$$

$$p_{\gamma,\mathbf{w}}(\mathbf{e}_s^i|\mathbf{f}_s) = \frac{\exp\left(\gamma \mathbf{w}^\top \cdot \mathbf{h}(\mathbf{e}_s^i, \mathbf{f}_s)\right)}{\sum_{i'} \exp\left(\gamma \mathbf{w}^\top \cdot \mathbf{h}(\mathbf{e}_s^{i'}, \mathbf{f}_s)\right)}$$

- γによるスムージング、エントロピーH(.)による正則化 、温度Tによる冷却(Smith and Eisner, 2006)
- ロスの計算?: BLEUはnon-linear

ティラー展開による近似

$$\log \text{Bleu} \approx \sum_{n=1}^{4} \frac{1}{4} \log \frac{c_n}{c_0} + \min\left(1 - \frac{r}{c_0}, 0\right)$$

$$\log \text{Bleu}' - \log \text{Bleu} \approx \sum_{n=0}^{4} (c'_n - c_n) \left. \frac{\partial \log \text{Bleu}'}{\partial c'_n} \right|_{c'_n = c_n}$$

$$= -\frac{c'_0 - c_0}{c_0} + \frac{1}{4} \sum_{n=1}^{4} \frac{c'_n - c_n}{c_n}$$

- ngramのカウント(c_n)の更新(c'_n)によるBleuへの貢献を近似(Tromble et al., 2008)

- Smith and Eisner (2006)ではBleuそのものを近似

Results

Training scheme	dev	test
MERT (Nbest, small)	42.6	47.7
MR (Nbest, small)	40.8	47.7
MR+DA (Nbest, small)	41.6	47.8
MR (hypergraph, small)	41.3	48.4
MR+DA (hypergraph, small)	41.9	48.3
MR (hypergraph, large)	42.3	**48.7**

(Li and Eisner, 2009)

