日本語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

下岡 和也[†] 内元 清貴^{††}
河原 達也[†] 井佐原 均^{††}

『日本語話し言葉コーパス (CSJ)』を対象として係り受け解析や文境界推定を自動で行なう手法について述べる。話し言葉の独話において、係り受け解析を行なう際に最も大きな問題となるのは、文境界が明示されていないことである。本論文では、文境界推定の精度を向上させる2つの手法を提案する。1つは係り受け情報を用いた統計的機械翻訳に基づく手法、もう1つはSVMを用いたテキストチャンキングに基づく手法である。提案手法により、文境界精度はF値で最大84.9となった。また、文境界推定精度が向上することにより、係り受け解析の精度も75.2%から77.2%に改善された。このように、自動推定した係り受け、文境界の情報を相互に利用することにより、係り受け解析精度、文境界推定精度ともに改善されることが示された。

キーワード: 話し言葉, 係り受け解析, 文境界推定, 機械学習

Interaction between Dependency Structure Analysis and Sentence Boundary Detection in Spontaneous Japanese

KAZUYA SHITAOKA[†], KIYOTAKA UCHIMOTO^{††}, TATSUYA KAWAHARA[†]
and HITOSHI ISAHARA^{††}

This paper describes methods to detect dependencies between Japanese phrasal units called *bunsetsus*, and sentence boundaries in a spontaneous speech corpus. In spontaneous monologues, the biggest problem with dependency structure analysis is that sentence boundaries are ambiguous. In this paper, we propose two methods for improving the accuracy of sentence boundary detection in spontaneous Japanese speech: One is based on statistical machine translation using dependency information and the other is based on text chunking using SVM. An F-measure of 84.9 was achieved for the accuracy of sentence boundary detection by using the proposed methods. The accuracy of dependency structure analysis was also improved from 75.2% to 77.2% by using automatically detected sentence boundaries. Furthermore, the accuracy of dependency structure analysis and that of sentence boundary detection were improved by interactively using the counterpart results.

KeyWords: *spontaneous speech, dependency structure analysis, sentence boundary detection, machine learning*

[†] 京都大学情報学研究科, School of Informatics, Kyoto University

^{††} 情報通信研究機構, National Institute of Information and Communications Technology

1 はじめに

『日本語話し言葉コーパス (CSJ)』(古井, 前川, 井佐原 2000) は, 学会講演や模擬講演などのモノローグを対象として収集・構築されたコーパスである. このコーパスには音声データだけでなく書き起こしも含まれており, 書き起こしの一部(コア)には, 人手により形態素・係り受け・談話構造など様々な情報が付与されている. しかし, コーパスの残りの大部分についてもこれらの情報を同様に人手で付与するとなると, 膨大な時間がかかってしまうため, 残りについては自動あるいは半自動で付与できることが望まれる. 本論文では, 係り受け情報に着目し, CSJ を対象として自動で係り受け解析を行なうことを考える.

一般に, 書き言葉の日本語係り受けは, 文節間の係り受けとして定義されることが多い. 同様に, 独話などの長い発話を対象とする場合, すべての文節に対して係り受け関係を定義しようとすると, いわゆる文間関係も文節の関係として定義する必要が生じる. しかし, 文間関係の特定は難しく, 人による揺れが顕著である. また, 重要文抽出の際に文圧縮をしたり格関係などを抽出する場合など, 必要となる係り受けの情報は文単位の係り受けであることが多い. したがって, 定義の難しさと実際の様々な応用とを考慮すると, 文節係り受けは文の単位で特定できれば十分であると考えられる. このような考えに基づき, CSJ においては, 書き言葉のコーパスと同様に係り受けは文の単位で付与されている. しかし, 書き言葉と異なり, 話し言葉では「文」の定義が明確ではない. CSJ では「節」を採用することで話し言葉における文の単位を定義しており(高梨, 丸山, 内元, 井佐原 2003), 本論文においても文の定義はこの定義にしたがうものとする.

