日本語テキストにおける読点位置の検出

村田 匡輝† 大野 誠寛‡ 松原 茂樹§

†名古屋大学大学院情報科学研究科 ‡名古屋大学大学院国際開発研究科
§名古屋大学情報基盤センター

1 はじめに

わかち書きされない日本語文において，読み手が文章を正確に理解する上で読点の働きは極めて大きい．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の出力テキストの可読性向上にも利用できる [1, 2]．

本論文では，日本語テキストにおける読点位置の検出手法を提案する．本手法では，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読点位置を検出する．京都コーパスを用いて読点位置の検出実験を行った．比較のために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．

2 関連研究

読点挿入に関する研究として，鈴木らは，機械翻訳システムで自然な日本語を生成するための句読点の挿入法を提案している [1]．この手法では，読点や句点で区切られた文字列の文字数や形態素に基づくルールによって読点の挿入位置を決定する．しかし，ルールの数は必ずしも十分ではなく，定量的な評価も行われていない．

林らは，日本語文の推敲支援システムを開発している [3]．この手法では，読点についての支援機能をもたせる目的で，ルールベースの読点挿入を行っている．しかし，このシステムでは，文を短くなるように分割しており，読点は比較的単純な位置に挿入され，一文に挿入される読点の数も最大二つに留まっている．

また，清水らは，音声翻訳単位境界の推定精度向上を目的に，形態素やポーズ，音声翻訳単位境界の情報に基づく句読点位置の推定を行っている [4]．句読点の推定性能は，再現率 82.6%，適合率 85.0%に達している．しかし，ポーズや単位境界の情報を推定に用いており，テキストを対象とする本研究とは前提が異なる．

3 読点位置の分析

読点に関しては，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．我々は文献 [5, 6] を調査し，読点の用法を表 1 のように分類した．

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	並列する単語・句の間に打たれる読点
6	先頭の接続詞の後に打たれる読点
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

表 2: 分析データのサイズ

文数	34,282
文節数	332,068
文字数	1,412,967
読点数	45,784
平均文長	41.22

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位置の検出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）[7] の 1 月 6 日から 17 日の全記事と 1 月から 12 月の社説記事を用いた．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[8] を用いて節境界情報を自動で付与した．分析データの規模を表 2 に示す．

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 2.16%(990/45,784) 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 297,786 箇所に対する読点挿入率は 15.04%(44,794/297,786) である．分析では，

表 3: 節境界への読点挿入率

節境界	読点挿入率 (%)	
主題ハ	20.11	(4,733/23,540)
連体節	1.09	(187/17,112)
連用節	84.75	(7,333/8,653)
補足節	19.04	(879/4,616)
テ節	25.67	(1,171/4,561)
談話標識	59.39	(2,559/4,309)
引用節	4.61	(191/4,142)
並列節ガ	94.82	(2,526/2,664)
並列節デ	81.91	(1,272/1,553)
連体節-形式名詞	0.00	(0/1,284)

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．

3.1 節間に打たれる読点

読点を節と節の間に打つことで文の構造が分かりやすくなる．また，文が連用形などで一時中止される場合にも打たれることが多い．以上のことから，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 83,865 箇所のうち 26,032 箇所に読点が挿入されており，節境界に対する挿入率は 31.04%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．

分析データに出現した 121 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「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-形式名詞」には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入されやすさが異なることを示している．

3.2 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．係り受け距離が近いなど，係り受け関係が分かりやすい文節の直後には読点が打たれにくく，離れた位置の文節に係る文節の直後には読点が打たれやすい．以下の例では，「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．

[1]節境界の種類として，節境界解析ツール CBAP[8] で定義されたものを用いた．

- 世界の成長センターとなったアジアで、急浮上する中国の存在は、希望にあふれていると同時に、困難な課題も山積している。

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 192,540 箇所に対して，読点が挿入されたのは 5,866 箇所で，挿入率は 3.04%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 36.99%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 44,794 個のうち，35,494 個 (79.11%) 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．

3.3 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 5,235 箇所のうち，92.13%(4,823/5,235) に，また，カタカナの場合は 98.55%(340/345) 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．