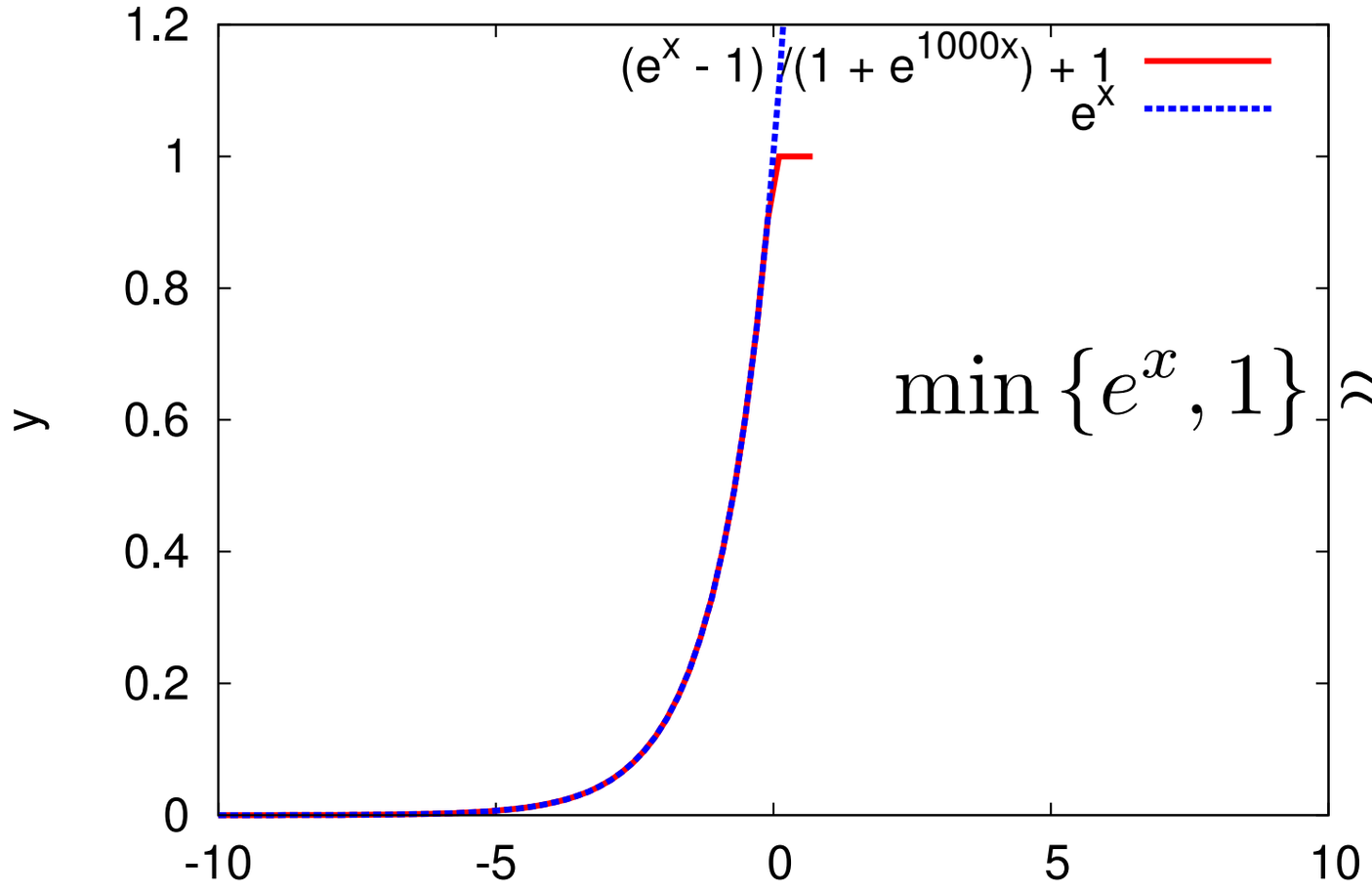
- MERTとほぼ同様な結果
- hypergraphで計算することにより向上

期待BLEU最大化

$$\prod_{n=1}^{4} \left(\frac{\min\left\{ \sum_s \sum_i \sum_{g_n \in \mathbf{e}_s^i} \mathbb{E}_{\gamma,\mathbf{w}}[c(g_n)], c^*(g_n) \right\}}{\sum_s \sum_i \sum_{g_n \in \mathbf{e}_s^i} \mathbb{E}_{\gamma,\mathbf{w}}[c(g_n)]} \right)^{\frac{1}{4}} \times \min\left\{ \exp\left(1 - \frac{\sum_s r_s}{\sum_s \sum_i \sum_{g_1 \in \mathbf{e}_s^i} \mathbb{E}_{\gamma,\mathbf{w}}[c(g_1)]}\right), 1 \right\}$$

- 期待BLEUを直接最大化 (Pauls et al., 2009; Rosti et al., 2010; Rosti et al., 2011)
  - ngram g_nの期待値E[.]を元に計算
  - Smith and Eisner (2006)のBLEU近似に近い

