節境界に基づく独話の漸進的係り受け解析

大野 誠寛† 松原 茂樹†§ 柏岡 秀紀§ 加藤 直人§ 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
§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

同時通訳や字幕生成のように，独話を入力と同時に処理するようなアプリケーションでは，話者による音声入力に従って順次，解析を行う漸進的解析手法が必要である．実際，漸進的構文解析に関していくつか研究されており（例えば，[1, 3]），そこでは，どのような言語単位を解析処理の単位として定めるかが問題となる．

一方，著者らは，1 文の長さが長く文の構造が複雑であるという独話文の特徴に着目し，節境界に基づく係り受け解析手法を提案している [5]．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，節境界をまたぐ係り受け関係を定め，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．

そこで本論文では，節を解析単位とする独話の漸進的係り受け解析手法を提案する．独話音声に対して，節が入力されるたびにその節の内部の係り受け構造を作り上げるとともに，すでに入力されている節の係り先を決定することを試みる．節の係り先となる文節の決定は，後続するいくつかの文節との係り受けの尤度を考慮した動的なタイミングで実行する．

本手法では，文境界が付与されていない独話データ全体に対してその係り受け構造を解析する．これは，独話には明示的な文末標識がなく，事前に文単位に区切ることは容易ではないという独話の特徴に対応している．独話データを用いた解析実験の結果，本手法により，従来の独話文解析手法 [5] と同等の解析性能を維持しつつ，漸進的な係り受け解析が実現できることを確認した．

本論文の構成は以下の通りである．次節で独話の解析単位について述べ，3 節で節境界に基づく係り受け解析手法を示す．4 節で独話の漸進的係り受け解析手法について説明し，5 節で解析実験について述べる．

2 独話の解析単位

本研究では，解析の処理単位として節を採用し，節が入力されるたびにその時点までの係り受け構造を出力する漸進的な独話係り受け解析システムを実現する．

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．本研究では，「文は

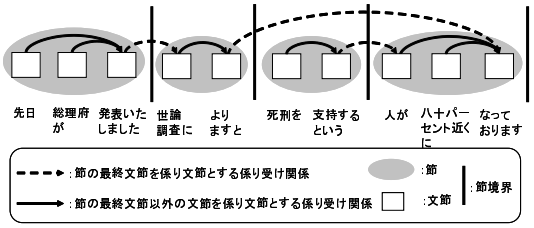

図 1: 節境界と係り受けの関係

一つ以上の節の連接であり，各節を構成する文節は，節の最終文節を除き，その節の内部の文節に係る」とみなす [5]．例として，独話文「先日総理府が発表いたしました世論調査によりますと死刑を支持するという人が八十パーセント近くになっております」の係り受け構造を図 1 に示す．なお，本来，文を節に一次元的に分割することは困難であるものの [2]，節境界解析により近似的に分割することは可能である [4]．本研究では，節境界解析により検出された節境界ではさまれた単位を**節境界単位**と呼び [2]，これを新たな解析単位と考える．

3 節境界に基づく独話文解析

本節では，著者らがこれまでに提案している節境界に基づく独話文の係り受け解析手法 [5] について概説する．この手法では，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し，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**
   一文中のすべての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**
   一文中のすべての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，一文を構成する節境界単位列を $C_1 \cdots C_m$，節境界単位 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 $dep(b^i_k)$，一文の係り受け構造を $\{dep(b^1_1), \cdots, dep(b^m_{n_m-1})\}$ と記す．

3.1 節レベルの係り受け解析

節レベルの係り受け解析では，節境界単位 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．ただし，節境界単位の最終文節 $b^i_{n_i}(1 \leq i \leq m)$ 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．

