

漸進的な日本語解析において出力する係り受け構造とその利用

大野 誠寛†

松原 茂樹†

†名古屋大学情報基盤センター

†名古屋大学大学院情報科学研究科

1 はじめに

同時通訳や音声対話、字幕生成などのアプリケーションでは、音声入力に対して同時に処理を行う必要があり、構文情報に関して、どのような情報をいつ取得できるのかが問題となる。しかし、これまでの漸進的な係り受け解析 [1, 2, 3] では、入力途中の段階において、係り先が未だ入力されていない文節の係り受けに関する情報をどのように出力するのかについては、ほとんど検討されていない。

本稿では、入力途中の任意の段階において、漸進的な日本語の係り受け解析が出力すべき係り受け構造を提案する。提案する係り受け構造では、係り先が未だ入力されていない文節の係り受け情報として、係り先が未入力文節であると明示することを解析器に課す。これにより、上位層のアプリケーションは、係り先が入力されていない文節について、その係り先が入力されるまで待たれることなく、その文節が入力済みの文節には係らないという情報を利用できるようになる。

また本稿では、提案した係り受け構造を、文節が入力されるごとに出力することができる漸進的係り受け解析を示す。解析実験を実施した結果、その実現可能性を確認した。さらに、漸進的係り受け解析の出力を、読みやすい字幕を生成するための漸進的な改行挿入手法に利用し、その応用における有用性を確認した。

2 漸進的な日本語解析において出力する係り受け構造

漸進的係り受け解析の利用が想定される、同時通訳や音声対話、字幕生成などのアプリケーションでは、係り受け情報を取得できるタイミングやその内容が、その後の処理に影響を与える。

これらのアプリケーションは音声入力に追従した処理が求められることから、随時、係り受け情報を取得することが望ましい。本研究では、文節間の関係を求める係り受け解析において、できる限り任意のタイミングで出力することを考え、漸進的係り受け解析は、文節が入力されるごとに解析結果を出力するものとする。

次に、漸進的係り受け解析の出力内容について考える。同時通訳や音声対話、字幕生成などのアプリケーションでは、どの文節がどの文節に係るかという情報が利用されていることになるが、ある文節列の係り受けが閉じているか否か、すなわち、構文的なまとまりに関する情報も有用である。実際、同時通訳における訳出タイミング [4] や、音声対話システムでのあいづち生成タイミング [5]、字幕生成での読みやすい改行位置 [6] などの決定において、構文的なまとまりに関する情報は重要な手掛かりとして利用されている。

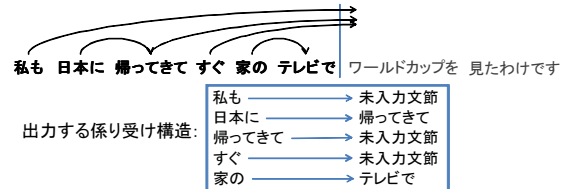


図 1: 文入力の途中段階で出力する係り受け構造の例

そこで本研究では、文入力の途中段階において漸進的係り受け解析が出力する係り受け情報として、係り先が未だ入力されていない場合は、その文節の係り先は未だ入力されていないと明示的に出力することとする。ある文節の係り先が未だ入力されていないことが分かれば、入力済み文節列内の構文的なまとまりを捉えることができるため、上位層のアプリケーションがその情報を利用できる。

以下では、本研究において、漸進的係り受け解析が文の入力途中の段階において出力する係り受け構造を示す。文節列 $b_1 \dots b_n$ からなる文 S を解析している際に、文節 $b_x (1 \leq x \leq n)$ が入力された時点で出力する係り受け構造 D_x を以下のように定義する。

$$D_x = \{ \langle b_1 \rightarrow uke(b_1) \rangle, \dots, \langle b_{x-1} \rightarrow uke(b_{x-1}) \rangle \}$$

(ただし, $uke(b_y) \in \{b_{y+1}, \dots, b_x, b_{over}\} (1 \leq y \leq x-1)$)