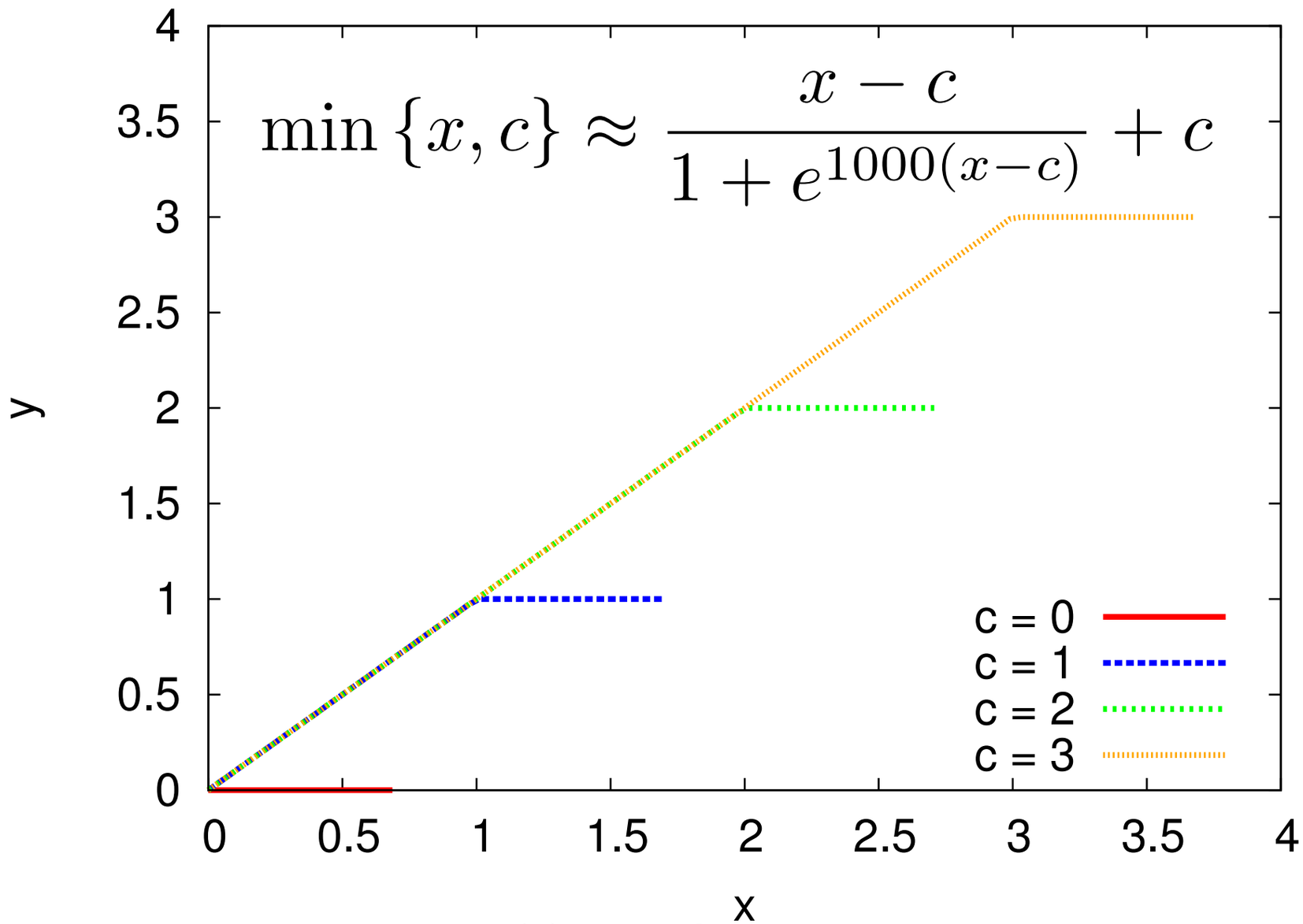
BPはどうする?



$$\min\{e^x, 1\} \approx \frac{e^x - 1}{1 + e^{1000x}} + 1$$

- matlabで色々試してたどり着いたらしい(Rosti et al., 2010; Rosti et al., 2011)
- BPを無視する(Tromble et al., 2008)
- minを無視する(Pauls et al., 2009)

clipはどうする?



- lattice/forestの計算で使用(Rosti et al., 2011)
- 注意: Rosti et al. (2011) の式(15)にバグ

Results

test	cz-en		de-en		es-en		fr-en	
System	TER	BLEU	TER	BLEU	TER	BLEU	TER	BLEU
worst	65.35	17.69	69.03	15.83	61.22	19.79	62.36	21.36
best	52.21	29.54	58.00	24.16	50.15	30.14	50.15	30.32
latBLEU	52.80	29.89	55.87	26.22	48.29	33.91	48.51	32.93
nbExpBLEU	52.97	29.93	55.77	26.52	48.39	33.86	48.25	32.94
latExpBLEU	52.68	29.99	55.74	26.62	48.30	34.10	48.17	32.91

