IBIS-2008
2008年10月29日

機械学習による自然言語処理の諸問題

松本裕治
奈良先端科学技術大学院大学
情報科学研究科
matsu@is.naist.jp

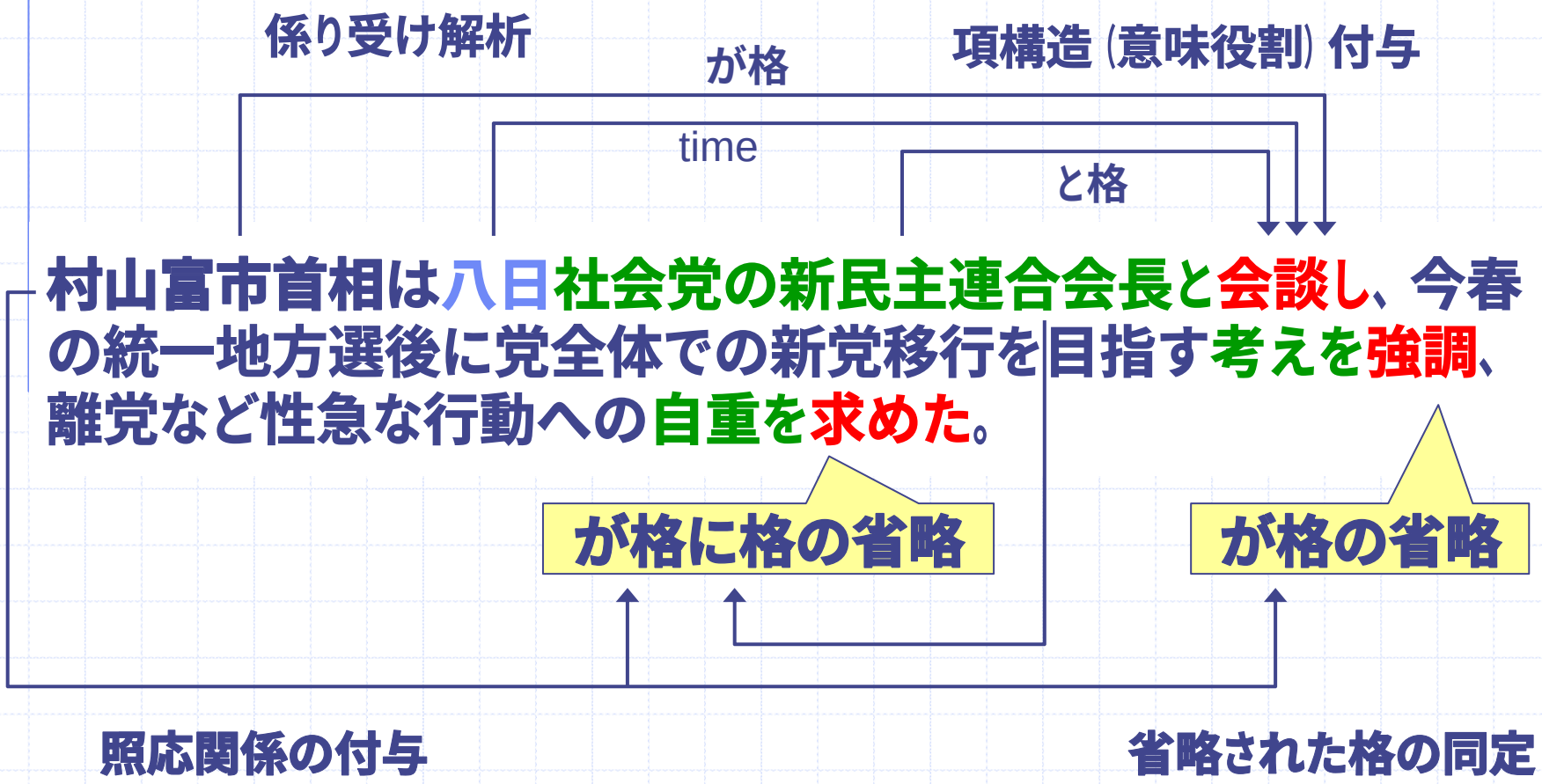
自然言語処理の主な研究項目

- 分類問題
  - 文書分類，語義曖昧性解消
- 構造解析
  - 形態素解析，統語解析，固有表現解析
- 複合処理
  - 述語項構造解析，共参照解析（照応解析）
  - 質問応答，機械翻訳，評判・意見情報抽出，知識獲得

典型的な言語処理の流れ

1. 形態素解析
   - 分かち書き，品詞付与
2. Chunking
   - 文節まとめ上げ，固有表現解析
3. 統語解析
   - 句構造解析，または，係り受け解析
4. 項構造解析
   - 述語（動詞，形容詞，動作名詞など）の項（意味的に必須な役割：主語，目的語など）の同定
   - 省略された項の補完（照応解析）

述語項構造の解析


係り受け解析
が格
項構造（意味役割）付与
time
と格
村山富市首相は八日社会党の新民主連合会長と会談し、今春の統一地方選後に党全体での新党移行を目指す考えを強調、離党など性急な行動への自重を求めた。
が格に格の省略
が格の省略
照応関係の付与
省略された格の同定

SynCha

Japanese predicate argument structure analyzer (demo)

メンバー
- 乾 健太郎
- 飯田 龍
- 小町 守

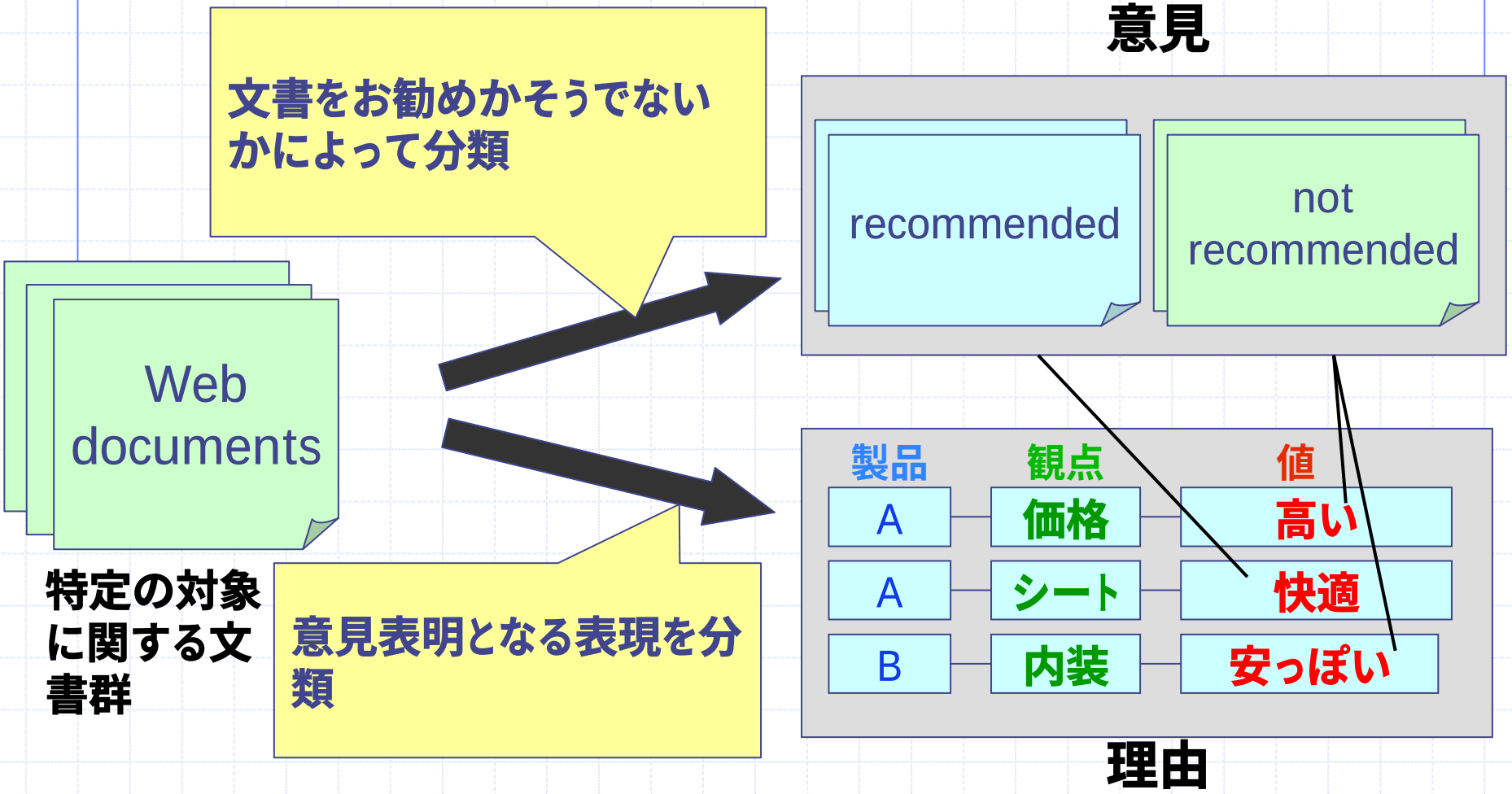
link
- タグの仕様
- 意見情報マイニング プロジェクト
- NAISTテキストコーパス

チェルノブイリ発電所の安全性については、Ｇ７サミット等において西側諸国が懸念を表すとともに早期の閉鎖を要望している。リトアニアのイグナリナ発電所を対象に調査を実施しており、これらの調査結果に基づき、課題の抽出及び対応策の取りまとめを行っている。

graph | table | clear

type	base form	nominative (ガ格)	accusitive (ヲ格)	dative (ニ格)
事態性名詞	懸念	西側諸国		
述語	表す	西側諸国	懸念	
事態性名詞	閉鎖			
述語	要望する			
事態性名詞	調査	イグナリナ発電所		
述語	実施する	西側諸国		
事態性名詞	調査結果	西側諸国		
述語	基づく	西側諸国		調査結果
事態性名詞	抽出及び	西側諸国		
述語	行う	西側諸国	取りまとめ	対象

応用例：Webからの意見・評判情報の抽出と分類



レストランの評判抽出システム

http://cl.naist.jp/~ryu-i/coe_demo/cgi/demo.cgi - Microsoft Internet Explorer

イル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)

レス(D) http://cl.naist.jp/~ryu-i/coe_demo/cgi/demo.cgi 移動

件に該当する7店が検索されました

今井リーガロイヤルホテル店
定休日: 無休
営業時間: 11:00～22:00
住所: 大阪府大阪市北区中之島5－3－68

肯定的な意見 (6)

