Vol.2010-NL-196 No.8
Vol.2010-SLP-81 No.8
2010/5/27

読点の用法的分類に基づく自動読点挿入

村田匡輝[†1]　大野誠寛[†2]　松原茂樹[†1]

本論文では，日本語テキストにおける読点の自動挿入手法を提案する．読点は，文中の語や句などの構成要素を区切るために重要な役割を果たす．本手法は，音声認識や機械翻訳などの文生成モジュールや日本語の非母語話者のための文作成支援ツールの要素技術として利用できる．本研究では，読点の用法を分類し，用法ごとに読点の出現傾向を調査した．本手法では，形態素，係り受け，節境界，読点間の距離などの情報を素性とする統計的手法によって読点の挿入位置を同定する．テキストコーパスを使用した実験によって本手法の有効性を確認した．

Automatic Comma Insertion based on Usage Classification of Commas

Masaki Murata,[†1] Tomohiro Ohno [†2]
and Shigeki Matsubara[†1]

This paper proposes a method of automatically inserting commas into Japanese texts. In Japanese, commas play important roles in explicitly separating constituents, such as word, phrase and clause, within a sentence. The method can be used as an elemental technology in natural language generation applications such as speech recognition and machine translation, or in writing-support tools for nonnative speakers. We categorized the usages of commas and investigated the appearance tendency for each category. In this method, the points into which commas should be inserted are decided based on a machine learning approach. We conducted a comma insertion experiment using a text corpus and confirmed the effectiveness of our method.

†1 名古屋大学大学院情報科学研究科
Graduate School of Information Science, Nagoya University
†2 名古屋大学大学院国際開発研究科
Graduate School of International Development, Nagoya University

1. はじめに

日本語では，文中の語や句といった構成要素を区切るために読点などの記号が用いられる．図 1 のように，読点が全く打たれていない日本語文は読みづらいばかりでなく，読み手が意味を取り違える可能性もある．それに比べて，適切な位置に読点が挿入された図 2 のような文では，読み手が意味を取り違えることも少なくなる．書き手が，読み手に対して，文の意味を正確に伝えるためには，読点を適切な位置に挿入することが重要である．

しかし，読点の挿入位置については明確な基準が存在しないため，留学生など日本語を母国語としない人々にとって適切な位置に読点を挿入することは難しい．そのような人々の文章作成を支援するために，読点の自動挿入技術が重要となる．またこの技術は，音声認識や機械翻訳など，自然言語生成の要素技術としても利用できる．

本論文では，日本語テキストにおける読点の挿入手法を提案する．読点挿入に関する研究としてこれまでに，機械翻訳の結果や留学生が書いた文書を対象に，ルールベースで挿入する手法が開発されているものの[1),2)]，ルールの数は必ずしも十分ではない．また，音声認識の出力結果を対象とした研究も行われているが[3)–6)]，いずれもポーズの情報を利用しており，テキストを対象とする本研究とは前提が異なる．

読点にはいくつかの用法が存在し，用法ごとに挿入位置が異なる．そのために，まず本研究では，読点の挿入位置，用法に関する文献[7)–9)] を調査し，読点の用法を 9 種類に分類した．機械学習に用いるための有効な素性について検討するため，新聞記事中の読点の挿入位置を，分類した読点の用法に従って分析した．分析では，読点の自動挿入に用いることが可能な情報として，形態素や係り受け，節境界，読点によって挟まれた文節列の文字数に注目した．

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，入力文中の各文節境界に対して，その位置に読点を挿入するか否かを同定する．新聞記事中の読点位置を分析した結果に基づき，形態素や係り受け，節境界などの情報を用いた統計的手法によって，一文中に挿入され得る読点の全ての組み合わせの中から，最適な組み合わせを決定する．

京都コーパス[10)] を用いた読点の挿入実験を行った．4,659 文に対して読点挿入を実行した結果，再現率で 69.13%，適合率で 84.13%を達成した．比較のための設定したベースライン手法と比べ性能が向上しており，本手法の有効性を確認した．

Vol.2010-NL-196 No.8
Vol.2010-SLP-81 No.8
2010/5/27

日本では都市銀行長期信用銀行信託銀行地方銀行信用金庫の一部外国銀行の国内支店などを対象としデリバティブ取引のほぼ一〇〇%を把握する。調査内容はデリバティブの対象となる為替金利エクイティコモディティの四つの大きなグループに分け今年三月末時点の想定元本と時価評価四月一カ月間の取引高を各金融機関から報告を受けて集計する。

図 1 読点が挿入されていないテキスト

日本では都市銀行、長期信用銀行、信託銀行、地方銀行、信用金庫の一部、外国銀行の国内支店などを対象とし、デリバティブ取引のほぼ一〇〇%を把握する。調査内容は、デリバティブの対象となる為替、金利、エクイティ、コモディティの四つの大きなグループに分け、今年三月末時点の想定元本と時価評価、四月一カ月間の取引高を各金融機関から報告を受けて集計する。

