情報処理学会論文誌 Vol. 50 No. 2 553–562 (Feb. 2009)

節の始境界検出に基づく独話文の係り受け解析

大野誠寛[†1] 松原茂樹[†2,†3]
柏岡秀紀[†3,†4] 稲垣康善[†5]

1 文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，節に分割し，節レベルと文レベルの 2 段階で係り受け解析を実行する枠組みが提案されており，その有効性が確認されている．しかし，上述の枠組みにおいては，節そのものに文を分割することはできないため，節の終境界で挟まれた単位（節境界単位）を解析の処理単位として用いており，そこでは解析単位の内部で係り受けが閉じない場合があることが問題となっていた．本論文では，節レベルと文レベルの 2 段階で解析を実行する枠組みに基づいて，節境界単位を拡張した完全に係り受けが閉じた単位を解析の処理単位とする係り受け解析手法を提案する．本手法では，ポーズや節境界タイプを考慮して，機械学習により節境界単位で閉じない係り受けの係り文節を検出し，この直後で節境界単位を再分割することにより，係り受けが閉じた単位を同定する．この単位を解析の処理単位として利用することにより，解析精度が改善されることを確認した．

Dependency Parsing of Spoken Monologue Based on Clause-start Identification

Tomohiro Ohno,[†1] Shigeki Matsubara,[†2,†3]
Hideki Kashioka[†3,†4] and Yasuyoshi Inagaki[†5]

A dependency parsing method based on sentence segmentation into clauses has been proposed and confirmed to be effective. In this method, dependency parsing is executed in two stages: at the clause level and the sentence level. However, since a sentence can not be segmented into complete clauses, in the past research, a unit sandwiched between two clause-end boundaries (**clause boundary unit**) is adapted as an approximate unit of the complete clause. There has been a problem that the dependency structure of the clause boundary unit is not necessarily closed. This paper proposes a method for dependency parsing based on sentece segmentation into units which corresponds to clauses and whose dependency structure is completely closed (**clause fragment**). Our method identifies such the unit by redividing a clause boundary unit at modifier bunsetsus of dependency relations over clause-end boundaries. As the results of the experiment, we confirmed the improvement of the dependency parsing accuracy by utilizing the clause fragment unit as a parsing unit.

1. はじめに

音声ドキュメント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，独話音声を構造化し蓄積することが望ましく，その要素技術の 1 つとして独話文の高性能な構文解析が必要となる．独話は，対話に比べて，文が長くなる傾向があり，また，一般に文が長くなればなるほど構文的あいまい性が増加するため，従来の文単位での構文解析手法を独話文に適用すると，解析時間の増加や解析精度の低下が問題となる．

このような問題は，独話文の解析だけでなく，翻訳や要約など，様々な言語処理タスクにおいて同様に発生する．このために，文を分割し，処理を実行する手法がいくつか提案されており，その有効性が示されている[1)–9)]．これらの研究では共通して，文より短く，構文的にも意味的にもまとまった言語単位である節への分割が試みられている．

一般に，節を同定するには，その始境界および終境界を検出する必要がある．それらの検出に関する研究は，英語を対象にしていくつか行われている．たとえば，人手で作成した規則に基づく検出手法[1),2)] や機械学習に基づく検出手法[10)–12)] が提案されており，いずれも高い検出精度を達成している．一方，日本語の場合，節の終境界は，述語句が節の終端に配置されるため，述語の活用形や接続助詞の種類などを考慮することにより，かなりの精度で検出できるものの[13)]，節の始境界は，英語における関係代名詞のように，節の先頭に特定の単語が現れるわけではないため，その検出は容易ではない．そのため，日本語文の分割を

†1 名古屋大学大学院国際開発研究科
Graduate School of International Development, Nagoya University
†2 名古屋大学情報連携基盤センター
Information Technology Center, Nagoya University
†3 情報通信研究機構音声コミュニケーショングループ
NICT Spoken Language Communication Group
†4 ATR 音声言語コミュニケーション研究所
ATR Spoken Language Communication Research Laboratories
†5 豊橋技術科学大学
Toyohashi University of Technology

行う従来研究のほとんどが節の終境界に基づいて文を分割している[4)–9),⋆1]．しかし，この方法で分割された単位は，埋め込み節が存在する場合に，文の構文構造と整合しないという問題がある．

そこで本論文では，節の終境界だけでなく，始境界も検出することにより，節に相当し，かつ，係り受けが完全に閉じた単位に分割する，日本語独話文の係り受け解析手法を提案する．分割された単位で必ず係り受けが閉じるため，その内部の係り受け構造を独立に計算することができ，解析時間の増加や解析精度の低下といった問題を軽減できる．

本手法では，まず，形態素列のパタンマッチングに基づく節境界解析[13)]により節の終境界を網羅的に検出し，その後，機械学習により節の始境界を検出する．これら両境界により文を分割し，分割された単位ごとに係り受け解析を行う．最終的に，各分割単位の最終文節の係り先を解析することにより，文全体の係り受け構造を作り上げる．独話データを用いた実験の結果，本手法により，節の終境界のみに基づいて文を分割した手法と比べて，解析精度が改善されることを確認した．