ここで、 $\langle b_y \rightarrow uke(b_y) \rangle$ は、文節 b_y を係り元の文節、 $uke(b_y)$ をその係り先の文節とする係り受け関係とする。 $uke(b_y)$ が取り得るものとして、文節 b_y 以降の文節 b_{y+1}, \dots, b_x の他に、 b_{over} を含めている。 $uke(b_y) = b_{over}$ とは、 b_y の係り先は、 b_{y+1}, \dots, b_x ではなく、未だ入力されていない何らかの文節であることを意味する。

図 1 に文入力の途中段階で出力する係り受け構造の例を示す。この例は、文「私も日本に帰ってきてすぐ家のテレビでワールドカップを見たわけです」のうち、文節「テレビで」まで入力された段階で出力する係り受け構造を示している。

3 漸進的係り受け解析

本節では、2 節で提案した係り受け構造を、文節が入力されるごとに出力する漸進的係り受け解析手法について述べる。本手法は、内元らの係り受け解析手法 [7] に改良を加えることにより実現する。

内元らの手法は、文 S が与えられたときに、文全体の係り受け構造が D となる確率 $P(D|S)$ が最も高くなるものを最適な係り受け構造として出力する手法である。内元らの手法の特徴は、着目している 2 つの文節の係り受け確率を求める際に、その 2 文節に対しては「係

る」確率，2文節の間の文節に対しては前文節がその文節を越えて後文節に係る確率（「越える」確率），後文節より文末側の文節に対しては前文節がその文節との間にある後文節に係る確率（「間」の確率）をそれぞれ計算し，それらをすべて掛け合わせた確率値を用いて係り受け確率を求める点にある。

本研究では，「越える」，「係る」，「間」の確率を全て掛け合わせて係り受け確率を求めている点を利用することにより，ある文節が入力済みの文節には係らない確率，すなわち，未入力文節（ b_{over} ）に係る確率を計算することができると考えた。

内元らの手法との違いは，内村らの手法が1文が入力された後に文全体の係り受け構造を求めるものであるのに対し，本手法は，文節列 $b_1 \dots b_n$ からなる文 S を解析する際，文節列 $B_x = b_1, \dots, b_x (1 \leq x \leq n)$ まで入力された時点で出力する係り受け構造 D_x を求める点にある。本手法では，文節列 B_x が与えられたとき，係り受け構造が D_x となる確率 $P(D_x|B_x)$ が最も高くなる係り受け構造を出力する。

ここで， D_x は各文節 $b_i (i = 1, \dots, x-1)$ を係り元の文節とする係り受け関係 d_i の順序付き集合 $D = \{d_1, \dots, d_{x-1}\}$ で表されるとする。さらに， $d_i = \{d_{i,i+1}, \dots, d_{i,x}\} (1 \leq i \leq x-1)$ で表されるとする。 $d_{i,i+j}$ は，文節 b_i と b_{i+j} の間の関係を表すフラグであり，以下のように「越える」，「係る」，「間」を表す0, 1, 2の3値をとるものとする。

$$d_{i,i+j} = \begin{cases} 0 & (1 \leq j < dep(i)) \\ 1 & (j = dep(i)) \\ 2 & (dep(i) < j \leq n-i) \end{cases}$$

ここで， $dep(i)$ は，文節 b_i の係り先の文節が $b_l (i < l \leq n)$ であるときに， $dep(i) = l - i$ と定義される。なお，本手法では，文節 b_i が未入力の何らかの文節 b_{over} に係ることも考慮するため， d_i を構成する要素として，未だ入力されていない文節 b_{x+1} の関係を表すフラグ $d_{i,x+1}$ をダミーとして入れている。

このとき，本手法では， $P(D_x|B_x)$ を以下のように計算する。

$$\begin{aligned} P(D_x|B_x)^2 &= \prod_{i=1}^{x-1} P(d_i|B_x) = \prod_{i=1}^{x-1} P(d_{i,i+1}, \dots, d_{i,x}|B_x) \\ &= \prod_{i=1}^{x-1} \left(\prod_{j=1}^{dep(i)-1} P(d_{i,i+j} = 0|B_x) \right. \\ &\quad \times P(d_{i,i+dep(i)} = 1|B_x) \\ &\quad \times \left. \prod_{j=dep(i)+1}^{x-i+1} P(d_{i,i+j} = 2|B_x) \right) \end{aligned}$$