図 2 読点が挿入されているテキスト

表 1 読点の用法の分類

番号	用法
1	節間に打たれる読点
2	係り受け関係を明確にする読点
3	難読・誤読を避ける読点
4	主題を示す読点
5	先頭の接続詞・副詞の後に打たれる読点
6	並列する単語・句の間に打たれる読点
7	時間を表わす副詞の後に打たれる読点
8	直前の語句を強調するための読点
9	その他

2. 読点位置の分析

読点位置に関しては，明治 39 年の文部大臣官房図書課草案の句読法（案）をはじめ，様々な議論が行われている．読点にはいくつかの用法が存在し，その用法によって文中での挿入位置が異なるため，挿入位置を機械的に決定する際に用いるべき情報もそれぞれ異なると考えられる．本研究では，まず読点の用法を分類することが重要であると考え，読点について書かれた文献[7)–9)] を調査した．これらの文献の調査結果から，本研究では読点の用法を表 1 のように分類した．

本研究では，表 1 の分類に従い，読点の用法に注目した読点の挿入手法の開発を目指す．適切な読点位置とは，いくつかの要因のバランスのもとに定まると考えられるため，本研究では統計的アプローチを採用する．機械学習のための有効な素性について検討するため，事前分析を与えた．分析には，京都テキストコーパス 4.0（以下，京都コーパス）[10)] の 1 月 1 日，及び，3 日から 11 日までの全記事を用いた．使用した分析データの規模を表 2 に示す．コーパス中のテキストには，形態素，文節境界，係り受け構造の構文的情報が，人手により付与されている．また，節境界解析ツール CBAP[11)] を用いて節境界情報を自動で付与した．

表 2 分析データの規模

文数	11,821
文節数	117,501
文字数	503,970
読点数	16,595
平均文長	42.63

読点のうち， 文節境界以外（すなわち，文節内）に挿入されているものは全体の 1.43% (238/16,595) に過ぎなかった．そこで，文節境界に挿入されている読点のみを分析の対象とした．文節境界 105,680 箇所に対する読点挿入率は 15.48% (16,357/105,680) である．分析では，それぞれの読点の用法ごとに，形態素や係り受け構造，節境界，読点によって挟まれた文節列の文字数の情報に注目し，それらと読点位置との関係について調査した．なお，分類 8 の「直前の語句を強調するための読点」については，執筆者の意図に依存するものであるため，本研究では対象としない．

2.1 節間に打たれる読点

読点を節と節の間に打つことにより文の構造が分かりやすくなる．このことから，節の境界は読点位置として有力であると考えられる．例えば以下の文

- 国連による対イラク制裁解除に向け、関係の深い仏に一層の協力を求めるのが狙いとみられる。

では，文節「向け」の直後に存在する節境界「連用節」に読点が挿入されている．

分析データでは，文末を除く節境界 29,278 箇所のうち 8,805 箇所に読点が挿入されており，節境界に対する挿入率は 30.01%であった．文節境界に対する挿入率よりも高いことから，節境界には読点が挿入されやすいといえる．

分析データに出現した 114 種類の節境界[⋆1]について，種類ごとに読点挿入率を調査した．出現数にして上位 10 種類の節境界とその読点挿入率を表 3 に示す．節境界「連用節」や「並列節ガ」,「並列節デ」の読点挿入率は 80%を越えているのに対して,「連体節」や「引用節」には 5%以下しか読点は挿入されていなかった．これらは，節境界の種類によって読点の挿

⋆1 節境界の種類として，節境界解析ツール CBAP[11)] で定義されたものを用いた．

Vol.2010-NL-196 No.8
Vol.2010-SLP-81 No.8
2010/5/27

表 3 節境界への読点挿入率

節境界	読点挿入率 (%)	
主題ハ	16.94	(1,446/8,536)
連体節	0.72	(43/5,960)
連用節	84.57	(2,685/3,175)
テ節	23.31	(394/1,690)
引用節	4.40	(74/1,680)
補足節	17.53	(245/1,398)
談話標識	60.13	(650/1,081)
並列節ガ	93.85	(946/1,008)
並列節デ	84.52	(606/717)
条件節ト	81.66	(423/518)

