文の分割に基づく独話の係り受け解析

大野誠寛 （名古屋大学）

背景

独話データへの効率的なアクセスや
その効果的な再利用を実現するために，
独話の構造解析技術の開発が望まれている

これまでに，**節境界に基づく独話文係り受け解析手法[Ohno et al. 2006]**を提案
- 1文の長さが長いという特徴を持つ独話文の高性能な係り受け解析を実現

動機

節境界に基づく独話文係り受け解析手法では
係り受け解析の処理単位に**節境界単位**を採用
- 節境界単位で係り受けが閉じていることを仮定して係り受け解析を実行

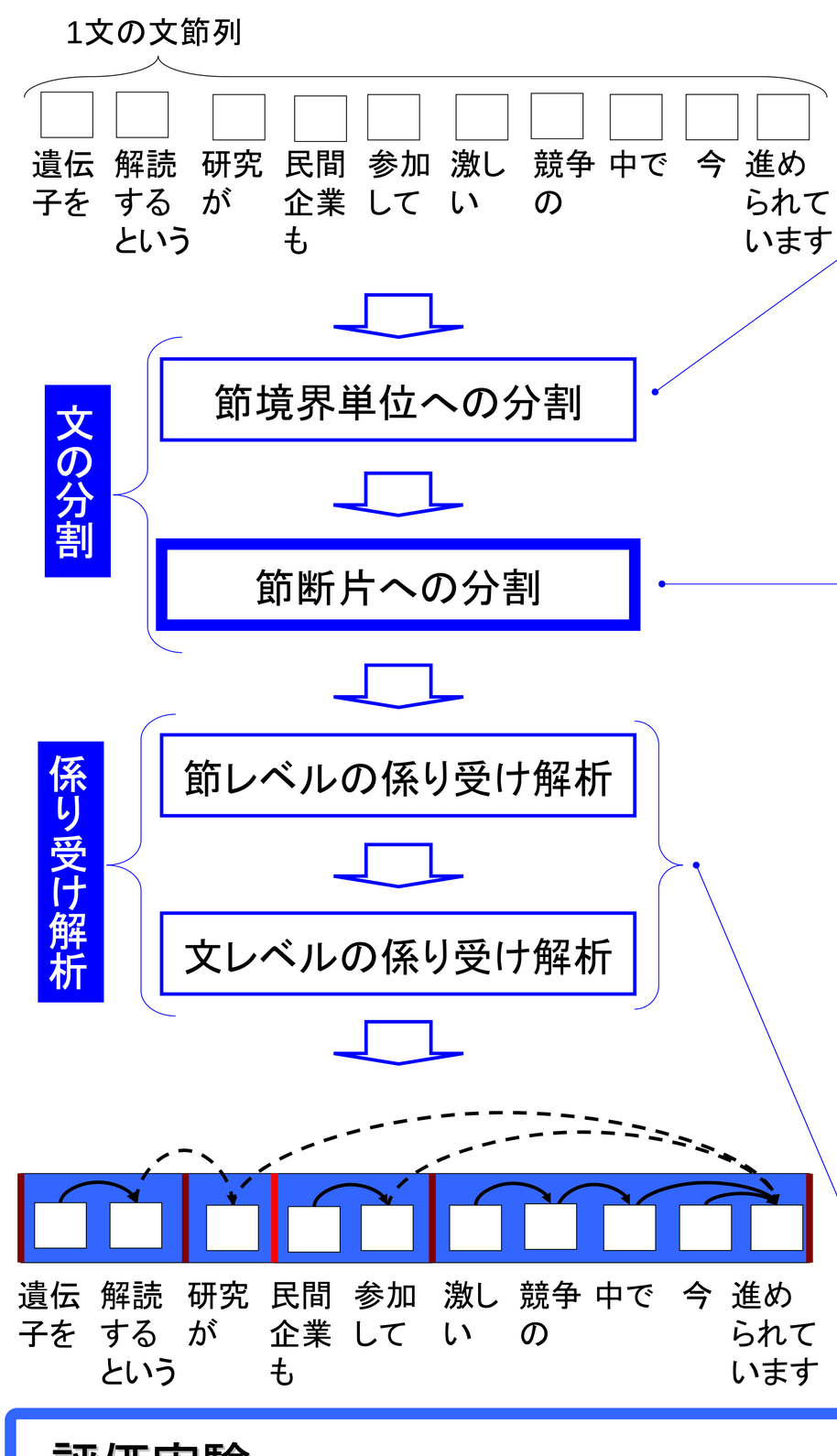
節境界単位で閉じていないものが実際には存在
これまでは解析できていなかった

目的

節境界に基づく独話文係り受け解析手法[Ohno et al. 2006]を拡張した，より高精度な係り受け解析手法を提案

係り受け解析の処理単位を修正
- 節境界単位で閉じてない係り受けを解析可能に