ここで，2文節間の関係（「越える」，「係る」，「間」）の確率 $P(d_{i,i+j}|B_x)$ は，次の値を使う。

- $1 \leq j \leq x-i$ のとき：内元らの手法 [7] と同様に最大エントロピー法により学習して推定した値
- $j = x-i+1$ のとき： $P(d_{i,i+j}=0|B_x)=0$, $P(d_{i,i+j}=1|B_x)=P(d_{i,i+j}=2|B_x)=1/2$

これは，文節 b_i と未だ入力されていない何らかの文節を表す b_{x+1} との間関係 $d_{i,x+1}$ は，「越える」関係になることはなく，「係る」，「間」のどちらかの関係になる

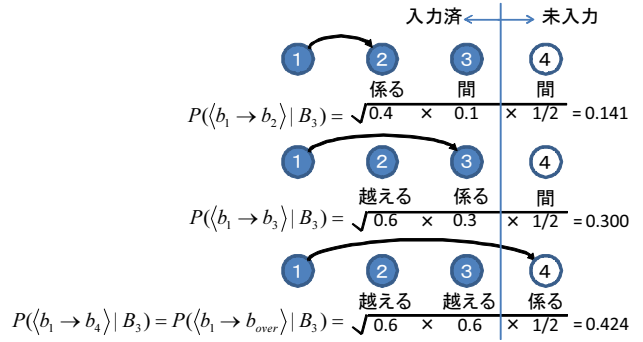


図 2: 本手法による係り受け確率の計算

が，その確率は不明であるため，等確率と仮定することによる。なお，最大エントロピー法において利用した素性は，文献 [8] と同様のものを利用した。

最後に， $P(D_x|B_x)$ を最大とする係り受け構造 D_x の中に，未入力の文節に係り先とする係り受け関係 $\langle b_i \rightarrow b_{x+1} \rangle$ がある場合，これらは全て $\langle b_i \rightarrow b_{over} \rangle$ に置き換えたものを出力する。

図 2 に，文節列 b_1, \dots, b_5 からなる文を解析する際に，文節列 b_1, \dots, b_3 まで入力された時点での， b_1 を係り元とする係り受け関係の係り受け確率 $P(d_1|B_3)$ の計算を示す。なお， $P(\langle b_1 \rightarrow b_4 \rangle | B_3) = P(\langle b_1 \rightarrow b_{over} \rangle | B_3)$ と捉えることができる。

4 評価実験

2節で提案した係り受け構造の漸進的な出力について，その実現可能性を評価するため，3節で提案した漸進的係り受け解析による解析実験を行った。

4.1 実験概要

実験データとして，同時通訳データベース [9] に収録されている日本語講演音声の書き起こしデータを使用した。すべてのデータに，形態素情報，文節境界情報，節境界情報，係り受け情報，改行情報が人手で付与されている [6]。

実験は，全 16 講演を用いた交差検定により実施した。すなわち，1 講演をテストデータとし，残りの 15 講演を学習データとして係り受け解析を実行した。ただし，16 講演のうち 2 講演は評価データから取り除き，残りの 14 講演（1,714 文，20,707 文節）に対する実験結果に基づいて評価した¹。なお，係り受け解析の入力として，形態素情報，文節境界情報，節境界情報は，人手で付与されたものを利用した。また，最大エントロピー法のツールとしては文献 [10] のものを利用した。オプションは，学習アルゴリズムにおける繰り返し回数を 1,000 に設定し，それ以外はデフォルトのまま使用した。

4.2 評価指標

本手法は，2節で定義した係り受け構造を文節が入力されるごとに出力できることを目指している。そのため，漸進的係り受け解析の正解率として，以下のような評価指標を導入し，解析精度を評価した。

¹5.2 節の改行挿入実験の評価データと合わせるためにこのような設定をした。なお，取り除いた 2 講演は，改行挿入確率の推定時の素性決定のための事前分析データとして使用した [6]。