これまでの日本語の係り受け解析の研究(藤尾, 松本 1997; 春野, 白井, 大山 1998; 内元, 関根, 井佐原 1999; 内元, 村田, 関根, 井佐原 2000; Kudo and Matsumoto 2000) は, ほとんどが書き言葉を対象としたものであり, 話し言葉を対象としているもの(Matsubara, Murase, Kawaguchi, and Inagaki 2002)についても, 一発話が短い対話が対象で, 文境界は明確であった. 一方, 我々は CSJ のような長い独話を対象とする. ここで一番問題となるのは, 「文」境界が明確でないことである. したがって, 係り受け解析を行なう前にあるいはそれと同時に「文」境界も推定する必要がある. 我々はこれまで自動で文境界推定をする研究を行ってきた(下岡, 南條, 河原 2004) が, この手法では文境界候補にポーズの存在を仮定し, 文字列のパターンマッチングで候補を検出しているため, 本稿で定義する「文」境界の候補を十分に検出することは難しいという問題がある.

本論文では, まず, CSJ のような長い独話における解析上の問題点を明らかにする. その中でも文境界が曖昧であるという問題に着目し, 文境界推定の精度を向上させる 2 つの手法を提案する. まず, 係り受けの情報を新たに文境界推定に用いることを考える. 実際に人手により文境界を検出する際にも係り受けの情報は用いられている. また従来手法は直接的に文境界の検出の学習を行っていなかったが, ここでは機械学習の手法として SVM を用いることも考え

る。さらに、提案手法による文境界推定で得られた結果を用いて、係り受け解析の精度を上げることも併せて考える。

2 話し言葉の係り受け解析と文境界推定

話し言葉は書き言葉と大きく異なる。そのため、話し言葉を対象として係り受け解析・文境界推定を行なう際には、書き言葉では見られない話し言葉特有の問題が生じる。以下では、それぞれの処理を行なう際に生じる話し言葉特有の問題について説明し、それに対してどのように対処するかについて述べる。

2.1 係り受け解析における問題点

(1) 文境界が明示されていない

話し言葉では文境界が明示されていない。そのため、全ての文節に対して係り受けを特定する際には、文間関係も文節の関係として特定する必要がある。しかし、予備実験で複数の被験者に係り受けを付与してもらったところ、文間関係に相当する係り受けは、被験者間の揺れが大きく、安定して係り受けを特定するのが難しいことがわかった。また、自動要約のための文圧縮において不要な要素を削除する場合などで実際に必要となるのは、文節間の修飾・被修飾関係や述語と格要素の関係といった文内の係り受け関係であることが多い。したがって、本研究では、文内の文節間係り受けを対象とし、同時に、文境界の推定も行なう。文境界の定義は次節で述べる CSJ のものに従う。

この問題が話し言葉の係り受け解析を困難にする最も大きな問題と考え、本論文では主にこの問題に着目しその対処法について述べる。

(2) 係り先がない文節がある

話し言葉では、途中で発話のプランが変わったために係り先が消失する場合、「あの一」「その一」などのフィラー、フィラー的な振る舞いをする副詞の「もう」、「はい」「うん」などの相槌、文頭の接続詞「で」、言いよどみなどのように、係り受け関係を特定しても用途はほとんど考えられず、係り受けを定義することに意味がない場合などがある。このような場合、CSJ における定義では係り受けを付与していない。ちなみに、今回実験に用いた CSJ の 188 講演の全文節数 170760 のうち、係り受けを付与していない文節は 14,988 箇所であった。

フィラーや相槌、言いよどみについては、話し言葉に数多く出現するので無視できないが、例えば、浅原らの手法(浅原, 松本 2003)を用いて係り受け解析の前にある程度特定できると考えており、本論文ではすべて削除して扱う。これらを削除するにあたり、本研究では形態素の品詞情報とラベル情報を用いることにした。CSJ では、フィラーや相

槌の品詞は「感動詞」であり、フィラーや言いよどみにはそれぞれ (F) や (D) といったラベルが付与されている。

それ以外の係り先を持たない文節 12295 箇所 (全体の 7.2%) については、便宜上、すべて直後の文節に係るものとして扱う。これらに関しては、本来、正しく「係り先なし」と推定すべきであるが、その推定については今後の課題とする。

(3) 係り受け関係が交差する

一般に、日本語の書き言葉においては「係り受け関係は互いに交差しない」という非交差条件が成り立つと言われている。しかし、話し言葉ではこの非交差条件が成り立たないことも多い。例えば「これが 私は 正しいと 思う」といった場合、

