読みにくい語順の文への読点の自動挿入

宮地 航太 †1　　大野 誠寛 †2　　松原 茂樹 †1
†1 名古屋大学　†2 東京電機大学
miyachi.kota@j.mbox.nagoya-u.ac.jp

1 はじめに

日本語テキスト生成は，機械翻訳や自動要約，音声筆記などの性能を決める重要な技術である．生成されたテキストが高い品質を備えているためには，読点が適切な位置に挿入されている必要がある．というのも，読点は文中の区切りを明示する記号であり，その挿入位置は，文の読みやすさや読み手による文の解釈に影響を与えるためである．

これまでに，日本語テキストに読点を挿入する手法が村田ら [1] により開発されている．村田らは，文構造を明確にする，並列する語の区切りを示す，など，読点の用法ごとにその挿入位置を分析し，その出現傾向を捉える特徴素を設定している．しかし，この手法では，新聞記事など，語順が比較的整った文を対象としている．読点は語順によって打ち方が大きく変わるため，村田らが用いた特徴素で，講演録など語順が整っていない文に対して，その読点位置を検出できるかは明らかでない．

本論文では，読みにくい語順の文に対応した読点挿入手法を提案する．本手法では，語順の適切さに注目する．すなわち，村田らの手法で提案された素性に対し，語順の適切さを表す素性を加えることで，適切な読点挿入位置を同定する．読みにくい語順の文データを用いて読点挿入実験を実施した結果，本手法の有効性を確認した．

本論文の構成は以下の通りである．2 節では，読点と語順の関係とその分析について述べる．3 節では，語順の適切さを考慮した読点挿入手法について説明する．4 節では，日本語テキストデータを用いた読点の挿入実験について報告する．

2 読点と語順の分析

2.1 読みにくい語順の文

以下の 2 文は互いに同一の文節から構成されているが，語順のみ異なる．

S1. 鈴木さんは都会に憧れ家を飛び出した。

S2. 鈴木さんは家を都会に憧れ飛び出した。

S2 は S1 に比べ，読みにくい語順の文といえる．日本語は語順が比較的自由であるものの，実際には選好が存在しているため，文法的には間違いではないものの，読みにくい語順の文が存在する．実際，十分に校閲されていないテキストや音声に，そのような文が多い．

読みにくい語順の文であっても，適切な位置に読点を挿入することにより，読みにくさを軽減することができる．例えば，S2 に読点を挿入した以下の文

S2′. 鈴木さんは家を、都会に憧れ、飛び出した。

は S2 よりも読みやすい.「都会に憧れ」という節の前後に読点を入れることにより，それが挿入節であることを明示することになり，文の構造が明確化されるためである．

しかし，読みにくい語順の文の読点位置は，読みやすい語順の文とは異なる．例えば，S1 に対する適切な読点挿入位置は，

S1′. 鈴木さんは都会に憧れ、家を飛び出した。

であり，S2′ の読点位置と異なり,「家を」の後に読点は挿入されない．

以上の通り，読みにくい語順の文は読みやすい語順の文と読点の入れ方が異なるため，その読点位置の検出を実現するために，読点が挿入された読みにくい語順の文を分析し，その特徴を明らかにする必要がある．

2.2 読みにくい語順の文の読点データ作成

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，読みにくい語順の文を新聞記事文から擬似的に作成し [2]，人手で読点を付与することにより読点付きの読みにくい語順の文データを作成した. 具体的には，

1. 京大テキストコーパス [4] に含まれる文の読点を除去し，係り受け構造を考慮して語順変更後に係り受け関係が交差することがないという制約の下，文節単位で語順をランダムに変更する．
2. 語順を変更した文のうち，母語話者が書くこともあると考えられる文だけを人手により選定する．
3. 選定した文に対して，それをできる限り自然に読めるように 3 名の作業者 (A，B，C) がそれぞれ読点を挿入する．

という手順で作成した．

上記の手順を，京大テキストコーパスに収録されている毎日新聞 1995 年 1 月 9 日の記事中の文に対して適用し，読みにくい語順の文データ 546 文を作成した．このうち 273 文とその 3 種類の読点挿入位置を**分析データ**とした. 以下の分析は全て，分析データと各文に対応した新聞記事の文との比較により行った。