本論文の構成は以下のとおりである．次章で独話文の解析単位について述べ，3 章で節の始境界の特徴について分析する．4 章で解析単位への分割手法について説明し，5 章で節境界に基づく係り受け解析手法を示す．6 章で解析実験について述べ，7 章で本論文のまとめと今後の課題を述べる．

2. 独話文の解析単位

本研究では，節の始境界および終境界に基づいて独話文を分割し，この分割単位を解析単位として 2 段階で係り受け解析を実行する．このため，解析単位は，文の構造的な情報を用いることなく検出できることが不可欠であり，そのためには，文を一次元的に分割できる必要がある．本章では，本研究で採用する解析単位について述べる．

2.1 節

節は，文より短く，その内部で係り受けが閉じているため，係り受け構造との親和性が高く，係り受け解析の解析単位として有望である．しかし，埋め込み節が存在する場合，文を節に一次元的に分割することはできない．例として，**図 1** に，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」の係り受け構造を示す．節「民間企業も参加して」が節「研究が激しい競争の中で今進められています」の中

⋆1 浜辺ら[14)]が節の始境界の検出を試みているものの，係り受け解析の結果を用いた手法となっている．

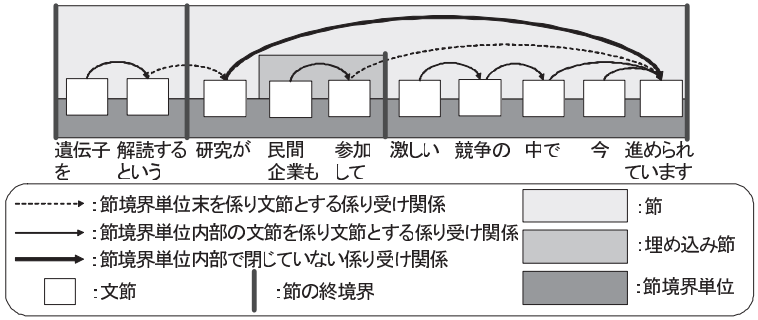

図 1　節と節境界単位
Fig. 1　Clauses and clause boundary units.

に埋め込まれている．節の埋め込みを検出するには，構造的な情報が不可欠であり，係り受け解析の前段階で節を同定することは難しい．

2.2 終境界のみに基づく解析単位

節の終境界を網羅的に検出し[13)]，これらの終境界によって挟まれた単位（以下，**節境界単位**）を解析単位として係り受け解析を実行する手法が提案されている[4)]．節境界単位は，節を近似した単位であるが，埋め込み節がある場合，節と一致せず，その内部で係り受けが閉じない．たとえば，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．このような単位を用いて係り受け解析を実行すると，「研究が」のように，係り先が単位の外側に位置する文節の解析を正しく行うことができない．

2.3 始境界と終境界を用いた解析単位

前節で述べたように，節境界単位は，埋め込み節がある場合，係り受けが閉じているという節の性質を満たさない．そこで本研究では，埋め込み節がある場合でも係り受けが閉じ，かつ，一次元的に分割可能な節相当単位を定義する．具体的には，節の終境界だけでなく，節の始境界によっても文を分割し，これら両境界によって挟まれた単位（以下，**節断片**）を新たな解析単位とする．**図 2** に節断片への分割例を示す．図 2 が示すように，埋め込み節がある場合でも，各節断片は，その内部で係り受けが必ず閉じる．これは，節の終境界だけでなく始境界も用いることにより，埋め込み節（「民間企業も参加して」）の範囲を特定で

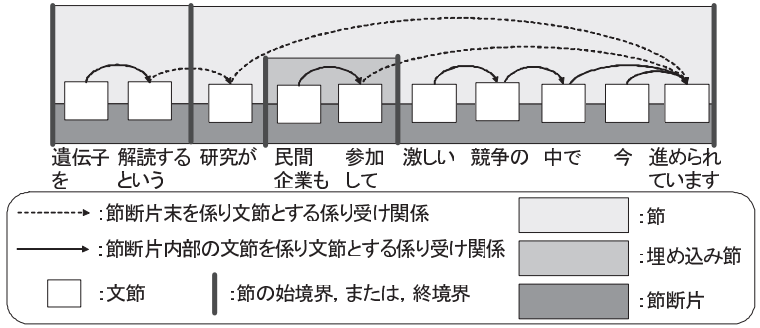

図 2　節と節断片
Fig. 2　Clauses and clause fragments.