例) これが―――
 私は―――
 正しいと――
 思う

のように「これが」が「正しいと」に係り、「私は」が「思う」に係るので係り受け関係が交差している。しかし、今回用いた 188 講演において、係り受け関係が交差している箇所は 689 箇所とそれほど多くないため、本論文では、係り受けの非交差条件が成り立つと仮定して係り受け解析を行なう。交差している場合については今後の課題である。

(4) 言い直しが多い

話し言葉ではしばしば言い直しが生じる。CSJ では、言い直し関係には、係り受け関係と同様の関係が付与され、さらに、D というラベルが付与されている。今回用いた 188 講演において、言い直し関係は 2,544 箇所であった。言い直し関係以外にも、並列関係・同格関係も係り受け関係と同様の関係が付与され、さらに、それぞれ P・A というラベルが付与されている。このうち、並列関係・同格関係については、書き言葉のコーパスである「京大コーパス」の基準に準拠している (Kurohashi and Nagao 1997)。

本来は、文節間の関係の推定のみではなくそれがどういった関係なのかまで推定すべきであるが、書き言葉を対象にした研究においても多くの場合は関係の有無の推定のみを対象としているため、本論文でも同様にする。

(5) 倒置表現がある

話し言葉ではしばしば倒置表現が用いられる。今回用いた 188 講演において、倒置が用いられているのは 172 箇所であった。CSJ では、倒置は左係りで表現されている。

本論文では、関係を特定することが重要と考え、CSJ における倒置に対して人手で修正を加え、便宜上、すべて右係りになるようにして用いた。具体的には、「ずっと 待ってるんですよ 大の 男が」という文に対して、実際には「男が」が「待ってるんですよ」に倒置に係るわけだが、

例) ずっと——
 待ってるんですよ——
 大の—————
 男が

のように「待ってるんですよ」が「男が」に係るように修正した。

2.2 文境界推定における問題点

日本語の話し言葉においては、文の定義が明確でない。CSJ では、文境界の候補として節境界を自動で検出した後 (丸山, 柏岡, 熊野, 田中 2003), 人手により文境界を特定することで、独自に話し言葉における文の単位を定義している (高梨他 2003)。

まず, CSJ では節境界として次の 3 種類を定義している。

絶対境界 : これはいわゆる文末表現で、述語の終止形・終助詞・「と文末」など。

強境界 : 並列節「ケレドモ」「ガ」「シ」・「ましテ」節・「でシテ」節など。

弱境界 : 理由節「カラ」「ノデ」・連用節・引用節・条件節「タラ」「ト」「ナラ」「レバ」など。これらの節境界は、節直後の切れ目の大きさという観点から区別される。絶対境界と強境界は基本的に文境界となり、弱境界は機能的に区切れていると判断される箇所のみが文境界となる。この判断は人手により行われるのだが、その際に弱境界に対する処理と並行して話し言葉特有の現象である「体言止め」や「言いさし」・「倒置」などの箇所に対する修正も行われる。本論文では、以上のような処理を経て検出された箇所を文境界として用いる。ちなみに、絶対境界と強境界の箇所を全て文境界とした場合、文境界推定精度は、今回用いた CSJ188 講演において、F 値でおよそ 81 であった。

従来の文境界推定手法 (下岡他 2004) では、主に上記の絶対境界にあたる箇所を対象としており、文境界候補にポーズ長を含めた文字列のパターンマッチングで検出してきた。しかし、上記 3 種類の節境界のうち、絶対境界以外は直後にポーズが置かれなくても多く、また、体言止めなどをパターンマッチングで検出するのは困難である。

そこで本論文では、文境界推定のための 2 つの手法を提案し、従来法と比較する。ひとつは、従来法に係り受け情報を利用した方法であり、もうひとつは、機械学習に基づく方法である。

3 係り受け解析と文境界推定のアプローチ

図 1 に本論文で行なう処理の概要を示す。以下では、係り受け解析・文境界推定のそれぞれについて本論文でどのように処理しているかを説明する。

図 1 係り受け解析・文境界推定の概要

3.1 係り受け解析

統計的日本語係り受け解析では、2 文節間の係りやすさは確率値で表される。この確率値は係り受け確率モデルから求められる。

入力文 S を一意に分割された n 個の文節の順序付き集合 $B = b_1, \dots, b_n$ で表す。そして、文全体の係り受け関係 D は、それぞれの文節 $b_i (i = 1, \dots, n-1)$ 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 $D = \{D_1, \dots, D_{n-1}\}$ で表されると仮定する。さらに文節の集合 B が決まると、それぞれ文節 $b_i (1 \leq i \leq n-1)$ と文節 $b_m (m > i, 2 \leq m \leq n)$ に関する素性 $F_{i,m}$ が一意に決まると仮定し、文節の集合 B を次のように素性の集合 F で表す。

