ポーズを考慮した文の分割に基づく独話文の係り受け解析

大野 誠寛† 松原 茂樹‡§ 柏岡 秀紀§♭ 稲垣 康善♮

†名古屋大学大学院情報科学研究科 ‡名古屋大学情報連携基盤センター
§情報通信研究機構 ♭ATR 音声言語コミュニケーション研究所 ♮愛知県立大学情報科学部
ohno@el.itc.nagoya-u.ac.jp

1 はじめに

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．独話構造解析の要素技術として，これまでに著者らは，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するために，節境界に基づく係り受け解析手法を提案している [1]．この手法では，節レベルと文レベルの二段階で係り受け解析を行う．まず，節境界解析により，文を節に分割し，各節に対して係り受け解析を行うことにより，節内の係り受け関係を同定する．次に，節末文節の係り先を定め，文全体の係り受け構造を作り上げる．解析実験により，独話解析において節を単位とする効果を確認している．

本稿では，この節境界に基づく独話文係り受け解析手法を拡張した，より高精度な係り受け解析手法を提案する．これまでの手法では，節に相当する単位として，節の終端境界により挟まれた単位（**節境界単位**）を解析の処理単位として利用してきた．具体的には，節境界単位は，節同様，その内部で係り受けが閉じていることを仮定し，この単位ごとに係り受け解析を実行していた．しかし，実際には節境界単位で閉じていない係り受けが存在しており，これらを解析することができなかった．本手法では，ポーズと節境界タイプを考慮して節境界単位で閉じていない係り受けの係り文節を事前に検出し，この文節の直後で再度，文を分割した単位を解析の処理単位とすることにより，このような係り受けを解析可能にする．独話データを用いた実験の結果，本手法により，節境界をまたぐ係り受けが同定できるようになり，解析精度が改善することを確認した．

2 独話文の解析単位

本研究では，文より短い単位である節を解析単位とすることにより，解析を効率化する．一文が長い独話文では，係り受け関係の探索範囲が狭められ，解析時間を短縮することができる．

2.1 節と節境界単位

節とは，述語を中心としたまとまりであり，複文や重文の場合，文は複数の節から構成される．さらに，節は，統語的，意味的にまとまった単位であるため，文に代わる解析単位として利用できる．しかし，複文において従属節が主節に埋め込まれる場合など，構文的な解析の前処理として節を検出することは必ずしも容易ではない．

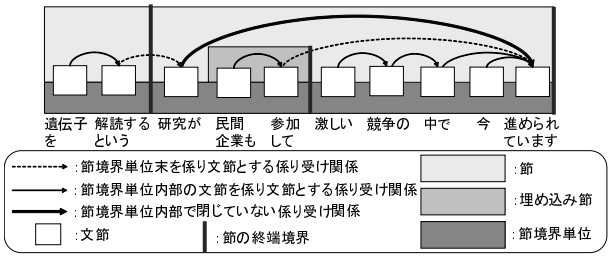

図 1: 節と節境界単位

そこで，本研究では，節境界解析 [2] を用いることにより，節への分割を近似的に実現してきた．節境界解析では，局所的な形態素列のみを手がかりとして，節の終端境界と種類を特定することができる．この解析により検出される節の終端境界により挟まれた単位を**節境界単位**[1]とよび，これを新たな解析単位として考えてきた．節境界単位は，ほとんどの場合，節と一致する．

しかし，図１に示すように，埋め込み節がある場合，節境界単位は節と一致しない．この例では，節「遺伝子を解読するという」と「民間企業も参加して」の終端境界に挟まれた「研究が民間企業も参加して」が節境界単位となるが，実際は「民間企業も参加して」で節を形成しており，節と節境界単位が一致していない．

2.2 節境界単位の拡張

これまで著者らは，節境界単位を係り受け解析の処理単位として採用してきたが [1]，節と節境界単位が一致しない場合，係り受けが節境界単位内部で閉じないという問題があった．これは，節の始端境界を無視して節境界単位が同定されるために，もともと異なる節を構成する文節同士が，同じ節境界単位にまとめられることにより生じていると考えられる．図１では，文節「研究が」と「進められています」の係り受け関係が節境界単位「研究が民間企業も参加して」の内部で閉じていない．これは，本来，節「研究が激しい競争の中で今進められています」を構成する文節である「研究が」が，埋め込み節「民間企業も参加して」と同じ節境界単位を構成することになったために生じている．