属性	評価	出現文脈	ID
店内	真新しい	**真新しい** **店内**のせいか割高な価格設定のせいか、あまり魅力的には思えなかった。	0
佇まい	上品	猥雑だが活気のある商店街の中でこちらのお**上品**な**佇まい**は少々浮いている気がする。	1
分煙	有難い	フロアごとで**分煙**になっているのは**有難い**が、うどん屋というよりは割烹といった雰囲気はあまり好みではないな。	2
ダシ	美味しい	肝心のうどんは、確かに**ダシ**は**美味しかっ**た。	3
ダシ	しっかり	油揚げのせいか少々甘めな**ダシ**は昆布のダシが**しっかり**出ていてなかなかの味わい。	4
味わい	なかなか	油揚げのせいか少々甘めなダシは昆布のダシがしっかり出ていて**なかなか**の**味わい**。	5

否定的な意見(2)

属性	評価	出現文脈	ID
設定	割高	真新しい店内のせいか**割高**な価格**設定**のせいか、あまり魅力的には思えなかった。	6
うどん	物足りない	ただ、私は美々卯系のやわい**うどん**はそれほど好みではないので正直言って**物足りなかっ**た。	7

麺匠の心つくし つるとんたん
定休日: 無休
営業時間: 11:00～20:00
住所: 大阪府大阪市中央区宗右衛門町3－17

肯定的な意見 (5)

属性	評価	出現文脈	ID
ダシ	美味しい	きつねは、ふっくらとしていますし、麺はコシがあり、**ダシ**はやや甘口で**美味しい**です。	8
雰囲気	良い	それは、うどん屋の常識をくつがえす程の**雰囲気**の**良**さです。	9
雰囲気	よい	広々とした店内に**雰囲気**がすごく**よく**、おうどんの種類もオリジナリティがあって、すごくおいしいです。	10
店内	広々	**広々**とした**店内**に雰囲気がすごくよく、おうどんの種類もオリジナリティがあって、すごくおいしいです。	11
おうどん	おいしい	広々とした店内に雰囲気がすごくよく、お**うどん**の種類もオリジナリティがあって、すごく**おいしい**です。	12

はがくれ梅田店
定休日: 日曜日
営業時間: [月～金] 11:00～14:45(L.O) 17:00～19:45(L.O) [土] 11:00～14:30
住所: 大阪府大阪市北区梅田1－1－3

肯定的な意見 (3)

属性	評価	出現文脈	ID

応用

意見・評判マイニング

意味解析
文脈解析

照応表現の解析

関係抽出

述語項構造解析

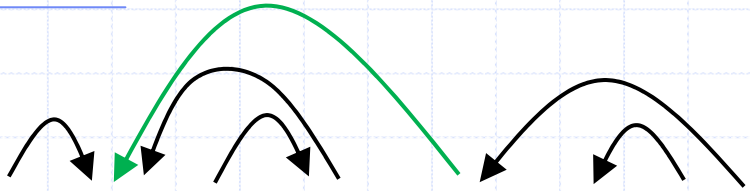
機械学習に基づく
基盤的言語解析
システム

文節係り受け
固有表現解析
品詞タグ付け
分かち書き

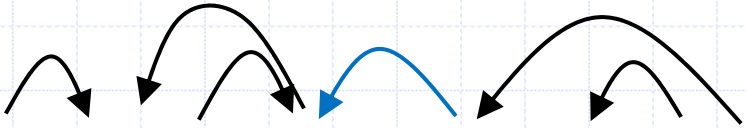
言語処理における2つの大きな問題

- 曖昧性
  - 言語処理のあらゆる段階に曖昧性があり，ただ一つの正しい解析にいたることが難しい
  - 曖昧性解消のためにどのような知識が必要かを特定することが難しい
  - **主な曖昧性の原因**
    - Attachment: **係り先（修飾先）の曖昧性**
    - **並列構造**(and, or, **で結ばれる等位句の範囲の曖昧性**)
- 頑健性 (システムのカバレージ)
  - どのような自然言語処理システムについても，開発者が想定していなかったような言語現象が存在する
  - 個々の現象に対して規則を書き並べても完全な解析システムにならない

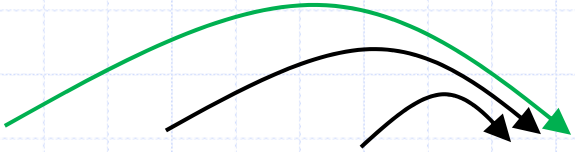
構文的曖昧性の例(attachment)



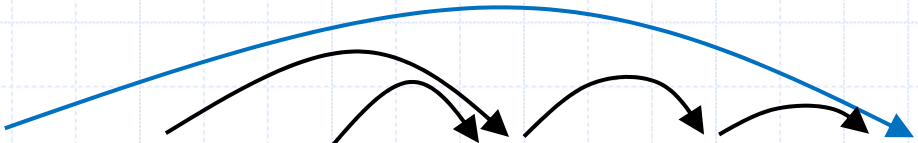
- I saw a girl with a telescope.



- I saw a girl with a scarf.

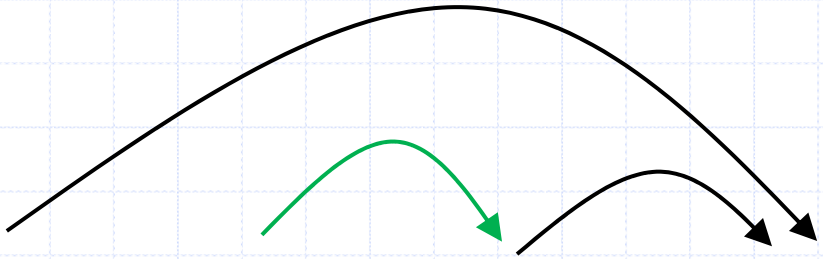


- 彼は　よく　本を　読む

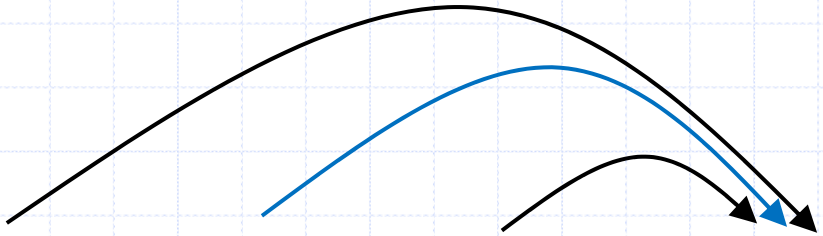


- 彼は　よく　本を　読む　人が　好きだ

構文的曖昧性の例(並列句)



- 私は　大阪と　京都へ　行きました



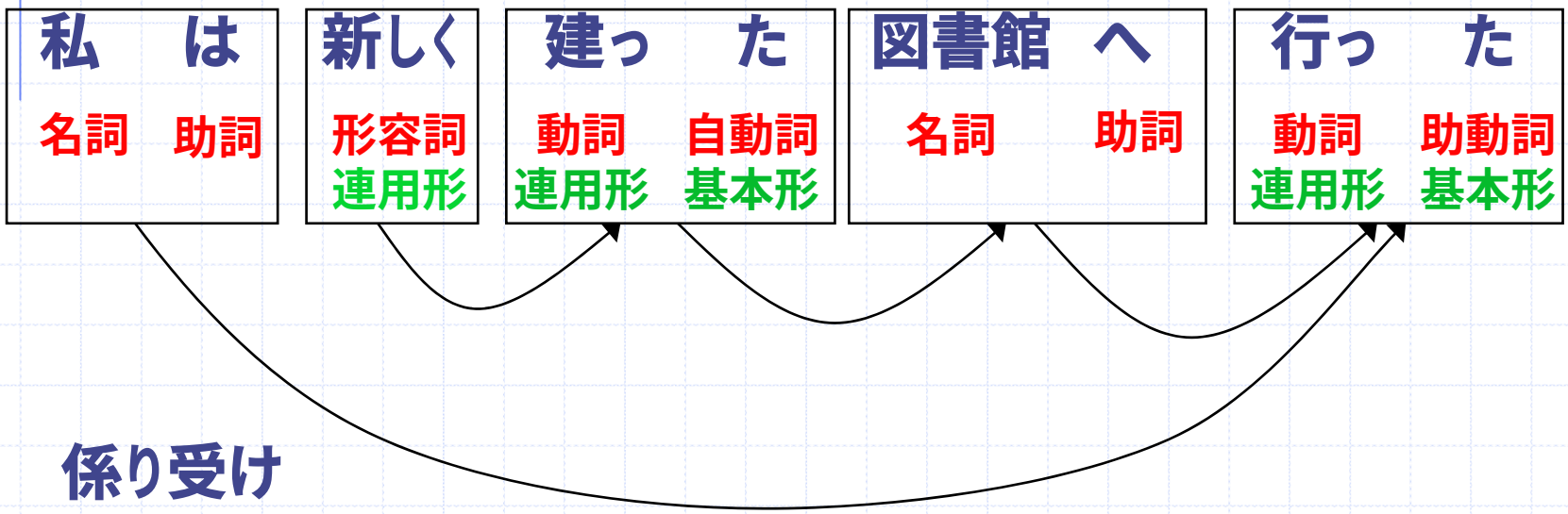
- 私は　友人と　京都へ　行きました

係り受けに基づく統語解析

- 単語(あるいは、文節/基本句)の間の統語的依存関係の解析
- 統語木は、単語だけからなり、句のような中間レベルがない
- 言語に依存しない統語解析が可能
- 明示的な文法規則を過程しない (係り受け関係のラベルを設定することはある) ので，文法の頑健さ (カバレージ) に関する問題が少ない
- 効率的な解析手法が多く提案されている

文節まとめ上げと係り受け解析

文節まとめ上げ



係り受け
解析

日本語係り受け解析システム「南瓜」
(Cascaded Chunking Model) [工藤，松本 02]

- 各文節が右側の文節に係ることができるかどうかを決定的に決めながら、文節のまとめあげを行う
- 文節間の係り受けの可否の判定をSVMを用いて学習する
- 用いているアルゴリズムは，局所最適化の繰り返し