$$F = \{F_{1,2}, F_{1,3}, \dots, F_{i,m}, \dots, F_{n-1,n}\}$$

統計的係り受け解析とは、入力文 S が与えられたときに確率 $P(D|S)$ が最も高くなるような文全体の係り受け関係 D_{best} を求める処理のことである。

その際に、前節の考察に基づいて以下の制約を課す。

- 1: 係り受けは前方から後方に向いている
- 2: 係り受け関係は交差しない
- 3: 係り要素は受け要素を一つだけ持つ

従来用いられている係り受け解析モデル (内元他 1999; 藤尾, 松本 1997; 春野他 1998; Collins 1996) では, 2つの文節の関係を「係る」か「係らない」かの2カテゴリとして学習し, 基本的には, 着目している2文節間の関係のみを考慮して2文節が係る確率を求めている. しかし, 本論文で用いる係り受け解析モデル (内元他 2000) では, 2つの文節間の関係を「間」「係る」「越える」の3カテゴリとして学習し, 着目している2文節の間にある文節や, それらよりも文末側にある文節との関係も考慮して2文節が係る確率を求めている. そのため, 従来のモデルに比べてより多くの情報を考慮できると考えられる.

このモデルを最大エントロピー (ME) モデルとして実装した. ME に与える素性としては, 単語の表層表現・品詞・活用形, 文節間距離 (およびそれらの組合せなど) を利用している. また, D_{best} を求めるために, 文末から文頭に向けて解析することにより, 効率良く組み合わせの数を減らしながら一文全体の係り受けを決定する方法を用いている. この方法では解の探索をビームサーチにより行っているが, 決定的に解析を行ってもビーム幅を広くしたときとほとんど同じ精度が得られることが実験によりわかっている (内元他 1999). したがって, 本論文でも文末から決定的に解析する.

3.2 言語尤度を利用した文境界推定 (従来手法)

ここでは話し言葉の文境界推定手法として提案された方法 (下岡他 2004) を従来手法として説明する.

句点を含まないがポーズ情報を含む文字列 X と, 句点を含む文字列 Y を別の言語と考え, 統計的機械翻訳により, 式 (1) に示すように $P(Y|X)$ を最大にする文字列 Y を求める問題として定式化する. 具体的には, ポーズが句点に変換されうる ($P(X|Y) = 1$ となる) 全ての箇所に対して, 句点を挿入する場合としない場合の言語モデル尤度 $P(Y)$ を比較し, 句点挿入の判定を行なう.

$$\max_Y P(Y|X) = \max_Y P(Y)P(X|Y) \quad (1)$$

変換モデル $P(X|Y)$ には, ポーズ前後の表現とポーズ長に依存するモデルを用いる. ポーズ前後の表現として2.2節で述べた節境界の表現を用いた. 話し言葉特有の文末表現「～と」「～ない」「で～」, および文末以外でも頻繁に用いられる文末表現「～た」においては平均ポーズ長以上の場合のみ挿入しうるとし, それ以外の表現では短いポーズ長でも挿入しうるとした. 言語モデル尤度 $P(Y)$ の計算には, 文境界が人手により付与された CSJ の書き起こしから学習された単語 3-gram モデルを用いる.

3.3 係り受け情報を用いた文境界推定 (提案手法 1)

この節では, 提案手法のひとつである, 係り受けを用いた文境界推定の手法について説明する. 従来手法では, 句点が挿入されうる箇所の候補をポーズ長を含めた文字列のパターンマッチ

図 2 文境界に関わる係り受けの条件

ングでのみ行ってきた。ここでは挿入されうる箇所の候補を、係り受け情報も用いて検出することを考える。

文境界直前の文節の係り受け関係について図 2 に示す 3 つの条件が考えられる。

(1) 自分に係る文節が 1 つ以上ある

係り受け関係は文内で閉じている。そのため、倒置文を除いては文境界直前の文節にはその 1 つ前の文節が必ず係ることになる。また本論文では、2.1 節で述べたように倒置で係っている箇所も人手で右係りになるように修正しているため、倒置文においても同様のことが成り立つ。よって、文境界直前の文節には必ず 1 つ以上自分に係る文節が存在すると仮定できる。