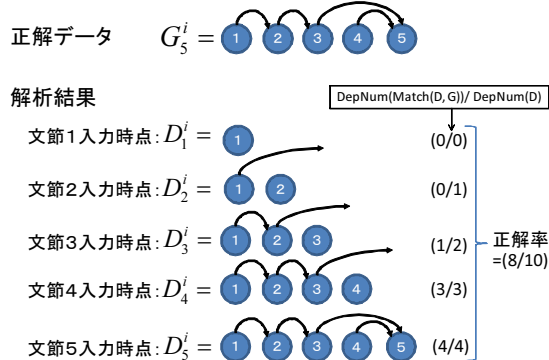


図 3: 漸進的係り受け解析における正解率の算出例

$$\text{正解率} = \frac{\sum_i^N \sum_{j=1}^{n_i} \text{DepNum}(\text{Match}(D_j^i, G_j^i))}{\sum_i^N \sum_{j=1}^{n_i} \text{DepNum}(D_j^i)}$$

ここで、 D_j^i と G_j^i はそれぞれ、文 S_i ($1 \leq i \leq N$) の解析において文節 b_j ($1 \leq j \leq n_i$) が入力された時点で解析器が出力する係り受け構造と、出力すべき正解の係り受け構造を表す。また、 $\text{DepNum}()$ は係り受け関係の集合を入力とし、その中の係り受け関係の数を返す関数、 $\text{Match}()$ は係り受け関係の集合 2 つを入力とし、両者で一致した係り受け関係の集合を返す関数である。

図 3 に 5 文節からなる文 S_i を解析した場合の正解率の算出例を示す。文節 3 が入力された時点について説明する。正解データの S_i 全体の係り受け構造から、 $G_3^i = \{\langle b_1 \rightarrow b_2 \rangle, \langle b_2 \rightarrow b_3 \rangle\}$ となる。一方、解析結果では、 $D_3^i = \{\langle b_1 \rightarrow b_2 \rangle, \langle b_2 \rightarrow b_{\text{over}} \rangle\}$ となっており、 $\text{DepNum}(\text{Match}(D_3^i, G_3^i)) = 1$ となる。

さらに、係り先が未入力の場合と既入力の場合に分けて、それぞれ再現率と適合率を測定した。係り先が未入力の場合は、以下のように計算する。

$$\text{再現率} = \frac{\sum_i^N \sum_{j=1}^{n_i} \text{DepNum}(\text{Over}(\text{Match}(D_j^i, G_j^i)))}{\sum_i^N \sum_{j=1}^{n_i} \text{DepNum}(\text{Over}(G_j^i))}$$

$$\text{適合率} = \frac{\sum_i^N \sum_{j=1}^{n_i} \text{DepNum}(\text{Over}(\text{Match}(D_j^i, G_j^i)))}{\sum_i^N \sum_{j=1}^{n_i} \text{DepNum}(\text{Over}(D_j^i))}$$

ここで、 $\text{Over}()$ は係り受け関係の集合を入力とし、その中で係り先が未入力文節である係り受け関係の集合を返す関数である。係り先が既入力の場合は、この Over 関数を NonOver 関数（係り受け関係の集合を入力とし、その中で係り先が既入力文節である係り受け関係の集合を返す関数）に置き換えて計算する。

4.3 実験結果

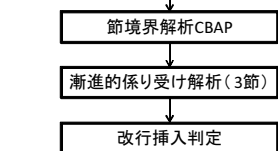
本手法の正解率は、73.9% (125,495/169,765) であった。参考のため、1 文の係り受け構造を求める内元らの手法 [7] で係り受け解析²を行った。その結果、従来の係り受け正解率 ($= \sum_i^N \text{DepNum}(\text{Match}(D_{n_i}^i, G_{n_i}^i)) / \sum_i^N \text{DepNum}(D_{n_i}^i)$) を測定したところ、75.8% (14,391/18,993) であった。