南瓜による文解析例

自民党は20日、衆院選小選挙区の第5次公認候補4人を発表した。

入力文

解析結果

```
        <ORGANIZATION>自民党</ORGANIZATION>は---------D
                                <DATE>２０日</DATE>、-------D
    <ORGANIZATION>衆院</ORGANIZATION>選小選挙区の-D    |
                                第５次公認候補-D |
                                        ４人を-D
                                        発表した。
EOS
＊ 0 5D 0/1 4.13561692
自民党      ジミントウ    自民党      名詞-固有名詞-組織                  B-ORGANIZATION
は          ハ           は          助詞-係助詞              0
＊ 1 5D 2/2 2.98545849
２          ニ           ２          名詞-数                  B-DATE
０          ゼロ         ０          名詞-数                    I-DATE
日          ニチ         日          名詞-接尾-助数詞                    I-DATE
、          、           、          記号-読点                0
＊ 2 3D 4/5 0.96527839
衆院        シュウイン   衆院        名詞-固有名詞-組織                  B-ORGANIZATION
選          セン         選          名詞-接尾-一般                  0
小          ショウ       小          接頭詞-名詞接続
選挙        センキョ     選挙        名詞-サ変接続
区          ク           区          名詞-接尾-地域                      0
の          ノ           の          助詞-連体化              0
＊ 3 4D 4/4 1.49313142
第          ダイ         第          接頭詞-数接続                0
５          ゴ           ５          名詞-数                  0
次          ジ           次          名詞-接尾-助数詞                    0
公認        コウニン     公認        名詞-サ変接続                  0
候補        コウホ       候補        名詞-一般                0
```

係り受け木

文節区切り

BIOタグによる固有表現同定

係り受け情報
3D: 3という文節にかかっているという意味

機械学習に基づく構造学習

代表的な
学習モデルと
用いられる素性

- 形態素解析
  - 隠れマルコフモデル, CRFs(Conditional Random Fields)
  - 1次または2次のマルコフ仮説
- 浅い統語解析(基本句解析, 固有表現認識)
  - Maximum Entropy modelによる線形chunking, CRFs
  - 前後±2単語前後の文脈情報
- 統語解析
  - 確率文法, 係り受け解析 (決定性解析, 最大全域木)
  - 部分木, または, 局所的な係り受け関係

典型的な目標関数

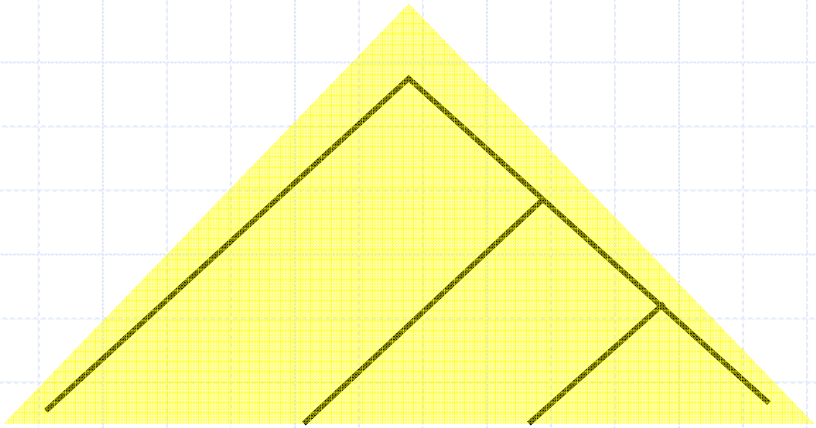
$$\arg\max_{y \in Y(x)} \theta \cdot f(x, y)$$

$$\arg\max_{y \in Y(x)} p(y \mid x)$$

機械学習に基づく構造解析の問題点

- 局所素性 vs 全域最適化のtrade-off
  - 構造全体にわたる最適化問題であっても，用いている素性は局所的
    - Window sizeの限られた素性の利用
  - 決定性局所処理のcascadingは解析の履歴(history)が使えるので，より広い素性を用いることができるが，未解析の部分は構造情報を用いることができない
    - 途中の解析エラーの伝播
- 異なるレベルの処理がパイプライン的に処理されることが多い(e.g. 形態素→文節→統語→．．．)
  - 同時解析：異なるレベルの解析の同時処理
    - 組み合わせ爆発を防ぐ必要がある

局所確率しか用いないことの問題点

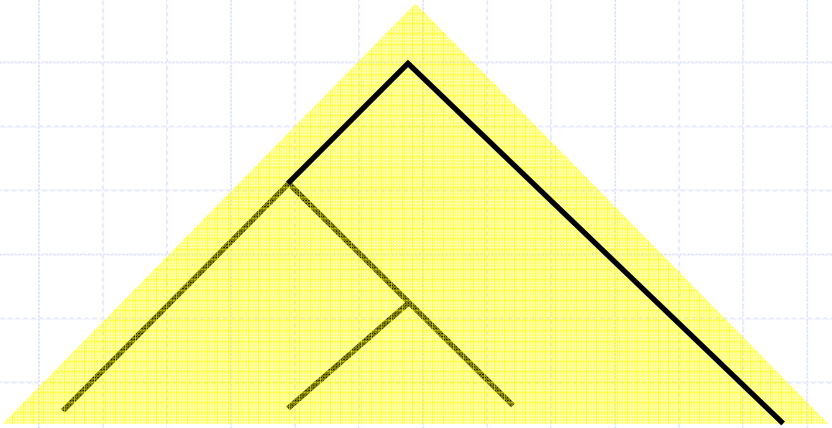


彼 は　　よく　　本 を　　読む

この局所的な文脈で
確率が最大の木

局所確率しか用いないことの問題点

この文脈では、この木の確率が高い



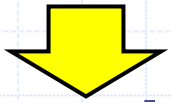
彼 は　よく　本 を　読む　人 が　好きだ

統計的学習による句構造解析

- 確率文脈自由文法
  - 各句構造規則への確率割り当て
- 語彙情報も用いた決定木学習による確率値計算
  - Magerman 1995
- 最大エントロピー法による決定的解析
  - Ratnaparkhi 1997
- 単語の共起確率と文法規則マルコフ化の導入
  - Collins 1996, 1997
- 確率文脈自由文法によるN-best解析＋より広い範囲の情報を用いた最大エントロピー法による確率値再推定
  - Charniak 2000
- Collins ModelによるN-best解析＋reranking(Tree Kernelを用いた統語木の部分木を素性とする識別学習)
  - Collins 2002

より広い文脈情報を考慮するにはどうすればよいか

- すべての可能な句構造木を生成して比較するのは非現実的
  - 可能な句構造木の数は文長の指数オーダー



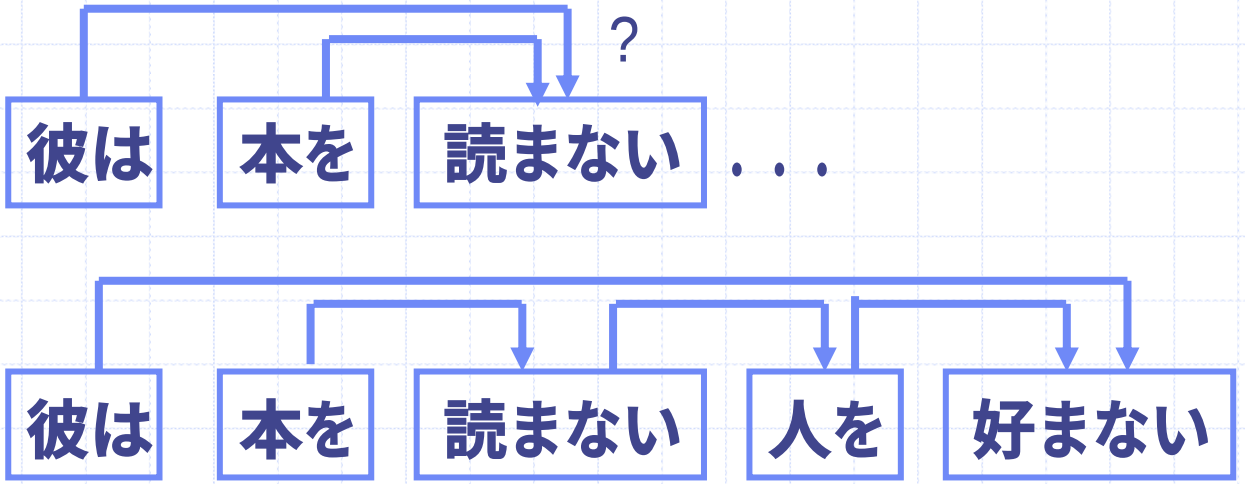
- 局所素性＋history:限定された前後文脈を見せることによって評価値(確率値)を推定できるようにする
- 簡単なモデルによるN-best解＋全域比較：大域素性を用いることができる．N-bestに正解がなければ，正しい解析が付加．効率性に難．

N-best解析との組み合わせ

- Maximum Entropy-inspired Parser[Charniak 00]
  - **確率文脈自由文法によるN-best解析＋**
  - **親の親や兄弟の情報を使った最大エントロピー法による確率推定**
- Reranking with Tree Kernel and Voted Percepton[Collins & Duffy 02]
  - Collins**の共起確率モデルによるN-best解析**
  - Tree Kernel(**句構造木のすべての部分木を要素とする内積)を用いた識別学習**
- Coarse-to-fine n-best and MaxEnt Reranking[Charniak & Johnson 05]
  - Charniak00 parser**による**n-best parse(2**段階**)
  - **句構造木全域を対象とする素性を用いた最大エントロピー法による確率推定**
  - **並列句のラベルや長さの情報，他の構造情報による性能向上**

