語順の読みにくさに頑健な係り受け解析のための節の始境界検出

中澤 貴大 †,a)　大野 誠寛 †,b)　松原 茂樹 ‡　絹川 博之 †

† 東京電機大学 未来科学部　‡ 名古屋大学 情報連携統括本部

a)nakazawa@cll.im.dendai.ac.jp　b)ohno@mail.dendai.ac.jp

1　はじめに

テキスト入力支援や作文推敲支援といったアプリケーションにおいて，高度な言語処理を実現するためには，構文情報の利用が不可欠となる．これらの入力は即興で生成された書き言葉であり，推敲されたものではないため，読みにくい語順をもった文が含まれることになる．しかし，従来の構文解析手法は，推敲された文（新聞記事など）を入力や学習の対象としているため，読みにくい語順をもった文に対しては解析精度が低下するという問題がある [1]．

一方，独話文の係り受け解析において，節の始境界と終境界を検出し，それらにより文を分割した単位の内部と，その分割単位間の各構造を 2 段階で解析する手法が提案されており，解析精度の向上が確認されている [2]．独話文も即興で生成されるものであり，読みにくい語順の性質をもつと考えられる．そのため，同様のアプローチをとることにより，読みにくい語順の書き言葉に対する解析精度も向上する可能性がある．

そこで本研究では，係り受け解析精度の向上を目指し，そのための前処理として，書き言葉の読みにくい語順をもった文に対する節の始境界検出を試みる．節の終境界については，日本語の場合，述語句が節の終端に配置されるため，高精度に検出できることが知られており，既存手法 [3] をそのまま活用する．

一方，日本語を対象とした節の始境界検出に関する研究はほとんど存在しないが，一部，従来研究 [2] の中において，独話文に対する最大エントロピー法を用いた手法（以下，従来手法）が提案されている．しかし，話し言葉が対象であるため，素性としてポーズ情報が利用されているなど，従来手法で設定された素性が読みにくい語順の書き言葉に対しても有効であるか否かは明らかではない．また，従来手法では，機械学習法として最大エントロピー法だけが試されており，その検出性能は必ずしも十分ではない．

そのため本稿では，まず，書き言葉の読みにくい語順をもった文が，節の始境界に関して，独話文と同様の特徴を備えていることを分析により示す．その上で，分析結果や従来手法 [2] を考慮して，機械学習法および学習データ，素性の 3 つの観点ごとに様々な手法を提案するとともに，書き言葉の読みにくい語順の文を対象として，節の始境界を検出する実験を実施し，各手法の有効性を検証する．

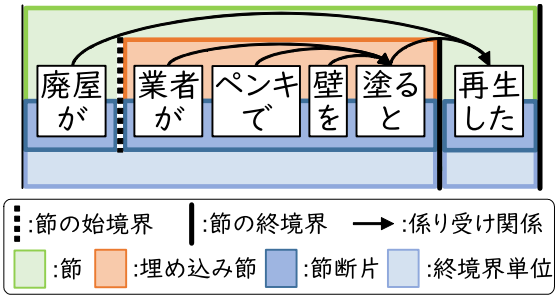


図 1: 終境界単位と埋め込み節

2　読みにくい語順をもった文

本節では，まず，節と読みにくい語順の関係について述べる．その後，読みにくい語順の文の収集方法について説明し，節境界という観点から，読みにくい語順の文の特徴を分析する．

2.1　節と読みにくい語順

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「廃屋が業者がペンキで壁を塗ると再生した」の係り受け構造を示す．節「業者がペンキで壁を塗ると」が節「廃屋が再生した」の中に埋め込まれている．このように埋め込み節が存在する場合，節の終境界だけで分割した単位（以下，終境界単位[1]）の内部では係り受けが閉じない．節の終境界だけでなく始境界でも分割することにより，係り受けが閉じ，かつ，一次元的に分割可能な単位（以下，節断片）となる．

