Cascaded Chunking Model における部分解析済み情報の利用

工藤 拓, 松本 裕治

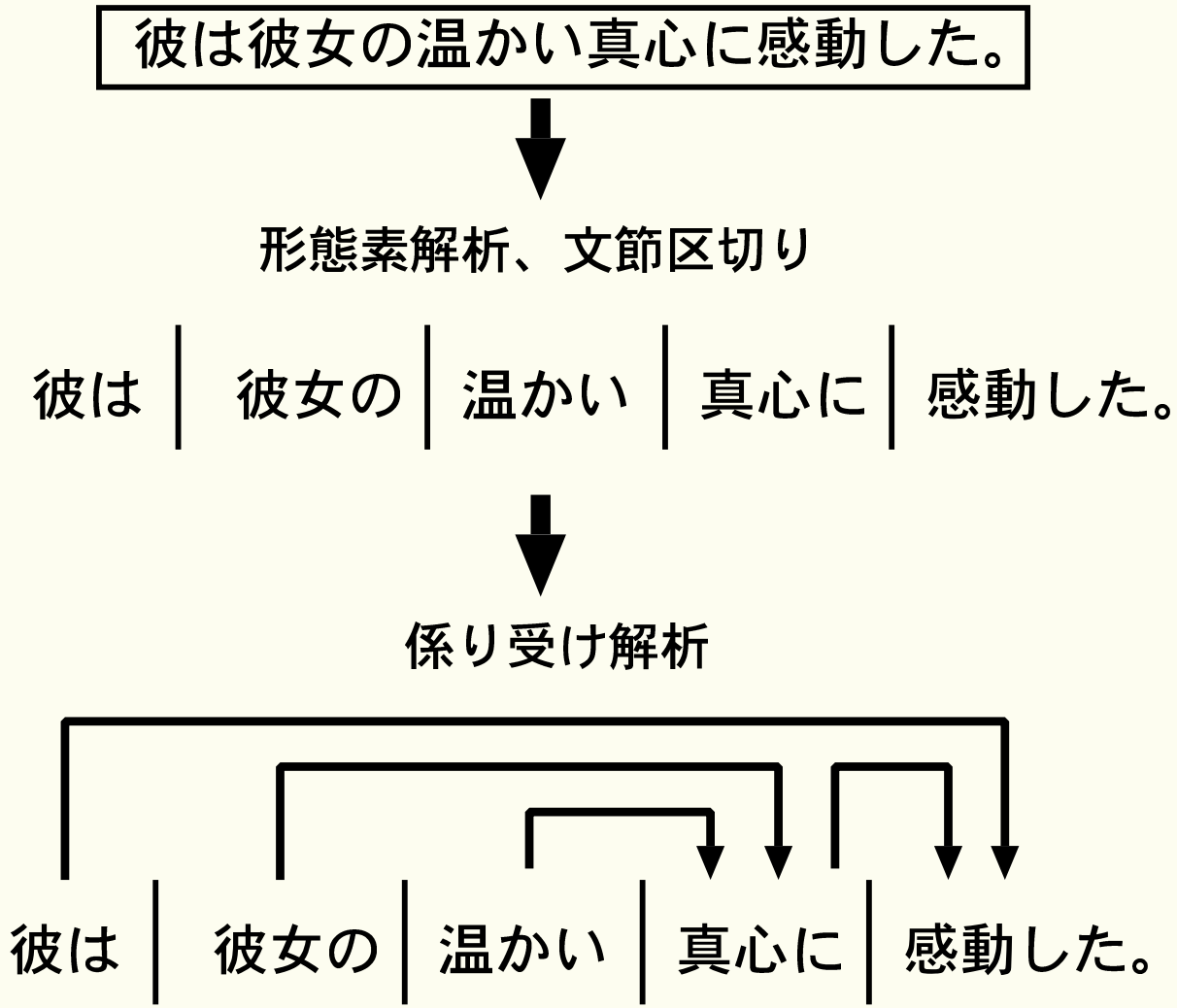
{taku-ku,matsu}@is.aist-nara.ac.jp.