(2) 自分を飛び越える文節がない

係り受け関係は文内で閉じていることから、係り受け関係が文をまたぐことは原則としてない。よって、文境界直前の文節には自分を飛び越えて係る文節は存在しないと仮定できる。

(3) 自分が係る確率は低い

本論文では、文境界直前の文節は便宜上直後の文節に係るものとして扱っているが、本来これらの文節はどこにも係らないものである。よって、3.1 節で述べた手法を用いて得られた係り受け解析結果において、文境界直前の文節が係り先の文節に係る確率は低いものになっていると考えられる。

以上 (1)~(3) の条件を全て満たす文節を、文境界直前の文節候補として検出する。そして、3.2 節において検出される文境界候補と統合した後、同様に言語モデル尤度により句点挿入判定を行なう。なお、(2) については、今回扱った 188 講演全てにおいて 50 文節以上離れて係る場合がなかったため、係り受け解析結果において 50 文節以上離れて係っている文節は無視することにした。また、(3) における閾値となる確率をパラメータ p とし、予備実験により最適な値を求める。

本手法により得られる文境界直前の文節候補数は全文節数のおよそ 1/3 程度になる。そのため、前後 2 単語内に文境界候補となる別の単語が含まれていることが十分に考えられる。本稿では、言語モデル尤度の計算に単語 3-gram モデルを用いるため、前後 2 単語内に別の境界候補が存在する場合、それらの処理を逐次的に行うのではなく、統合的に行う必要がある。したがって、前後 2 単語内に文境界候補となる表現が存在しなくなる範囲において、その全ての変換パターンの尤度を比較して最終的な出力を決定することとした。場合によっては相当数の変換パターンが生成される可能性もあるので、ビームサーチを導入した。なお、ビーム幅は 10 で固定した。

3.4 機械学習を利用した文境界推定 (提案手法 2)

この節では、もうひとつの提案手法である、機械学習を用いた文境界推定の手法について説明する。機械学習に基づく方法については、最大エントロピー (ME) モデルに基づく方法 (Reynar and Ratnaparkhi 1997) が、以前、書き言葉を対象に提案されているが、話し言葉ではピリオドなどの句点の情報が利用できないため、この手法を単純には適用できない。

本研究では、Support Vector Machine(SVM) を用いることとした。すなわち、文境界推定の問題をテキストチャンキングの問題として扱い、テキストチャンカとして、SVM に基づく YamCha(Kudo and Matsumoto 2001) を用いることとした。YamCha では、カーネル関数には多項式カーネルが用いられており、現在位置の単語のチャンクタグを推定する際に、前後数単語の単語情報を静的素性として、推定により得られた前数単語のチャンクタグを動的素性として用いている。

本論文では、SVM に与える素性としては以下のものを用いた。

- (1) 前後 3 単語の単語情報 (表層表現・読み・品詞情報・活用の種類・活用形)
- (2) 1 講演で正規化したポーズ長
- (3) 2.2 節で述べたどの文境界候補にマッチしたか
- (4) 自分が係る確率
- (5) 自分に係る文節の個数とその確率

また、YamCha における多項式カーネルの次数は 3、解析方向は Left to Right とし、ラベリングスキームには IOE を用いた。

4 実験と評価

ここでは、係り受け解析と文境界推定の実験結果と考察について述べる。実験に用いたコーパスは CSJ の 188 講演の書き起こしである。テストデータは、全ての実験を通じて同一の 10 講演を用いた。なお、係り受け解析については、講演の最後の文節を除く残り全ての文節に対して係り先を正しく推定できた文節の割合 (係り受け正解率) で、文境界推定については F 値で評価を行なう。また、係り受け解析を行なう際にはテストデータが closed な場合と open な場合の 2 通りを行い精度を比較する。以下の表 1~3 において、それぞれの場合を (closed な場合)・(open な場合) と表記する。