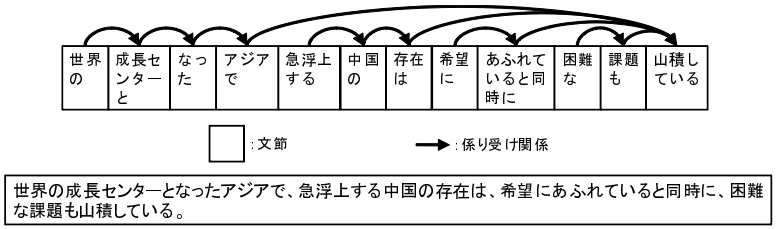

図 3 係り受け関係を明確にする読点

入されやすさが異なることを示している．

2.2 係り受け関係を明確にする読点

読点には係り受け関係を明確にする働きがある．図 3 では,「アジアで」という文節は「山積している」という文節に係るため，その係り受け関係を明確にするため「アジアで」の直後に読点が挿入されている．実際，分析データを調査したところ，係り受け関係にある隣接文節間 66,984 箇所に対して，読点が挿入されたのは 2,302 箇所で，挿入率は 3.44%に過ぎなかった．一方，係り受け関係にない隣接文節間への挿入率は 36.32%であった．

また，係り受け構造と読点との関係，すなわち，読点によって挟まれた文節列内で係り受けが閉じているかどうかを調べた．ここで，係り受けが閉じている文節列とは，文節列外の文節に係る文節が，文節列末の文節以外に存在しない文節列のことをいう．読点に挟まれた文節列 16,357 個のうち，12,496 個 (76.40%) で係り受けが閉じていた．この結果も，係り受け距離が遠くなる文節の直後には読点が挿入されやすい傾向を反映している．

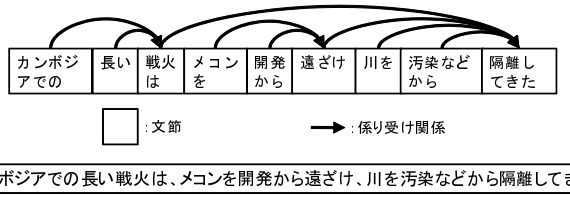

図 4 隣接する文節に係らない文節の直後に存在する節境界「主題ハ」への読点挿入

ある文節の係り先が節末の文節よりも遠い場合，その文節の係り先を明確にするため読点が挿入されやすいと考えられる．例えば，以下の例では，文節「響き」の直後に節境界「連用節」が存在しており，文節「中心に」が節境界の直前の文節「響き」よりも遠くの文節「占拠」に係っている．そのため，文節「中心に」の直後に読点が挿入されている．

- モガディシオからの報道によると、市内のベルムダ地区を中心に、ロケット弾の爆発音が響き、通りを武装兵士が占拠。

隣接する文節が所属する節の節末文節よりも遠くの文節に係る文節の境界に対する読点挿入率を測定した．挿入率は 54.24% (7,478/13,786) であり，文節境界全体と比べて読点が挿入されやすい．

2.3 難読・誤読を避ける読点

漢字やカタカナが続けて出現すると，読み手が誤読をしたり，読みづらさを感じたりすることがある．それを避けるために，このような文節境界には読点が挿入される．以下の例では「営業マン」と「マイケル・スタメンソン氏とともに」の間の読点が，誤読・難読を避けるために挿入されている．

- 出納責任者ロバート・L・シトロン氏は、アドバイザーでもあったメリル・リンチ証券の営業マン、マイケル・スタメンソン氏とともに、米証券取引委員会から事情聴取を受けている模様。

文節にまたがって漢字が出現するような文節境界 2,409 箇所のうち 90.83%(2,188/2,409) に，また，カタカナの場合は 97.69%(211/216) に読点が挿入されていた．文節にまたがって漢字やカタカナが連続する場合，そのほとんどの文節境界に読点が打たれる傾向にある．

2.4 主題を示す読点

文の主題を示すような文節の直後には，主題を明確にする目的で読点が打たれやすいと考えられる．例えば，

- その最大の理由は、香港町が低空飛行を続けるカナダ・トロント経済を活性化している

Vol.2010-NL-196 No.8
Vol.2010-SLP-81 No.8
2010/5/27

からだ。

という文では，主題を示す文節「理由は」を明確にするために，その直後に読点が挿入されている．そこで，節境界「主題ハ」に注目して分析を行った．節境界「主題ハ」への読点挿入率は 16.94% (1,446/8,536) であり，文節境界に対する読点挿入率との差は小さい．これは，単純に主題を表わす文節の直後に読点を挿入すると，読点の数が多くなり文が読みにくくなることから，読点が挿入されないことも多いためであると考えられる．しかし，節境界「主題ハ」に挿入されている読点は，読点全体の 8.84% (1,446/16,357) を占めている．

例えば，図 4 の文節「戦火は」のように，隣接する文節に係らない文節の直後に存在する節境界「主題ハ」への読点挿入率は 20.71%であり,「主題ハ」全体に対する読点挿入率よりも高い．隣接しない文節に係る文節の直後に存在する「主題ハ」には読点が挿入されやすいといえる．また，以下の例の「報道では」のように，節境界「主題ハ」の直前の文字列が「では」であった場合，読点挿入率は 35.82% (254/709) であり,「主題ハ」への読点挿入率よりも高い．