一方，埋め込み節の存在は，読みにくい語順となる要因の一つといえる．図 1 では，埋め込み節の存在により，「廃屋が」とその係り先「再生した」が遠く離れているため，「廃屋が」の係り先が分かりにくくなっている．読みにくい語順の文には，このような埋め込み節が頻出している可能性が高いと考えられ，文を分割して 2 段階で係り受け解析を実行するためには，節の始境界を検出することが不可欠となる．

[1] 文献 [2] が節境界単位と定義した単位と同一である．

表 1: 各データの基礎統計及び分析結果

データ	SIDB	KC	読みにくい語順の文
文	1,935	28,803	546
文節	23,598	250,475	4,735
EOC	9,664	56,510	1,083
BOC	1,405	6,985	380
EOC/文節	0.41	0.23	0.23
BOC/判定対象	0.10	0.04	0.10

2.2 読みにくい語順をもった文の収集

推敲支援への応用を念頭に置くと，日本語母語話者が自然に作成した文を収集することが考えられる．しかし，その収集は容易ではなく，また，そのような文には語順以外にも読みやすさを低下させる要因が含まれるため，語順に焦点を絞ることは難しい．

そこで本研究では，新聞記事中の文は読みやすい語順で書かれていることを前提に，新聞記事文から，読みにくい語順の文を擬似的に作成した．具体的には，まず,（1）京都大学テキストコーパス [4] に含まれる文の読点を除去し，語順をランダムに変更する．次に，（2）母語話者が書くこともあると考えられる，読みにくい語順の文だけを人手により選定する．最後に，（3）選定した文に対して，読みにくい語順を固定した上で，できる限り自然に読めるように読点を人手で挿入することにより，読みにくい語順の文を作成した．なお，手順（1）と（2）の詳細は文献 [1] と同じである．

2.3 読みにくい語順をもった文の特徴分析

本節では，独話文や，書き言葉の読みやすい文との比較により，2.2 節で収集した読みにくい語順の文の特徴を分析する．特に本研究では，節の始境界検出を試みるため，節境界に着目して分析する．

比較のための独話文データとして，同時通訳データベース（以下，SIDB）の 16 講演分 [5] を，また，書き言葉の読みやすい文のデータとして，新聞記事が収録されている京都大学テキストコーパス Version 4.0[4]（以下，KC)[2]を使用した．これら両データには，係り受け情報が人手で付与されている．節の終境界は節境界解析プログラム CBAP[3] を用いて自動的に付与した．節の始境界は係り受け情報と節の終境界の情報を用いて自動的に付与した．

各データの基礎統計および分析結果を表 1 に示す．ここで，EOC は節の終境界を，BOC は埋め込み節の始境界を示す．まず，6 行目の「EOC/文節」，すなわち，総文節数に対する節の終境界の割合に着目する．書き言葉の読みにくい語順の文は，KC から語順を変更して作成した文であるため，これら両者はほぼ同じ割合であった．一方，独話文（SIDB）は書き言葉と比べ，EOC の割合が約 2 倍となった．

次に，7 行目の「BOC/判定対象」，すなわち，節の始境界を検出する際の判定対象となる文節（＝総文節数 - EOC 数）のうち，節の始境界が直後にくるものの割合に着目する．書き言葉の読みにくい語順の文における割合は，SIDB と同程度に高いことがわかる．一方，KC における割合は，他の 2 つと比べ低い．節の始境界の出現割合に関して，書き言葉の読みにくい語順の文は，書き言葉の読みやすい文よりも，独話文に近い性質を持つことが確認できる．

[2]なお，読みにくい語順の文を作成する際に使用した文は，KC から除いた．

3 節の始境界検出

本研究では，基本的に従来手法 [2] と同様の枠組みにより埋め込み節の始境界検出を行う．すなわち，形態素解析，文節まとめ上げ，節の終境界検出が施された 1 文の文節列を入力とする．また，終境界単位の最終文節でない文節に対して，埋め込み節の始境界の直前の文節であるか否か，より詳述すると，終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機械学習を用いて文節ごとに繰り返す．ただし本研究では，書き言葉の読みにくい語順の文に対してより高精度な検出を実現するために，機械学習法と学習データ，素性の 3 つの観点ごとに，様々な手法を用意し，その有効性を検証する．以下では，各観点において検証する手法を説明する．