き，埋め込み節を覆っている節（「研究が激しい競争の中で今進められています」）を，埋め込み節の前側の断片（「研究が」）と後側の断片（「激しい競争の中で今進められています」）に分割することができるためである．

なお，本研究では，「独話文は 1 つ以上の節断片の連接であり，各節断片を構成する文節は，節断片の最終文節を除き，その節断片の内部の文節に係る」として，係り受け解析を実行する．図 2 の独話文の場合，4 つの節断片「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，各節断片が係り受け構造を形成し，それらが節断片の最終文節からの係り受け関係によってつながっていると見なす．

3. 節境界単位で閉じていない係り受けの特徴

本研究では，文を節断片に分割するため，節の始境界，ならびに，終境界を検出する必要がある．このうち，節の終境界については，形態素列のパタンマッチングに基づく節境界検出プログラム CBAP [13] によって高い精度で網羅的に検出することができる．一方，節の始境界は，埋め込み節以外については，同時に左隣の節の終境界でもあるため，節の終境界が検出されれば自動的に検出できる．ただし，埋め込み節の始境界は，同時に節の終境界となることはなく，特定の単語を手がかりとした単純なパタンマッチングでは検出できない．

そこで本研究では，「節境界単位で閉じていない係り受けの係り文節」を同定することにより，埋め込み節の始境界の検出を試みる．これは，節境界単位で閉じていない係り受けは，埋め込み節が存在する場合にのみ生じ，その係り文節の直後が埋め込み節の始境界となるためである．以下では，そのような係り受けの特徴について分析した．

表 1　分析データ（「あすを読む」）
Table 1　Analysis data (“Asu-Wo-Yomu”).

項目	数値
番組数	6
文数	315
節数	1,612
文節数	4,017
形態素数	9,973
節境界単位で閉じていない係り受け数	140

3.1　分析に用いたデータ

分析には，NHK の解説番組「あすを読む」の書き起こしデータ[⋆1]に対して形態素解析，文節まとめあげ，節境界解析，係り受け解析を自動的に行い，人手で修正したものを用いた．ここで，形態素解析には ChaSen [15] を，文節まとめあげ，係り受け解析には CaboCha [16] を，節境界解析には CBAP [13] を用いた．なお，形態素解析は ChaSen [15] の IPA 品詞体系[17] に，文節まとめあげは CSJ 作成基準[18] に，節境界解析は丸山らの基準[13] に，係り受け文法は京大コーパスの作成基準[19] にそれぞれ準拠して人手により修正している．ただし，話し言葉特有の現象については，新たに作成基準を設けた．具体的には，話し言葉特有のいい回し表現（「こっから」，「という」など）については，新たな辞書項目を設けて，形態素ごとに品詞を定めた．また，文節まとめあげでは，形式名詞の前で一律に文節を区切る仕様とした．分析データの基礎統計を**表 1** に示す．以下では，節境界単位で閉じていない係り受けの係り文節を機械学習により検出する際に有益な素性を調査するため，節境界単位で閉じていない係り受け 140 個について詳しく分析する．

3.2　ポーズ情報

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．このため，そのときの係り受け関係は，係り文節と受け文節の距離が長くなる．このような構文的関係を示唆するため，埋め込み節の直前にポーズが入りやすいと考えられる．

実際，節境界単位で閉じていない係り受けの係り文節 140 個のうち，45.0%（63/140）は，

⋆1 ATR と NHK の共同研究において使用した．

直後に 200 ms 以上のポーズ[⋆1]が存在した．節境界単位で閉じている係り受けの係り文節（節境界単位末を除く）のうち，その後にポーズが存在する割合はたかだか 7.6%（172/2,265）であり，ポーズを考慮することにより，節境界単位で閉じていない係り受けの係り文節を検出できる可能性がある．

3.3　節境界単位の種類と受け文節の位置

節境界単位で閉じていない係り受けは，それが生じる節境界単位の種類によって，異なる傾向を持つことが分かっている[4]．そのため，節境界単位の種類ごとに，機械学習により検出する際の有益な素性は異なる．以下では，係り受けが閉じていない節境界単位を，その種類（節の終境界のラベル名）ごとに分類し，全体の 70.6%を占める上位 3 つの節境界単位，“連体節”，“主題ハ”，“テ節” を対象に分析を行う，節境界単位で閉じていない係り受けの傾向の違いを見るため，特に，その受け文節に着目し，特徴を分析した．

3.3.1　節境界単位の種類 “連体節”

節境界単位で閉じていない係り受けのうち 52 個は，節境界単位 “連体節” で生じていた．これらを調べてみると，大きく以下の 3 つに分類できることが分かった．

（1）「節境界単位内部の文節が，節境界単位内の述語に係らず，直後の述語に係る現象」が 25 個の係り受けで見られた．

例　税の公平という見地から痛みも伴う/連体節/
税の構造改革に踏み込む/連体節/
段階を向かえようとしていると…
「見地から」が，「伴う」に係らず，直後の述語「踏み込む」に係っており，節境界単位 “連体節” で閉じていない．

（2）「節境界単位内部の文節がこの連体節が修飾する文節と並列関係や同格関係になっている現象」が 7 個の係り受けで見られた．

例　大陸を統治する/連体節/
国と台湾を統治する/連体節/
国が存在している
「（大陸を統治する）国と」と，「（台湾を統治する）国が」が並列関係としての係り受け関係にあり，節境界単位 “連体節” で閉じていない．

（3）「節境界単位内部の文節がこの連体節が修飾する文節を修飾している現象」が 7 個の係り受けで見られた．

例　不服審査審査庁のような独立した/連体節/
機関を作っては…
「不服審査審査庁のような」が，「独立した」が修飾する「機関を」を修飾しており，節境界単位 “連体節” で閉じていない．