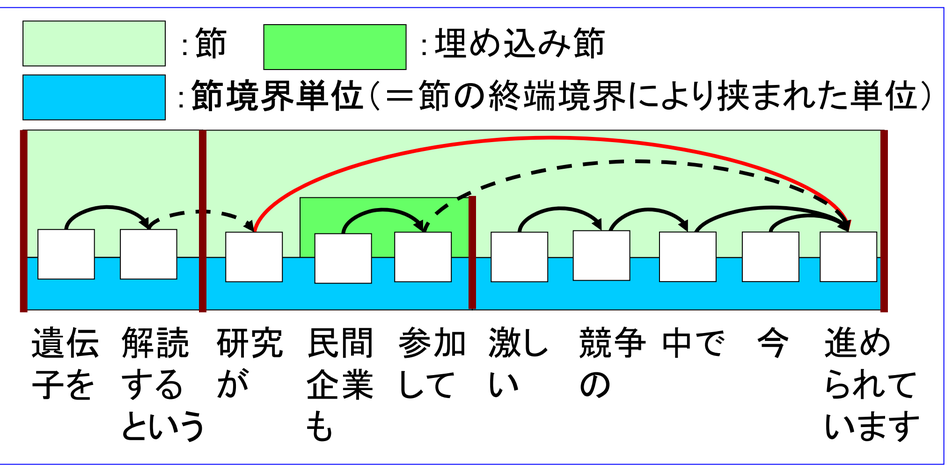
文の分割に基づく係り受け解析



節境界単位への分割

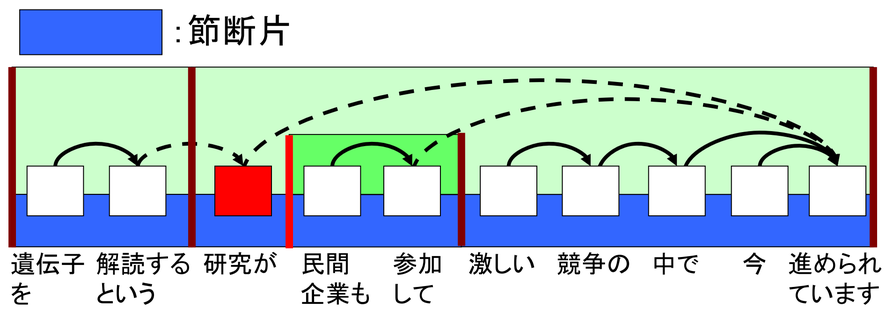
節の終端境界により文を分割
- 節境界解析ツールCBAP [丸山ら 2004]により検出

CBAP: 局所的な形態素列のみを手がかりに，節の終端境界と種類を特定



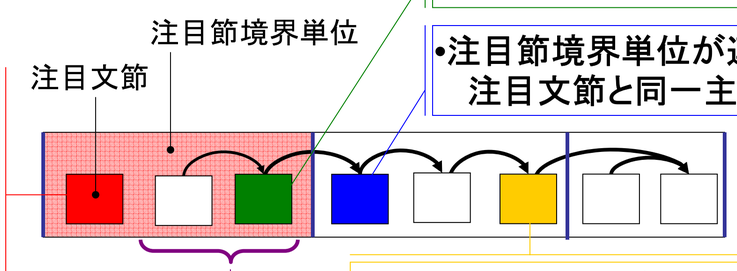
節断片への分割

最大エントロピー法により，係り先が節境界単位外に存在する文節を検出し，その直後で節境界単位を再度分割．
（この分割単位を節断片と呼ぶ）



最大エントロピー法で利用した素性

- 主辞の基本形
- 主辞の品詞(大分類,細分類)
- 語形の出現形
- 語形の品詞(大分類,細分類)
- 助詞1の出現形
- 助詞1の品詞細分類
- 助詞2の出現形
- 助詞2の品詞細分類
- 直後にポーズがあるか否か
- 所属する節境界単位の種類