まず、係り受け解析および文境界推定精度のベースラインを求めた。文境界推定のベースラインの手法としては 3.2 節で述べた手法を用いた (図 1 の処理 1)。 $P(Y)$ の推定に使用する言語モデルはテストデータを除く 178 講演で学習した。その結果、再現率 64.5%，適合率 94.2%，F 値 75.6 であった。次に、文境界推定のベースライン手法により文境界を推定し得られた文ごとに係り受け解析を行った結果を係り受け解析のベースラインとした (図 1 の処理 3)。結果は、open テストで 75.2%，closed テストで 80.7% であった。

なお、上記の係り受け解析では文単位で係り受け解析を行なうため、文境界直前の文節の係り受けは推定されない。したがって、係り受けの情報を文境界推定に利用する際には、文境界直前の文節は隣に係るとし、係り受けの確率は、他の文節との整合性を考慮して全て 0.5 に固定した。

4.1 係り受け解析結果を用いた文境界推定 (提案手法 1)

ベースラインの係り受け解析の結果を用いて、3.2 節で述べた手法により文境界推定を行った (図 1 の処理 5)。

まず、3.2 節で述べたパラメータ p についてチューニングを行った。チューニングデータとしてテストデータと異なる 15 講演を用いた。その結果、係り受け解析が open テストの場合、 $p = 0.9$ の時に F 値 78.3、係り受け解析が closed テストの場合、 $p = 0.8$ の時に F 値 78.6 で最大となった。よって、以降の実験ではこの値を用いる。

得られた結果を表 1 に示す。係り受け解析が open な場合は F 値で 1.4、closed な場合は F 値で 2.0 上昇した。用いた係り受け解析の精度が、上記にあるように open な場合と closed な場合とで約 5.5% 異なるにもかかわらず、文境界精度は約 0.6 しか変わらない。これは、文境界と関係している係り受けに関しては open な場合でも closed な場合と同等の精度が得られているためと考えられる。

言語モデル尤度による判定を行わずに、検出された候補を全て文境界であるとした場合の精度は、ベースラインの手法では再現率 68.2%(769/1,127)、適合率 81.5%(769/943)、提案手法

表 1 係り受け情報を用いた文境界精度

	再現率	適合率	F 値
係り受け情報利用 (open な場合)	74.1 % (835/1,127)	82.5 % (835/1,012)	78.0
係り受け情報利用 (closed な場合)	74.2 % (836/1,127)	83.5 % (836/1,001)	78.6
ベースライン	64.5 % (727/1,127)	94.2 % (727/772)	76.6

で係り受け解析が open な場合では再現率 87.2%(983/1,127), 適合率 27.7%(983/3,544) であった。これは, 係り受け情報を使うことで新たに 214 箇所を正しく文境界候補として検出できていることを示している。しかし, 言語モデル尤度による判定を行った結果, これらのうち 108 箇所しか選ばれていない。選ばれていない箇所を調べてみると, 体言止めの箇所, 「～と思う」「～は難しい」といった動詞や形容詞で終わっている箇所, あるいは「～というのは」「～としては」といった箇所であった。これらの箇所, 特に体言止めの箇所以外については, 言語モデルの学習コーパスを増加することである程度対処できると考えられる。

一方, ベースライン・提案手法 1 とともに, 誤って文境界が挿入されている箇所の多くは「～が」「～まして」「～けれども」あるいは「～て」といった箇所である。最初の 3 つの箇所は 2.2 節における絶対境界の表現で, 基本的には文境界となる。また, 「～て」は 2.2 節における弱境界の表現で, 基本的には文境界とはならない。これらの表現はある特殊な状況でのみ文境界とはならない, あるいはある特殊な状況でのみ文境界になる。言語モデル尤度のみではその微かな違いを区別するのは難しいと考えられる。

4.2 SVM を用いた文境界推定 (提案手法 2)

次に, SVM を用いて文境界推定を行った (図 1 の処理 6)。学習にはテストデータを除く 178 講演を用いた。なお, 比較のため係り受け情報を使用しない場合 (図 1 の処理 2) の精度も評価した。