3.1 機械学習法

機械学習法は以下の 3 つを検証する．

1. 最大エントロピー法（ME）
   従来手法 [2] では，最大エントロピー法により推定された確率値が 0.5 を超えたときのみ，その文節の後に節の始境界があると判定していた．本研究では，閾値 θ を設け，より高精度に検出できる閾値 θ を実験的に定める．

2. サポートベクターマシン（SVM）
   各文節が節の始境界の直前の文節であるか否かを，正例と負例の 2 値で判定する．

3. 条件付き確率場（CRF）
   終境界単位ごとに系列ラベリングを実施する．ラベルは，節の始境界の直前文節であるか否かの 2 種類とする．ただし素性は，1 文の文節列から得られる情報を用いている．

3.2 学習データ

2.3 節の分析結果より，書き言葉の読みにくい語順をもった文は，独話文と書き言葉の両方の性質を持ち合わせている可能性がある．そこで，機械学習に用いる学習データとして，SIDB の単独データ，KC の単独データ，SIDB と KC を併せたデータ（以下，SIDB+KC），の 3 種類のデータの利用をそれぞれ検証する．

3.3 素性

本研究では，従来手法 [2] で設定された素性が書き言葉の読みにくい語順の文に対しても有効であるか否かを明らかにするため，従来手法の素性を基本的にはそのまま採用する．ただし，本研究の対象は書き言葉であるため，従来手法 [2] の素性のうち，ポーズの有無に関する素性は読点の有無に置き換えることとした．以下に本研究で用いた素性を示す．

1) 注目文節（判定を行っている文節）
   - 主辞の基本形(1)，品詞（大分類(2)，細分類(3)）
   - 語形の出現形(4)，品詞（大分類(5)，細分類(6)）
   - 助詞 1 の出現形 (7)，品詞細分類 (8)
   - 助詞 2 の出現形 (9)，品詞細分類 (10)
   - 直後にポーズがあるか否か (11)
2) 注目終境界単位（注目文節が属している終境界単位）
   - ラベル名 (12)
   - 注目文節以降に同一格の文節があるか否か (13)
3) 注目終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 (14)（注目終境界単位が “連体節”の場合のみ）
4) 注目終境界単位の直後の述語文節
   - 注目終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が高いか否か (15)（注目終境界単位が “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）

上述した主辞や語形，助詞 1，助詞 2 の定義，また，ある文節が述語に係りうる文節か否かの判定方法は，従来研究 [2] の通りである．なお，各素性の右肩につけた番号 (x) は，4.2 節において参照する．

4 評価実験

3 節で示した 3 つの観点について検証するため，書き言葉の読みにくい語順の文に対して，節の始境界を検出する実験を行う．まず，機械学習法と学習データの各組み合わせの有効性を検証する実験を実施し，次に，各素性の有効性を検証する実験を実施した．

4.1 機械学習法と学習データに関する実験

機械学習 3 種（ME，SVM，CRF）および，学習データ 3 種（SIDB，KC，SIDB+KC）の組み合わせによる 9 通りの方法について，検出精度を検証する．なお，いずれの組み合わせにおいても，3.3 節で挙げた全素性を共通して用いる．

4.1.1 実験概要

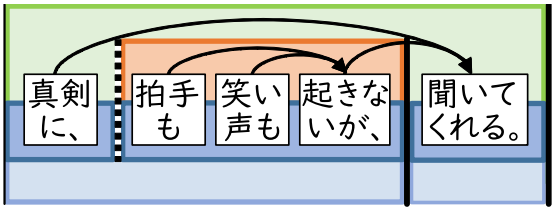
テストデータには 2.2 節の読みにくい語順の文を，学習データには，SIDB，KC，SIDB+KC をそれぞれ利用する．いずれも表 1 に記載のものである．