- システムコンビネーションおよびlattice での期待値の計算(Rosti et al., 2011)
- 期待semiringによる効率のよい計算

まとめ

- MERTが標準: 目的関数などを工夫することで他の最適化アルゴリズムを適用可能
- 根本的な問題
  - BLEU(あるいはそれ以外の尺度)の近似
  - n-best結合あるいはオンラインによる近似

内容

- 木構造に基づく機械翻訳
  - 背景: CFG, hypergraph, deductive system
  - 同期文脈自由文法 (synchronous-CFG)
    - 同期文法: {string,tree}-to-{string,tree}
    - 二言語の構文解析(biparsing)
    - 同期から非同期
- 最適化

最後に

他にも...

- pre-reordering: デコード前に並び替え: 統語論的構造(Collins et al., 2005; Isozaki et al., 2010)、木構造(Tromble and Eisner, 2009; DeNero and Uszkoreit, 2011)
- 構文解析木による言語モデル(Shen et al., 2008; Mi and Liu, 2010; Shwartz et al., 2011)
- モデルコンビネーション(Lieu et al., 2009; DeNero et al., 2010)

- 機械翻訳は「応用分野」ではなく「基礎研究」
- SMTにより、問題分割が容易
- 全体を把握した上で要素技術の研究開発

参考文献

- Hiyan Alshawi, Srinivas Bangalore, and Shona Douglas. 2000. Learning dependency translation models as collections of finite state head transducers. *Computational Linguistics*, 26(1):45-60.
- Abhishek Arun, Chris Dyer, Barry Haddow, Phil Blunsom, Adam Lopez, and Philipp Koehn. 2009. Monte carlo inference and maximization for phrase-based translation. In *Proc. of CoNLL-2009*, pages 102-110, Boulder, Colorado, June.
- Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proc. of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65-72, Ann Arbor, Michigan, June.
- Taylor Berg-Kirkpatrick, Alexandre Bouchard-Côté, John DeNero, and Dan Klein. 2010. Painless unsupervised learning with features. In *Proc. of NAACL-HLT 2010*, pages 582-590, Los Angeles, California, June.
- Phil Blunsom, Trevor Cohn, and Miles Osborne. 2008. A discriminative latent variable model for statistical machine translation. In *Proc. of ACL-08: HLT*, pages 200-208, Columbus, Ohio, June.
- Phil Blunsom, Trevor Cohn, Chris Dyer, and Miles Osborne. 2009. A gibbs sampler for phrasal synchronous grammar induction. In *Proc. of ACL/IJCNLP 2009*, pages 782-790, Suntec, Singapore, August.
- Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proc. of EMNLP-CoNLL 2007*, pages 858--867.
- Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79-85.

- Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263-311.
- David Burkett, John Blitzer, and Dan Klein. 2010. Joint parsing and alignment with weakly synchronized grammars. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127-135, Los Angeles, California, June.
- Daniel Cer, Dan Jurafsky, and Christopher D. Manning. 2008. Regularization and search for minimum error rate training. In *Proc. of SMT 2008*, pages 26-34, Columbus, Ohio, June.
- Daniel Cer, Michel Galley, Daniel Jurafsky, and Christopher D. Manning. 2010. Phrasal: A statistical machine translation toolkit for exploring new model features. In *Proc. of of the NAACL HLT 2010 Demonstration Session*, pages 9-12, Los Angeles, California, June.
- Colin Cherry and Dekang Lin. 2006. Soft syntactic constraints for word alignment through discriminative training. In *Proc. of the COLING/ACL 2006*, pages 105-112, Sydney, Australia, July.
- David Chiang, Steve DeNeefe, Yee Seng Chan, and Hwee Tou Ng. 2008a. Decomposability of translation metrics for improved evaluation and efficient algorithms. In *Proc. of EMNLP 2008*, pages 610-619, Honolulu, Hawaii, October.
- David Chiang, Yuval Marton, and Philip Resnik. 2008b. Online large-margin training of syntactic and structural translation features. In *Proc. of EMNLP 2008*, pages 224-233, Honolulu, Hawaii, October.
- David Chiang, Kevin Knight, and Wei Wang. 2009. 11,001 new features for statistical machine translation. In *Proc. of NAACL-HLT 2009*, pages 218-226, Boulder, Colorado, June.