3.4 主題を示す読点

文の主題を示すような文節の直後には読点が打たれやすいと考えられる．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」の読点挿入率は 20.11%(4,733/23,540) であり，文節境界に対する読点挿入率よりも高くなっている．しかし，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．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 24.75%であり,「主題ハ」全体に対する読点挿入率よりも高い．ある程度離れた文節に係る文節の直後に存在する「主題ハ」へは読点が挿入されやすいと言える．

4 読点位置の検出手法

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

表 4: 最大エントロピー法で用いた素性

形態素情報	b_k^j の主辞（品詞，活用形）と語形（品詞）
	b_k^j の語形の品詞が「助詞」である場合，語形の表層文字
	b_{k+1}^j の第一形態素 (品詞)
節間に打たれる読点	b_k^j の直後に節境界があるか否か
	b_k^j 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_k^j が直後の文節に係るか否か
	b_k^j が節末文節に係るか否か
	b_k^j が直前の文節から係られるか否か
	直前の読点から b_k^j 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_k^j の最終形態素，かつ，b_{k+1}^j の第一形態素が漢字であるか否か
	b_k^j の最終形態素，かつ，b_{k+1}^j 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_k^j の直後が節境界「主題ハ」であり，かつ，b_k^j が直後の文節に係るか否か
	b_k^j 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_k^j の直後が節境界「主題ハ」であり，かつ， b_k^j と係り先が同一である文節が存在し，その文節の主辞の品詞が動詞であるか否か
先頭の接続詞の後に打たれる読点	b_k^j が文頭の文節で，かつ，その最終形態素の品詞が「接続詞」か否か
その他	b_k^j の主辞の品詞が動詞で文末の述語に係り，かつ， b_k^j 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
	直前の読点から b_k^j 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 17 文字以下，18 文字以上)

かを同定する．3 章の分析において，読点の 97.84%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，読点位置の推定結果を $R = r_1 \cdots r_n$ 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 $L_j = b_1^j \cdots b_{n_j}^j (1 \leq j \leq m)$ とした場合，$1 \leq k < n_j$ のとき $r_k^j = 0$，$k = n_j$ のとき $r_k^j = 1$ となる．

4.1 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，$P(R|B)$ を最大にする読点位置の推定結果 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．

$$
\begin{aligned}
& P(R|B) \qquad (1)\\
= \;& P(r_1^1 = 0, \cdots, r_{n_1-1}^1 = 0, r_{n_1}^1 = 1, \cdots,\\
& \quad r_1^m = 0, \cdots, r_{n_m-1}^m = 0, r_{n_m}^m = 1|B)\\
\cong \;& P(r_1^1 = 0|B) \times \cdots\\
& \times P(r_{n_1-1}^1 = 0|r_{n_1-2}^1 = 0, \cdots, r_1^1 = 0, B)\\
& \times P(r_{n_1}^1 = 1|r_{n_1-1}^1 = 0, \cdots, r_1^1 = 0, B) \times \cdots\\
& \times P(r_1^m = 0|r_{n_{m-1}}^{m-1} = 1, B) \times \cdots\\
& \times P(r_{n_m-1}^m = 0|r_{n_m-2}^m = 0, \cdots, r_1^m = 0, r_{n_{m-1}}^{m-1} = 1, B)\\
& \times P(r_{n_m}^m = 1|r_{n_m-1}^m = 0, \cdots, r_1^m = 0, r_{n_{m-1}}^{m-1} = 1, B)
\end{aligned}
$$

ここで，$P(r_k^j = 1|r_{k-1}^j = 0, \cdots, r_1^j = 0, r_{n_{j-1}}^{j-1} = 1, B)$ 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_k^j の直後に読点が挿入される確率を表す．同様に，$P(r_k^j = 0|r_{k-1}^j = 0, \cdots, r_1^j = 0, r_{n_{j-1}}^{j-1} = 1, B)$ は，文節 b_k^j 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．

4.2 最大エントロピー法で用いた素性

本研究では，$P(r_k^j = 1|r_{k-1}^j = 0, \cdots, r_1^j = 0, r_{n_{j-1}}^{j-1} = 1, B)$ ならびに $P(r_k^j = 0|r_{k-1}^j = 0, \cdots, r_1^j = 0, r_{n_{j-1}}^{j-1} = 1, B)$ を最大エントロピー法により推定する際，3 章の分析に基づき，表 4 に示す素性を用いた．