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

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．係り文節における自立語の原形を h^i_k，その品詞を t^i_k，係りの種類を r^i_k とし，受け文節における自立語の原形を h^i_l，その品詞を t^i_l とする．また，受け文節が節境界単位の最終文節であるか否かを e^i_l とし，文節間距離を d^{ii}_{kl} とする．ここで，係りの種類とは，係り文節が付属語を伴うときはその付属語の語彙，品詞，活用形であり，そうでない場合は文節末の形態素の品詞，活用形である．以上の属性を用いて，確率 $P(b^i_k \overset{rel}{\rightarrow} b^i_l|B_i)$ を以下のように計算する．

$$\begin{aligned} & P(b^i_k \overset{rel}{\rightarrow} b^i_l|B_i) \quad (2) \\ \cong\ & P(b^i_k \overset{rel}{\rightarrow} b^i_l|h^i_k, h^i_l, t^i_k, t^i_l, r^i_k, e^i_l, d^{ii}_{kl}) \\ =\ & \frac{F(b^i_k \overset{rel}{\rightarrow} b^i_l, h^i_k, h^i_l, t^i_k, t^i_l, r^i_k, e^i_l, d^{ii}_{kl})}{F(h^i_k, h^i_l, t^i_k, t^i_l, r^i_k, e^i_l, d^{ii}_{kl})} \end{aligned}$$

ただし，F は共起頻度関数である．

3.2 文レベルの係り受け解析

節境界単位の最終文節の受け文節を同定する．一文の文節列を $B(=B_1 \cdots B_m)$ とし，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 $\{dep(b^1_{n_1}), \cdots, dep(b^{m-1}_{n_{m-1}})\}$ 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．

$$P(S_{last}|B) = \prod_{i=1}^{m-1} P(b^i_{n_i} \overset{rel}{\rightarrow} b^j_l|B) \quad (3)$$

ここで，$P(b^i_{n_i} \overset{rel}{\rightarrow} b^j_l|B)$ は，一文の文節列 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，式 (2) と同様に計算する．最尤の係り受け構造は，式 (3) 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．

4 独話の漸進的係り受け解析

本節では，節境界単位に基づく漸進的係り受け解析について述べる．本手法では，音声入力に対して節境界を随時判定し，節境界単位が同定されると，その時点までの入力に対して係り受け解析を実行する．節境界の判定は節境界解析 [4] により実行する．係り受け解析は，入力された節境界単位の内部の係り受け構造を解析するとともに，すでに入力された節境界単位の最終文節の係り先を可能であれば決定する．

このうち，内部の係り受け解析は，3.1 節で述べた方法により解析すればよい．それに対して最終文節に対する係り受け解析は，その受け文節がいつ入力されるかは明らかではないため，それを決定するタイミングが問題となる．本研究では，文節間の係り受け関係が文をまたぐことはなく，また，その距離が格段に長くなることはないということに着目し，節境界単位の最終文節が入力されてからある程度解析が進んだ時点でその受け文節を決定することとした．具体的には，節境界単位が入力されるたびにその時点での最尤の係り受け構造を 3.2 節で述べた方法により解析し，ある最終文節の係り受け関係が一定の入力回数（以下，固定値）変わらなかった場合，その受け文節を係り先として決定する．

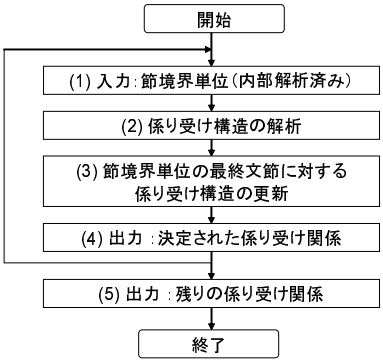

図 2: 漸進的係り受け解析の流れ

本節の以下では，節境界単位の最終文節に対する係り受け解析について説明する．

4.1 漸進的係り受け解析アルゴリズム

係り受け解析の流れを図 2 に示す．解析では，節境界単位 C_i が入力されるごとに，すでに入力された節境界単位を $C_1, \cdots, C_i$ の各最終文節 $b^i_{n_1}, \cdots, b^i_{n_i}$ に対する係り受け構造 $D = \{(dep(b^j_{n_j}), k)|\ 1 \leq j \leq i\}$ を更新することにより実行する．ここで k は $dep(b^j_{n_j})$ の不変回数を示す．以下に係り受け解析アルゴリズムを示す．なお，固定値を σ とする．

(1) 内部の係り受け構造が決定された節境界単位 C_i を入力する．

(2) 節境界単位の最終文節のうち，係り先が未決定な文節に対して，それを係り文節とする係り受け関係を 3.2 節で説明した方法により求める．

(3) (2) で生成された係り受け関係 $dep(b^j_{n_j})$ に基づき，最終文節に対する係り受け関係 D を更新する．ここで $dep(b^j_{n_j})$ が同一の場合は不変回数を $k+1$ とし，異なる場合は 1 とする．