- David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2): 201-228.
- Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proc. of ACL 2011*, pages 176-181, Portland, Oregon, USA, June.
- Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proc. of ACL'05*, pages 531-540, Ann Arbor, Michigan, June.
- Steve DeNeefe and Kevin Knight. 2009. Synchronous tree adjoining machine translation. In *Proc. of EMNLP 2009*, pages 727-736, Singapore, August.
- John DeNero and Dan Klein. 2007. Tailoring word alignments to syntactic machine translation. In *Proc. of ACL 2007*, pages 17-24, Prague, Czech Republic, June.
- John DeNero and Jakob Uszkoreit. 2011. Inducing sentence structure from parallel corpora for reordering. In *Proc. of EMNLP 2011*, pages 193-203, Edinburgh, Scotland, UK., July.
- John DeNero, Alexandre Bouchard-Côté, and Dan Klein. 2008. Sampling alignment structure under a Bayesian translation model. In *Proc. of EMNLP 2008*, pages 314-323, Honolulu, Hawaii, October.
- John DeNero, Shankar Kumar, Ciprian Chelba, and Franz Och. 2010. Model combination for machine translation. In *Proc. of NAACL-HT 2010*, pages 975-983, Los Angeles, California, June.
- Yuan Ding and Martha Palmer. 2005. Machine translation using probabilistic synchronous dependency insertion grammars. In *Proc. of ACL '05*, pages 541-548, Morristown, NJ, USA.
- George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *In Proc. ARPA Workshop on Human Language Technology*.

- Chris Dyer and Philip Resnik. 2010. Context-free reordering, finite-state translation. In *Proc. of NAACL-HLT 2010,* pages 858-866, Los Angeles, California, June.
- Chris Dyer, Jonathan H. Clark, Alon Lavie, and Noah A. Smith. 2011. Unsupervised word alignment with arbitrary features. In *Proc. of ACL-HLT 2011,* pages 409-419, Portland, Oregon, USA, June.
- Chris Dyer. 2010. Two monolingual parses are better than one (synchronous parse). In *Proc. of NAACL-HLT 2010,* pages 263-266, Los Angeles, California, June.
- Jason Eisner. 2003. Learning non-isomorphic tree mappings for machine translation. In *Proc. of ACL 2003,* pages 205-208, Sapporo, Japan, July.
- Heidi Fox. 2002. Phrasal cohesion and statistical machine translation. In *Proc. of EMNLP 2002,* pages 304-3111, July.
- Michel Galley and Christopher D. Manning. 2010. Accurate non-hierarchical phrase-based translation. In *Proc. of NAACL-HLT 2010,* pages 966-974, Los Angeles, California, June.
- Michel Galley and Chris Quirk. 2011. Optimal search for minimum error rate training. In *Proc. of EMNLP 2011,* pages 38-49, Edinburgh, Scotland, UK., July.
- Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proc. of HLT-NAACL 2004,* pages 273-280, Boston, Massachusetts, USA, May 2 - May 7.
- Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proc. of ACL/COLING 2006,* pages 961-968, Sydney, Australia, July.
- Kuzman Ganchev, João V. Graça, and Ben Taskar. 2008. Better alignments = better translations? In *Proceedings of ACL-08: HLT,* pages 986-993, Columbus, Ohio, June.