- グロズヌイからの報道では、ロシア軍は激しい空爆と砲撃を加えた後、装甲軍部隊が大統領官邸付近に進出。

2.5 先頭の接続詞・副詞の後に打たれる読点

以下の例の「しかし」のように，文頭に出現する接続詞や副詞の直後には前置きの語を区切るという目的で読点が挿入されることが多い．

- しかし、旧民社党は大半の議員が新進党に参加し、さきがけとの連携も流動的で連携相手は不確定だ。

分析データ中で，最終形態素が「接続詞」である文頭の文節の直後には，71.65% (498/695) の確率で読点が挿入されていた．また，最終形態素が「副詞」である場合では，挿入率は 30.97% (140/452) であった．文節境界に対する読点挿入率よりも高い挿入率であることから，文頭に出現する前置きの語の直後には読点が挿入されやすいといえる．

2.6 並列する単語・句の間に打たれる読点

読点には，対等の関係で並列された同じ種類の語や句を区切るという働きがある．以下に例を示す．

- むしろ地球規模の環境、人口、食糧など広範に国連の果たさなければならない役割は大きい。

この例では,「環境」「人口」「食糧」と並列された名詞を区切るためにその間に読点が挿入されている．最終形態素が名詞である文節が連続する場合，その文節境界への読点挿入率は

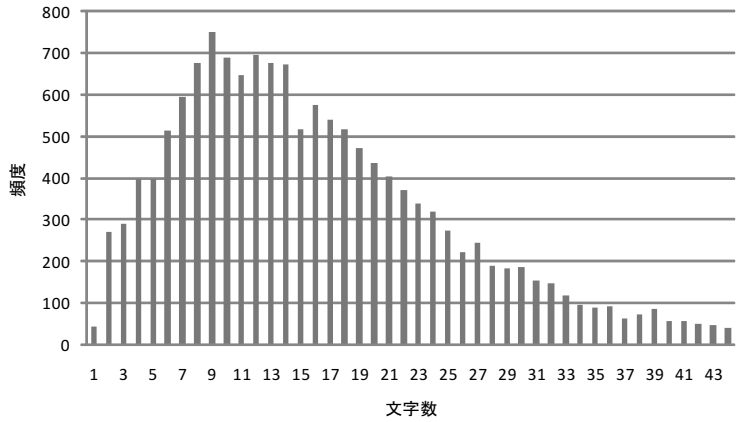

図 5 読点によって挟まれた文節列の文字数と頻度

59.39% (3,330/5,607) であった．

また，語が並列される以外に句が並列される場合がある．以下の例

- メニューは前夜、首相が何を食べたかを調べて同じ献立を避けたり、和食と洋食のバランスを考えたりして決める。

では,「同じ献立を避けたり」という動詞句と「和食と洋食のバランスを考えたり」という動詞句が並列されているため，動詞句の並列を明確にするために，文節「避けたり」の直後に読点が打たれている．分析データ中で，文末の述語に係る文節 b（動詞）が存在し，文節 b の後方に，文末の述語に係る文節（動詞）が存在する場合の文節 b への読点挿入率を調査した．例えば，上記の例文では，文節「避けたり」は文末「決める」に係り，かつ，その後方に同じく「決める」に係る文節「考えたりして」が存在するため，調査対象に該当する．調査の結果，読点挿入率は 79.89% (751/940) であり，文節に対する読点挿入率と比較して大幅に高い値であった．すなわち，節が並列された場合には読点が打たれやすいといえる．

2.7 読点によって挟まれた文節列の文字数

読点には表 1 に示したように様々な用法があるため，むやみに多く打たれる訳ではない．そのため，読点によって挟まれた文節列の文字数が短かくなりすぎることはない．読点によって挟まれた文節列の文字数とその頻度を図 5 に示す．

図 5 から，文字数が 1 文字である文節列の出現頻度は 50 回以下である．文節列の文字数

Vol.2010-NL-196 No.8
Vol.2010-SLP-81 No.8
2010/5/27

が 4 文字，5 文字の場合は頻度が約 400 回であり，6 文字から 21 文字までは頻度が 400 回を超えていた．22 文字以降は頻度が下がっていくという結果が得られた．この分析結果に従って，読点によって挟まれた文節列の文字数を 4 分類し，素性として用いた．

3. 統計的な読点挿入手法

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，入力文中の各文節境界に対して，その位置が読点位置であるか否かを同定する．**2.** の分析において，読点の 98.57%が文節境界に挿入されていたことから，本手法では文節境界のみを読点位置の候補とした．入力文に対する適切な読点位置を同定するために，一文において考えうる読点位置の全ての組み合わせの中から，最適な組み合わせを確率モデルを用いて決定する．