係り受け解析における曖昧性の問題

- 係り先 (attachment) の曖昧性

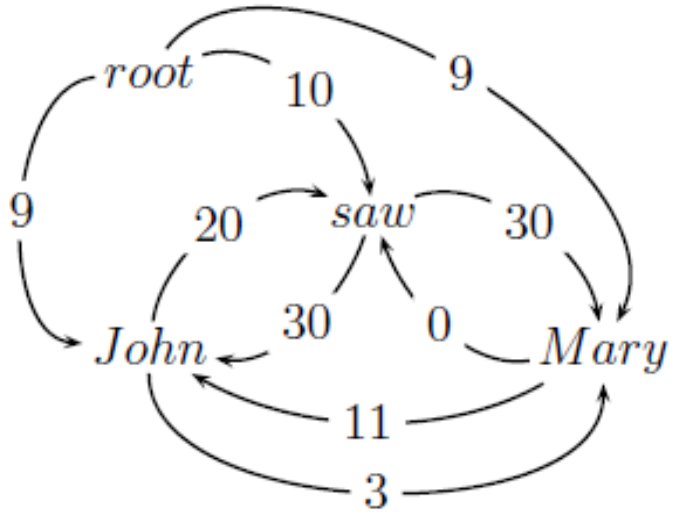


  - このように同じような単語列が異なる解析結果をもつ場合、学習法を工夫しなければ、学習が破綻する
  - 広い文脈を考慮する必要があるが、その決定は困難

Statistical dependency parsers

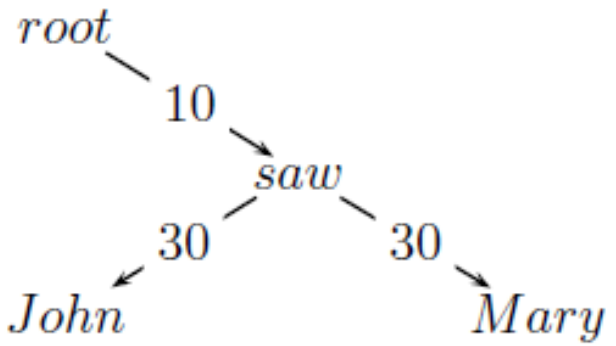
- Eisner (1996)
  - **確率モデル＋CKY法(動的計画法)に基づく解析**
- Kudo & Matsumoto (VLC 00, CoNLL 02):**日本語**
  - SVM**出力の擬似確率値利用＋CKY法**
  - SVM**による**cascade chunking
- Yamada & Matsumoto (IWPT03)
  - SVM**による**cascade chunking
- Nivre (IWPT03, COLING04, …)
  - **決定性**Shift-reduce parsing
- McDonald-Crammer-Pereira (ACL05)
  - **最大全域木アルゴリズム**
- Reidel & Clarke (EMNLP06)
  - **整数(線形)計画法による全域最適化**

最大全域木による係り受けのイメージ



“John saw Mary”
に対するすべての単語対の
コスト(rootはダミーの根ノード)

最大コストを
もつ全域木



Chu-Liu-Edmonds algorithm:$O(n^2)$

Notation and definition

$\mathrm{T} = \{(\boldsymbol{x}_t, \boldsymbol{y}_t)\}_{t=1}^{T}$ 訓練データ

$\boldsymbol{x}_t$ t番目の文

$\boldsymbol{y}_t = \{(i, j)\}_{i=1}^{n}$ t番目の文の依存構造木(枝の集合)

$f(i, j)$ 素性関数

$\mathrm{dt}(\boldsymbol{x})$ xに対する可能な依存木の集合

基本アイデア

依存木に対するスコア関数

$$s(\boldsymbol{x}, \boldsymbol{y}) = \sum_{(i,j)\in \boldsymbol{y}} s(i, j) = \sum_{(i,j)\in \boldsymbol{y}} \mathbf{w} \cdot \mathbf{f}(i, j)$$

最適化の目標関数

$$\min \|\mathbf{w}\|$$

$$\text{s.t.} \quad s(\boldsymbol{x}, \boldsymbol{y}) - s(\boldsymbol{x}, \boldsymbol{y}') \geq L(\boldsymbol{y}, \boldsymbol{y}')$$

$$\forall (\boldsymbol{x}, \boldsymbol{y}) \in T, \quad \boldsymbol{y}' \in \mathrm{dt}(\boldsymbol{x})$$

L(y,y')は，正しい木 y に対する y' のloss．
loss = 親が間違っている単語の数と定義する

weight vectorのupdate

$$\min \left\| \mathbf{w}^{(i+1)} - \mathbf{w}^{(i)} \right\|$$

$$\text{s.t.} \quad s(\boldsymbol{x}, \boldsymbol{y}) - s(\boldsymbol{x}, \boldsymbol{y}') \geq L(\boldsymbol{y}, \boldsymbol{y}')$$

$$\forall \boldsymbol{y}' \in \mathrm{dt}(\boldsymbol{x}_t)$$

しかし，これでは，すべての可能な依存木を求める必要があるため，次のようにk-bestだけを対象とする

$$\min \left\| \mathbf{w}^{(i+1)} - \mathbf{w}^{(i)} \right\|$$

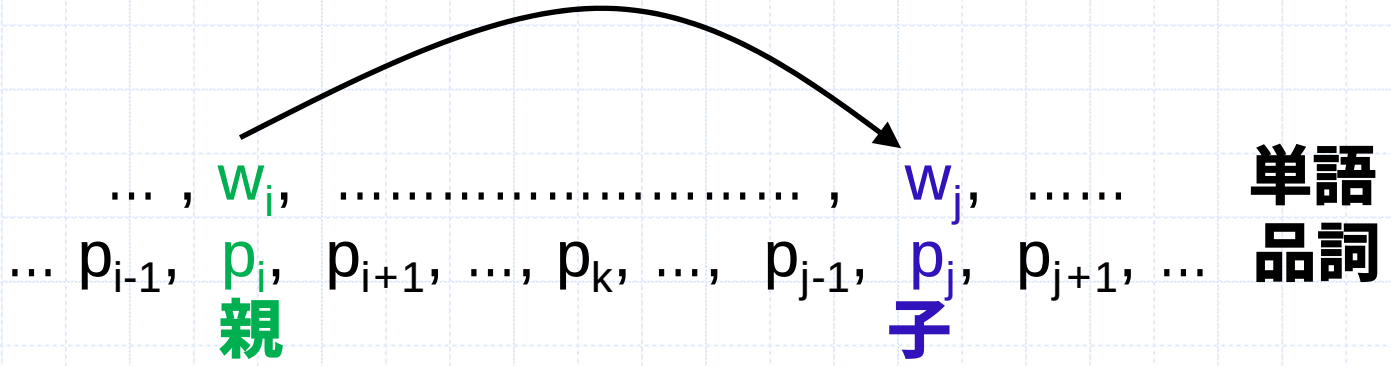
$$\text{s.t.} \quad s(\boldsymbol{x}, \boldsymbol{y}) - s(\boldsymbol{x}, \boldsymbol{y}') \geq L(\boldsymbol{y}, \boldsymbol{y}')$$

$$\forall \boldsymbol{y}' \in \mathrm{best}_{\mathrm{k}}(\boldsymbol{x}_t; \mathbf{w}^{(i)})$$

(k=5〜20)

学習に用いられる素性

- Unigram素性：親の単語，品詞，子の単語，品詞(5文字以上の語の単語素性は5文字prefix)
- Bigram素性：親の品詞・単語と子の品詞・単語の組合せ
- Trigram素性：親子の品詞とその間に存在する語の品詞の3つ組
- 文脈素性：親子の品詞と前後に現れる語の品詞．4種類の4-gram素性を構成(trigramでbackoff)



機械学習に基づく英語(多言語)の依存構造解析(代表的な手法)

- 上昇型決定性解析
  - SVMによる決定性動作学習[Yamada, Matsumoto 03]
    - 解析の動作をSVMによって学習
  - 決定性Shift-reduce parsing[Nivre 03]
    - 解析の動作をmemory-based modelにより学習、後にSVMを使用
- 全体最適化
  - 最大全域木(MST)に基づく手法[McDonald, et al 05]
    - 単語間の依存関係を素性の重みとしてパーセプトロンにより学習、全体の依存関係の和を最大化するspanning treeを求める
  - 広い範囲の素性を利用した最適化[Nakagawa 07]
    - 文全体を対象とする素性(global feature)を用いる確率学習(sampling法を利用)と最大全域木を利用
- 混合モデル
  - Nivre法とMST法のそれぞれの結果を素性として，相手の入力として利用[Nivre & McDonald 08]
  - MST法+整数計画法による大域的な制約[Reidel & Clarke 06]

日本語係り受け解析手法の変遷(1)

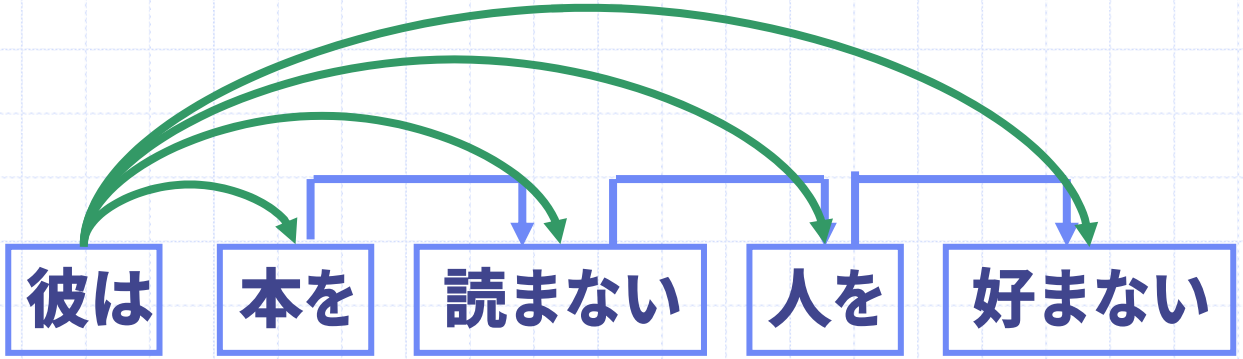
- 人手による係り受け規則の記述
- 文節係り受け確率の学習
  - 単語の共起確率[Fujio, Matsumoto 97]
  - 最大エントロピー法による確率推定[Uchimoto et al 99]
  - Support Vector Machinesによる擬似確率推定[Kudo, Matsumoto 00]
- 上昇型決定性解析
  - SVMによる決定性動作学習[Kudo, Matsumoto 02]
  - 決定性Shift-reduce parsing[Sassano 04]

日本語係り受け解析手法の変遷(2)

- 競合する係り受け関係の比較を考慮した解析
  - 文法制約により係り先を絞った上で選択[金山他 00]
    - 文法を用いた解析で係り先候補を絞る
    - 候補が2つの場合と3つの場合の選択モデルを学習
  - 相対的な係りやすさを考慮した確率推定[工藤、松本 05]
    - 係り受け確率の学習時に、正しい係り受けの確率が正しくないどの係り受け確率より大きな値をもつように学習
    - 解析時は、通常の解析(ビーム幅を考慮、しかし、決定性解析で充分な精度)
  - トーナメントモデルを利用した最適な係り先選択 [Iwatate, Asahara, Matsumoto 08]
    - 係り先候補を2つずつ比較し、最適なものを選択

言語解析における曖昧性の問題

- 可能な係り先をいかに選択するか



トーナメントモデルを用いた日本語係り受け解析

[Iwatate 08]