- Andrea Gesmundo and James Henderson. 2010. Faster Cube Pruning. In *Proc. of IWSLT 2010*, pages 267-274.
- Kevin Gimpel and Noah A. Smith. 2009. Feature-rich translation by quasi-synchronous lattice parsing. In *Proc. of EMNLP 2009*, pages 219-228, Singapore, August.
- Kevin Gimpel and Noah A. Smith. 2011. Quasi-synchronous phrase dependency grammars for machine translation. In *Proc. of EMNLP 2011*, pages 474-485, Edinburgh, Scotland, UK., July.
- Jonathan Graehl, Kevin Knight, and Jonathan May. 2008. Training tree transducers. *Computational Linguistics*, 34:391-427, September.
- Aria Haghighi, John Blitzer, John DeNero, and Dan Klein. 2009. Better word alignments with supervised itg models. In *Proc. of ACL/IJCNLP 2009*, pages 923-931, Suntec, Singapore, August.
- Katsuhiko Hayashi, Taro Watanabe, Hajime Tsukada, and Hideki Isozaki. 2009. Structural Support Vector Machines for Log-Linear Approach in Statistical Machine Translation. In *Proc. of IWSLT 2009*, pages 144-151, Tokyo, Japan.
- Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proc. of EMNLP 2011*, pages 1352-1362, Edinburgh, Scotland, UK., July.
- Liang Huang and David Chiang. 2005. Better k-best parsing. In *Proc. of IWPT'05*, pages 53-64, Vancouver, British Columbia, October.
- Liang Huang and David Chiang. 2007. Forest rescoring: Faster decoding with integrated language models. In *Proc. of ACL 2007*, pages 144-151, Prague, Czech Republic, June.
- Liang Huang and Haitao Mi. 2010. Efficient incremental decoding for tree-to-string translation. In *Proc. of EMNLP 2010*, pages 273-283, Cambridge, MA, October.

- Liang Huang, Kevin Knight, and Aravind Joshi. 2006. Statistical syntax-directed translation with extended domain of locality. In *In Proc. AMTA 2006*, pages 66-73.
- Liang Huang, Hao Zhang, Daniel Gildea, and Kevin Knight. 2009. Binarization of synchronous context-free grammars. *Computational Linguistics*, 35:559-595, December.
- Rebecca Hwa, Philip Resnik, Amy Weinberg, and Okan Kolak. 2002. Evaluating translational correspondence using annotation projection. In *Proc. of ACL 2002*, pages 392-399, Philadelphia, Pennsylvania, USA, July.
- Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010a. Automatic evaluation of translation quality for distant language pairs. In *Proc. of EMNLP 2010*, pages 944-952, Cambridge, MA, October.
- Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh. 2010b. Head finalization: A simple reordering rule for sov languages. In *Proc. of SMT-MetricsMATR 2010*, pages 244-251, Uppsala, Sweden, July.
- Dan Klein and Christopher D. Manning. 2001. Parsing and hypergraphs. In *Proc. of IWPT-2001*, pages 123-134.
- Kevin Knight and Jonathan Graehl. 2005. An overview of probabilistic tree transducers for natural language processing. In *Proc. of CICLing 2005*, pages 1-24.
- Kevin Knight. 1999. Decoding complexity in word-replacement translation models. *Computational Linguistics*, 25:607-615, December.
- Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical pharse-based translation. In *Proc. of HLT-NAACL 2003*, pages 48-54, Edmonton, May-June.
- Philipp Koehn. 2009. Statistical Machine Translation. Cambridge University Press.