以下では，n 個の文節からなる入力文を $B = b_1 \cdots b_n$ とするとき，読点の挿入結果を $R = r_1 \cdots r_n$ と記す．ここで，r_i は，文節 b_i の直後が読点位置であるか（$r_i = 1$）否か（$r_i = 0$）のいずれかの値をとる．入力文を読点によって m 個に分割した j 個目の文節列を $L_j = b^j_1 \cdots b^j_{n_j} (1 \leq j \leq m)$ とした場合，$1 \leq k < n_j$ のとき $r^j_k = 0$，$k = n_j$ のとき $r^j_k = 1$ となる．

3.1 読点位置検出のための確率モデル

本手法では，入力文の文節列を B とするとき，$P(R|B)$を最大にする読点の挿入結果 R を求める．各文節境界が読点位置であるか否かは，直前の読点位置を除く，他の読点位置とは独立であると仮定すると，$P(R|B)$ は次のように計算できる．

$$
\begin{aligned}
& P(R|B) \qquad (1)\\
= \; & P(r^1_1 = 0, \cdots, r^1_{n_1-1} = 0, r^1_{n_1} = 1, \cdots,\\
& \qquad r^m_1 = 0, \cdots, r^m_{n_m-1} = 0, r^m_{n_m} = 1 | B)\\
\cong \; & P(r^1_1 = 0|B) \times \cdots\\
& \times P(r^1_{n_1-1} = 0 | r^1_{n_1-2} = 0, \cdots, r^1_1 = 0, B)\\
& \times P(r^1_{n_1} = 1 | r^1_{n_1-1} = 0, \cdots, r^1_1 = 0, B) \times \cdots\\
& \times P(r^m_1 = 0 | r^{m-1}_{n_{m-1}} = 1, B) \times \cdots\\
& \times P(r^m_{n_m-1} = 0 | r^m_{n_m-2} = 0, \cdots, r^m_1 = 0, r^{m-1}_{n_{m-1}} = 1, B)\\
& \times P(r^m_{n_m} = 1 | r^m_{n_m-1} = 0, \cdots, r^m_1 = 0, r^{m-1}_{n_{m-1}} = 1, B)
\end{aligned}
$$

ここで，$P(r^j_k = 1 | r^j_{k-1} = 0, \cdots, r^j_1 = 0, r^{j-1}_{n_{j-1}} = 1, B)$ は，1 文の文節列 B が与えられ，$j-1$ 個目の読点位置が同定されているときに，文節 b^j_k の直後に読点が挿入される確率を表す．同様に，$P(r^j_k = 0 | r^j_{k-1} = 0, \cdots, r^j_1 = 0, r^{j-1}_{n_{j-1}} = 1, B)$ は，文節 b^j_k の直後に読点が挿入されない確率を表す．これらの確率を最大エントロピー法により推定した．最尤の読点位置の推定結果は，式 (1) の確率を最大とする読点位置の推定結果であるとして動的計画法を用いて計算する．

表 4 最大エントロピー法で用いた素性

形態素情報	b^j_k の主辞（品詞，活用形）と語形（品詞）
	語形の品詞が「助詞」である場合，語形の表層文字
	文節 b^j_{k+1} の第一形態素（品詞）
節間に打たれる読点	b^j_k の直後に節境界があるか否か
	b^j_k の直後に節境界がある場合，節境界のラベル
係り受け関係を明確にする読点	b^j_k が直後の文節に係るか否か
	b^j_k が b^j_k の隣接文節が所属する節の節末文節より遠くに係るか否か
	b^j_k が直前の文節から係られるか否か
	直前の読点から b^j_k までの文節列で係り受けが閉じているか否か
難読・誤読を避ける読点	b^j_k の最終形態素，かつ，b^j_{k+1} の第一形態素が漢字であるか否か
	b^j_k の最終形態素，かつ，b^j_{k+1} の第一形態素がカタカナであるか否か
主題を示す読点	b^j_k の直後が節境界「主題ハ」であり，かつ，b^j_k が直後の文節に係るか否か
	b^j_k の直後が節境界「主題ハ」であり，かつ，直前の文字列が「では」であるか否か
	b^j_k の直後が節境界「主題ハ」である場合，主題を示す語句[⋆2]の文字数
	b^j_k の直後が節境界「主題ハ」であり，かつ，b^j_k と係り先が同一で，主辞の品詞が「動詞」である文節が存在するか否か
先頭の接続詞・副詞の後に打たれる読点	b^j_k が文頭の文節で，かつ，その最終形態素の品詞が「接続詞」であるか否か
	b^j_k が文頭の文節で，かつ，その最終形態素の品詞が「副詞」であるか否か
並列する単語・句の間に打たれる読点	b^j_k，かつ，b^j_{k+1} の最終形態素の品詞（大分類）が「名詞」であるか否か
	b^j_k の主辞の品詞が「動詞」で文末の述語に係り，かつ，b^j_k より後方に文末の述語に係る文節（主辞の品詞が「動詞」）が存在するか否か
読点によって挟まれた文節列の文字数	直前の読点から b^j_k までの文節列の文字数が以下の 4 分類のいずれであるか（1 文字，2 文字以上 3 文字以下，4 文字以上 21 文字以下，22 文字以上）