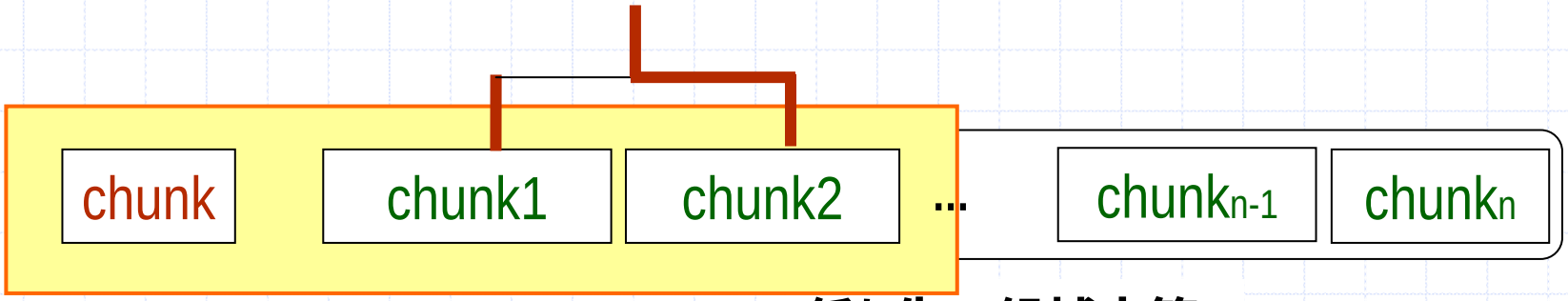
- 係り受け解析を文末から解析する
- 各文節の係り先の最適な候補をトーナメントによって選択する

係り先の候補文節

トーナメントモデルを用いた日本語係り受け解析

[Iwatate 08]

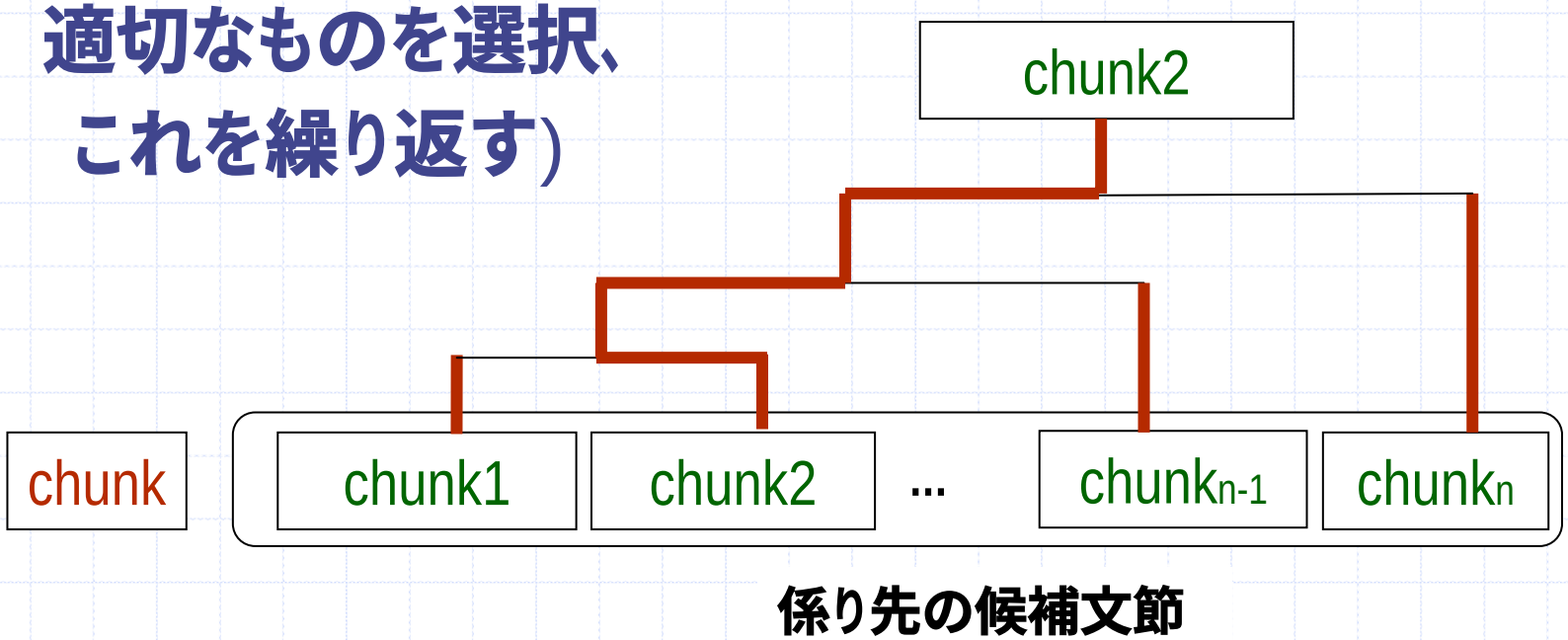
- 係り受け解析を文末から解析する
- 各文節の係り先の最適な候補をトーナメントによって選択する



トーナメントモデルを用いた日本語係り受け解析

[Iwatate 08]

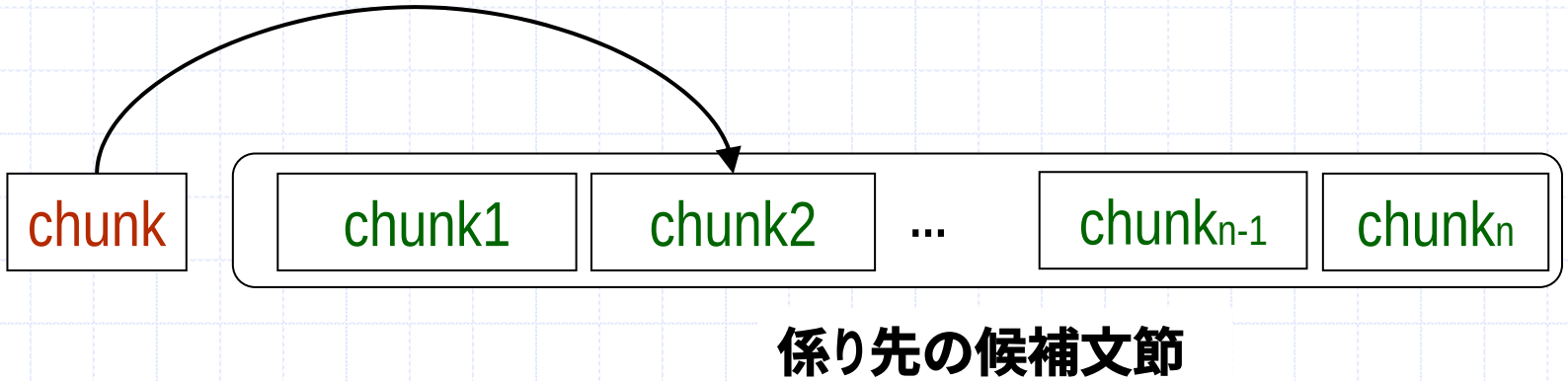
- 係り受け解析を文末から解析する
- 各文節の係り先の最適な候補をトーナメントによって選択する(係り先候補を2つずつ比較し、より適切なものを選択、
  これを繰り返す)



トーナメントモデルを用いた日本語係り受け解析

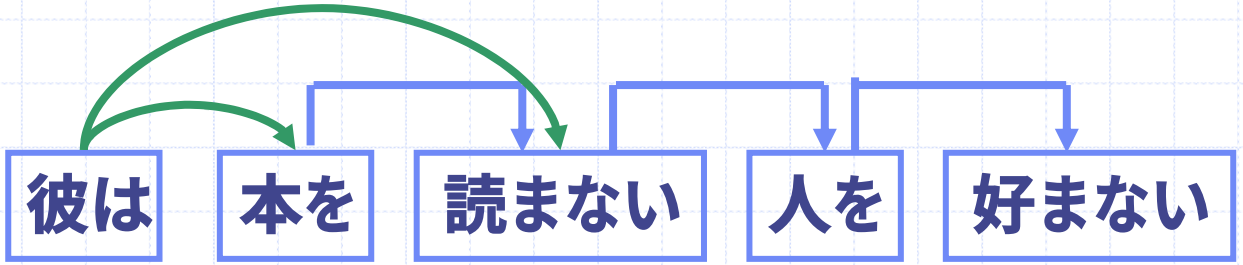
[Iwatate 08]

- 係り受け解析を文末から解析する
- 各文節の係り先の最適な候補をトーナメントによって選択する



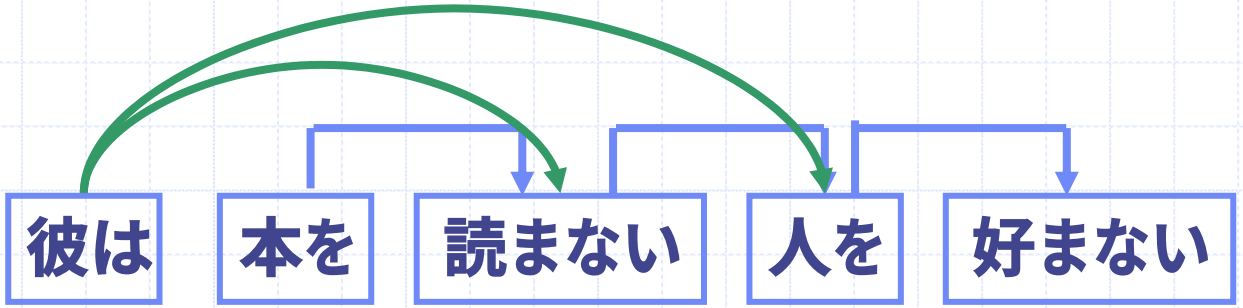
言語解析における曖昧性の問題

- 係り先の曖昧性に対応する解析法
  - トーナメントモデル



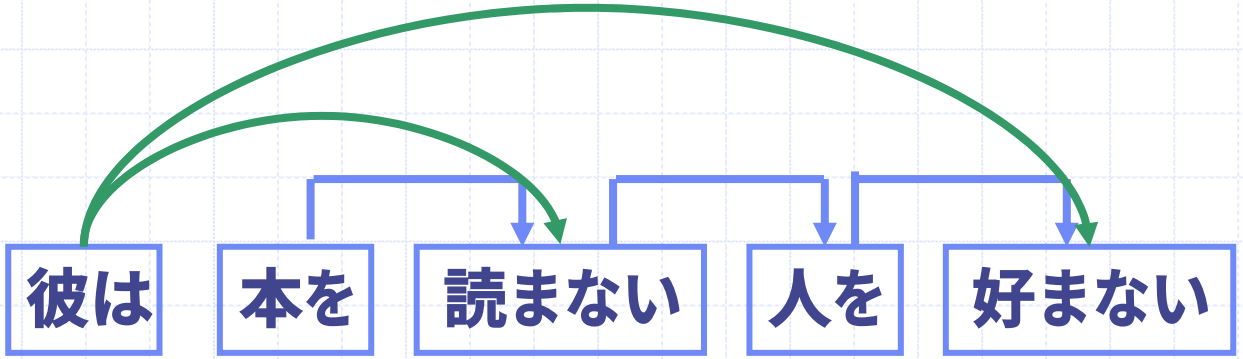
言語解析における曖昧性の問題

- 係り先の曖昧性に対応する解析法
  - トーナメントモデル



言語解析における曖昧性の問題

- 係り先の曖昧性に対応する解析法
  - トーナメントモデル



日本語係り受け解析性能の比較

- **南瓜** [Kudo & Matsumoto 02]
  - Cascaded chunking model
    - **隣り合う文節間のまとめ上げの繰り返し**
- tournament model [Iwatate, Asahara, Matsumoto 08]
  - **係り先候補を2つずつ比較しながら，最適の候補を選択**

	test1	test2	test3
CaboCha	89.22/47.90	89.80/47.94	89.53/49.79
Tournament	90.09/47.79	90.11/49.02	90.35/52.59