そこで，本手法では，節境界単位を拡張し，節の終端境界だけでなく，節の始端境界でも文を分割し，これらの境界によって挟まれた単位を**拡張節境界単位**と呼び，

[1]節境界単位の終端境界に付与されたラベル名をその節境界単位の種類とする．

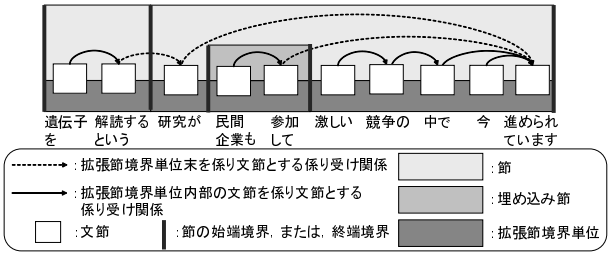

図 2: 節と拡張節境界単位

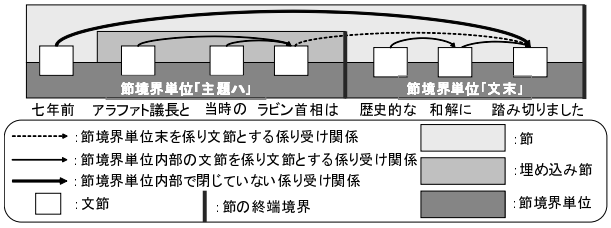

図 3: 節境界単位「主題ハ」で閉じていない係り受け

これを新たな解析単位として採用する．拡張節境界単位では，その内部で係り受けが必ず閉じている．本手法では,「独話文は一つ以上の拡張節境界単位の連接であり，各拡張節境界単位を構成する文節は，節境界単位の最終文節を除き，その節境界単位の内部の文節に係る」とみなして，係り受け解析を実行する．

例として，独話文「遺伝子を解読するという研究が民間企業も参加して激しい競争の中で今進められています」の係り受け構造を図 2 に示す．この文は 4 つの拡張節境界単位「遺伝子を解読するという」,「研究が」,「民間企業も参加して」,「激しい競争の中で今進められています」から構成され，各拡張節境界単位が係り受け構造を形成し，それらが拡張節境界単位の最終文節からの係り受け関係でつながっている．

3 拡張節境界単位への分割

形態素解析及び文節まとめ上げが施された独話文を入力とし，文中の全ての節の始端境界と終端境界を検出することにより，拡張節境界単位を以下の手順により同定する．

1. 節境界単位の同定
   節境界解析ツール CBAP[2] を用いて入力文に対して節の終端境界を検出し[2]，節境界単位を同定する．

2. 節境界単位の分割
   節境界単位で閉じていない係り受けの係り文節を検出し，この文節の直後で節境界単位を再度分割することにより，拡張節境界単位を同定する．

以下では，拡張節境界単位を同定するため，節境界単位で閉じていない係り受けの係り文節を検出する方法について説明する．検出は以下の手順により行う．

1. ポーズによる再分割
   200ms 以上のポーズがある文節境界で節境界単位を再分割する．

2. 述語なし節境界単位の再分割
   述語を持たない節境界単位内で閉じていない係り受けの係り文節を検出するルールにより，述語を持たない節境界単位を再分割する．

[2]CBAP では，統語的に大きな切れ目になると考えられる「主題ハ」や「談話標識」など,「述語を中心としたまとまり」という節の定義に逸脱した境界も一部検出する [2]．本研究ではこれらも節の終端境界として扱う．

表 1: 述語に係る文節の最終形態素の品詞

品詞	品詞細分類
助詞	格助詞-一般，格助詞-引用，格助詞-連語，係助詞 副助詞，副詞化
名詞	副詞可能，非自立-副詞可能，非自立-助動詞語幹 接尾-副詞可能，接尾-助数詞
副詞	一般，助詞類接続

3.1 ポーズによる分割

2.2 節で述べたように，節境界単位で閉じていない係り受けは，係り受け関係が埋め込み節をまたぐことにより生じるため，その係り受け距離が長くなる．したがって，埋め込み節の直前の文節の直後には，この文節の係り先が遠く離れることを示唆するため，ポーズが入りやすいと考えられる．そこで，節境界単位内部に存在する，直後に 200ms 以上のポーズがある文節を節境界単位内で閉じていない係り受けの係り文節として検出し，このポーズがある文節境界により節境界単位を再分割する．なお，新たに分割された，ポーズの直前の拡張節境界単位の種類名を「PAU」とした．