(4) $k = \sigma$ を満たす係り受け関係 $(dep(b^j_{n_j}), k) \in D$ に対して，文節 $b^j_{n_j}$ の係り先が決定したとして $dep(b^j_{n_j})$ を出力する．

(5) すべての節境界単位が入力された時点で，$k < \sigma$ の $(dep(b^j_{n_j}), k) \in D$ に対して，その係り受け関係 $dep(b^j_{n_j})$ を出力する．

なお，本手法では，文末は係り先がないとして解析する．そのため，節境界単位末の解析では係り先なしを候補に含める．具体的には，(3) 式において，係り先のな

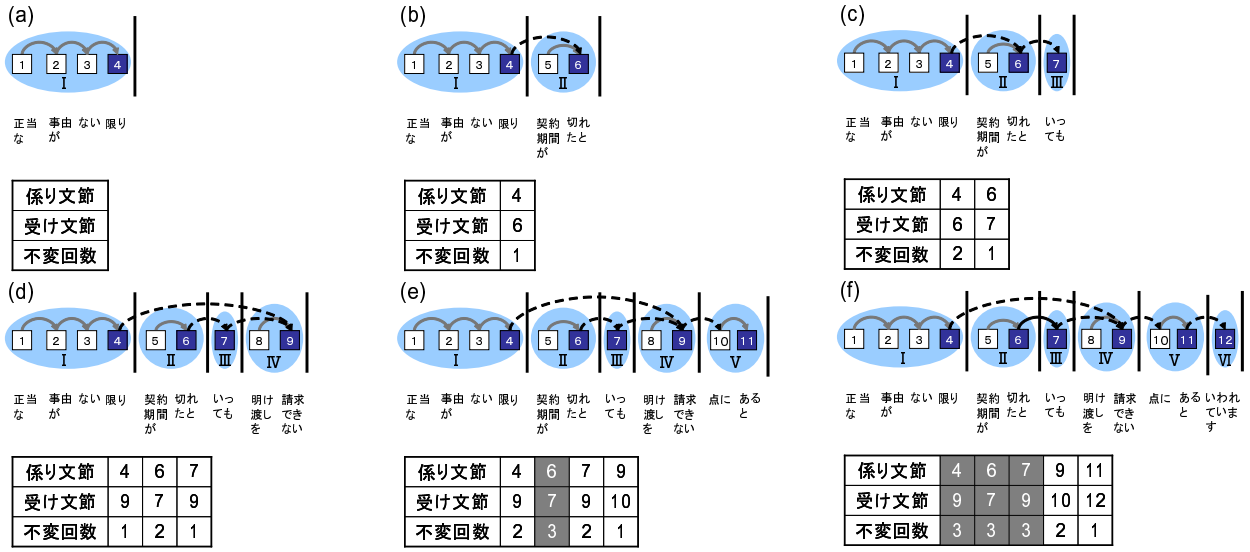

図 3: 漸進的係り受け解析の例 (固定値 3 の場合)

い文節はそれ自身に係る（すなわち，$b^i_{n_i} = b^j_l$）とし，係り先なしとなる確率も計算する．

4.2 解析例

独話「正当な事由がない限り契約期間が切れたといっても明け渡しを請求できない点にあるといわれています」の節境界単位末の文節の係り先を解析する様子を図 3 に示す．(a)〜(f) の 6 過程から構成され，それぞれ上部に係り受け構造を，下部に節境界単位末の文節の係り受け構造を示す．すなわち，$(dep(b^j_{n_j}), k) \in D$ の $dep(b^j_{n_j})$ が係り文節及び受け文節に，k が不変回数に相当する．なお，ここでは固定値 3 として説明する．

(a) は，最初の節境界単位 I が入力された状態を，**(b)** は，節境界単位 II が入力され，係り受け構造 $\{dep(\text{限り})\}$ が解析された状態を示す．$dep(\text{限り})$ は上部の点線矢印に相当し，「限り」の係り先が「切れた」であり，不変回数は 1 であることが下部に記録される．同様にして，**(c)**，**(d)** は，それぞれ節境界単位 III，IV が入力されたときの最尤の係り受け構造 $\{dep(\text{限り}), dep(\text{切れた})\}$，$\{dep(\text{限り}), dep(\text{切れた}), dep(\text{いっても})\}$ が解析された状態を示す．