以上の分析結果から，節境界単位 “連体節” で閉じていない係り受けの係り文節は，直後の述語，もしくは，直後の名詞に係りやすいことが分かった．

3.3.2　節境界単位の種類 “主題ハ”

節境界単位で閉じていない係り受けのうち 31 個は，節境界単位 “主題ハ” で生じていた．なお，節境界単位 “主題ハ” は「述語を中心としたまとまり」という節の定義を逸脱しているが，統語的に大きな切れ目になると考え[13]，本研究ではこれについても節境界単位とした．分析の結果，「節境界単位 “主題ハ” 内に述語が存在しないために，述語に係ると考えられる文節が直後の述語に係る現象」が全体の 54.8%（17/31）を占めた．

例　キリシタン文化の流入にマカオは/主題ハ/
深く関わってきました
「流入に」が「関わってきました」（述語）に係っており，節境界単位 “主題ハ” で閉じていない．

以上の結果から，節境界単位 “主題ハ” で閉じていない係り受けの係り文節は，直後の述語に係りやすいことが分かった．

3.3.3　節境界単位の種類 “テ節”

節境界単位で閉じていない係り受けのうち 16 個は，節境界単位 “テ節” で生じていた．これらの中で，「節境界単位内部の文節が，節境界単位内の述語に係らず，直後の述語に係る現象」が最も多く見られ，10 個存在した．

例　イギリスが中国に迫って/テ節/
割譲させた/連体節/
植民地であります
文全体の係り受け構造を考えた場合，「イギリスが」は，「迫って」を飛び越えて，「割譲させた」に係るため，節境界単位 “テ節” で係り受けが閉じていない．

以上の結果から，節境界単位 “テ節” で閉じていない係り受けの係り文節は，直後の述語に係りやすいことが分かった．

⋆1 分析に用いたデータには 200 ms 以上のポーズに対して時間情報が付与されている．

4. 節断片への分割

節断片を係り受け解析の処理単位として利用するためには，係り受け解析の前処理として文を節断片に分割する必要がある．本章では，節断片の同定手法について述べる．本手法では，形態素解析および文節まとめあげが施された独話文を入力とし，以下の手順により，文中のすべての節の終境界，および，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，節断片を同定する．

(1)　節境界単位の同定
節境界解析ツール CBAP [13] を用いて入力文に対して節の終境界を検出し[⋆1]，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，節断片を同定する．

以下では，節境界単位で閉じていない係り受けの係り文節の検出について詳述する．

4.1 最大エントロピー法による検出

節境界単位で閉じていない係り受けの係り文節の検出アルゴリズムでは，1 文の文節列を入力とし，節境界単位の最終文節でない文節に対して，節境界単位で閉じていない係り受けの係り文節であるか否かの判定を先頭の文節から順に繰り返す．

節境界単位で閉じていない係り受けの係り文節であるか否かの判定は，最大エントロピー法に基づくモデルにより行う．すなわち，ある文脈において，その文節が節境界単位で閉じていない係り受けの係り文節である確率を推定し，この確率値が閾値 α 以上であれば，この文節は節境界単位で閉じていない係り受けの係り文節であると判定する．

4.2 検出モデルに利用した素性

ここでは，最大エントロピー法に基づくモデルにおいて利用した素性について説明する．本研究では，3 章の分析結果に基づき，1）現在，節境界単位で閉じていない係り受けの係り文節であるか否かの判定を行っている文節（以下，**注目文節**），2）注目文節が属している節境界単位（以下，**注目節境界単位**），3）注目節境界単位の直後の文節，4）注目節境界単位の後方，かつ，最も近隣に位置する述語を含む文節（以下，注目節境界単位の直後の述語文節），に着目した．**図 3** に 1）〜4）の言語単位の位置関係を示す．本手法では，これら 4 つの言語単位において以下に示す素性を利用した．

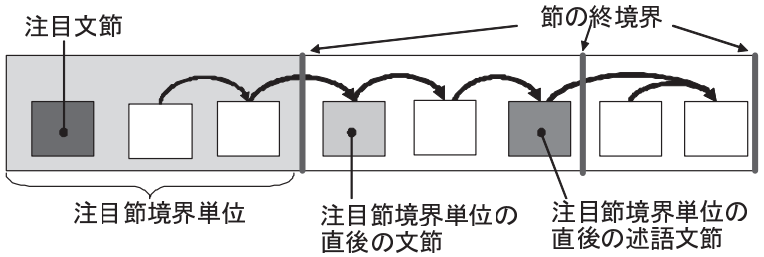

図 3　節境界単位で閉じていない係り受けの係り文節の検出において着目した言語単位
Fig. 3　Language units which are focused in detecting a modified bunsetsu of a dependency relation which is not closed in a clause boundary unit.