3.2 述語なし節境界単位の分割

節境界単位「主題ハ」など述語を持たない節境界単位[3]では，図 3 に示すように，述語に係りやすい文節（「七年前」）は，それが所属する節境界単位の外側に位置する述語（「踏み切りました」）に係る現象が多く見られた．したがって，これらを検出するには，述語なし節境界単位の中で，述語を修飾する文節を検出すればよい．

そこで，文節が述語に係るか否かを判定するために，文献 [3, 4] を参考に，述語に係る文節の最終形態素の品詞を定めた．その一覧を表 1 に示す[4]．表 1 の品詞と，文節の最終形態素の品詞が一致するとき，その文節は述語に係る文節であると判定し，この文節の直後で述語なし節境界単位を再分割する．このルールを再帰的に適用し，最終的に拡張節境界単位が同定する．なお，この文節を含む拡張節境界単位の種類名を「述語修飾」とし，この拡張節境界単位も述語を持たないものとして扱う．

[3]述語を持たない節境界単位には,「主題ハ」,「体言止」,「感動詞」,「談話標識」,「間投句」,「PAU」,「述語修飾」がある．

[4]この他の品詞として，感動詞や接続詞などがあるが，これらは CBAP により別の節境界単位になるので，ここには含めていない．

4 節境界に基づく係り受け解析

本手法では，形態素解析，文節まとめ上げ，及び節境界解析が施された文を入力とする．また，この手法では，係り受けの後方修飾性，係り先の唯一性，非交差性の 3 つの性質を絶対的制約とする．解析の手順は以下の通りである．

1. 節レベルの係り受け解析
   1 文中のすべての拡張節境界単位に対して，その内部の係り受け構造を解析する．
2. 文レベルの係り受け解析
   1 文中のすべての拡張節境界単位に対して，その最終文節の係り先を解析する．

なお，以下では，1 独話を構成する拡張節境界単位列を $C_1 \cdots C_m$，拡張節境界単位 C_i を構成する文節列を $b_1^i \cdots b_{n_i}^i$，文節 b_k^i を係り文節とする係り受け関係を $dep(b_k^i)$，1 独話の係り受け構造を $\{dep(b_1^1), \cdots, dep(b_{n_m-1}^m)\}$ と記す．

4.1 節レベルの係り受け解析

節レベルの係り受け解析では，拡張節境界単位 C_i 中の文節列 $b_1^i \cdots b_{n_i}^i$ を B_i とするとき，$P(S_i|B_i)$ を最大にする係り受け構造 $S_i(=\{dep(b_1^i), \cdots, dep(b_{n_i-1}^i)\})$ を求める．ここでは，拡張節境界単位の最終文節 $b_{n_i}^i (1 \leq i \leq m)$ の受け文節は決定しない．

係り受け関係は互いに独立であると仮定すると，$P(S_i|B_i)$ は以下の式で計算できる．

$$P(S_i|B_i) = \prod_{k=1}^{n_i-1} P(b_k^i \overset{rel}{\rightarrow} b_l^i|B_i) \quad (1)$$

ここで，$P(b_k^i \overset{rel}{\rightarrow} b_l^i|B_i)$ は，入力文節列 B_i が与えられたときに，文節 b_k^i が b_l^i に係る確率を表す．最尤の係り受け構造は，式 (1) の確率を最大とする構造であるとして動的計画法を用いて計算する．

次に，$P(b_k^i \overset{rel}{\rightarrow} b_l^i|B_i)$ の計算について述べる．$P(b_k^i \overset{rel}{\rightarrow} b_l^i|B_i)$ は，内元ら [5] の係り受け確率モデルを用いて最大エントロピー法により推定した．用いた素性は，内元らの手法 [5] とほぼ同様であるが，話し言葉を対象としているため，読点や括弧の素性は取り除いている．また，節レベルの解析では，内元らの手法で利用されている句点の情報を節末か否かという情報に置き換えた．