- Shankar Kumar, Wolfgang Macherey, Chris Dyer, and Franz Och. 2009. Efficient minimum error rate training and minimum bayes-risk decoding for translation hypergraphs and lattices. In *Proc. of ACL/IJCNLP 2009,* pages 163-171, Suntec, Singapore, August.
- Zhifei Li and Jason Eisner. 2009. First- and second-order expectation semirings with applications to minimum-risk training on translation forests. In *Proc. of EMNLP 2009,* pages 40-51, Singapore, August.
- Zhifei Li, Jason Eisner, and Sanjeev Khudanpur. 2009. Variational decoding for statistical machine translation. In *Proc. of ACL-IJCNLP 2009,* pages 593-601, Suntec, Singapore, August.
- Percy Liang, Alexandre Bouchard-Côté, Dan Klein, and Ben Taskar. 2006a. An end-to-end discriminative approach to machine translation. In *Proc. of ACL/COLING 2006,* pages 761-768, Sydney, Australia, July.
- Percy Liang, Ben Taskar, and Dan Klein. 2006b. Alignment by agreement. In *Proc. of NAACL/HLT 2006,* pages 104-111, New York City, USA, June.
- Hui Lin and Jeff Bilmes. 2011. Word alignment via submodular maximization over matroids. In *Proc. of ACL-HLT 2011,* pages 170-175, Portland, Oregon, USA, June.
- Yang Liu, Qun Liu, and Yajuan Lü. 2011. Adjoining tree-to-string translation. In *Proc. of ACL-HLT 2011,* pages 1278-1287, Portland, Oregon, USA, June.
- Wolfgang Macherey, Franz Och, Ignacio Thayer, and Jakob Uszkoreit. 2008. Lattice-based minimum error rate training for statistical machine translation. In *Proc. of EMNLP 2008,* pages 725-734, Honolulu, Hawaii, October.
- Daniel Marcu and William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proc. of EMNLP-2002,* Philadelphia, PA, July.
- Haitao Mi and Liang Huang. 2008. Forest-based translation rule extraction. In *Proc. of EMNLP 2008,* pages 206-214, Honolulu, Hawaii, October.

- Haitao Mi and Qun Liu. 2010. Constituency to dependency translation with forests. In *Proc. of ACL 2010*, pages 1433-1442, Uppsala, Sweden, July.
- Haitao Mi, Liang Huang, and Qun Liu. 2008. Forest-based translation. In *Proc. of ACL-08: HLT*, pages 192-199, Columbus, Ohio, June.
- Robert C. Moore and Chris Quirk. 2008. Random restarts in minimum error rate training for statistical machine translation. In *Proc. of Coling 2008*, pages 585-592, Manchester, UK, August.
- Markos Mylonakis and Khalil Sima'an. 2011. Learning hierarchical translation structure with linguistic annotations. In *Proc. of ACL-HLT 2011*, pages 642-652, Portland, Oregon, USA, June.
- Graham Neubig, Taro Watanabe, Eiichiro Sumita, Shinsuke Mori, and Tatsuya Kawahara. 2011. An unsupervised model for joint phrase alignment and extraction. In *Proc. of ACL-HLT 2011*, pages 632-641, Portland, Oregon, USA, June.
- Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proc. of ACL 2002*, pages 295-302, Philadelphia, Pennsylvania, USA, July.
- Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19-51, March.
- Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of ACL 2003*, pages 160-167, Sapporo, Japan, July.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL 2002*, pages 311-318, Philadelphia, Pennsylvania, USA, July.

- Adam Pauls, John Denero, and Dan Klein. 2009. Consensus training for consensus decoding in machine translation. In *Proc. of EMNLP 2009,* pages 1418-1427, Singapore, August.
- Adam Pauls, Dan Klein, David Chiang, and Kevin Knight. 2010. Unsupervised syntactic alignment with inversion transduction grammars. In *Proc. of NAACL-HLT 2010,* pages 118-126, Los Angeles, California, June.
- Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: syntactically informed phrasal smt. In *Proc. of ACL '05,* pages 271-279, Morristown, NJ, USA.
- Antti-Veikko Rosti, Bing Zhang, Spyros Matsoukas, and Richard Schwartz. 2010. Bbn system description for wmt10 system combination task. In *Proc. of SMT-MetricsMATR 2010,* pages 321-326, Uppsala, Sweden, July.
- Antti-Veikko Rosti, Bing Zhang, Spyros Matsoukas, and Richard Schwartz. 2011. Expected bleu training for graphs: Bbn system description for wmt11 system combination task. In *Proc. of SMT 2011,* pages 159-165, Edinburgh, Scotland, July.
- Markus Saers, Joakim Nivre, and Dekai Wu. 2009. Learning stochastic bracketing inversion transduction grammars with a cubic time biparsing algorithm. In *Proc. of IWPT'09,* pages 29-32, Paris, France, October.
- Lane Schwartz, Chris Callison-Burch, William Schuler, and Stephen Wu. 2011. Incremental syntactic language models for phrase-based translation. In *Proc. of ACL-HLT 2011,* pages 620-631, Portland, Oregon, USA, June.
- Libin Shen, Jinxi Xu, and Ralph Weischedel. 2008. A new string-to-dependency machine translation algorithm with a target dependency language model. In *Proceedings of ACL-08: HLT,* pages 577-585, Columbus, Ohio, June.