1）注目文節
- 主辞の基本形，品詞（大分類，細分類）
- 語形の出現形，品詞（大分類，細分類）
- 助詞 1 の出現形，品詞細分類
- 助詞 2 の出現形，品詞細分類
- 直後にポーズがあるか否か

2）注目節境界単位
- ラベル名
- 注目文節以降に同一格の文節が存在するか否か

3）注目節境界単位の直後の文節
- 注目文節と同一主辞の基本形を持つか否か（注目節境界単位が “連体節” の場合のみ）

4）注目節境界単位の直後の述語文節
- 注目節境界単位内のどの文節よりも注目文節との係り受け確率[⋆2]が高いか否か（注目節境界単位が “主題ハ”or“連体節”or“テ節”，かつ，注目文節が述語に係りうる文節の場合のみ）

ここで，**主辞**は，各文節内で，品詞の大分類が記号，助詞，名詞-接尾となるものを除き，

⋆1 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[13]．本研究ではこれらも節の終境界として扱う．

⋆2 5.2 節の係り受け確率 $P(b^i_{n_i} \xrightarrow{rel} b^j_l|B)$ と同様に計算する．

表 2　述語に係る文節の最終形態素の品詞
Table 2　Part-of-speech of the rightmost morpheme in the bunsetsu which depends on a predicate.

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

最も文末に近い形態素を，**語形**は，各文節内で，記号を除き最も文末に近い形態素をそれぞれ意味し，各文節内で，一番文末に近い助詞を**助詞 1**，その次に文末に近い助詞を**助詞 2**として表記している．また，4）で条件としてあげた，ある文節が述語に係りうる文節か否かは，文献 17)，20) を参考にして，文節の最終形態素の品詞により決定した．**表 2** の品詞[⋆1]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する．

5. 係り受け解析

本手法では，形態素解析，文節まとめあげ，および節断片同定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とおりである．

(1) **節断片の係り受け解析**　1 文中のすべての節断片に対して，その内部の係り受け構造を解析する．

(2) **文の係り受け解析**　1 文中のすべての節断片に対して，その最終文節の係り先を解析する．

なお，以下では，1 文を構成する節断片列を $C_1 \cdots C_m$，節断片 C_i を構成する文節列を $b^i_1 \cdots b^i_{n_i}$，文節 b^i_k を係り文節とする係り受け関係を $dep(b^i_k)$，1 独話の係り受け構造を $\{dep(b^1_1), \cdots, dep(b^m_{n_m-1})\}$ と記す．

5.1 節断片の係り受け解析

節断片の係り受け解析では，節断片 C_i 中の文節列 $b^i_1 \cdots b^i_{n_i}$ 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（$= \{dep(b^i_1), \cdots, dep(b^i_{n_i-1})\}$）を求める．ここでは，節断片の最終文節 $b^i_{n_i}$（$1 \leq i \leq m$）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

$$P(S_i|B_i) = \prod_{k=1}^{n_i-1} P(b^i_k \overset{rel}{\rightarrow} b^i_l|B_i) \tag{1}$$

ここで，$P(b^i_k \overset{rel}{\rightarrow} b^i_l|B_i)$ は，入力文節列 B_i が与えられたときに，文節 b^i_k が b^i_l 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，$P(b^i_k \overset{rel}{\rightarrow} b^i_l|B_i)$ の計算について述べる．$P(b^i_k \overset{rel}{\rightarrow} b^i_l|B_i)$ は，内元ら[21] 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[21]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，節断片の解析では，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．

5.2 文の係り受け解析

節断片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，節断片の最終文節を係り文節とするような係り受け構造 $\{dep(b^1_{n_1}), \cdots, dep(b^{m-1}_{n_{m-1}})\}$ 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．

$$P(S_{last}|B) = \prod_{i=1}^{m-1} P(b^i_{n_i} \overset{rel}{\rightarrow} b^j_l|B) \tag{2}$$

ここで，$P(b^i_{n_i} \overset{rel}{\rightarrow} b^j_l|B)$ は，1 文の文節列 B が与えられたときに，C_i の最終文節 $b^i_{n_i}$ が b^j_l に係る確率を表し，5.1 節と同様に最大エントロピー法を用いて計算する．文の係り受け解析では，節断片の係り受け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．

6. 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．

6.1 実験に使用したデータ

実験で使用したデータを**表 3** 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，文節まとめあげを施した 500 文を用いた．正解の

⋆1 このほかの品詞として，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．

表 3　実験で使用したデータ（あすを読む）
Table 3　Experimental data (“Asu-Wo-Yomu”).