奈良先端科学技術大学院大学 情報科学研究科
自然言語処理学講座

係り受け解析 (1/2)

- 係り受け解析 = 文節の修飾関係の同定
- 2 つの制約
  1. ある**文節** は後方にある一つの文節に係る **(後方参照)**
  2. 係り関係は交差しない **(非交差条件)**

係り受け解析 (2/2)


彼は彼女の温かい真心に感動した。
形態素解析、文節区切り
彼は | 彼女の | 温かい | 真心に | 感動した。
係り受け解析
彼は | 彼女の | 温かい | 真心に | 感動した。

従来手法 (確率モデル) (1/3)

- 二つの文節の係りやすさを示す**確率**を計算
  $M[i,j] = P(Dep(i) = j|\mathbf{f}_{i,j})$ (文節 i が j に係る確率)
- 係り関係はすべて独立と仮定, 文の生成確率は個々の確率の積

$$P(D|S) = \prod_i P(Dep(i) = j|\mathbf{f}_{i,j})$$

- 非交差条件を考慮しながら $P(D|S)$ を最大にする係り関係を探索

確率モデル (2/3)

Modifiee

彼は　彼女の 温かい 真心に 感動した。

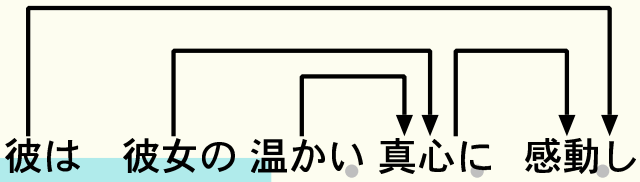
Modifier	彼は	彼女の	温かい	真心に	感動した。
彼は	0.0	0.1	0.2	0.1	0.6
彼女の	0.0	0.0	0.3	0.5	0.2
温かい	0.0	0.0	0.0	0.8	0.2
真心に	0.0	0.0	0.0	0.0	1.0
感動した。	0.0	0.0	0.0	0.0	0.0

$$M[i, j] = P(D(i) = j \mid f_{ij})$$

確率モデル (3/3)

Sekine's backward beam search method (beam width=3)

	彼は	彼女の	温かい	真心に	感動した。		
ID	1	2	3	4	5		
Cand1				5(1.0)		Prob. 1.0	
Cand1			4(0.8)	5(1.0)		Prob. 0.8	
Cand2			5(0.2)	5(1.0)		Prob. 0.2	
Cand1		4(0.5)	4(0.8)	5(1.0)		Prob. 0.4	
Cand2		3(0.3)	4(0.8)	5(1.0)		Prob. 0.24	
Cand3		3(0.2)	4(0.8)	5(1.0)		Prob. 0.16	
Cand1	5(0.6)	4(0.5)	4(0.8)	5(1.0)		Prob. 0.24	◀ Best
Cand2	3(0.6)	3(0.3)	4(0.8)	5(1.0)		Prob. 0.14	
Cand3	5(0.6)	3(0.2)	4(0.8)	5(1.0)		Prob. 0.08	