- 直後にポーズがあるか否か
- 注目節境界単位が連体節である場合のみ，注目文節と同一主辞であるか否か
- 注目節境界単位が主題ハorテ節，かつ，注目文節が述語に係りえる文節の場合のみ，注目節境界単位の最終文節よりも高い係り受け確率を持つか否か
- 注目文節と同じ格を持つ文節が 存在するか否か

節境界に基づく係り受け解析

節レベルと文レベルの二段階で解析
- 各レベルの解析ではMEに基づく統計的手法[内元ら 1999]を利用

1. 節レベルの係り受け解析
   - 節断片の内部の係り受け構造を解析
2. 文レベルの係り受け解析
   - 節断片の最終文節の係り先を解析

評価実験

- NHKの解説番組「あすを読む」の書き起こしデータを用いて，以下の3つの手法により解析
  - 節断片に基づく係り受け解析手法
  - 節境界単位に基づく係り受け解析手法
  - 文単位の係り受け解析手法

実験で使用したデータ「あすを読む」

	テストデータ	学習データ
文数	500	5,532
節境界単位数	2,237	26,318
文節数	5,298	65,762
形態素数	13,342	165,173

各手法の係り受け解析結果

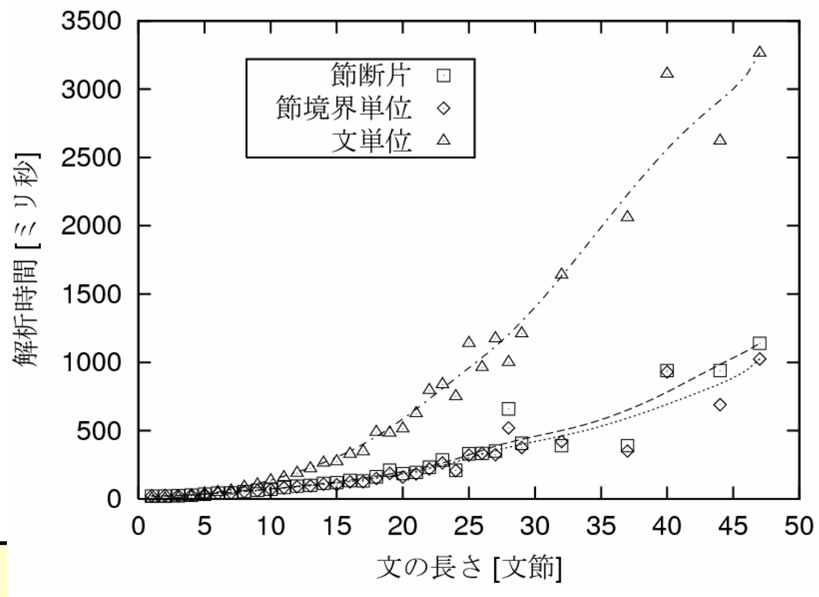
	節断片	節境界単位	文単位
節内部	**91.4% (2,798/3,061)**	90.2% (2,762/3,061)	90.1% (2,759/3,061)
節末文節	**75.8% (1,317/1,737)**	75.8% (1,317/1,737)	75.4% (1,309/1,737)
全体	**85.8% (4,115/4,798)**	85.0% (4,079/4,798)	84.8% (4,068/4,798)

節境界単位で閉じていない係り受けに対する各手法の係り受け解析結果

	節断片	節境界単位	文単位
再現率	28.9% (44/152)	1.3% (2/152)	40.8% (62/152)
適合率	53.7% (44/82)	11.8% (2/17)	40.3% (62/153)

係り先が節境界単位外に存在する文節の検出結果

再現率	48.7% (74/152)
適合率	58.7% (74/126)



文の長さと解析時間の関係

解析単位に節断片を用いることにより，節境界単位を用いた場合と同程度の解析速度を維持しつつ，解析精度を改善できることを確認