なお，本手法では，素性として係り受け情報を用いている．読点を取り除いたテキストで学習した係り受け解析器 CaboCha[10] を用いて読点が挿入されていないテキストの係り受け解析を行ったところ，正解率は 88.51%であり，同じテキストを読点が入った状態で学習し，係り受け解析を行った場合の解析正解率 89.42%と比較して，正解率の低下はそれほど見られなかった．この結果から，読点位置の検出に係り受け情報を使用することは妥当であると考え，素性として用いている．

[2]主題を示す語句とは，係り受け関係を係りから受けに辿って，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．

表 5: 実験結果

	再現率	適合率	F 値
提案手法	68.70% (3,902/5,680)	83.02% (3,902/4,700)	75.18
ベースライン	55.12% (3,131/5,680)	73.26% (3,131/4,274)	62.91

5 実験

5.1 実験概要

実験には京都コーパス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1日，1月3日から5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．

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 [9] 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，

$$\text{再現率} = \frac{\text{正しく挿入された改行数}}{\text{正解の改行数}}$$

$$\text{適合率} = \frac{\text{正しく挿入された改行数}}{\text{挿入された改行数}}$$

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）と隣接文節の第一形態素（品詞，表層）を用いた．

5.2 実験結果

提案手法ならびにベースラインの再現率と適合率を表 5 に示す．提案手法は，再現率で 68.70%，適合率で 83.02%を達成した．これらの調和平均である 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．

正解の読点位置のうち，検出されなかった箇所は 1,778 箇所であった．そのうち，426 箇所は節境界「主題ハ」であった．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く，節境界「主題ハ」に関する素性が有効に働いていなかったと言える．実際，テストデータ中で，「主題ハ」に挿入されている読点は 525 箇所存在したが，そのうち正しく検出できた箇所は 99 箇所であった.「主題ハ」への読点挿入をより正しく検出するための素性の検討が今後の課題となる．

また，読点位置であると検出された箇所のうち，正解の読点位置と異なるものは 798 箇所存在した．そのうち，文節にまたがって漢字が出現する位置を誤って読点位置であると推定した箇所が 121 箇所存在した．漢字が文節にまたがって出現する場合でも読点が挿入されない場合があるが，そのような箇所も読点位置であると誤検出してしまったと言える．

6 おわりに

本論文では，日本語テキストにおける読点位置の検出手法を提案した．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位置の検出を実現する．京都コーパスを用いた読点位置の検出実験では F 値で 75.18 を示しており，本手法の有効性を確認した．

我々はこれまでに，読みやすい字幕を生成するための改行挿入手法を開発している [11]．より読みやすい字幕を生成するために，改行挿入手法と今回開発した読点挿入手法を組み合わせ，改行と読点を同時的に挿入できる手法を開発することが今後の課題である．

謝辞 本研究は，一部，科研費（若手研究 (B)）(No. 21700157), (財）旭硝子財団研究助成により実施した．

参考文献

[1] 鈴木，島田，近藤，佐藤，“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] 遠山，永田，“音声認識支援システムにおける句読点挿入方法の提案” 信学技報. NBC, Vol.100, No.100, pp.25-32 (2000).

[3] 林，“技術文章向けの日本文推敲支援システムの実現と評価” 信学論 (D)，Vol.J77-D-II, No.6, pp.1124-1134 (1994).

[4] 清水，中村，河原，“音声翻訳単位の推定における句読点情報の効果” 情処研報. SLP, Vol.2008, No.123, pp.127-131 (2008).

[5] 小学館辞典編集部編，句読点、記号・符号活用辞典。, 小学館 (2007).

[6] 本多，日本語の作文技術, 朝日新聞社出版局 (1982).

[7] 河原，黒橋，橋田，“「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).

[8] 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).

[9] L. Zhang: Maximum entropy modeling toolkit for python and c++, http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html (2007).

[10] 工藤，松本，“チャンキングの段階適用による日本語係り受け解析” 情報処理学会論文誌，Vol.43, No.6, pp.1834-1842 (2002).

[11] 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 (D)，Vol.J92-D, No.9, pp.1621-1631 (2009).