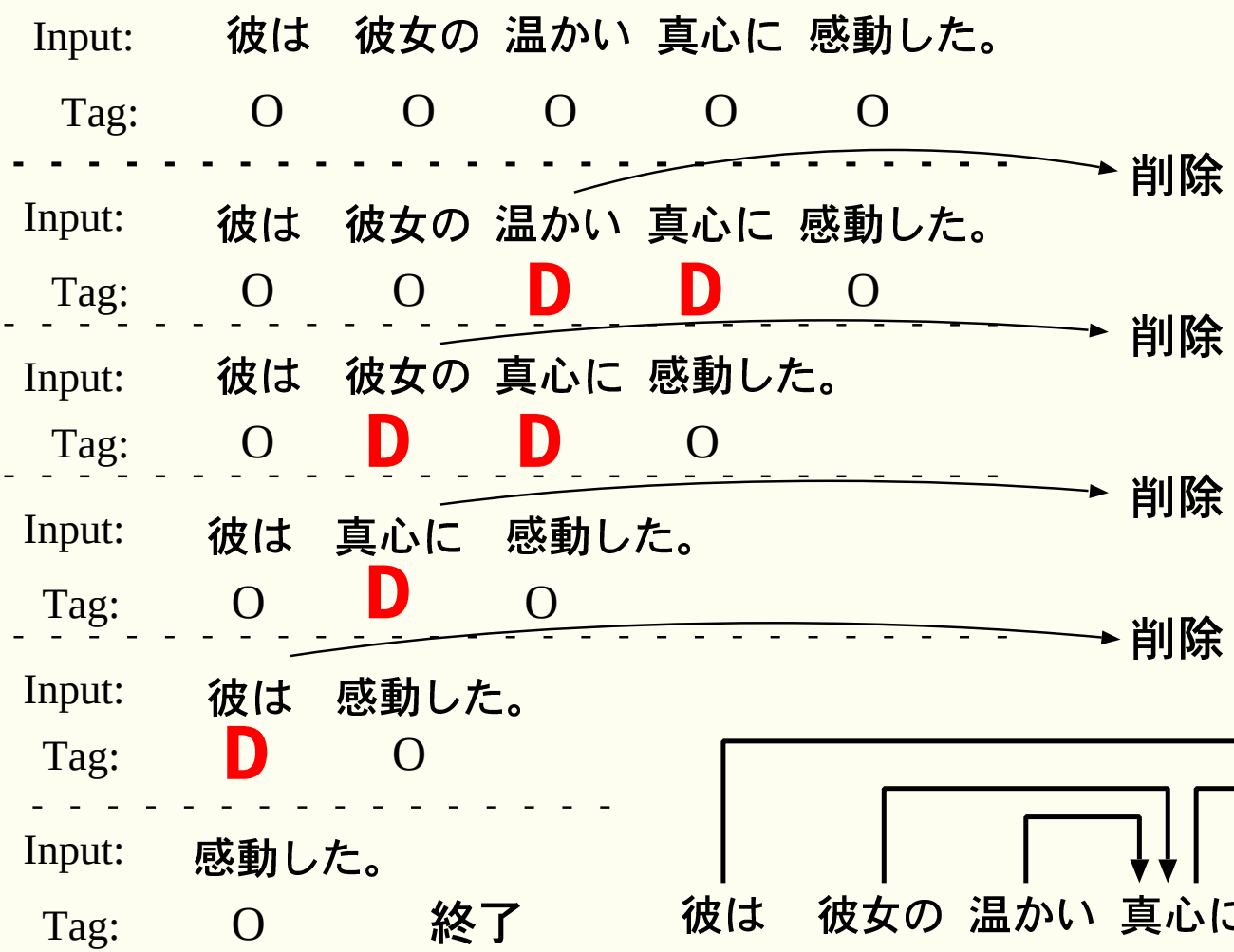


カスケードモデル (1/2)

[工藤 松本 2001]

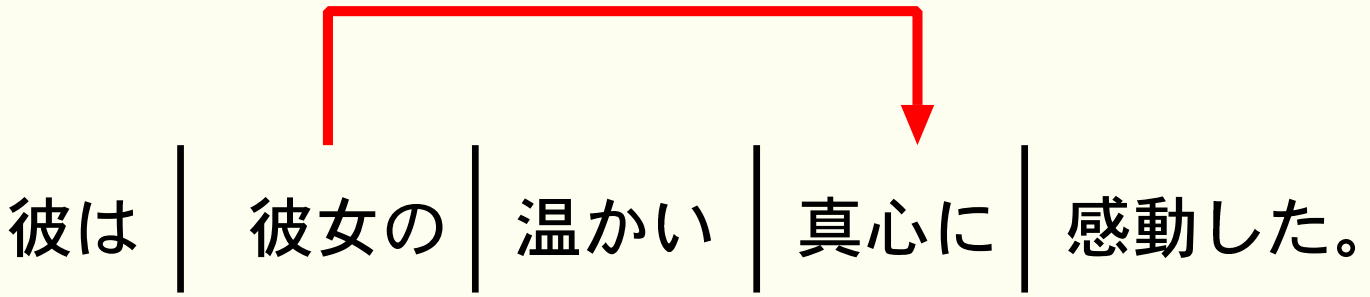
- 直後の文節に係るか係らないかという観点で決定的に解析
- アルゴリズムが簡潔, 実装が容易, 高効率
- 従来法以上の性能 (係り受け正解率 **89 - 90%**)
- 確率値や尤度は必要ない. 二値分類が行なえる学習アルゴリズムならあらゆるものが適用可能 (実際には, SVM を使用)
- 文頭からの自然な解析

カスケードモデル (2/2)