	テストデータ	学習データ 1	学習データ 2
文数	500	5,532	2,274
節境界単位数	2,237	26,318	10,852
文節数	5,298	65,762	27,027
形態素数	13,342	165,173	67,183

学習データ 1：係り受け解析時
学習データ 2：節境界単位で閉じていない係り受けの係り文節検出時

節境界，および，係り受けは人手で付与した[4]．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，係り受け解析時の学習データには，形態素解析，文節まとめあげ，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．このうち，時間情報が付与されている 2,274 文を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ときの学習データとして利用した．

6.2　実験の概要

本手法の有効性を評価するために，上述したデータを用いて以下の 2 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比較した[⋆1]．

- **節断片に基づく係り受け解析手法**　4 章，5 章でそれぞれ述べた，節断片への分割，係り受け解析を順に行う．なお，閾値 α は 0.5 とした．これは，ある文節が節境界単位で閉じていない係り受けの係り文節である確率が，そうでない確率より高くなれば，その文節を節境界単位で閉じていない係り受けの係り文節と判定するという考えに基づいて設定した．
- **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，4 章で述べた節断片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．

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 22) のものを利用し，オプションなどはデフォルトのまま使用した．

6.3　実 験 結 果

各手法の解析時間を**表 4** に示す．節断片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．係り受け解析の前処理として，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ことにより解析全体に与える時間的な影響はほとんどないことが分かった．

表 4　平均解析時間 [ミリ秒/文]
Table 4　Average parsing time [millisecond/sentence].

節断片 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
93.08	88.62	204.20

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

表 5　係り受け正解率
Table 5　Dependency accuracy.

節断片 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
85.74% (4,114/4,798)	84.93% (4,075/4,798)	84.74% (4,066/4,798)

表 6　節境界単位で閉じていない係り受けに対する精度
Table 6　Accuracy for dependency relations which are not closed in a clause boundary unit.

	節断片 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	29.61% (45/152)	1.32% (2/152)	38.82% (59/152)
適合率	55.56% (45/ 81)	11.76% (2/ 17)	39.07% (59/151)

次に，各手法の係り受け正解率[⋆2]を**表 5** に示す．節断片の解析手法は，節境界単位の解析手法と比べ，正解率が 0.8%増加した．マクネマ検定を行った結果，本手法は，節境界単位の解析手法と比較して有意差（有意水準 1%）があることが分かった．

節境界単位で閉じていない係り受けに対する係り受け解析結果を**表 6** に示す．ここで，再現率とは，正解データにおいて節境界単位で閉じていない係り受けのうち，正しく解析できたものの割合を，適合率とは，解析結果において節境界単位で閉じていない係り受けのうち，正しく解析できたものの割合をそれぞれ示す．なお，係り受けが節境界単位で閉じているか否かの判断は，再現率と適合率の両者とも，正解データ上の節境界に基づいて行った．母比率の差の検定では，本手法の再現率と適合率はともに，節境界単位の係り受け解析手

⋆1 参考までに，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 1 度に求める手法（以下，文単位の係り受け解析手法）の解析結果についても記載する．

⋆2 係り受け正解率とは，文末文節を除くすべての文節のうち，正しく受け文節を同定できたものの割合である．

表 7　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
Table 7　Results of detecting a modified bunsetsu of a dependency relation which is not closed in a clause boundary unit.

再現率	50.00% (76/152)
適合率	52.78% (76/144)

法と比べて，有意（有意水準 1%）に上回っており，節断片を解析の処理単位とすることによって，節境界単位で閉じていない係り受けを解析できるようになることが分かった．なお，節境界単位に基づく係り受け解析手法において，2 つの節境界単位で閉じていない係り受けが同定されているが，これらは，節境界の検出誤りにより，偶然同定されたものである．

次に，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 7** 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における再現率・適合率はそれほど高くなく，節断片をより正確に同定することが望まれる[⋆1]．しかし，節断片の同定の精度がこの程度であっても，上述したように，節断片の解析手法は，節境界単位の解析手法と比べ，係り受け正解率が改善した．節断片の同定が多少不正確であっても，係り受け解析の段階でそのミスが吸収されるためだと考えられる．

以上の結果から，本手法によって，節境界単位の解析手法の解析時間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．

6.4 考　　察

6.4.1 節境界単位で閉じていない係り受けの係り文節検出

上述したように，係り受け正解率を向上させるためには，節断片をより正確に同定することが望まれる．そこで，節境界単位で閉じていない係り受けの係り文節の検出について，実験結果をさらに分析した．

表 8 に，節境界単位の種類ごとの，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示す．3.3 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，その節境界単位の種類によって性質が異なっており，節境界単位をまたぐ係り受けの係り文節の検出の難易度は異なっていることが分かる．節境界単位 “主題ハ” で閉じていない係り受けの係り文節の検出は再現率・適合率ともに高い．節境界単位 “主題ハ” の場合は，その内部に述語が存在しないため，述語に係るような文節を検出すればよい．一方，節境界単位 “連体節” や “テ節” で閉じていない係り受けの係り文節の検出は再現率・適合率ともに低い．特に，3.3.1 項の（1）や，3.3.3 項であげた「節境界単位内部の文節が，節境界単位内の述語に係らず，直後の述語に係る現象」に対処するためには，節境界単位内部の文節が節境界単位内の述語に係る場合もあることを考慮しなければならない．本手法では，文節間の係り受け確率に基づいて，どちらの文節に係る確率が高いかという素性を取り入れているが，節境界単位内部の文節が，節境界単位内の述語に係るか，それとも，直後の述語に係るかは，文全体，もしくは，文脈から決まるため，大域的な情報が必要となると考えられる．

⋆1 節断片を 100%正確に検出できることを仮定して係り受け解析実験を行った場合の係り受け正解率は 86.5%（4,152/4,798）であった．

表 8　節境界単位の種類ごとの節境界単位で閉じていない係り受けの検出結果
Table 8　Results of detecting dependency relations which are not closed in a clause boundary unit on each type of clause boundary units.