4.2 文レベルの係り受け解析

拡張節境界単位の最終文節の受け文節を同定する．1 文の文節列を $B(=B_1 \cdots B_m)$ とし，拡張節境界単位の最終文節を係り文節とするような係り受け構造 $\{dep(b_{n_1}^1), \cdots, dep(b_{n_{m-1}}^{m-1})\}$ を S_{last} とするとき，$P(S_{last}|B)$ を最大とする S_{last} を求める．$P(S_{last}|B)$ は以下の式で計算できる．

$$P(S_{last}|B) = \prod_{i=1}^{m-1} P(b_{n_i}^i \overset{rel}{\rightarrow} b_l^j|B) \quad (2)$$

表 2: 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
節数	2,237	26,318
文節数	5,298	65,762
形態素数	13,342	165,173

表 3: 3 手法の実験結果（平均解析時間:ミリ秒/文）

	拡張節境界単位係り受け解析	節境界単位係り受け解析	文単位係り受け解析
解析時間	100.4	85.7	205.1

注) 実装言語:LISP, 使用計算機:Pentium 4 2.4GHz, Linux

ここで，$P(b_{n_i}^i \overset{rel}{\rightarrow} b_l^j|B)$ は，1 文の文節列 B が与えられたときに，C_i の最終文節 $b_{n_i}^i$ が b_l^j に係る確率を表し，4.1 節と同様に最大エントロピー法を用いて計算する．文レベルの解析では，節レベルの解析で利用した素性に，文末か否かの素性を付け加えた素性を利用した．最尤の係り受け構造は，式 (2) の確率を最大とする構造であるとして動的計画法を用いて計算する．

5 解析実験

独話文の係り受け解析における本手法の有効性を評価するため，解析実験を行った．

5.1 実験に使用したデータ

実験で使用したデータを表 2 に示す．テストデータとして，NHK の解説番組「あすを読む」の書き起こしデータに形態素解析，文節まとめ上げを施した 500 文を用いた．正解の節境界，及び，係り受けは人手で付与した [1]．なお，節境界単位で閉じていない係り受け関係は，テストデータの正解中に 152 個存在した．一方，学習データには，形態素解析，文節まとめ上げ，節境界解析，係り受け解析が施された「あすを読む」の書き起こし 5,532 文を用いた．

5.2 実験の概要

本手法の有効性を比較評価するために，上述したデータを用いて以下の 3 つの手法で解析を行い，それぞれの解析時間と解析精度を求めた．

- **拡張節境界単位に基づく係り受け解析手法**
  3 節，4 節でそれぞれ述べた，拡張節境界単位への分割，係り受け解析を順に行う．
- **節境界単位に基づく係り受け解析手法**
  上述の手法のうち，3 節で述べた拡張節境界単位への再分割は行わず，節境界単位を解析単位として，係り受け解析を行う．
- **文単位の係り受け解析手法**
  上述の手法のうち，節境界単位への分割を行わず，文を解析単位として，文全体の係り受け構造を一度に求める．

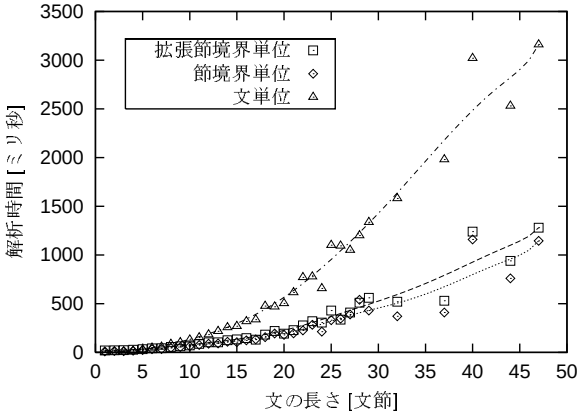

図 4: 文の長さと解析時間の関係

表 4: 3 手法の実験結果（係り受け正解率）

	拡張節境界単位 係り受け解析	節境界単位 係り受け解析	文単位 係り受け解析
節内部	91.5% (2,801/3,061)	90.7% (2,775/3,061)	90.0% (2,755/3,061)
節末文節	75.2% (1,307/1,737)	75.2% (1,307/1,737)	75.0% (1,303/1,737)
全体	85.6% (4,108/4,798)	85.1% (4,082/4,798)	84.6% (4,058/4,798)