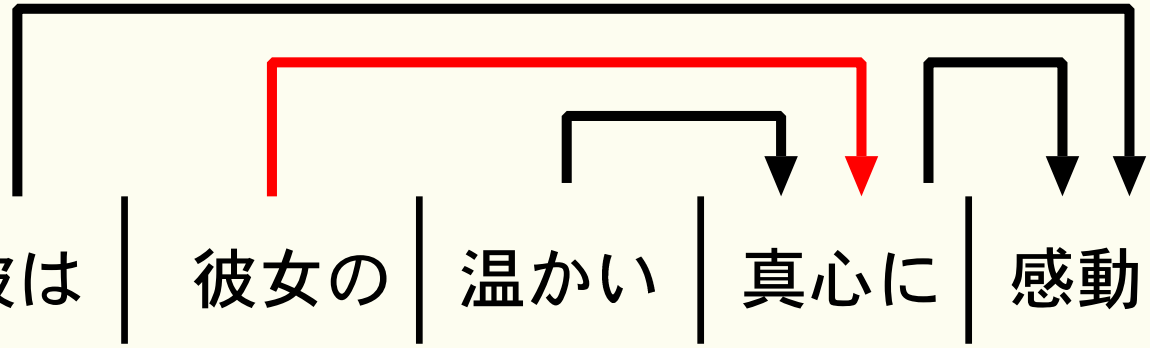

Input: 彼は 彼女の 温かい 真心に 感動した。
Tag: O O O O O
削除
Input: 彼は 彼女の 温かい 真心に 感動した。
Tag: O O D D O
削除
Input: 彼は 彼女の 真心に 感動した。
Tag: O D D O
削除
Input: 彼は 真心に 感動した。
Tag: O D O
削除
Input: 彼は 感動した。
Tag: D O
Input: 感動した。
Tag: O
終了
彼は 彼女の 温かい 真心に 感動した。

カスケードモデルの拡張 (1/2)

部分的な係り関係を制約として付与



制約を考慮しながら解析



カスケードモデルの拡張 (2/2)

拡張目的: 「他のシステムとの柔軟な結合」

- カスケードモデルの弱点
  - トップダウンの情報を考慮できない
  - 並列構造に弱い
- これらの弱点を補完するシステムと融合
  (ルールベースシステム, トップダウンパーザ, 人手処理)
- 複文を前処理で切り分け, 文法による制約..

拡張の具体的な方法 (1/2)

- 各文節毎にスタックを作成
  - 各スタックは, その文節が係りうる文節を近い順に保持
  - 制約有りと無しで, スタックの内容が変わる
- スタックトップが直後の文節の場合, 推定
- 係る場合は D タグを付与, スタックを空にする
- トップして空になる場合は, 無条件に係ける
- 削除の際に, スタックの中の候補も同時に削除

拡張の具体的な方法 (2/3)

Input: 彼は 彼女の 温かい 真心に 感動した。
Tag: O O O O O

Input: 彼は 彼女の 温かい 真心に 感動した。
Tag: O O **D** **D** O

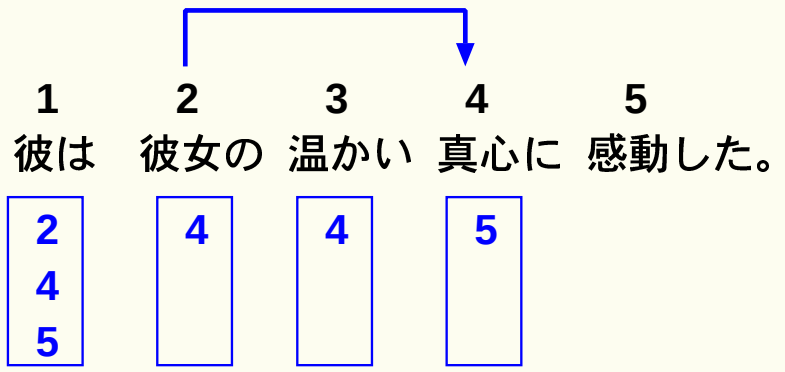
Input: 彼は 彼女の 真心に 感動した。
Tag: O **D** **D** O