² 2 文節間の関係を ME 法で推定する際の素性は 3 節と同じものをを用いた。

表 1: 漸進的係り受け解析実験の結果

	再現率	適合率	F 値
係り先が ³	81.3%	84.4%	82.8%
未入力	(41,721/51,299)	(41,721/49,406)	
係り先が ³	70.7%	69.6%	70.1%
既入力	(83,774/118,466)	(83,774/120,359)	

形態素解析、文節まとめ上げが施された
1 文の文節列が 1 文節ずつ入力される



入力された文節とその直前の文節の間に、
改行を挿入するか否かを出力する

図 4: 漸進的改行挿入の流れ

次に、表 1 に、係り先が未入力の場合と既入力の場合に分けて、それぞれ再現率と適合率を測定した結果を示す。本手法は、係り先が既入力文節である係り受け関係を同定する精度は必ずしも高くない。係り先が既入力文節である場合は、係り先の文節を具体的に同定する必要があり、係り先が未入力文節であると同定することよりも難易度が高い。一方、係り先が未入力文節である係り受け関係に対しては、比較的高精度に同定できていることが分かる。提案した係り受け構造を文節が入力されるごとに出力することの実現可能性を確認した。

5 漸進的改行挿入への利用

本節では、3 節で述べた漸進的係り受け解析を利用した漸進的改行挿入について述べる。

5.1 漸進的改行挿入

本手法では、形態素解析、文節まとめ上げが施された 1 文の文節列が 1 文節ずつ入力されるごとに、入力された文節とその直前の文節の間の文節境界に対して、その位置に改行を挿入するか否かを決定的に同定する。

本研究における漸進的改行挿入の流れを図 4 に示す。最初に、節境界解析ツール CBAP[11] を用いて、その時点で入力された文節 b_x とその直前の文節 b_{x-1} の文節境界に節境界が存在するか否かを判定する。次に、3 節で提案した漸進的係り受け解析手法を用いて、2 節の係り受け構造 D_x を文節が入力されるごとに解析する。

最後に、改行挿入判定では、その時点で入力された文節 b_x とその直前の文節 b_{x-1} の間の文節境界に対して、その位置に改行を挿入するか否かを最大エントロピー法を用いて決定的に判定する。最大エントロピー法では、それまでに入力された文節列の形態素情報、及び、文節境界情報、節境界情報、係り受け情報、また、それまでに決定した改行位置の情報を利用して、その位置に改行を挿入する確率（改行挿入確率）を推定する。その確率が 0.5 より大きい場合に改行を挿入すると判定する。

文節 b_{x-1} と b_x の間の文節境界に対する改行挿入確率を ME 法で推定する際に用いた素性は、村田らの手法 [6] (1 文が入力された後に、文全体の最尤な改行位

表 2: 改行挿入実験の結果

	再現率	適合率	F 値
本手法	79.6% (4,375/5,497)	70.2% (4,375/6,228)	74.6%
従来手法	73.7% (4,052/5,497)	74.4% (4,052/5,447)	74.0%

置を求める改行挿入手法) で用いた素性のうち、係り受けに関する以下の 3 つの素性を除いたものである。

- b_{x-1} が、節末文節に係るか否か
- b_{x-1} が、行頭から文字数が最長文字数 (20 文字) 以内の位置にある文節に係るか否か
- b_{x-1} の右側で、かつ、行頭からの文字数が最大表示文字数以内の位置にある文節の中で、 b_{x-1} と同じ係り先をもつ文節があるか否か

上述の素性は、 b_{x-1} の係り先の文節が b_x でない場合、その係り先の文節が入力されるまで素性の値を決定できないため、入力に追従して決定的に改行挿入位置を同定する本手法では利用していない。

5.2 改行挿入実験

2 節で提案した係り受け構造の利用可能性を評価するため、5.1 節で述べた漸進的改行挿入手法による改行挿入実験を行った。

5.2.1 実験概要

実験は、4.1 節と同じデータを用いて、同様の交差検定により行った。なお、図 4 と同じ入力とするため、テストデータ中の人手で付与された節境界情報、係り受け情報、改行情報は削除した。