(約1万文による実験の数値，数字は「係り受け精度/文解析の精度)」

統語解析に関するこれまでの考え方の変遷

- **局所的かつ文脈非依存の尺度による大域的最適化**
  - (HMM), PCFG, MST
- **限定された文脈を用いた決定性解析**
  - Cascaded chunking, Shift-reduce parsing
- **他候補との比較**
  - **全体比較**：N-best**解析**＋reranking
  - **局所的な比較**：Tournament model

曖昧性解消のためのその他の手法

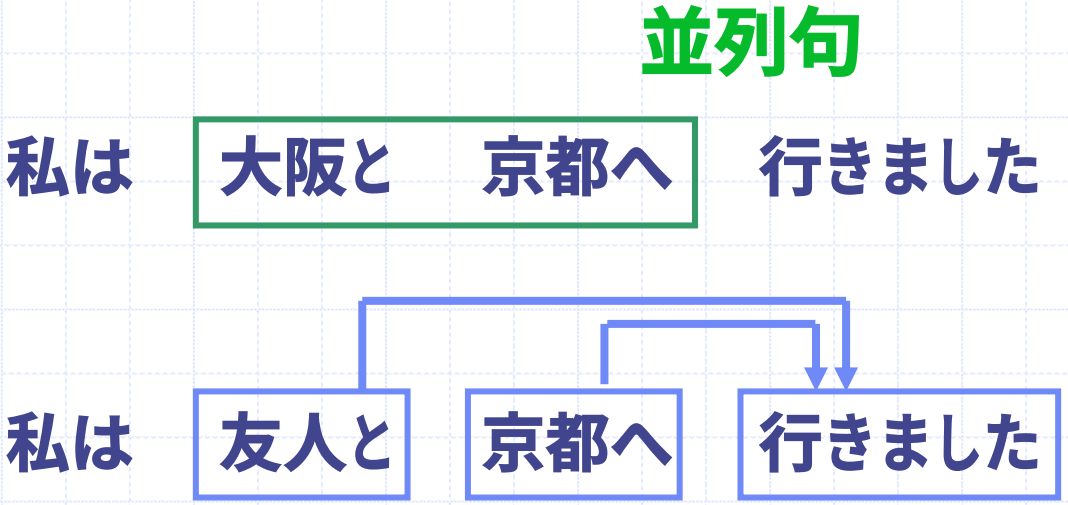
- 制約との融合
  - 言語に関する文法的知識を制約として同時にモデル化
  - 制約を確率的にモデル化
    - e.g., probabilistic inductive logic programming, Markov logic networks
- 同時推定(joint inference) vs pipeline
  - 異なるレベル (e.g.,形態素解析と統語解析) を同時にモデル化し学習する
  - これまでの多くの手法は, 個々のレベルを独立にモデル化し, pipeline的につなげていた

制約との融合の例

- 形態素解析における種々の文法的制約
  - 文には少なくとも一つ述語（動詞，形容詞など）が存在しなければならない
  - （英語）：助動詞の後には数単語以内に必ず動詞の原形が存在しなければならない
  - このような制約を確率論理プログラムで実装[]
- 並列句解析における整合性に関する制約
  - 文中に複数の並列構造が出現する場合，それらの開始･終了位置が交差していてはならない
  - これを系列アラインメントと文法制約により融合[原，他 08]

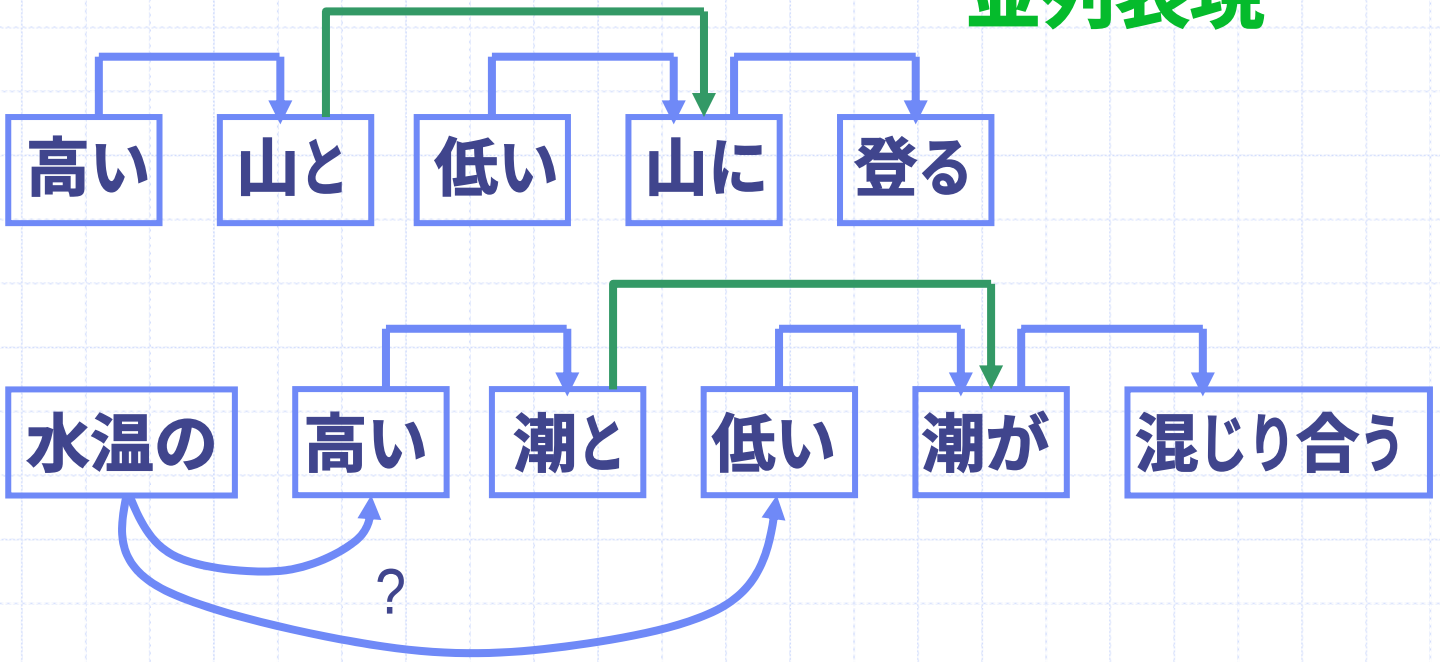
係り受け解析における並列句の問題

◆単純な並列句の場合



係り受け解析における並列句の問題

- 並列句にまつわる係り受けの表現の問題

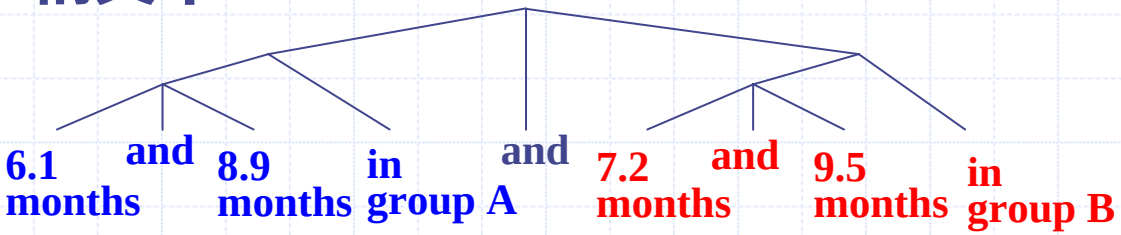


「水温の」が「高い」と「低い」の両方に
係ることをどう表現すればよいか

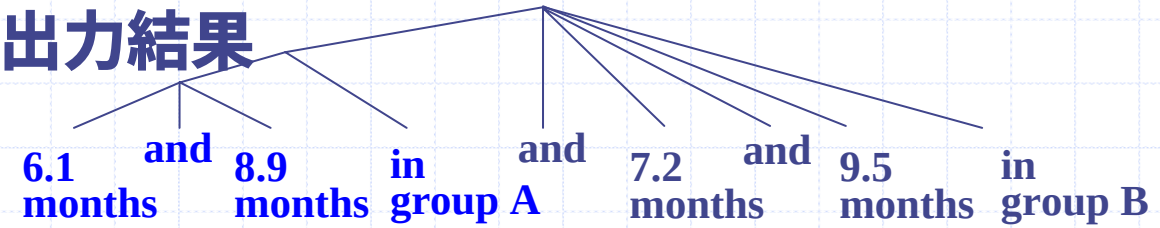
並列句は、既存の統語解析法では正しく解析するのが困難

“Median times to progression and median survival times were **6.1 months and 8.9 months in group A and 7.2 months and 9.5 months in group B**” の、下線部分に対する、