節境界単位名	再現率	適合率
主題ハ	92.86% (39/42)	81.25% (39/48)
連体節	28.95% (11/38)	37.93% (11/29)
テ節	48.15% (13/27)	46.43% (13/28)
連体節トイウ	12.50% (1/ 8)	25.00% (1/ 4)
補足節	28.57% (2/ 7)	66.67% (2/ 3)
体言止	25.00% (1/ 4)	50.00% (1/ 2)

*頻度 4 以上を抜粋

6.4.2 節情報を考慮した文単位係り受け解析との比較

節境界単位に基づく従来手法[4] では，文を節境界単位に分割し，2 段階で係り受け解析を実行している．これにより，解析時間を短縮することができるものの，節境界単位で係り受けが閉じるという制約のもとで解析することとなり，この制約に逸脱した係り受けを解析できないという問題があった．それに対し，本手法では，解析時間を同程度に維持したうえでの解析精度の改善を実現するため，文を節境界単位ではなく節断片に分割することにより，2 段階解析の枠組みを残しつつ，節境界単位で係り受けが閉じるという制約を緩和した解析を行っている．上記の実験結果は，この制約の緩和が係り受け正解率の向上に大きく寄与した可能性を示唆している．

一方，節境界単位で係り受けが閉じるという制約を緩和するという観点で考えると，文単位の係り受け解析において，節境界情報を素性として用いることにより，節境界単位で係り受けが閉じるという制約を確率的に取り入れるということも考えられる．そこで，以下の 2 つの解析手法を実装し，同じデータを用いて実験した．

- **節境界単位情報を用いた文単位の係り受け解析手法**
  （文単位の係り受け解析手法に以下の素性を追加した係り受け解析手法）

表 9　係り受け正解率（節情報を用いた文単位係り受け解析手法との比較）
Table 9　Dependency accuracy (comparison with sentence-based method using clause information).

節断片 係り受け解析	文単位（節境界単位情報） 係り受け解析	文単位（節断片情報） 係り受け解析
85.74% (4,114/4,798)	85.66% (4,110/4,798)	84.85% (4,071/4,798)

表 10　平均解析時間 [ミリ秒/文]（節情報を用いた文単位係り受け解析手法との比較）
Table 10　Average parsing time [millisecond/sentence] (comparison with sentence-based method using clause information).

節断片 係り受け解析	文単位（節境界単位情報） 係り受け解析	文単位（節断片情報） 係り受け解析
93.08	240.66	274.62

- 係り文節が，節境界単位の最終文節であるか否か
- 受け文節が，節境界単位の最終文節であるか否か
- 係り受けが節境界単位で閉じているか否か

- **節断片情報を用いた文単位の係り受け解析手法**

（文単位の係り受け解析手法に以下の素性を追加した係り受け解析手法）

- 係り文節が，節断片の最終文節であるか否か
- 受け文節が，節断片の最終文節であるか否か
- 係り文節と受け文節がともに節断片の最終文節であるか否か

上記の両手法と本手法の係り受け正解率を**表 9** に，解析時間を**表 10** にそれぞれ示す．本手法は，上記の両手法と比較して，係り受け正解率を低下させることなく，解析時間を改善しており，本手法の有効性を確認した．

7. おわりに

本論文では，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も解析可能な手法を提案した．解析実験の結果，本手法によって，節境界単位に基づく係り受け解析手法と同程度の解析時間を維持しつつ，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，節境界単位 “連体節” や “テ節” で閉じてい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，一部，通信・放送機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」，科学研究費補助金（基盤（B））「入力文の分割・翻訳・連結による同時通訳システム」（課題番号 20300058），ならびに，財団法人中部電力基礎技術研究所研究助成により実施したものである．