なお，学習のための最大エントロピー法のツールとしては，文献 [6] のものを利用した．

5.3 実験結果

各手法の解析時間を表 3 に示す．拡張節境界単位と節境界単位の両解析手法の解析速度は文単位解析手法に比べて，平均して約 2 倍向上した．また，拡張節境界単位と節境界単位の両解析手法の解析時間にはほとんど差がなかった．文の長さと解析時間の関係を図 4 に示す．文単位解析手法では文の長さが 10 文節を超えたあたりから，急激に解析時間が上昇するのに対し，拡張節境界単位と節境界単位の両解析手法の解析時間の変化は小さい．実験で使用した 6032 文の平均文節数は 11.8 であり，平均以上の長さをもつ独話文に対する拡張節境界単位と節境界単位の両解析手法の効果を確認した．

各手法の係り受け正解率を表 4 に示す．表 4 の第 1 行は，節末文節を除く節内の全ての文節に対する正解率を，第 2 行は，文末を除く全ての節末文節に対する正解率を示す．節内部，節末文節とも，文単位解析手法に劣らない解析精度を，節境界単位，拡張節境界単位の両解析手法が備えていることがわかる．ここで，節境界単位で閉じていない係り受けの係り文節の検出結果を表 5 に示す．節境界単位で閉じていない係り受けの係り文節の検出精度はあまり高くなく，拡張節境界単位をより正確に同定することが望まれる．しかし，節境界単位を現状のルールで拡張することによっても，拡張節境界単位の解析手法は，節境界単位の解析手法と比べ，正解率がわずかに増加した．節境界をまたぐ係り受けに対する解析結果を表 6 に示す．節境界単位を拡張することによって，節境界単位で閉じていない係り受けに対する解析精度も改善している．

表 5: 節境界単位で閉じていない係り受けの係り文節の検出結果

再現率	63.8% (97/152)
適合率	31.2% (97/311)

表 6: 節境界単位で閉じていない係り受けに対する 3 手法の実験結果

	拡張節境界単位 係り受け解析	節境界単位 係り受け解析	文単位 係り受け解析
再現率	28.9% (44/152)	1.3% (2/152)	40.8% (62/152)
適合率	55.7% (44/ 79)	16.7% (2/ 12)	38.3% (62/162)

以上の結果から，拡張節境界単位の解析手法によって，節境界単位の解析手法の解析速度を同程度に維持しつつ，解析精度を改善できることを確認した．

6 おわりに

本稿では，節境界単位に基づく係り受け解析を拡張し，これまでは解析できなかった節境界単位で閉じていない係り受け関係をも解析可能な手法を提案した．解析実験の結果，本拡張手法によって，節境界単位に基づく係り受け解析手法の解析速度を同程度に維持しつつ，解析精度を改善できることを確認した．今後は，節境界単位で閉じていない係り受けの係り文節をより正確に検出するため，述語なし以外の節境界単位で閉じてない係り受けの係り文節を検出する手法について検討したい．

謝辞 本研究は，総務省戦略的情報通信研究開発推進制度の研究委託「講演など独話データの知的構造化に関する研究開発」，ならびに，科学研究費補助金（特別研究員奨励費）「大規模音声言語コーパスを用いた独話データの構造化とその応用に関する研究」（課題番号 18・6433）により実施したものである．

参考文献

[1] T. Ohno, S. Matsubara, H. Kashioka, T. Maruyama, and Y. Inagaki : Dependency Parsing of Japanese Spoken Monologue Based on Clause Boundaries, *Proc. of COLING/ACL2006*, pp. 169–176 (2006).

[2] 丸山 岳彦，柏岡 秀紀，熊野 正，田中 英輝: 日本語節境界検出プログラム CBAP の開発と評価，自然言語処理，Vol. 11，No. 3，pp. 39–68 (2004).

[3] 浅原 正幸，松本 裕治: IPADIC ユーザーズマニュアル，version2.5.1 (2002).

[4] 益岡 隆志，田窪 行則: 基礎日本語文法 - 改訂版 -，くろしお出版 (1992).

[5] 内元 清貴，関根 聡，井佐原 均: 最大エントロピー法に基づくモデルを用いた日本語係り受け解析，情報処理学会論文誌，Vol.40, No.9, pp. 3397–3407 (1999).

[6] L. Zhang: Maximum Entropy Modeling Toolkit for Python and C++, http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html