正しい構文木



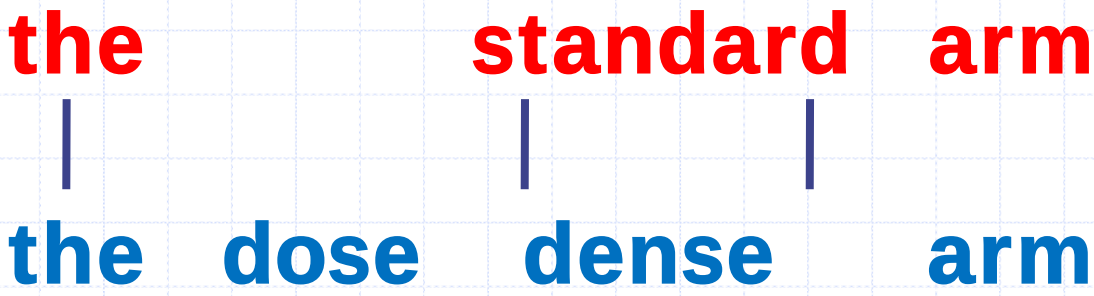
誤りの構文木 [Charniak and Johnson, 2005] による出力結果



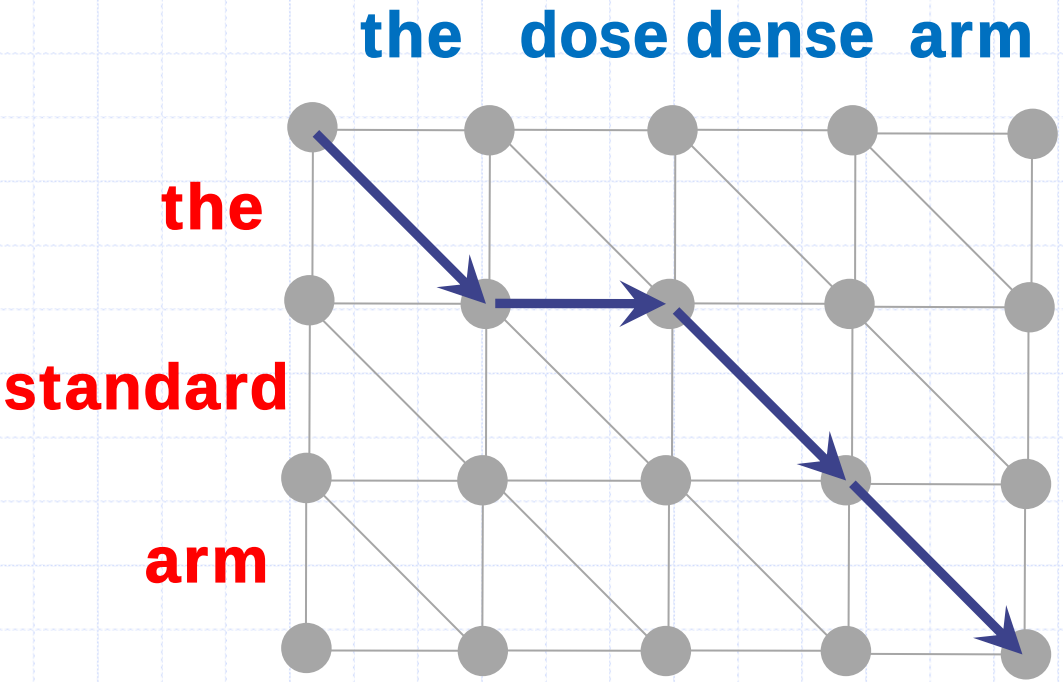
並列句構成要素対のアラインメント

- 相同する部分に対応を付ける。

“the standard arm and the dose dense arm”



アラインメントは編集グラフ上の経路に対応



the standard arm

the dose dense arm

入れ子を伴う複雑な並列構造

- 系列アラインメントによる手法では入れ子になった並列構造を解析するのが困難
- 複数の並列句が1文に存在する場合，それらが交差しないことを保証するのも困難

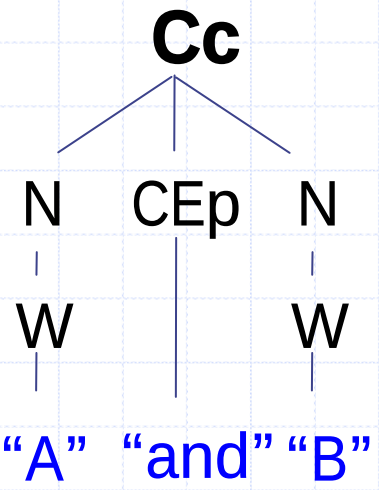
“((Median times to progression) and (median survival times)) were
((((6.1 months) and (8.9 months)) in group A) and
(((7.2 months) and (9.5 months)) in group B))”

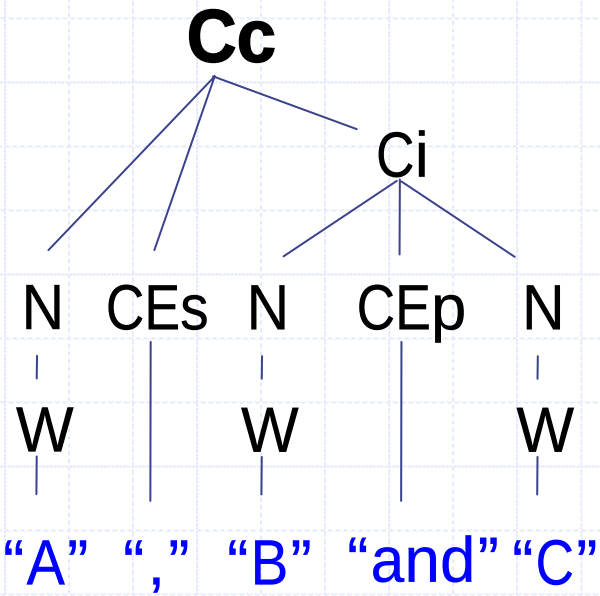
並列句範囲同定へのアラインメントと文法制約の融合 [原, 新保, 大熊, 松本 08]

1. 並列句構成要素対にスコアを定義する
   - アラインメント素性の重みに基づく
2. 並列句に特化した文法により，通常の句構造解析を行う (並列句構造の整合性制約)
   - それぞれの構文木は、並列句範囲の異なる決め方に対応する
3. 訓練データを用い、素性の重みを最適化する
   - CKY統語解析アルゴリズムとパーセプトロンを組み合わせたアルゴリズム

Cc (完全な並列句) の生成規則

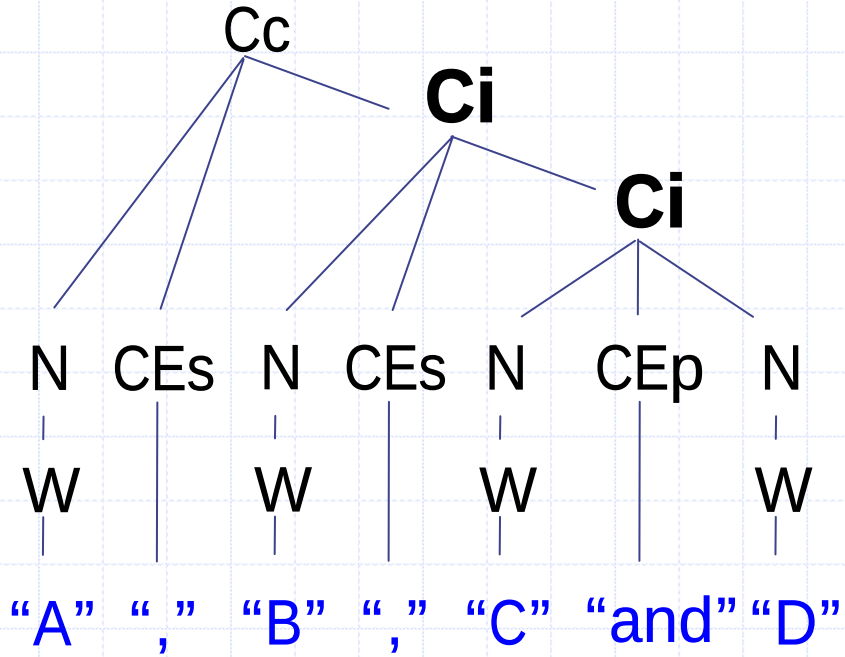
- Cc → {N, Cc} CEp {N, Cc}
- Cc → {N, Cc} CEs Ci