3.2 最大エントロピー法で用いた素性

本研究では，$P(r^j_k = 1 | r^j_{k-1} = 0, \cdots, r^j_1 = 0, r^{j-1}_{n_{j-1}} = 1, B)$ ならびに $P(r^j_k = 0 | r^j_{k-1} = 0, \cdots, r^j_1 = 0, r^{j-1}_{n_{j-1}} = 1, B)$ を最大エントロピー法により推定する際，**2.** の分析に基づき，表 4 に示す素性を用いた．

4. 実　　験

本手法の有効性を評価するため，日本語テキストデータを用いて読点の挿入実験を実施した．

Vol.2010-NL-196 No.8
Vol.2010-SLP-81 No.8
2010/5/27

表 5 テストデータの規模

文数	4,659
文節数	46,511
文字数	198,899
読点数（文節境界）	6,549
平均文長	42.69

表 6 実験結果

	再現率	適合率	F 値
提案手法	69.13% (4,527/6,549)	84.13% (4,527/5,381)	75.90
ベースライン	51.38% (3,365/6,549)	70.90% (3,365/4,746)	59.58

4.1 実験概要

実験には京都コーパス[10] に収録されている日本語テキストデータを用いた．テストデータには 1 月 14 日から 17 日の全記事を，学習データには分析データと同一のテキストを使用した．テストデータの規模を表 5 に示す．

なお，実験のための最大エントロピー法のツールとしては，文献[12] のものを利用した．オプションに関しては，学習アルゴリズムにおける繰り返し回数を 2,000 に設定し，それ以外はデフォルトのまま使用した．

評価は，正解の読点位置に対する再現率と適合率，及び，それらの調和平均である F 値により行った．再現率，適合率はそれぞれ以下を測定した．

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

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

を測定した．

比較のために，節や係り受けの情報などを考慮せず，形態素情報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た．ベースライン手法では素性に，文節の主辞（品詞，活用形），語形（品詞，表層文字）と隣接文節の第一形態素（品詞，表層文字）を用いた．

しかし、激しい雪の影響で復旧作業は進まず、直江津発上野行きの特急「あさま３８号」が妙高高原駅で停車したままとなったのをはじめ、計七本が長野、新潟両県内の駅で立ち往生。
九、十の両日行われたジュネーブ会談の成果として、当事者などを含めた拡大会談の開催が決まったが、アラタス外相は「この拡大会談は、国連事務総長のリーダーシップで行われるものではなく、東ティモール人自身の前向きな姿勢で可能となるものである」と強調し、五月十九日にニューヨークで開く次の会談にホルタ氏の参加を促した。
村山富市首相は十六日昼、首相公邸で、さきがけ代表の武村正義蔵相と会談、今後の社会、さきがけ、両党による連携強化や社会党の新民主連合による新会派結成問題について協議する。

図 6 提案手法による読点挿入結果のテキストの例

4.2 実験結果

提案手法ならびにベースラインの再現率，適合率及び F 値を表 6 に示す．提案手法は，再現率で 69.13%，適合率で 84.13%を達成した．F 値の比較において, 提案手法はベースラインと比較して高い性能を示しており，提案手法の有効性を確認した．

提案手法による読点挿入結果のテキストの例を図 6 に示す．節間や並列する名詞の間，文頭の接続詞の直後や主題を示す「は」の直後などに正しく読点が挿入されていることがわかる．すべての読点位置が一致した文数の割合を示す文一致率は 55.81% (2,600/4,659) であり，半数以上の文で，一文中の全ての読点位置に正しく読点を挿入できた．

提案手法による読点挿入結果のテキストとベースラインによる読点挿入結果のテキストを図 7 に示す．ベースライン手法では，文節「浮かんでいるが」の直後や「決まらないため」の直後に読点が挿入されていなかったり,「名乗る」と「副司令官」の間のように不自然な位置に読点が挿入されている．一方，提案手法ではそのような位置に正しく読点が挿入できている．

正解の読点位置のうち，ベースライン手法では読点が挿入されず，提案手法のみ挿入した箇所は 1,603 箇所であった．一方，ベースライン手法によってのみ正解の読点位置に読点を挿入できた箇所は 441 箇所であった．ベースラインで挿入できていなかった種類の読点が提案手法で挿入できるようになったというわけではなく，それぞれの読点の用法に関する素性を用いることによって，読点挿入の性能が全体的に向上したといえる．