表 1: データごとの読点の総数

新聞記事	読みにくい語順の文		
	A	B	C
309	479	464	494

2.3 読点の挿入頻度

新聞記事コーパスと分析データとの間で読点数を比較した．表 1 に読点の総数を示す．読みにくい語順の文に対して，各作業者が挿入した読点の総数はいずれも，新聞記事の文の総数よりも多い．これは，語順が読みにくくなることによって文構造の把握が難しくなり，より多くの読点が必要になることを示している．

2.4 語順の変化と読点有無の関係

語順は文を構成する要素（本研究では文節）の集合に対して定義されるものであり，読点も 1 文全体を考慮して決まるものである．そのため，語順の変化による読点への影響を分析するには，文を分析の単位として，その変化を調査する必要があるが，データスパースネスを考慮し，ここでは，文節を単位に分析する．

具体的には，新聞記事文（読みやすい語順の文）と，その文から作成した分析データの文（読みにくい語順の文）とを比較することにより，両者の語順において，ある 1 つの文節の語順位置に変化があるか否かと，分析データにおいて，その文節の直後に読点が挿入されているか否かとの関係を調査した．

ここで，ある文節 b_i の語順位置に変化があるか否かは，文節 b_i と同じ係り先をもつ文節の集合 $Sibling(b_i)$（ただし，$b_i \notin Sibling(b_i)$）を考え，その集合内の各文節 $b_x \in Sibling(b_i)$ と文節 b_i との間の順序関係が，語順変更前後で変化しているか否かにより判定する．$Sibling(b_i)$ 内の全ての文節との間で順序関係に変化がない場合は，b_i の語順位置は変化無とし，順序関係に変化があるものが $Sibling(b_i)$ 内に 1 文節でもあれば，b_i の語順位置は変化有とする．

例えば，2.1 節の文 S1′ と S2′ をもとに，文節「家を」の語順位置に変化があるか否かの判定を考える．「家を」と同じ係り先をもつ文節の集合 $Sibling$(「家を」) は {「鈴木さんは」,「憧れ」} となる.「家を」と「鈴木さんは」の順序関係は，語順変更前後（S1′ と S2′）で変化していないが,「家を」と「憧れ」の順序関係は，語順変更前後（S1′ と S2′）で変化しているため，文節「家を」の語順位置は“ 変化有 ”と判定することになる．なお，この例では，S2′ において「家を」の直後に読点が有ることになる．

分析データ中の文末文節を除く全文節（計 2,107 文節）に対して，語順位置の変化の有無と，直後への読点有無の 2 軸で分類し集計した結果を表 2 に示す．括弧内の割合は，語順位置の変化が有る場合と無い場合のそれぞれにおける，読点の有無の割合を表す．語順位置に変化が無い場合に読点が挿入される割合（12.28%）と比較して，語順位置に変化が有る場合に読点が挿入される割合（37.40%）は高い．これは，同じ係り先を持つ文節との間で適切でない順序関係で文節を配置すると，その文節の直後に読点を打つ必要があることを示唆している．

表 2: 各文節の語順位置の変化と読点有無の関係

読点		有	無
語順位置の変化	有	328 (37.40%)	549 (62.60%)
	無	151 (12.28%)	1,079 (87.72%)

3 読みにくい語順の文への読点挿入

本節では，読みにくい語順の文への読点挿入手法について述べる．本手法では，形態素解析，文節まとめ上げ，節境界解析，係り受け解析が与えられた文を入力とし，入力文中の各形態素境界に対して，その位置が読点位置であるか否かを SVM により同定する．

3.1 読点の用法に基づく素性

本手法では，SVM で用いる素性として，村田ら [1] の素性を使用する．村田らは,「文構造を明確にする」「並列する語の区切りを示す」などの用法に応じて読点を分類し，用法ごとに読点の挿入位置を分析することにより，その出現傾向を捉えた特徴素を設定している．以下では，これを**基本素性**と呼ぶ．