- Stuart M. Shieber, Yves Schabes, and Fernando C. N. Pereira. 1995. Principles and implementation of deductive parsing. *Journal of Logic Programming*, 24(1-2):3-36, July-August.
- David Smith and Jason Eisner. 2006a. Quasi-synchronous grammars: Alignment by soft projection of syntactic dependencies. In *Proc. of SMT 2006*, pages 23-30, New York City, June.
- David A. Smith and Jason Eisner. 2006b. Minimum risk annealing for training log-linear models. In *Proc. of the COLING/ACL 2006*, pages 787-794, Sydney, Australia, July.
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proc. of AMTA 2006*, pages 223-231.
- Colin Cherry and Dekang Lin. 2007. Inversion transduction grammar for joint phrasal translation modeling. In *Proc. of SSST 2007*, pages 17-24, Rochester, New York, April.
- Roy Tromble and Jason Eisner. 2009. Learning linear ordering problems for better translation. In *Proc. of EMNLP 2009*, pages 1007-1016, Singapore, August.
- Roy Tromble, Shankar Kumar, Franz Och, and Wolfgang Macherey. 2008. Lattice Minimum Bayes-Risk decoding for statistical machine translation. In *Proc. of EMNLP 2008*, pages 620-629, Honolulu, Hawaii, October.
- Taro Watanabe, Hajime Tsukada, and Hideki Isozaki. 2006. Left-to-Right Target Generation for Hierarchical Phrase-Based Translation. In *Proc. of ACL/COLING 2006*, pages 777-784, Sydney, Australia, July.
- Taro Watanabe, Jun Suzuki, Hajime Tsukada, and Hideki Isozaki. 2007. Online Large-Margin Training for Statistical Machine Translation. In *Proc. of EMNLP-CoNLL 2007*, pages 764-773, Prague, Czech Republic, June.

- Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics,* 23(3):377-403.
- Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In *Proc. of ACL 2001,* pages 523-530, Toulouse, France, July.
- Richard Zens and Hermann Ney. 2003. A comparative study on reordering constraints in statistical machine translation. In *Proc. of ACL 2003,* pages 144-151, Sapporo, Japan.
- Richard Zens, Hermann Ney, Taro Watanabe, and Eiichiro Sumita. 2004. Reordering Constraints for Phrase-Based Statistical Machine Translation. In *Proc. of COLING 2004,* pages 205-211, Geneva, Switzerland, Aug 23-Aug 27.
- Hao Zhang and Daniel Gildea. 2005. Stochastic lexicalized inversion transduction grammar for alignment. In *Proc. of ACL '05,* pages 475-482, Stroudsburg, PA, USA.
- Hao Zhang, Chris Quirk, Robert C. Moore, and Daniel Gildea. 2008. Bayesian learning of non-compositional phrases with synchronous parsing. In *Proc. of ACL-08: HLT,* pages 97-105, Columbus, Ohio, June.
- Hao Zhang, Licheng Fang, Peng Xu, and Xiaoyun Wu. 2011. Binarized forest to string translation. In *Proc. of ACL-HLT 2011,* pages 835-845, Portland, Oregon, USA, June.
- Shaojun Zhao and Daniel Gildea. 2010. A fast fertility hidden markov model for word alignment using MCMC. In *Proc. of EMNLP 2010,* pages 596-605, Cambridge, MA, October.
- Andreas Zollmann and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *Proc. of StatMT '06,* pages 138-141, Morristown, NJ, USA.