(e) は，節境界単位 V が新たに入力され，最尤の構造 $\{dep(\text{限り}), dep(\text{切れた}), dep(\text{いっても}), dep(\text{請求できない})\}$ が求まった状態を示している．このとき，係り受け関係 $dep(\text{切れた})$ の不変回数が 3 に達したため，この関係を決定し出力する．

(f) は，節境界単位 VI が新たに入力され，最尤の係り受け構造 $\{dep(\text{限り}), dep(\text{切れた})\}$，$\{dep(\text{限り}), dep(\text{切れた}), dep(\text{いっても}), dep(\text{あると})\}$ が求まった状態を示す．(e) と同様に不変回数が固定値に達している係り受け関係を決定し出力する．

5 解析実験

独話の漸進的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．

表 1: 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
番組数	7	95
文数	470	5,532
節境界単位数	2,140	26,318
文節数	5,054	65,821
形態素数	12,753	165,129

5.1 実験の概要

実験には，NHK の解説番組「あすを読む」（番組あたりの長さは約 10 分）を使用した．使用したデータの概要を表 1 に示す．テストデータとして，書き起こしデータ [5] に形態素解析，文節まとめ上げを施した 7 番組 (470 文) を用いた[1]．節境界，及び，係り受けの正解は人手で作成した．一方，学習データとしては，形態素，文節まとめ上げ，節境界，係り受けに関する情報が与えられた 95 番組 (5,532 文) を用いた．

これらのデータを用いて解析を行い，係り受け正解率と解析時間を求めた．解析システムは，GNU Common LISP で実装し，CPU が Pentium4 2.40GHz，メモリが 2GB の Linux PC 上で実行した．なお，4.1 節で説明した固定値を 1 から 12 まで変化させて，計 12 回実験した．

5.2 実験結果

各固定値ごとの係り受け正解率を表 2 に示す．表 2 の第 1 列は，番組末を除く全ての節境界単位末に対する正解率を，第 2 列は，番組末を除く全ての文節に対する正解率を示す．固定値が 2 及び 3 のときに，節境界単位末に対する正解率が最も高く，全体の正解率は 76.2%となった．なお，節境界単位末を除く節境界単位内に対する解析の正解率は 87.5%であった．表 3 に，CBAP の節境界解析の精度について，ラベルを無視して節境界の

[1]書き起こしデータでは，句点により文が区切られている．実験では，句点を取り除き 1 番組分の発話を連結している．

表 2: 固定値ごとの係り受け正解率

固定値	節境界単位末	全体
1	57.6% (1,228/2,133)	74.9% (3,778/5,047)
2	**60.8%**(1,296/2,133)	**76.2%**(3,847/5,047)
3	**60.8%**(1,296/2,133)	**76.2%**(3,847/5,047)
4	60.4% (1,289/2,133)	76.1% (3,840/5,047)
5	59.8% (1,276/2,133)	75.8% (3,827/5,047)
6	59.4% (1,268/2,133)	75.7% (3,819/5,047)
7	58.8% (1,254/2,133)	75.4% (3,805/5,047)
8	58.6% (1,251/2,133)	75.4% (3,803/5,047)
9	58.7% (1,253/2,133)	75.4% (3,805/5,047)
10	58.4% (1,245/2,133)	75.2% (3,797/5,047)
11	57.6% (1,229/2,133)	74.9% (3,780/5,047)
12	57.9% (1,235/2,133)	75.0% (3,786/5,047)

表 3: CBAP の節境界解析結果（ラベルは無視）

再現率	97.6% (2,088/2,140)
適合率	99.1% (2,088/2,106)

位置のみで評価した結果を示す．適合率，再現率ともに高く，後に行われる解析への影響はあまりない．

固定値と解析時間の関係を図 4 に示す．固定値を大きくするにしたがって，解析時間が増加している．解析時間が最も短かかったのは，固定値が 3 のときで，全 7 番組で 12.5 秒，1 番組あたり 1.8 秒だった．なお，この解析時間には，CBAP による節境界解析の時間も含まれている．節境界解析の平均解析時間は 1 番組あたり 0.3 秒程度である．