基本素性は，読みやすい語順の典型である新聞記事コーパスを用いた実験によりその有用性が示されている．しかし，2.1 節でも論じた通り，読みにくい語順の文では読点の挿入傾向が異なり，また，2.3 節で示した通り，挿入される読点間の間隔も狭いため，その学習に基本素性のみで十分であるかは明らかではない．そこで，基本素性のみを用いた場合の読みにくい語順の文に対する挿入性能を調査するため，予備実験を行った．

3.2 基本素性による読点挿入性能の調査

学習には，京大テキストコーパス [4] に収録されているテキストから，2.2 節の分析データの作成に用いた文を除く 37,854 文のデータを用いた．テストには分析データ (読みにくい語順の文) と，それに対応した京大テキストコーパスの文 (新聞記事文) を用い，それぞれに対する読点挿入性能を比較する．

実験では，読点を除いた文を入力とし，形態素，及び，係り受けは京大テキストコーパスのデータを，節境界は解析ツール CBAP[5] で付与したものをそれぞれ使用した．また，SVM のツールとして LIBSVM[1] を利用した．LIBSVM では，スケーリングでは“ -l 0 ”，学習では“ -h 0 -b 1 ”, テストでは“ -q -b 1 ”のオプションをそれぞれ付与した．その他の設定はデフォルトと同じものを用いた．

評価は，3 名の作業者による読点のうち，2 名以上によって付与された読点位置を正解の読点位置とし，

[1] http://www.csie.ntu.edu.tw/ cjlin/libsvm

表 3: 語順の違いによる読点挿入性能の比較

	再現率	適合率	F 値
新聞記事	58.53% (175/299)	67.31% (175/260)	62.61
読みにくい語順の文	36.33% (174/479)	74.68% (174/233)	48.88

正解に対する再現率，及び，適合率により行った．再現率，適合率はそれぞれ，

$$\text{再現率} = \frac{\text{正しく挿入された読点数}}{\text{正解の読点数}}$$

$$\text{適合率} = \frac{\text{正しく挿入された読点数}}{\text{挿入された読点数}}$$

を測定した．

京大テキストコーパス (新聞記事文)，及び，分析データ (読みにくい語順の文) の各テストデータにおいて，SVM の素性として基本素性のみを用いた場合の読点挿入性能を調査した．

各データにおける再現率，適合率とその F 値を表 3 に示す．新聞記事文と比べ，読みにくい語順の文に対する F 値は大きく低下した．新聞記事文を学習データとして用いているため，読みにくい語順の特徴を捉えきれていないことが原因として考えられる．

3.3 語順の適切さに関する素性

本手法では，読みにくい語順の文への適切な読点挿入を実現するため，2.4 節の分析に基づいて，各文節の語順位置が適切であるか否かを示す素性を新たに導入する．具体的には，形態素 m_j の直後の形態素境界に読点を挿入するか否かを推定する際に，

- m_j が文節の最終形態素であり，かつ，m_j が属する文節 b_i と同じ係り先に係る文節が 1 つ以上存在する場合，文節 b_i の語順位置が適切であるか否か

という素性を導入する．

ここで，文節 b_i の語順位置が適切であるか否かは，2.4 節の分析と同様に，文節 b_i と同じ係り先文節 $b_h (h \neq i)$ に係る文節の集合 $Sibling(b_i)$（ただし，$b_i, b_h \notin Sibling(b_i)$）を考え，その集合内の各文節 $b_x \in Sibling(b_i)$ と文節 b_i との間の順序関係（2 文節間の順序関係）が適切であるか否かにより判定する．$Sibling(b_i)$ 内の全ての文節との間で順序関係が適切であれば，b_i の語順位置は適切であるとし，順序関係が適切でないものが $Sibling(b_i)$ 内に 1 文節でもあれば，b_i の語順位置は適切でないとする．