表 2 に結果を示す。4.1 節の結果と比べて F 値が約 6.9 程度高い。これは, 教師つき機械学習の効果と考えられる。また, 係り受け情報を用いることによってわずかに精度が上昇しているが, 提案手法 1 に比べて上昇の度合いが小さくなっている。その原因として, 係り受け解析において用いている素性と SVM で用いている素性が重複していることが考えられる。つまり, 係り受け解析の際に素性として用いられている単語情報は SVM に与える素性 (1) とほとんど同じであるため, 素性 (1) からすでに素性 (4)(5) の情報が得られているのではないかと考えられる。ただし, SVM では前後 3 単語しか見ていないため, それより離れた文節の係り受け情報は素性 (1) からでは得られない。それにもかかわらず精度が変わらない理由として, 離れて係る文節と文境界があまり関係していない, あるいは離れて係る文節は精度が悪いため素性 (5) の情報が

表 2 SVM を用いた文境界精度

	再現率	適合率	F 値
係り受け情報利用 (open な場合)	80.0 % (902/1,127)	90.3 % (902/999)	84.9
係り受け情報利用 (closed な場合)	79.9 % (900/1,127)	90.5 % (900/994)	84.9
係り受け情報 利用せず	79.3 % (894/1,127)	90.1 % (894/992)	84.4

表 3 文境界推定結果を用いた係り受けの再推定結果

	open な場合	closed な場合
4.1 節の結果	75.8 %	81.2 %
4.2 節の結果	77.2 %	82.5 %
ベースライン	75.2 %	80.7 %

生かされていない, といったことが考えられる.

4.3 文境界推定結果を用いた係り受け解析

上記の 2 つの手法により得られた文境界推定の結果を用いて, 再度, 係り受け解析を行った (図 1 の処理 7). 表 3 にその結果を示す. 4.2 節の結果は係り受け情報を用いた場合のものを使用した. open テスト, closed テストともに, 最も精度が高かった文境界推定結果を用いることで約 2 % 程度上昇している. これは, 文境界精度が上昇することでより多くの文末の文節を特定でき, また, それによって別の文の文節に誤って係っていた箇所などが改善されたためである.

ここで, 文境界推定の影響を調べるため, 精度が 100 % であると仮定して実験を行った. 結果は open テストで 80.6 %, closed テストで 86.1 % であった. つまり, 完全に文境界が推定されたとしても, closed テストでさえ約 14 % 誤りがあり, 書き言葉 (新聞記事) を対象とした場合よりも 8 % 近く精度が低い. これは, 話し言葉には書き言葉のように読点がなく, また, 挿入構造があるため離れた文節に係る場合も多いことなどが原因であると考えられる. 挿入構造に対する対処は今後の課題としたい.

5 まとめ

本論文では, CSJ を対象にして, 係り受け解析と文境界推定を自動で行なう手法について述べた. 長い独話では, 書き言葉や対話に比べて係り受けを推定するのが難しい. その一番の原因は文境界が曖昧なことである. 本論文では, 長い独話を対象に, その文境界を推定するための二つの手法を提案し, また文境界推定結果を利用することで係り受け解析の精度を向上させることができることを示した. さらに, 係り受け解析の結果を文境界推定に利用することによ

て双方の精度を向上させることができることも示した。

文境界推定については、従来手法による精度がF値で76.6であるのに対し、係り受け情報を用いた場合で78.0となり、またSVMを用いることで84.9と改善された。係り受け解析については、従来手法による文境界推定結果を用いた場合75.2%に対し、提案手法による推定結果(SVM)を用いた場合で77.2%となり、2%程度向上した。

今後の課題としては、実験の結果明らかになった問題点をふまえてさらに改善を図ることや、2.1節で述べたように、今回対象としなかった話し言葉の係り受け解析における問題点に対処することなどが挙げられる。

参考文献

- 浅原正幸, 松本裕治 (2003). “形態素解析とチャンキングの組み合わせによるフィラー/言い直し検出.” 言語処理学会 第9回年次大会 発表論文集, pp. 651–654.
- Collins, M. (1996). “A New Statistical Parser Based on Bigram Lexical Dependencies.” In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pp. 184–191.
- 藤尾正和, 松本裕治 (1997). “統計的手法を用いた係り受け解析.” 情報処理学会 自然言語処理研究会 NL117-12, pp. 83–90.
- 古井貞熙, 前川喜久雄, 井佐原均 (2000). “科学技術振興調整費開放的融合研究推進制度—大規模コーパスに基づく『話し言葉工学』の構築—.” 日本音響学会誌, **56** (11), pp. 752–755.
- 春野雅彦, 白井諭, 大山芳史 (1998). “決定木を用いた日本語係り受け解析.” 情報処理学会論文誌, **39** (12), pp. 3177–3186.
- Kudo, T. and Matsumoto, Y. (2000). “Japanese Dependency Structure Analysis Based on Support Vector Machines.” In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 18–25.
- Kudo, T. and Matsumoto, Y. (2001). “Chunking with Support Vector Machines.” In *Proceedings of the 2nd Meeting of the North American Chapter of the Association of Computational Linguistics*, pp. 192–199.
- Kurohashi, S. and Nagao, M. (1997). “Building a Japanese Parsed Corpus while Improving the Parsing System.” In *Proceedings of the Natural Language Processing Pacific Rim Symposium*, pp. 451–456.
- 丸山岳彦, 柏岡秀紀, 熊野正, 田中英輝 (2003). “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 発表論文集, pp. 517–520.
- Matsubara, S., Murase, T., Kawaguchi, N., and Inagaki, Y. (2002). “Stochastic Dependency