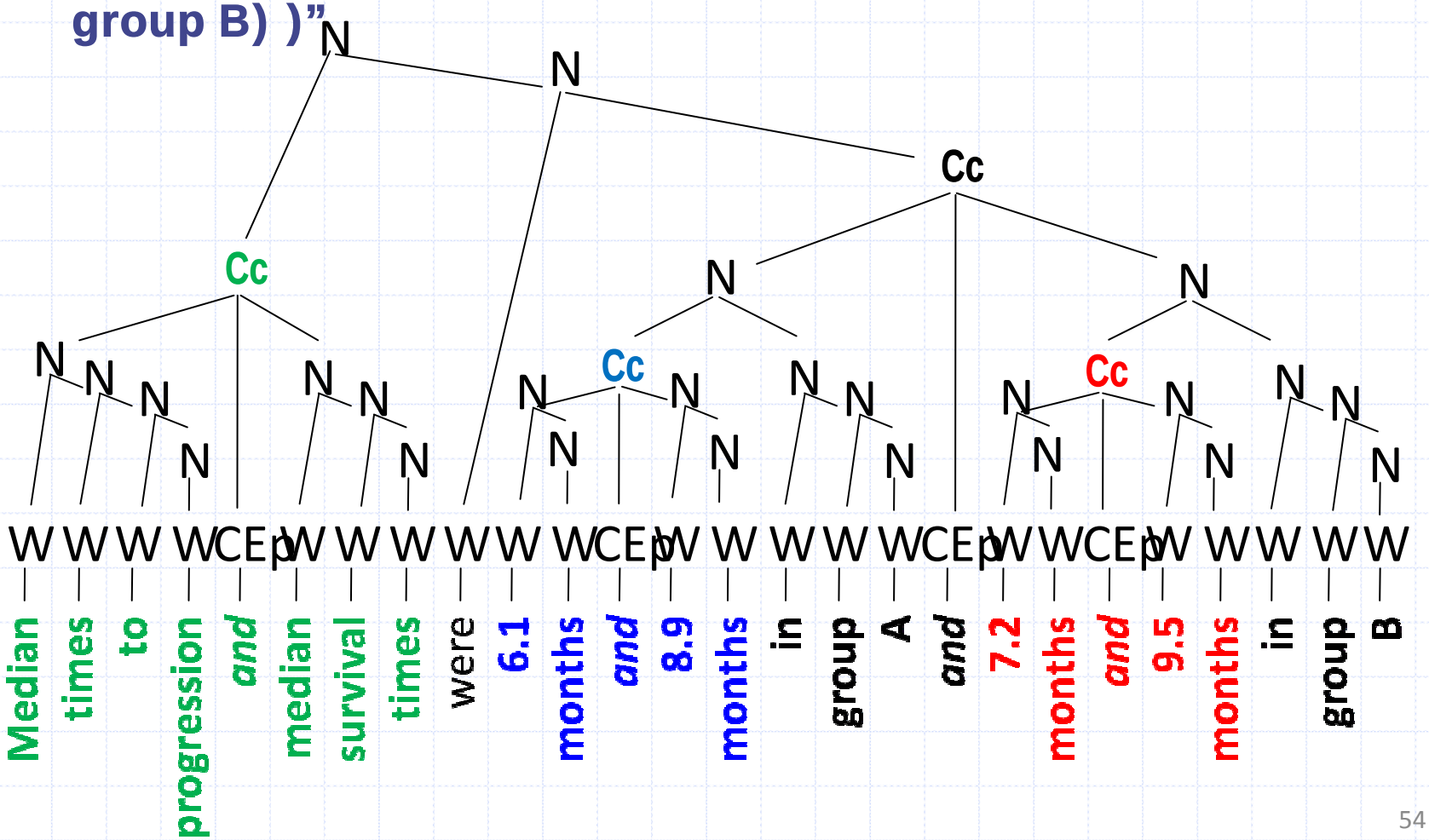
Ci (不完全な並列句) の生成規則

- Ci → {N, Cc} CEp {N, Cc}
- Ci → {N, Cc} CEs Ci



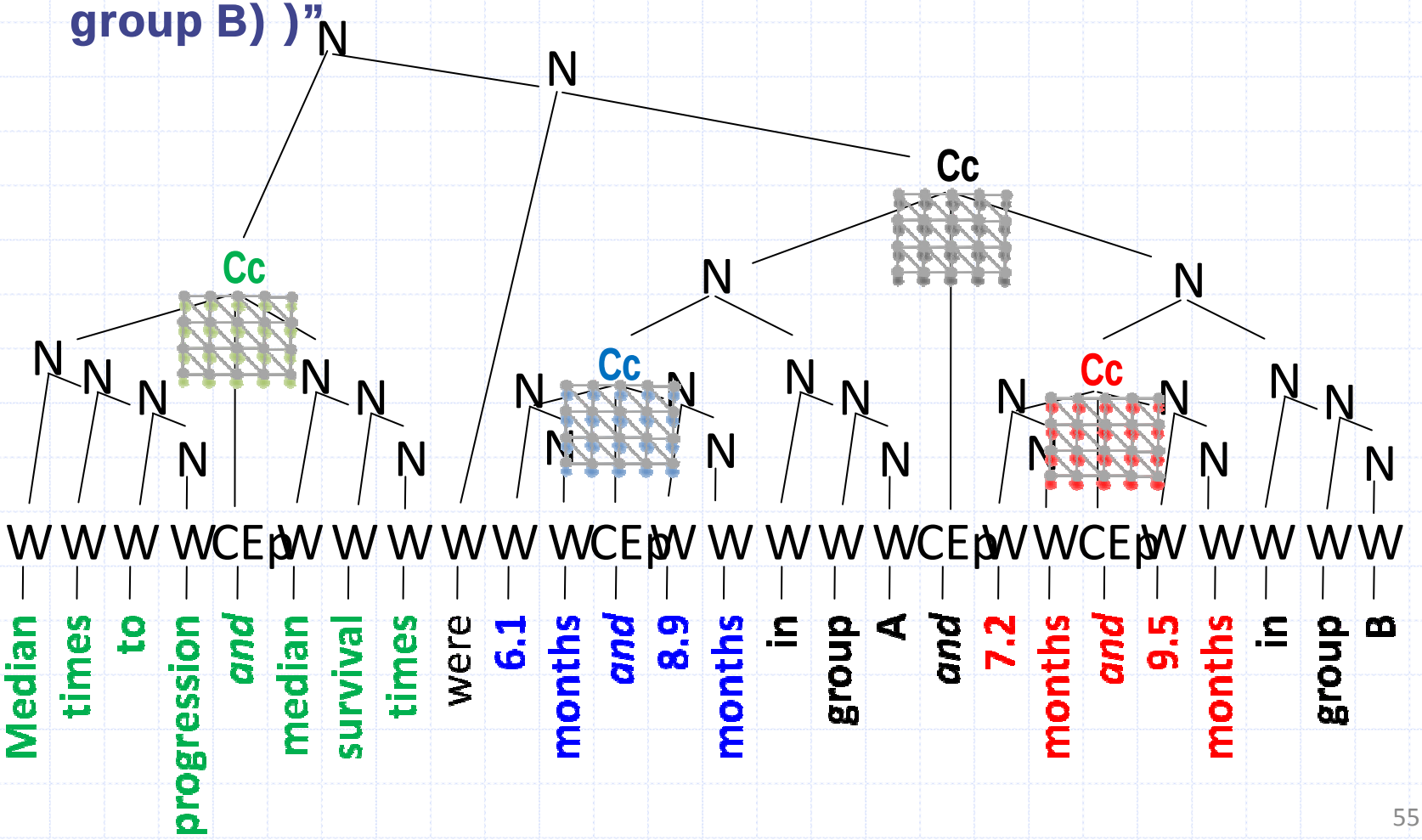
文とその並列句の範囲が与えられたとき、これを一意の構文木に表現できる

"((Median times to progression) and (median survival times)) were ((((6.1 months) and (8.9 months)) in group A) and (((7.2 months) and (9.5 months)) in group B))"



構文木全体のスコアは、並列句のアラインメントスコアの総和として制約を保証しつつ最適化

"((Median times to progression) and (median survival times)) were ((((6.1 months) and (8.9 months)) in group A) and (((7.2 months) and (9.5 months)) in group B))"



評価実験

- GENIA コーパスを使用する
- “and”を接続詞としてもつ並列句 2,948 個を解析対象とする
  - GENIA コーパスの COOD タグ (4,129個) の約7割
- このうちの、名詞並列句 (1,976個) の並列範囲を同定するタスクに対し、提案手法を評価する

実験結果

手法	適合率	再現率	F値
提案手法	**56.8%**	**54.5%**	**55.6**
Charniak and Johnson, 2005	45.6%	42.9%	44.2
Bikel, 2004	43.9%	44.4%	44.1

同時処理の例

- 品詞タグ付けと固有表現の同時処理
  - Factorial CRFによる同時処理[Sutton, Rohanimanesh, McCallum 04]
- 形態素解析と統語解析の同時処理
  - 両者が生成モデルとして確率値計算するなら統合は容易：形態素lattice入力のparsing
  - CRFによるヘブライ語の形態素と確率文法による句構造解析の融合[Cohen & Smith 07]
  - 下位のモデルから得られる素性の確率値を上位モデルで利用[Bunescu 08]

同時モデルの事例

- “Learning with Probabilistic Features for Improved Pipeline Models,” Razvan Bunescu (EMNLP 08)
  - **ヘブライ語の形態素解析と係り受け解析の同時処理**
  - **英語の品詞タグ付け，係り受け解析，固有表現認識の同時処理**

2つの階層的なモデルの融合

- Pipeline inference model
  - z(**下位のモデル**)**の最適解のみを上位で利用**

$$\hat{y}(x) = \arg\max_{y \in Y(x)} w \cdot \phi(x, y, \hat{z}(x))$$

$$\hat{z}(x) = \arg\max_{z \in Z(x)} p(z \mid x)$$

- Joint inference model
  - **下位モデルの全解を考慮した素性の確率値を利用**

$$\hat{y}(x) = \arg\max_{y \in Y(x)} w \cdot \varphi(x, y)$$

$$\varphi(x, y) = \sum_{z \in Z(x)} p(z \mid x) \cdot \phi(x, y, z)$$

2つの階層的なモデルの融合

- 中間的なモデル
  - z(下位のモデル)の最適解の確率だけを利用

$$\hat{y}(x) = \arg\max_{y \in Y(x)} w \cdot \varphi(x, y)$$

$$\varphi(x, y) = \sum_{z \in \hat{z}(x)} p(z \mid x) \cdot \phi(x, y, z)$$

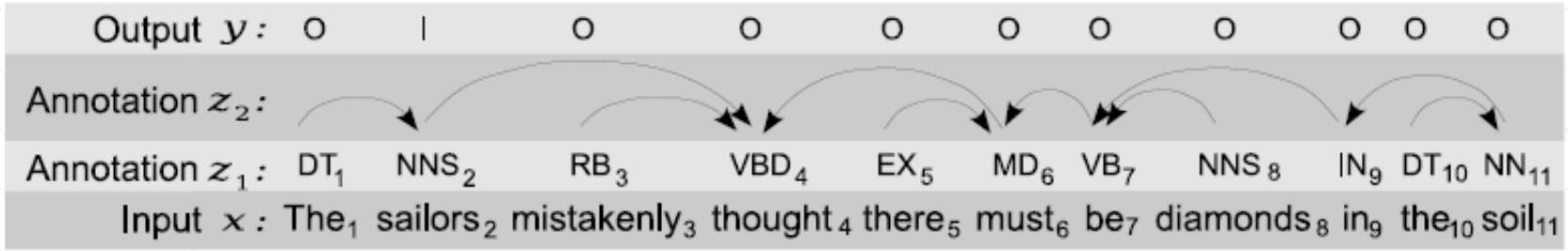


Figure 2: Named Entity Recognition Example.

Bunescu (EMNLP 08)より引用

まとめ

- 構造学習の諸問題を統語解析を例として次のような話題を説明
  - 局所素性のみによる大域的最適化
  - 大域素性の利用
    - N-best解の利用による木全体の比較
    - 競合対象の直接比較
  - 異なるレベルの情報の利用
    - 上位の制約の利用
    - 同時学習 (異なるレベルの)

言わなかったこと

- 未解析データの利用
  - 半教師あり学習(semi-supervised learning)
  - 新領域への適用(domain adaptation)
  - クラス数推定：語義曖昧性解消（語義数）
  - Bootstrapping
- 素性選択
  - 特に，素性の組み合わせ
  - 構造をもつ素性（系列，木構造）のマイニング
- 制約との融合
  - 語彙知識の活用
  - 素性としての制約の学習(素性マイニングと同義)
- 耐規模性，あるいは，そのための近似手法
- 部分アノテーションデータの構築と利用
  - アノテーション作業の動的な選択・指定
  - 部分情報をもったデータの解析