⋆2 主題を示す語句とは，係り受け関係を係りから受けにたどって，b_k^j に到達可能な全ての文節と b_k^j からなる文節列のことである．

Vol.2010-NL-196 No.8
Vol.2010-SLP-81 No.8
2010/5/27

（提案手法）

マルコスと名乗る副司令官が表に出てくるが、実際の司令官は不明。

（ベースライン）

マルコスと名乗る、副司令官が表に出てくるが、実際の司令官は不明。

（提案手法）

候補者として石原信雄内閣官房副長官や岩國哲人・島根県出雲市長、鳩山邦夫前労相、作家の堺屋太一氏らの名前が浮かんでいるが、前提となる政党の枠組みが決まらないため、調整は難航。

（ベースライン）

候補者として石原信雄内閣官房副長官や岩國哲人・島根県出雲市長鳩山邦夫前労相、作家の堺屋太一氏らの名前が浮かんでいるが前提となる政党の枠組みが決まらないため調整は難航。

図 7 提案手法とベースラインによる読点挿入結果の比較

表 7 節境界「主題ハ」に対する読点挿入結果

再現率	適合率	F 値
23.46% (141/601)	59.49% (141/237)	33.65

5. 考　　察

5.1 読点挿入誤りの分析

正解の読点位置のうち，読点が挿入されなかった箇所は 2,022 箇所であった．正解の読点位置に挿入されなかった箇所のうち，862 箇所は節境界であり，節境界「主題ハ」がその 53.36% (460/862) を占めていた．節境界「主題ハ」は出現数が多く，読点が挿入される数も多くなる．しかし，読点挿入率自体はそれほど高くない．本手法では，節境界「主題ハ」に関する素性を 4 種類導入しているが，それらが必ずしも有効に働いていなかったといえる．表 7 に節境界「主題ハ」に対する読点挿入の再現率及び適合率を示す．実際，テストデータ中で,「主題ハ」に挿入されている読点は 601 箇所存在したが，そのうち正しく挿入できた箇所は 141 箇所であった.「主題ハ」への読点挿入をより正しく行うための素性の検討が今後の課題となる．読点が挿入できなかった節境界のうち,「主題ハ」に次いで多かったのは節境界「テ節」であり，108 箇所であった．

節境界以外では，連続する名詞間に読点が挿入されなかった箇所が 130 箇所存在した．以下にそのような読点挿入結果の例を示す．

- （正解）
  ボウルに豚の背脂、ニンニク、ショウガ、ネギのみじん切りを入れ、彩りの赤ピーマンも加えます。
- （提案手法）
  ボウルに豚の背脂ニンニク、ショウガ、ネギのみじん切りを入れ、彩りの赤ピーマンも加えます。

上記の例で，正解データでは，文節「背脂」と「ニンニク」の間に読点が挿入されている．提案手法では，文節「背脂」の直後に読点が挿入された場合，読点間の距離が短くなることから，読点によって挟まれた文節列の文字数に関する素性が悪影響を及ぼした可能性がある．一方，文節「ニンニク」と「ショウガ」の間には正しく読点が挿入されているが，これは名詞が連続していることの他に，カタカナが文節にまたがって出現しているため，読点が正しく挿入されたと考えられる．

5.2 不自然な読点挿入

読点位置の中には，正解の読点位置とは異っても許容できる読点位置が存在する．しかし，明らかに不自然な位置に読点が挿入された場合，文の意味が変わったり，読み手が係り受け構造を誤って認識したりするなど，その読点挿入誤りが与える影響は大きい．そこで，提案手法によって明らかに不自然な位置に挿入されている読点位置を調査した．1 月 14 日の記事中の 217 文（2,349 文節）に対する読点挿入結果のうち，提案手法が正解と異なる位置に挿入した読点 47 箇所とした．読点が不自然であるか否かの判定は，3 名の作業者による協議のもと決定した．

調査の結果，明らかに不自然な読点挿入位置と判定したのは，47 箇所のうち 4 箇所であった．以下で，その 4 箇所の読点挿入位置と，不自然と判定した要因について述べる．不自然と判定した読点挿入位置を下線で示す．

- 政党助成を行う主要国の例は、ドイツが年間約百五十億円で、国民一人当たり百八十四円▽フランス、約百五億円同百八十三円▽スウェーデン、約十九億円同二百十七円――などだ。

この文では，文節「フランス」の直後と「スウェーデン」の直後に挿入されている読点を不自然と判断した．この 2 箇所の位置に読点を挿入することによって,「約百五億円同百八十三円▽スウェーデン」が一つのまとまりに見えてしまうためである．