本手法では，文末は係り先がないとして解析を実行している．すなわち，係り先なしと判定された文節を文末であるとみなしている．このような観点から，本手法の文末判定性能を評価した．表 4 に文末判定の適合率，再現率，F 値を示す．固定値 3 のときに最も高い F 値を示した．独話の文境界判定手法はすでにいくつか提案されているが（例えば，[6]），本手法では漸進的係り受け解析と同時的に文末を判定できるという特徴がある．

以上の結果から，本実験においては固定値が 3 のとき，最も高い性能を示しており，文単位を入力とする従来の係り受け解析手法（正解率で 79.0%，処理時間で 1 番組あたり約 2.1 秒）[5] と比較しても，同程度の解析精度と解析時間を達成していることを確認した．

6 おわりに

本論文では，節境界単位での漸進的な独話係り受け解析手法を提案した．本手法は，文境界が同定されていない独話発話全体に対して，漸進的に係り受け関係を同定する．文末は係り先がないとして解析する．本手法の有効性を評価するために，独話コーパスを用いて係り受け解析実験を行った．実験の結果，文単位が特定されていない独話全体を入力とする本手法が，文単位を入力とする従来の係り受け解析手法と同程度の性能を備えていることを確認した．

今後は，ポーズ情報を効果的に利用することなどにより，漸進的係り受け解析の性能向上を図る予定である．

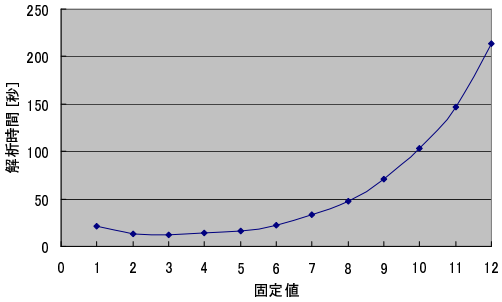

図 4: 固定値と解析時間の関係

表 4: 文末判定の適合率・再現率・F 値

固定値	適合率	再現率	F 値
1	54.6% (337/617)	72.1% (334/463)	62.1
2	69.8% (312/447)	67.4% (312/463)	68.6
3	**74.6%**(296/397)	**63.9%**(296/463)	**68.8**
4	75.7% (278/367)	60.0% (278/463)	66.9
5	76.8% (271/353)	58.5% (271/463)	66.4
6	76.9% (260/338)	56.2% (260/463)	64.9
7	78.5% (252/321)	54.4% (252/463)	64.3
8	78.7% (247/314)	53.3% (247/463)	63.6
9	81.1% (249/307)	53.8% (249/463)	64.7
10	80.3% (245/305)	52.9% (245/463)	63.8
11	79.9% (238/298)	51.4% (238/463)	62.6
12	79.8% (233/292)	50.3% (233/463)	61.7

謝辞　独話文係り受けコーパスの作成に御協力いただいた名古屋大学大学院国際言語文化研究科の大学院生のみなさまに感謝致します．本研究は，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」，ならびに，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」により実施したものである．

参考文献

[1] J. Nivre : Incrementality in Deterministic Dependency Parsing, Proceedings of the ACL Workshop Incremental Parsing: Bringing Engineering and Cognition Together, pp. 50–57 (2004).

[2] 柏岡秀紀，丸山岳彦: 節境界単位による翻訳 - 連体節について -，言語処理学会第 10 回年次大会論文集，pp. 460–463 (2004).

[3] 加藤 芳秀，松原 茂樹，外山 勝彦，稲垣 康善 : 主辞情報付き文脈自由文法に基づく漸進的な依存構造解析，電子情報通信学会論文誌，Vol.86-D-II, No. 1, pp. 86-97 (2003).

[4] 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).

[5] 大野誠寛，松原 茂樹，丸山 岳彦，柏岡 秀紀，田中 英輝，稲垣 康善 : 節境界に基づく独話文係り受け解析の効率化，情報処理学会研究報告，NL-162，pp. 213–220 (2004).

[6] 下岡和也，内元清貴，河原達也，井佐原均 : 話し言葉の係り受け解析と文境界推定の相互作用による高精度化，話し言葉の科学と工学ワークショップ，pp. 119–126 (2004).