- Parsing of Spontaneous Japanese Spoken Language.” In *Proceedings of the 19th International Conference on Computational Linguistics*, pp. 640–645.
- Reynar, J. C. and Ratnaparkhi, A. (1997). “A Maximum Entropy Approach to Identifying Sentence Boundaries.” In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pp. 16–19.
- 下岡和也, 南條浩輝, 河原達也 (2004). “講演の書き起こしに対する統計的手法を用いた文体の整形.” 自然言語処理, 11 (2), pp. 67–83.
- 高梨克也, 丸山岳彦, 内元清貴, 井佐原均 (2003). “話し言葉の文境界—CSJ コーパスにおける文境界の定義と半自動認定—.” 言語処理学会 第9回年次大会 発表論文集, pp. 521–524.
- 内元清貴, 村田真樹, 関根聡, 井佐原均 (2000). “後方文脈を考慮した係り受けモデル.” 自然言語処理, 7 (5), pp. 3–17.
- 内元清貴, 関根聡, 井佐原均 (1999). “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, 40 (9), pp. 3397–3407.

略歴

- 下岡 和也:** 2002 年京都大学工学部情報学科卒業. 2005 年同大学院情報学研究科修士課程修了. 同年豊田中央研究所入所. 話し言葉処理の研究に従事.
- 内元 清貴:** 1994 年京都大学工学部電気工学第二学科卒業. 1996 年同大学院修士課程修了. 博士 (情報学). 同年郵政省通信総合研究所入所. 現在, 独立行政法人情報通信研究機構主任研究員. 自然言語処理の研究に従事. 言語処理学会, 情報処理学会, ACL, 各会員.
- 河原 達也:** 1987 年京都大学工学部情報工学科卒業. 1989 年同大学院修士課程修了. 1990 年同博士後期課程退学. 同年京都大学工学部助手. 1995 年同助教授. 1998 年同大学情報学研究科助教授. 2003 年同大学学術情報メディアセンター教授. 現在に至る. この間, 1995 年から 1996 年まで米国ベル研究所客員研究員. 1998 年から ATR 客員研究員. 1999 年から 2004 年まで国立国語研究所非常勤研究員. 2001 年から 2005 年まで科学技術振興事業団さがけ研究 21 研究者. 音声言語処理, 特に音声認識・理解に関する研究に従事. 京大博士 (工学). 1997 年度日本音響学会栗屋潔学術奨励賞受賞. 2000 年度情報処理学会坂井記念特別賞受賞. 情報処理学会連続音声認識コンソーシアム代表. IEEE SPS Speech TC 委員. 言語処理学会理事. 日本音響学会, 人工知能学会各評議員. 情報処理学会, 電子情報通信学会, IEEE 各会員.
- 井佐原 均:** 1978 年京都大学工学部電気工学第二学科卒業. 1980 年同大学院修士課程修了. 博士 (工学). 同年通商産業省電子技術総合研究所入所. 1995 年郵政省通信総合研究所. 現在, 独立行政法人情報通信研究機構けいはんな

情報通信融合研究センター自然言語グループリーダーおよびタイ自然言語ラボラトリー長. 自然言語処理, 語彙意味論の研究に従事. 言語処理学会, 情報処理学会, 人工知能学会, 日本認知科学会, ACL, 各会員.

(2004 年 7 月 21 日 受付)

(2005 年 3 月 9 日 再受付)

(2005 年 4 月 11 日 採録)