2 文節間の順序関係が適切であるか否かは，コーパスを用いて統計的に判定する．2.4 節の分析では，語順が読みやすい文と読みにくい文との対応データ（語順だけが変化したもの）を使って，各文節の語順位置の変化を判定していたが，実際に読みにくい文に読点を挿入する際には，入力文の語順を読みやすく整えた文のデータの存在を仮定することはできない．そこで，新聞記事コーパス（読みやすい語順の文の集合とみなす）を用いて，文節 b_i と b_x が b_h に係るときに ($D^h_{i,x}$ で表す)，b_i が b_x よりも前方にあるという順序関係 $O(b_i, b_x)$ となる確率 $P(O(b_i, b_x)|D^h_{i,x})$ を以下の式 (1) により推定する．

$$P(O(b_i,b_x)|D^h_{i,x}) = \frac{freq(O(b_i,b_x), D^h_{i,x})}{freq(O(b_i,b_x), D^h_{i,x}) + freq(O(b_x,b_i), D^h_{i,x})} \cong \frac{freq(O(w^f_i, w^f_x), O(w^f_x, w^c_h)}{freq(O(w^f_i, w^f_x), O(w^f_x, w^c_h)) + freq(O(w^f_x, w^f_i), O(w^f_i, w^c_h))} \quad (1)$$

ここで，$freq(O(b_i, b_x), D^h_{i,x})$ は，コーパスにおいて，文節 b_i，b_x が b_h に係り，b_i が b_x よりも前方にある文の数を表す．ただし，この計算は近似的に $freq(O(w^f_i, w^f_x), O(w^f_x, w^c_h))$ により求める．w^f_k は文節 b_k の語形の見出しを，w^c_k は文節 b_k の主辞の見出しをそれぞれ表し，$freq(O(w^f_i, w^f_x), O(w^f_x, w^c_h))$ は，コーパスにおいて，w^f_i が w^f_x より前方にあり，かつ，w^f_x が w^c_h より前方にあるという順序関係で並んでいる文の数である．なお，語形とは各文節内で，品詞の大分類が特殊となるものを除き最も文末に近い形態素であり，主辞とは各文節内で，品詞の大分類が特殊，助詞，接尾辞となるものを除き，最も文末に近い形態素を指す [3]．例えば，2.1 節の文 S2′ で考える場合，$freq(O(b_1, b_2), D^5_{1,2}) \cong freq(O(w^f_1, w^f_2), O(w^f_2, w^c_5))$ は文節 b_1 が「鈴木さんは」，b_2 が「家を」，b_5 が「飛び出した.」となるので，コーパス中において，「は」が「を」より前方にあり，かつ，「を」が「飛び出す」より前方にあるという順序関係で並んでいる文の数となる．

式 (1) の $P(O(b_i, b_x)|D^h_{i,x})$ は，読みやすい語順の文を集めたコーパスを用いて推定するため，$P(O(b_i, b_x)|D^h_{i,x}) \geq 0.5$ であるとき，文節 b_i が b_x の前方にあるという順序関係 $O(b_i, b_x)$ は順序関係が適切であると判定する．

4 実験

本手法で新たに導入した語順の適切さに関する素性の有効性を評価するため，読みにくい語順の文に対する読点の挿入実験を実施した．

4.1 実験概要

学習データ，解析器，SVM の設定，及び，評価指標は 3.2 節の予備実験と同様である．テストデータには 2.2 節で作成した 546 文の読点付き読みにくい語順の文データのうち，分析データ以外の 273 文を用いた．なお，2.4 節で示した語順の適切さに関する素性は，学習データを用いて算出している．

3.3 節で新たに導入した素性の有効性を評価するため，村田ら [1] が提案した基本素性のみを用いて読点を挿入する手法をベースラインとして設定し，本手法と性能を比較した．

4.2 実験結果

提案手法並びにベースライン手法の再現率，適合率，及び，F 値を表 4 に示す．提案手法は，再現率で

表 4: 実験結果

	再現率	適合率	F 値
提案手法	41.26% (203/492)	71.73% (203/283)	52.39
ベースライン	39.02% (192/492)	73.28% (192/262)	50.93

提案手法

ヒントを今川焼きに得て、大阪から上京した初代当主、故神戸清次郎が創案したのがそもそもの始まり

ベースライン

ヒントを今川焼きに得て大阪から上京した初代当主、故神戸清次郎が創案したのがそもそもの始まり

図 1: 新聞記事の文「大阪から上京した初代当主、故神戸清次郎がヒントを今川焼きに得て創案したのが、そもそもの始まり。」の読みにくい語順の文に対する提案手法とベースライン手法との読点挿入結果の比較