Input: 彼は 真心に 感動した。
Tag: O **D** O

Input: 彼は 感動した。
Tag: **D** O

Input: 感動した。
Tag: O 終了

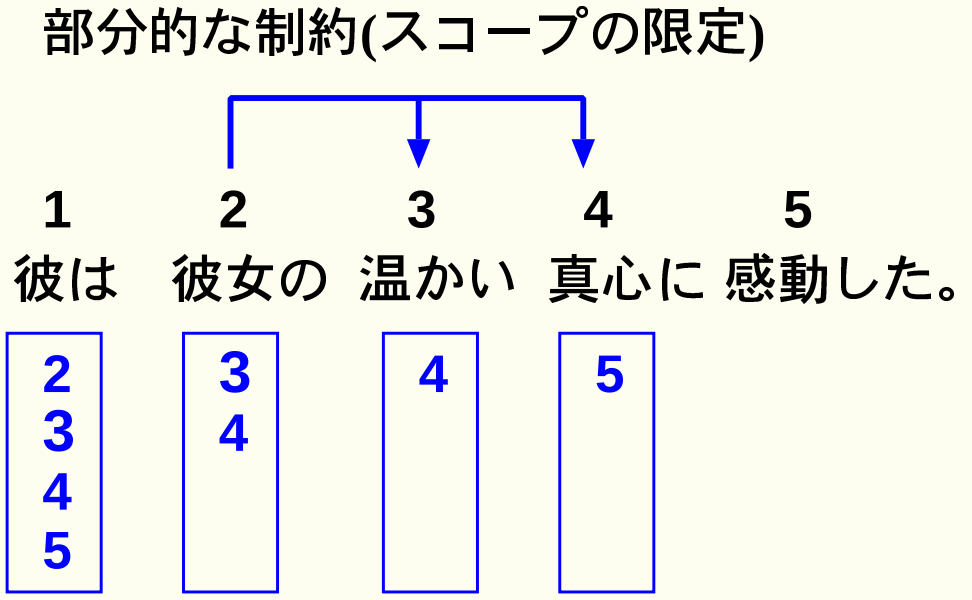
部分的な制約



候補を保持したスタックを作成

拡張手法の特徴

- 元のカスケードモデルを完全に包含
- 係り先スコープを限定するような制約でも動作可能



部分解析済み情報の利用 (1/2)

任意の係り関係を制約として追加できるが...

- どの係り関係を優先的に追加すればよいか?
- 追加することで, どの程度精度向上するか?

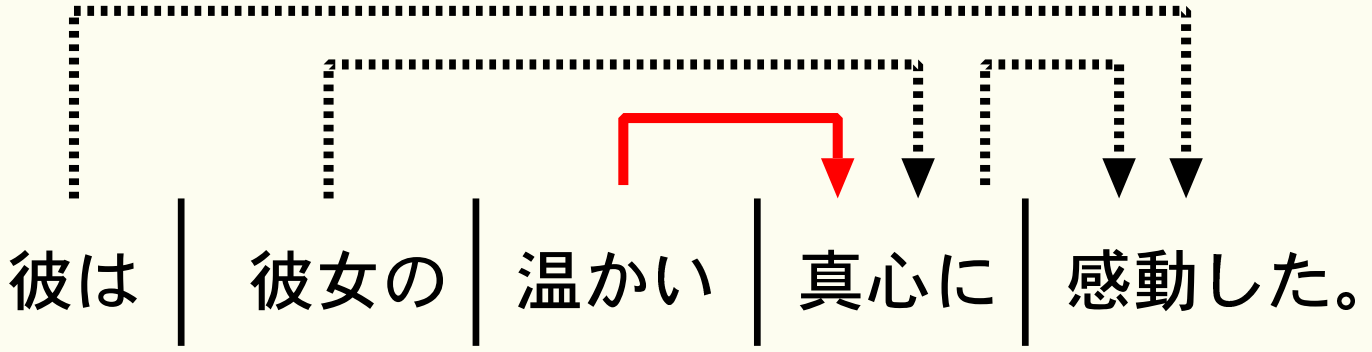
これらの検証実験を行う

部分解析済み情報の利用 (2/2)

具体的な実験手法

- 係り関係の推定が**困難**であると判定された文節を１つだけ決定し, その文節の真の係り先を制約として与える
- 困難さの基準
  - 無作為 (ベースライン)
  - 係り先の分散

