節境界に基づく独話文の係り受け解析

大野 誠寛† 松原 茂樹‡§ 丸山 岳彦§ 柏岡 秀紀§ 田中 英輝§ 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@matubara.net

1 はじめに

話し言葉は，一人の話者のみが話す「独話」と複数の話者が交替で話す「対話」に分類できる．これまでの話し言葉解析の研究は，対話文を対象としたものがほとんどであり，非文法性に対して頑健に対処する手法が提案されてきた (例えば，[3, 4])．しかしその一方で，独話文を対象とした研究はほとんどないのが現状である．

独話文は，対話文に比べ，1 文の長さが長く文の構造が複雑であるといった特徴をもつ．そのような文に対して解析を実行すると，一般に，解析時間が長くなるうえ，高い解析精度を達成することが難しくなる．高い性能を備えた独話文解析を実現するために，適切なユニットに文を分割し，簡単化することが効果的な方法である．

そこで本稿では，文分割に基づく独話文の係り受け解析手法を提案する．本手法では，節レベルと文レベルの二段階で係り受け解析を実行する．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．独話文係り受け解析実験により節境界解析に基づく本手法の有効性を確認した．

2 節境界と係り受け構造

節とは，述部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．本研究で提案する手法では，「文は一つ以上の節の連接であり，各節を構成する文節は，節の最終の文節を除き，その節内の文節に係る」とみなす（図 1 参照）．このような仮定を設ける理由は以下の通りである．

- 一文が長い独話文では，文を短く分割することにより係り受け関係の探索範囲が狭められ，解析時間を短縮できる．
- 節は単文に相当する言語的単位であり，その内部で係り受けがまとまりやすい．そのため，上述の仮定に逸脱する係り受けは少なく，解析精度の低下への影響は小さい．実際，独話文の 90 %の節が上述の性質を満たすという概算が報告されている [1]．

Dependency Parsing of Japanese Spoken Monologue based on Clause Boundaries: Tomohiro Ohno, Shigeki Matsubara (Nagoya University), Takehiko Maruyama, Hideki Kashioka, Hideki Tanaka (ATR) and Yasuyoshi Inagaki (Aichi Prefectural University)

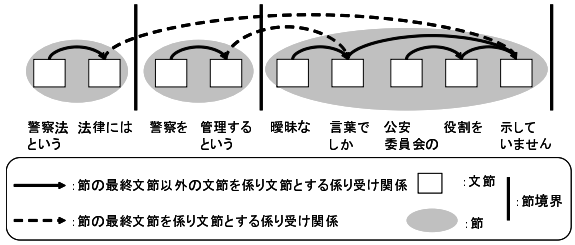

図 1: 節境界と係り受けの関係

なお，独話文の節への分割は，節境界解析により実現できる [2]．この手法では，形態素列パタンに関する節境界検出ルールをもとに，局所的な形態素解析結果のみから高い精度での検出が可能である．

3 節境界に基づく係り受け解析手法

本手法では，前節で設けた仮定に基づき，形態素解析，節境界解析及び文節まとめ上げが施された文を入力とする．以下の手順で解析を実行する．

1. 一文中の各節に対して，内部の係り受け構造を解析
2. 節の最終文節を係り文節とする係り受け構造を解析

なお，以下では，一文中の節列を $C_1, ..., C_m$，節 C_i 中の文節列を $b^i_1, ..., b^i_{n_i}$，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 $\{dep(b^1_1), ..., dep(b^m_{n_m-1})\}$ と記す．

本手法では，まず，節 C_i が入力されるごとに，節内部の係り受け構造 $\{dep(b^i_1), ..., dep(b^i_{n_i-1})\}$ を求める．その後，文末まで，節の最終文節の係り受け構造 $\{dep(b^1_{n_1}), ..., dep(b^{m-1}_{n_{m-1}})\}$ を求める．なお，いずれの解析においても，係り受けは非交差性，後方修飾性，係り先の唯一性を満たすものとする．

3.1 節内部の係り受け解析

節内部の係り受け解析は，入力節 C_i 中の文節列 $b^i_1, ..., b^i_{n_i}$ を B_i とし，$P(S_i|B_i)$ を最大にする係り受け構造 $S_i(=\{dep(b^i_1), ..., dep(b^i_{n_i-1})\})$ を求める．なお，節内部の係り受け解析では，節の最終文節 $b^i_{n_i}$ の受け文節は決定しない．

それぞれの係り受け関係は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．

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

表 1: 実験で使用したデータ

	テストデータ	学習データ
データ名	あすを読む	京大コーパス
文数	200	7,758
節数	951	27,060
文節数	2,430	72,393
形態素数	6,017	205,731