評価は、正解の改行点に対する再現率 (= 正しく挿入された改行数/正解の改行数) 及び適合率 (= 正しく挿入された改行数/挿入された改行数) により行った。

比較のため、文ごとに 1 文中の改行位置を求める村田らの手法 [6] (以下、従来手法) で改行挿入を行った。なお、従来手法への入力としては、形態素情報と文節境界情報は正解データのもの、節境界情報と係り受け情報はそれぞれ、節境界解析ツール CBAP、内元らの係り受け解析手法 [7] によって付与したものを利用した。

5.2.2 改行挿入実験の結果

本手法と従来手法の適合率と再現率を表 2 に示す。本手法の F 値は 74.6% であり、文ごとに 1 文中の改行位置を求める従来手法を若干上回る結果となった。本手法は、従来手法と同程度の改行挿入精度を維持しつつ、漸進的に出力できることを確認した。

ただし、1 文中の改行位置がすべて一致した文の割合を調査したところ、本手法は 35.8% (614/1,714)、従来手法は 46.2% (792/1,714) であった。従来手法は文全体を見て最尤の改行位置を決めるのに対して、本手法は 1 文節が入力されるごとに決定的に改行するか否かを判定しており、このような結果につながったと考えられる。また、本手法は従来手法と比べてより多く改行する傾向にあった。正解データとの一致による評価だけでなく、被験者による主観的評価も実施し、実際の読みやすさが本手法により低下していないことを確認したい。

6 おわりに

本稿では、入力済みの文節列に対して、漸進的係り受け解析が出力する係り受け構造として、係り先が未だ入力されていない文節については、係り先は未入力文節であると明示することを提案した。提案した係り受け構造を出力可能な漸進的係り受け解析を開発し、評価実験を行った結果、その実現可能性を確認した。さらに、読みやすい字幕生成のために適切な改行位置を漸進的に同定する手法に利用することを試みた。今後は、提案した係り受け構造の利用可能性をより詳細に評価する予定である。

謝辞 本研究は一部、科研費基盤研究 (B)「同時的な発話理解のための話し言葉処理に関する研究」(No. 22300051) により実施した。

参考文献

- [1] 加藤, 松原, 外山, 稲垣. 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. 信学論, Vol. J86-DII, No. 1, pp. 86–97, 2003.
- [2] 大野, 松原, 柏岡, 加藤, 稲垣. 節境界に基づく独話の漸進的係り受け解析. 信学論, Vol. J90-D, No. 2, pp. 556–566, 2007.
- [3] J. Nivre. Algorithms for deterministic incremental dependency parsing. *Computational Linguistics*, Vol. 34, No. 4, pp. 513–553, 2008.
- [4] 笠, 松原, 稲垣. 英日同時翻訳のための依存構造に基づく訳文生成手法. 信学論, Vol. J92-D, No. 6, pp. 921–933, 2009.
- [5] 大野, 神谷, 松原. 安定性を備えたあいづちコーパスの設計と評価. 信学論, Vol. J94-D, No. 3, pp. 623–627, 2011.
- [6] 村田, 大野, 松原. 読みやすい字幕生成のための講演テキストへの改行挿入. 信学論, Vol. J92-D, No. 9, pp. 1621–1631, 2009.
- [7] 内元, 村田, 関根, 井佐原. 後方文脈を考慮した係り受けモデル. 自然言語処理, Vol. 7, No. 5, pp. 3–18, 2000.
- [8] T. Ohno, S. Matsubara, H. Kashioka, T. Maruyama, H. Tanaka, and Y. Inagaki. Dependency parsing of Japanese monologue using clause boundaries. *Language Resources and Evaluation*, Vol. 40, No. 3-4, pp. 263–279, 2007.
- [9] S. Matsubara, A. Takagi, N. Kawaguchi, and Y. Inagaki. Bilingual spoken monologue corpus for simultaneous machine interpretation research. In *Proc. 3rd Language Resources and Evaluation Conference*, pp. 153–159, 2002.
- [10] Le Zhang. Maximum entropy modeling toolkit for Python and C++ (online). (accessed 2007-09-06).
- [11] 丸山, 柏岡, 熊野, 田中. 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.