- 同省の特殊法人は計十三、あり、所管省庁別では運輸省に次いで多い。

上記の文では,「十三あり」が一つのまとまりであるにも関わらず,「十三」と「あり」の間

Vol.2010-NL-196 No.8
Vol.2010-SLP-81 No.8
2010/5/27

表 8 人間による読点挿入

	再現率	適合率	F 値
作業者	78.30% (249/318)	80.58% (249/309)	79.42
提案手法	71.07% (226/318)	82.78% (226/273)	76.48

に読点が挿入されている．人間はこのように明らかに不自然な位置に読点を挿入することはないと考えられる．

- 日米首脳会談のため、訪米していた村山富市首相は十三日午後二時前、政府専用機で羽田空港に到着した。

上記の文において，文節「ため」は直後の文節「訪米していた」に係る．しかし,「ため」の直後に読点が挿入されることによって,「ため」が遠くの文節に係ると読み手が錯覚する可能性がある．文の係り受け構造を誤って認識する可能性があることから，この読点は不自然であると判定した．

不自然と判断された箇所が 47 箇所のうち 4 箇所であったことから，提案手法は，ある程度自然な位置に読点を挿入できているといえる．

5.3 人間による読点挿入の一致率

実験では，正解データとの比較によって読点挿入の結果を評価した．しかし，実験結果の値がどの程度の値を示せば十分なのかは定かではない．そこで，人間による読点挿入作業を行い，その結果を一つの指標とし，提案手法の読点挿入性能を評価した．

5.2 で用いたデータと同様の 217 文に対して，日本語の文章作成に精通する作業者 1 名が読点挿入を行った．正解データに対する作業者，及び，同様のデータに対する提案手法の再現率，適合率とその F 値を表 8 に示す．人間の作業においても F 値は 79.42 であり，読点挿入作業は人間でも揺れが生じるタスクであることが分かる．提案手法は F 値において，作業者の読点挿入性能の 96.30% (76.48/79.42) を達成している．また，提案手法は精度で 82.78%を示していることから，揺れが生じる読点作業において，適切な読点挿入が行えていることがわかる．

6. おわりに

本論文では，日本語テキストにおける読点の挿入手法を提案した．本手法では，読点の用法に注目し， 形態素や係り受け，節境界等の情報に基づき，統計的手法によって一文中の適切な読点の挿入位置を同定する．京都コーパス[10] を用いた読点位置の検出実験では再現率と適合率の F 値で 75.90 を示しており，本手法の有効性を確認した．

実験結果の分析から，特定の用法の読点が挿入できていないことが分かった．今後は，その用法の読点に関する，より有効な素性を発見・利用し，本手法の再現率を向上することが課題となる．また，本研究では対象としなかった「直前の語句を強調するための読点」に関しても，今後検討する必要がある．

謝辞 本研究は，一部，科学研究費補助金（若手研究 (B)）(No. 21700157)，ならびに，財団法人電気通信普及財団研究調査助成により実施したものである．

参 考 文 献

1) 鈴木英二，島田静雄，近藤邦雄，佐藤尚，“日本語文章における句読点自動最適配置，” 情報処理学会全国大会講演論文集，Vol.50, No.3, pp.185-186 (1995).
2) 林良彦，“技術文章向けの日本文推敲支援システムの実現と評価，” 電子情報通信学会論文誌，Vol.J77-D-II, No.6, pp.1124-1134 (1994).
3) J. Kim and P. C. Woodland, “The use of prosody in a combined system for punctuation generation and speech recognition, ” In Proceedings of 7th European Conference on Speech Communication and Technology (Eurospeech-2001), pp.2757-2760 (2001).
4) H. Christensen, Y. Gotoh, and S. Renal, “Punctuation annotation using statistical prosody models, ” In Proceedings of ISCA Workshop on Prosody in Speech Recognition and Understanding, pp.35-40 (2001).
5) Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, “Enriching speech recognition with automatic detection of sentence boundaries and disfluencies, ” IEEE Transactions on Audio, Speech, and Language Processing, Vol.14, No.5, pp.1526-1540 (2006).
6) 清水徹，中村哲，河原達也，“音声翻訳単位の推定における句読点情報の効果，” 情報処理学会研究報告，2008-SLP-74, Vol.2008, No.123, pp.127-131 (2008).
7) 本多勝一，日本語の作文技術, 朝日新聞社出版局 (1982).
8) 犬飼隆，文字・表記探究法 （シリーズ < 日本語探究法 > 5），朝倉書店 (2002).
9) 小学館辞典編集部編，句読点、記号・符号活用辞典。, 小学館 (2007).
10) 河原大輔，黒橋禎夫，橋田浩一，“「関係」タグ付きコーパスの作成，” 言語処理学会第 8 回年次大会発表論文集，pp.495-498 (2002).
11) 丸山岳彦，柏岡秀紀，熊野正，田中英輝，“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).
12) L. Zhang: Maximum entropy modeling toolkit for python and c++, http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html (2008) [Online; accessed 1-March-2008].