ME のツールには NLTK[3] を利用した．3.1 節で述べた閾値 θ の値は，$0 \leq \theta \leq 1$ の範囲を 0.1 ずつ変化させ，学習データに SIDB と KC をそれぞれ単独で用いたとき（以下，SIDB 学習時，KC 学習時）は共に $\theta = 0.2$，SIDB+KC を用いたとき（以下，SIDB + KC 学習時）は $\theta = 0.1$ と実験的に決定した．

SVM のツールには LIBSVM[4] を用い，カーネル関数に RBF を採用した．なお，本研究で判定する正例と負例は，出現頻度が大きく異なる不均衡データであるため，通常のパラメータによる SVM では節の始境界を検出できない．そこで，出現頻度の高い負例に対する重み付けは変えず，出現頻度の低い正例に対する重み付けを 1 ずつ変化させ，SIDB 学習時は 5，KC 学習時は 8，SIDB+KC 学習時は 9 と実験的に決定した．

CRF のツールには CRF++[5] を用い，素性については，事前に列挙した上で，素性テンプレートにおいて Unigram 素性として記述した．なお，Bigram 素性を含めるか否かについては，事前実験の結果，KC 学習時および SIDB+KC 学習時のみ含めることとした．

評価指標には，再現率，適合率，F 値を使用する．再現率は，正解データにおける節の始境界のうち，正しく検出できたものの割合を，適合率は，検出結果における節の始境界のうち，正しく検出できたものの割合を示す．F 値は，再現率と適合率の調和平均である．



※各図形の意味は図 I と同様.

図 2: SVM による検出成功例

4.1.2 実験結果

実験結果を表 2 に示す．いずれの学習データを用いた場合でも，機械学習法として SVM を用いた場合が最も高い F 値を達成した．なお，CRF を用いた場合は，他の 2 つと比べ，約 10%～15%低くなった．

一方，学習データに着目すると，いずれの機械学習法を用いた場合でも，KC 学習時に F 値が最も高く，SIDB を用いると低くなる傾向にあった．

9 通りの組み合わせにおいて最も高い F 値が確認できたのは，SVM に学習データとして KC あるいは SIDB+KC を用いる組み合わせとなった．なお，従来手法 [2]（ME，$\theta = 0.5$）の F 値は，SIDB 学習時に 50.58，KC 学習時に 57.89，SIDB+KC 学習時に 53.67 となった．本研究で用いたいずれの機械学習法も，最大の F 値を達成した KC 学習時において，従来手法を上回ることを確認した．

図 2 に，KC 学習時において，SVM では成功し，他の機械学習法では失敗した例を示す．この例では，文節「真剣に,」からの係り受けが終境界単位「真剣に，拍手も笑い声も起きないが,」で閉じずに，後ろの文節に係っているため，埋め込み節が生じている．SVM では，文節「真剣に,」の直後に存在する節の始境界を検出できていたが，ME や CRF では検出できなかった．

[3] http://www.nltk.org/

[4] https://www.csie.ntu.edu.tw/~cjlin/libsvm/

[5] https://taku910.github.io/crfpp/

表 2: 各手法の実験結果

学習データ	ME			SVM			CRF		
	再現率	適合率	F 値	再現率	適合率	F 値	再現率	適合率	F 値
SIDB	77.63% (295/380)	48.51% (295/608)	59.70	58.42% (222/380)	77.08% (222/288)	66.46	40.00% (152/380)	67.25% (152/226)	50.16
KC	63.94% (243/380)	77.14% (243/315)	69.92	64.21% (244/380)	79.73% (244/306)	71.13	45.78% (174/380)	89.23% (174/195)	60.51
SIDB+KC	72.89% (277/380)	63.09% (277/439)	67.63	64.21% (244/380)	79.73% (244/306)	71.13	41.31% (157/380)	90.22% (157/174)	56.67