参 考 文 献

1) Papageorgiou, H.V.: Clause Recognition in the Framework of Alignment, *Recent Advances in Natural Language Processing*, MitKov, R. and Nicolov, N. (Eds.), John Benjamins Publishing Company (1997).
2) Leffa, V.J.: Clause Processing in Complex Sentences, *Proc. 1st LREC*, pp.937–943 (1998).
3) Kim, M.-Y. and Lee, J.-H.: Syntactic Analysis of Long Sentences Based on S-clauses, *Proc. 1st IJCNLP*, pp.518–526 (2004).
4) Ohno, T., Matsubara, S., Kashioka, H., Maruyama, T., Tanaka, H. and Inagaki, Y.: Dependency Parsing of Japanese Monologue Using Clause Boundaries, *Language Resources and Evaluation*, Vol.40, No.3–4, pp.263–279 (2007).
5) 宇津呂武仁，西岡山滋之，藤尾正和，松本裕治：コーパスからの日本語従属節係り受け選好情報の抽出およびその評価，自然言語処理，Vol.6, No.7, pp.29–60 (1999).
6) 白井　諭，池原　悟，横尾昭男，木村淳子：階層的認識構造に着目した日本語従属節間の係り受け解析の方法とその精度，情報処理学会論文誌，Vol.36, No.10, pp.2353–2361 (1995).
7) 金　淵培，江原暉将：日英機械翻訳のための日本語長文自動短文分割と主語の補完，情報処理学会論文誌，Vol.35, No.6, pp.1018–1028 (1994).
8) 福島孝博，江原暉将，白井克彦：短文分割の自動要約への効果，自然言語処理，Vol.6, No.6, pp.131–147 (1999).
9) 武石英二，林　良彦：接続構造解析に基づく日本語複文の分割，情報処理学会論文誌，Vol.33, No.5, pp.652–663 (1992).
10) Carreras, X. and Màrquez, L.: Boosting Trees for Clause Splitting, *Proc. CoNLL-2001*, pp.73–75 (2001).
11) Molina, A. and Pla, F.: Clause Detection Using HMM, *Proc. CoNLL-2001*, pp.70–72 (2001).
12) Sang, E.F.T.K.: Memory-Based Clause Identification, *Proc. CoNLL-2001*, pp.67–69 (2001).
13) 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 の開発とその評価，自然言語処理，Vol.11, No.3, pp.517–520 (2004).
14) 浜辺良二，内元清貴，河原達也，井佐原均：話し言葉における引用節・挿入節の自動認定結果を利用した係り受け解析，言語処理学会第 12 回年次大会発表論文集，pp.133–136

(2006).
15) 松本裕治，北内　啓，山下達雄，平野善隆，松田　寛，高岡一馬，浅原正幸：形態素解析システム『茶筌』version2.2.9 使用説明書 (2002).
16) 工藤　拓，松本裕治：チャンキングの段階適用による係り受け解析，情報処理学会論文誌，Vol.43, No.6, pp.1834–1842 (2002).
17) 浅原正幸，松本裕治：IPADIC ユーザズマニュアル version2.5.1，奈良先端科学技術大学院大学 (2002).
18) 前川喜久雄，籠宮隆之，小磯花絵，小椋秀樹，菊池英明：日本語話し言葉コーパスの設計，音声研究，Vol.4, No.2, pp.51–61 (2000).
19) Kurohashi, S. and Nagao, M.: Building a Japanese Parsed Corpus While Improving the Parsing System, *Proc. 1st LREC*, pp.719–724 (1998).
20) 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).
21) 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).
22) Zhang, L.: Maximum Entropy Modeling Toolkit for Python and C++ (online). available from http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html (accessed 2007-09-06)

(平成 20 年 6 月 4 日受付)
(平成 20 年 11 月 5 日採録)

大野　誠寛（正会員）

2003 年名古屋大学工学部電気電子・情報工学科卒業．2007 年名古屋大学大学院情報科学研究科博士後期課程修了．博士（情報科学）．同年より同大学院国際開発研究科助教．この間，日本学術振興会特別研究員．自然言語処理，音声言語処理の研究に従事．電子情報通信学会，言語処理学会各会員．

松原　茂樹（正会員）

1993 年名古屋工業大学工学部電気情報工学科卒業．1998 年名古屋大学大学院工学研究科情報工学専攻博士後期課程修了．博士（工学）．同年同大学言語文化部助手．2002 年より名古屋大学情報連携基盤センター助教授（現在，准教授）．この間，日本学術振興会特別研究員，ATR 音声言語コミュニケーション研究所客員研究員，情報通信研究機構専攻研究員．自然言語処理，音声言語処理，情報検索，ディジタル図書館の研究に従事．IEEE，ACM，電子情報通信学会，言語処理学会等各会員．

柏岡　秀紀（正会員）

1993 年大阪大学大学院基礎工学研究科博士後期課程修了．博士（工学）．同年 ATR 音声翻訳通信研究所入社．1998 年同研究所主任研究員（現 ATR 音声言語コミュニケーション研究所）．1999 年奈良先端科学技術大学院大学情報科学研究科客員助教授（兼任）．2006 年情報通信研究機構専門研究員（兼任）．2006 年 ATR 音声言語コミュニケーション研究所音声言語処理研究室室長．主に音声対話，自然言語処理，機械翻訳，音声言語処理の研究に従事．

稲垣　康善（フェロー）

1962 年名古屋大学工学部電子工学科卒業．1967 年名古屋大学大学院博士課程修了．工学博士．同大学助教授，三重大学教授を経て，1981 年名古屋大学工学部教授．1997 年同工学部長・工学研究科長．2003 年名古屋大学名誉教授，愛知県立大学情報科学部教授．2007 年愛知県立大学名誉教授，愛知工業大学経営情報科学部教授．2008 年より豊橋技術科学大学理事・副学長．この間，コンピュテーションとコミュニケーションの理論，オートマトン言語理論，ソフトウェア基礎論，自然言語処理に関する研究に従事．本学会名誉会員．電子情報通信学会名誉員・フェロー．IEEE，SM．日本ソフトウェア科学会，人工知能学会，言語処理学会，ACM，EATCS 各会員．