ここで，$P(b_k^i \stackrel{rel}{\rightarrow} b_l^i|B_i)$ 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，$P(b_k^i \stackrel{rel}{\rightarrow} b_l^i|B_i)$ の計算について述べる．まず，係り文節における自立語の原形を h_k^i，その品詞を t_k^i，係りの種類を r_k^i とし，受け文節における自立語の原形を h_l^i，その品詞を t_l^i とする．また，文節間距離を d_{kl}^{ii} 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，そうでない場合は一番最後の形態素の品詞，活用形である．

以上の属性を用いて，確率 $P(b_k^i \stackrel{rel}{\rightarrow} b_l^i|B_i)$ を以下のように計算する．

$$P(b_k^i \stackrel{rel}{\rightarrow} b_l^i|B_i) = \frac{F(b_k^i \rightarrow b_l^i, h_k^i, h_l^i, t_k^i, t_l^i, r_k^i, d_{kl}^{ii})}{F(h_k^i, h_l^i, t_k^i, t_l^i, r_k^i, d_{kl}^{ii})} \quad (2)$$

ただし，F は共起頻度関数である．

3.2 節の最終文節の係り受け解析

節の最終文節の受け文節を同定する．一文の文節列を $B(= B_1, ..., B_m)$ とし，節の最終文節を係り文節とするような係り受け構造 $\{dep(b_{n_1}^1), ..., dep(b_{n_{m-1}}^{m-1})\}$ 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．

$$P(S_{last}|B) = \prod_{i=1}^{m-1} P(b_{n_i}^i \stackrel{rel}{\rightarrow} b_l^j|B) \quad (3)$$

ここで，$P(b_{n_i}^i \stackrel{rel}{\rightarrow} b_l^j|B)$ は，一文の文節列 B が与えられたときに，文節 $b_{n_i}^i$ が b_l^j に係る確率を表す．本手法では，先に解析した節内部の係り受け構造を前提として決定する．

なお，$P(b_{n_i}^i \stackrel{rel}{\rightarrow} b_l^j|B)$ は，式 (2) と同様に計算する．

4 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．

4.1 実験の概要

実験で使用したデータを表 1 に示す.「あすを読む」の書き起こしに形態素解析，節境界解析，文節まとめ上げを施したデータ 200 文をテストデータとして用いた．正解の係り受けは人手で付与した．一方，本手法では，節の最終文節を係り文節とする係り受け関係を除いて，係り受け関係は節境界をまたがないことを前提としているが，この前提を満たさない係り受け関係は，テストデータの正解中に 94 個 (全体の約 3.9%) 存在した．

表 2: 実験結果（係り受け正解率）

	本手法	従来手法
節の内部	83.6%(1,236/1,479)	82.6%(1,222/1,479)
節の最終文節	58.1%(436/ 751)	55.0%(413/ 751)
合計	75.0%(1,672/2,230)	73.3%(1,635/2,230)

一方，学習データとして,「あすを読む」に対する十分な量の係り受けデータは存在しないため，新聞記事である京大コーパス 7,758 文を用いた．

なお， 節に分割することなく文の係り受け構造を一度に求める手法（以下，従来手法）によっても係り受け解析を行い，本手法と比較した．

4.2 実験結果

一文あたりの平均解析時間は，本手法が 0.012 秒，従来手法が 0.055 秒であり．本手法の解析速度が従来手法に比べて，約 5 倍向上した．

一方，両手法の係り受け正解率を表 2 に示す．表 2 の第 1 行は，節の最終文節を除く節内の全ての文節に対する正解率を示す．第 2 行は，文末を除く全ての節の最終文節に対する正解率を示す．本手法では節境界をまたぐ係り受け関係を正しく同定することはできないが，その一方で，受け文節となる文節の候補を適切に絞ることができるといった利点がある．実験により，本手法が，従来手法に劣らない解析精度を備えていることを確認した．

5 おわりに

本稿では，節境界に基づく独話文の係り受け解析手法を提案した．本手法の有効性を評価するために，独話文係り受け解析実験を行った．実験の結果，本手法の解析精度と解析時間はともに従来手法を上回り，独話文係り受け解析における本手法の有効性を確認した．

参考文献

[1] 柏岡秀紀，丸山岳彦，田中英輝: 節境界と係り受け解析，言語処理学会第 9 回年次大会論文集，pp. 117-120 (2003).

[2] 丸山岳彦，柏岡秀紀，熊野正，田中英輝: 節境界自動検出ルールの作成と評価，言語処理学会第 9 回年次大会論文集，pp. 517-520 (2003).

[3] Matsubara, S., Murase, T., Kawaguchi, N. and Inagaki, Y.: Stochastic Dependency Parsing of Spontaneous Japanese Spoken Language, *Proc. of 19th COLING*, Vol. 1, pp. 640-645 (2002).

[4] 大野誠寛，松原 茂樹，河口 信夫，稲垣 康善：日本語音声対話文の統計的係り受け解析とその評価，情報処理学会第 65 回全国大会講演論文集，Vol. 2，pp. 1-2 (2003).