表 3: 利用した各素性の効果（削除後の F 値増減）

素性名	ME	SVM	CRF
(1) 主辞基本形	−0.45	−2.83	−0.51
(2) 主辞品詞大分類 (3) 主辞品詞細分類	−0.59	−0.09	−0.17
(4) 語形出現形 (5) 語形品詞大分類 (6) 語形品詞細分類	+0.04	±0.00	+0.70
(7) 助詞 1 出現形 (8) 助詞 1 品詞細分類	−0.58	±0.00	−0.38
(9) 助詞 2 出現形 (10) 助詞 2 品詞細分類	−0.39	−0.09	−1.01
(11) 直後に読点があるか	−20.73	−29.14	−22.89
(12) ラベル名	−1.30	±0.00	−6.56
(13) 同一格文節が存在するか	−0.41	−0.09	−0.41
(14) 同一主辞の基本形を持つか	−0.19	−0.59	−0.38
(15) 係り受け確率が高いか	−0.19	−0.09	−0.41

4.2 素性に関する実験

3.3 節に示した各素性が検出精度に与える影響について検証する．

4.2.1 実験概要

3.3 節の全素性を用いた場合に対して，素性をあるまとまりごとに取り除いた場合にどの程度，検出精度が変動するかを確認する．

機械学習の学習データとしては，4.1 節の実験において，いずれの機械学習法でも高精度に検出することができていた KC を用いることとし，4.1 節と同じ読みにくい語順の文を検出対象とする．機械学習法については，ME，SVM，CRF の場合をそれぞれ検証する．利用するデータおよびツール，評価指標の詳細は，4.1 節と同一である．

4.2.2 実験結果

表 3 に，各素性を削除したことによる F 値の増減を示す．なお，表 3 の素性番号は 3.3 節のものに対応する．基本的にいずれの素性も精度向上に貢献しており，特に，直後の読点の有無が最も効果が大きいと考えられる．また，機械学習法により異なるが，次点で，注目終境界単位のラベル名や主辞の基本形の貢献が大きいと考えられる．

読点に関する素性を取り除くと，SVM では，適合率が 30.24%（261/863）になった．節の始境界があるとして出力された数が 863 と，素性を取り除く前と比べ，約 3 倍になっており，誤って節の始境界があると判定することが増えたため，適合率，ひいては，F 値を下げていることが分かった．

5 おわりに

本稿では，読みにくい語順の文における節の始境界を高精度に検出するために，機械学習法や利用するデータを検討し，様々なバリエーションで検証を行った．実験の結果，SVM を用いた手法が最も高い F 値を達成し，従来手法 [2] を上回ることを確認した．また，従来手法 [2] の素性の有効性を検証したところ，ポーズの有無を読点の有無に置き換える必要があるものの，それ以外はいずれの素性も，書き言葉における節の始境界検出に有効であることを確認した．

今後は，独話文データを学習データに利用した際に検出精度が低下した原因について詳しく調査する．具体的には，読点情報が重要であることが判明した点を考慮し，読点とポーズとの関係について調査する．また，その結果に基づいて，独話文データを有効に利用する方法を検討したい．

謝辞 本研究は一部，科研費基盤研究 (C)（No. 16K00300）により実施した．

参考文献

[1] 大野, 吉田, 加藤, 松原. 係り受け解析との同時実行に基づく日本語文の語順整序. 信学論, Vol. J99-D, No. 2, pp. 201–213, 2016.

[2] 大野, 松原, 柏岡, 稲垣. 節の始境界検出に基づく独話文の係り受け解析. 情処学論, Vol. 50, No. 2, pp. 553–562, 2009.

[3] 丸山, 柏岡, 熊野, 田中. 日本語節境界検出プログラム CBAP の開発とその評価. 自然言語処理, Vol. 11, No. 3, pp. 517–520, 2004.

[4] 黒橋, 長尾. 京都大学テキストコーパス・プロジェクト. 言語処理学会第 3 回年次大会発表論文集, pp. 115–118, 1997.

[5] 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.