41.26%, 適合率で 71.73%となり，F 値で 52.39 を達成した．F 値において，ベースライン手法よりも，高い性能を示しており，提案手法の有効性を確認した．

提案手法とベースライン手法による読点挿入結果の例を図 1 に示す．ベースライン手法では，「得て」の直後に読点が挿入できていないが，提案手法では正解データと同様の読点が挿入できている．

4.3 閾値に基づく読点挿入

読みにくい語順の文データを大量に収集することは困難であるため，学習データとして京大テキストコーパスを用いている．しかし，2.3 節の分析が示すように，学習データはテストデータと比べ，読点挿入頻度が少ない．実際，実験結果においても，表 4 からわかるように，出力される読点数は正解データと比べて少なくなっている．そのため，学習器の設定を変更し，より多くの読点を出力した場合の読点挿入性能の比較を行った．LIBSVM の出力値のうち，読点を挿入する確率を表す値を参照し，読点挿入確率の閾値を下げることにより，より多くの読点を出力することを試みた．

各手法における閾値と F 値の関係を図 2 に，また，デフォルトの閾値 (0.5) と F 値が最大となる閾値における各手法の再現率・適合率・F 値を表 5，表 6 にそれぞれ示す．いずれもデフォルトの 0.5 から閾値が下がっていくにつれ F 値が上がり，ベースライン手法では閾値が 0.08 のときに，提案手法では閾値が 0.12 のときに F 値が最大となった．ベースライン手法に比べ提案手法の方が，F 値が最大となる閾値，及び，各閾値での F 値が高いことから，正解の読点位置により高い確率が与えられていることが分かる．

5 まとめ

本論文では，読みにくい語順の文における読点の自動挿入手法を提案した．本手法では，語順の適切さに注目し，先行研究で示された基本素性に語順の適切さを表現する素性を加えることで，適切な読点挿入位置

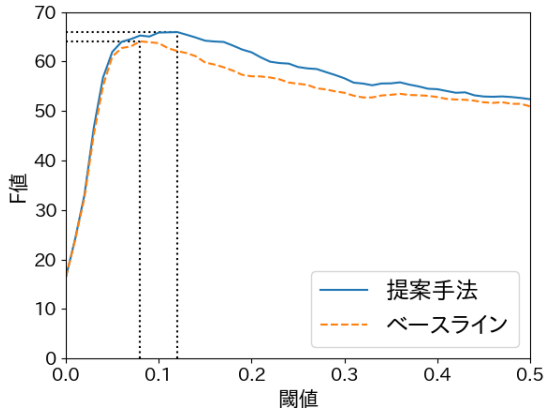

図 2: 各手法における閾値と F 値の関係

表 5: ベースラインにおける閾値ごとの読点挿入性能

閾値	再現率	適合率	F 値
0.5	39.02% (192/492)	73.28% (192/262)	50.93
0.08	66.87% (329/492)	61.50% (219/535)	64.07

表 6: 提案手法における閾値ごとの読点挿入性能

閾値	再現率	適合率	F 値
0.5	41.26% (203/492)	71.73% (203/283)	52.39
0.12	67.68% (333/492)	64.29% (333/518)	65.94

を同定する．読みにくい語順の文データを用いた読点の挿入実験の結果，ベースライン手法より高い F 値を達成しており，本手法の有効性を確認した．

謝辞 本研究は，一部，科学研究費補助金 基盤研究（B）（No. 26280082）及び（C）（No. 16K00300）により実施した．

参考文献

[1] 村田，大野，松原，“読点の用法的分類に基づく日本語テキストへの自動読点挿入，” 電子情報通信学会論文誌，Vol. J95-D，No. 9，pp. 1783-1793，2012.

[2] 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，” 電子情報通信学会論文誌，Vol. J99-D，No. 2，pp. 201-213，2016.

[3] 内元，村田，馬，内山，関根，井佐原，“コーパスからの語順の学習，” 自然言語処理，Vol. 7，No. 4，pp. 163-180， 2000.

[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] 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.