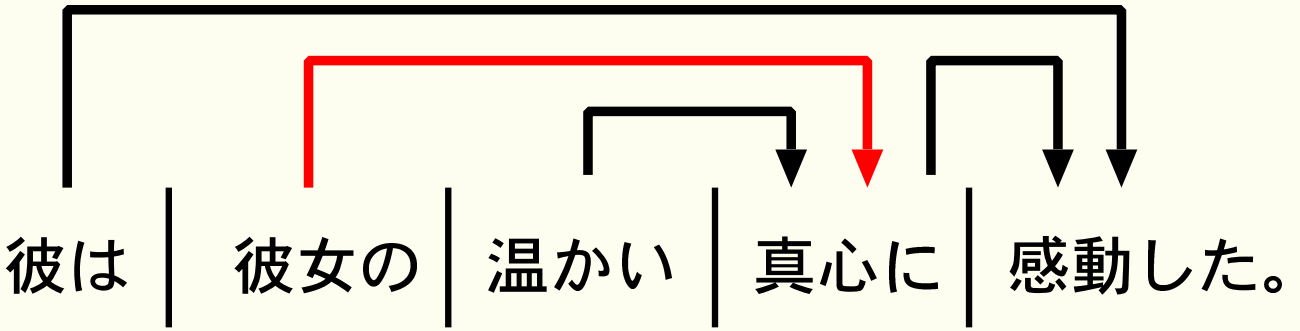
無作為に選択



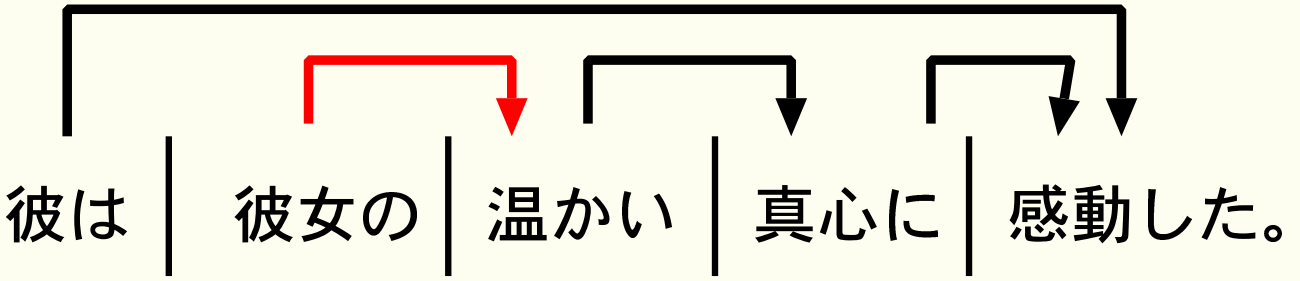
ランダムに一つ選択

係り先の分散

1



2



…

冗長解析の結果、
係り先の分散が一番大きい文節を選択

実験設定

- 京大コーパス Version 2.0
  - 学習データ: 1/1 - 1/8 (7958 文)
  - テストデータ: 1/9 (1246 文)
- 冗長係り受け解析器に [工藤 松本 2000] の SVM Parser を使用
  (確率推定に SVM を適用した Parser)

実験結果

	追加前	無作為	分散 (2)	分散 (10)
文節正解率 (%)	89.29	90.38	92.57	92.23
文正解率 (%)	47.53	52.08	61.56	57.86
正解に変更	-	116	337	290
他に好影響	-	11	38	46
他に悪影響	-	5	6	5

考察

- 無作為に選択するより, 係り先の分散が大きい文節を選択する方が良い
- 89.29% → 92.57%, 十分な精度向上
- 係り受けの困難さの判定は, 上位 2 位の結果を使うだけで充分
- 正解を与えなかった他の係り受け関係に対しても, 良い影響を与える

まとめ

- 部分的に係り関係が付与された状態から解析できるようカスケードモデルを拡張
- 冗長解析結果を用い, 係り先の推定が困難だと思われる文節を自動的に決定し, 制約として与える実験を行う
- 無作為選択より, 係り先の分散に基づく選択が精度向上の度合が大きい

今後の課題

- 個々の事例の詳細な調査
- 学習事例の選択に有効か実験, 検証
- 人手によるタグ付与の負荷を軽減できるか検証
- カスケードモデルの弱点を補完する具体的なシステムの提案(トップダウンパーザ など)
- Co-Training の可能性

本システムは, フリーソフトとして公開しています

http://cl.aist-nara.ac.jp/~taku-ku/