文の長さと読点生成確率を用いた読点挿入システム

坂 祥太朗　　山村 毅
愛知県立大学 情報科学部
{is151066@cis, yamamura@ist}.aichi-pu.ac.jp

1 はじめに

読点は日本語の文章を構成する上でとても重要な役割を果たす．文の終わりに挿入すればよい句点と異なり，読点の挿入位置については挿入位置に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 [3]．また近年では，SNS などの急速な普及により短文でのコミュニケーションが増加し，読点を使って文章を作成する機会が減ったため，特に若者を中心に，読点挿入に対してのハードルが上がってきている．我々が普段，使用している Word のようなワードプロセッサには「スペルチェック」という自動校正機能が働いており，スペルミスなどを自動で教えてくれる．しかし，読点についての校正システムはまだ実用化されていないようである．

本研究では，文に読点を挿入するシステムの開発を行う．

林 [1] は，日本語文の推敲支援の一部として，読点・語順の調整を行う方法を提案している．「容易な理解，正確な理解を助ける文を再生成する」ことを目的としたルールベースの方法であり，読点の挿入だけでなく，読点の除去についても取り扱っている．また，鈴木ら [2] は，読点や句点で区切られる文字列の長さに注目した読点の挿入規則を提案している．

村田ら [3, 4, 5] は，文節境界を対象に，形態素や係り受け，節境界などの情報を用いて，一文中に挿入されうる読点の全ての組合せの中から最適なものを決定する統計的手法を提案している．京都コーパスを用いた評価実験で，再現率 70.66%，適合率 84.65%という高い精度を実現している．

一般に読点の使用には，意味の切れめ，文の構造，文の長さ，表記，リズムなどさまざまな要因が絡む [6]．したがって，本来は，こういった要因を考慮して読点を挿入するべきであるが，これには，(特に意味の取り扱いにおいて) 困難が生じる．

そこで，本研究では，簡略化したモデルを考え，それでどの程度正しく読点を挿入することができるのかを調べる．具体的には，文の長さに依存して決まる読点の個数の確率と形態素と形態素の間に読点が入る確率とを用いて読点を挿入する方法を提案する．特に，文の長さの測り方として形態素数，文字数，文節数の 3 つを考え，それらの違いによって読点挿入精度がどう異なるかを考察する．

2 提案手法

n 個の形態素列からなる文を $S = w_1 w_2 \dots w_n$ とする．各形態素 $w_i (i < n)$ の直後に読点が入るか否かを表す確率変数を q_i とするとき，$Q = q_1 q_2 \dots q_{n-1}$ が S に対する読点挿入の一つの結果を表すことになる．これを用いれば，最適な読点挿入の結果 $\hat{Q}$ は，以下によって求めることができる．

$$\hat{Q} = \arg\max_{Q} P(Q|S) \tag{1}$$

しかし一般には，$P(Q|S)$ を直接求めることは難しい．

そこで，各 q_i はその前後の形態素 w_i，w_{i+1} にのみ依存すると仮定し，式 (1) を以下のように近似する．

$$\begin{aligned} P(Q|S) &= P(q_1, \dots, q_{n-1}|w_1, \dots, w_n) \\ &\simeq \prod_{i=1}^{n-1} P(q_i|w_i, w_{i+1}) \end{aligned} \tag{2}$$

だだし，この近似により，文の長さに無関係に読点を挿入してしまうことになるため，「文が長いので読点を入れる」というような場合をうまく表現できなくなる．

そこで，文の長さによって読点を挿入する個数を決める確率 (以下の $P_l(j)$) を式 (2) にかけたものを考え，これを最大化するように各 q_i を決めるようにする．

すなわち，

$$\widehat{q_1 \dots q_{n-1}}, j = \arg\max_{q_1 \dots q_{n-1}, j} P_l(j) \times \prod_{i \in R_j} P(q_i = 1 | w_i, w_{i+1}) \times \prod_{i \in Q - R_j} P(q_i = 0 | w_i, w_{i+1}) \quad (3)$$

ここで，$P_l(j)$ は文の長さが l であった場合に読点を j 個挿入する確率，R_j は j 個の読点を挿入する位置を表す ($Q - R_j$ は読点を挿入しない位置を表す)．また，$q_i = 1/q_i = 0$ は読点が入る/入らないを表す．

3 評価実験

2 で述べた方法を実装し，精度を評価した．評価に先立って，まずは毎日新聞の 2008 年，2009 年の記事を用いて，式 (3) における，$P_l(j)$，$P(q_i = 1|w_i, w_{i+1})$，$P(q_i = 0|w_i, w_{i+1})$ を計算した．形態素解析には MeCab を，依存構造解析には CaboCha を用いた．

$P_l(j)$ の計算においては，文の長さとして，形態素数，文字数，文節数の 3 つを考え，それぞれにおいて，挿入される読点の個数を調べた．

$P(q_i = 1|w_i, w_{i+1})$ と $P(q_i = 0|w_i, w_{i+1})$ の計算においては，ゼロ頻度問題に対処するため，固有表現は特別な記号に置き換えて処理を行なった．具体的には，MeCab の品詞 ID の「名詞, 固有名詞, 一般」,「名詞, 固有名詞, 人名」,「名詞, 固有名詞, 組織」,「名詞, 固有名詞, 地域, 一般」にあたる形態素を，それぞれ特別な記号に置き換えて $P(q_i = 1|w_i, w_{i+1})$ と $P(q_i = 0|w_i, w_{i+1})$ の計算を行なった．

こうして求めた $P_l(j)$，$P(q_i = 1|w_i, w_{i+1})$，$P(q_i = 0|w_i, w_{i+1})$ を用いて，評価対象の文に対して，式 (3) を計算し，読点の挿入を行なった．毎日新聞の 2010 年の記事からランダムに選んだ 1000 文を用い (ただし,「」や () の入っている文は除外)，以下の再現率と適合率を求めた．

$$\text{再現率} = \frac{\text{正しく挿入された読点数}}{\text{原文の読点数}} \quad (4)$$

$$\text{適合率} = \frac{\text{正しく挿入された読点数}}{\text{挿入された読点数}} \quad (5)$$

4 実験結果

表 1–3 に $P_l(j)$ の結果の一部を表す．また，評価実験の結果を表 4 に示す．

表 1: 形態素数による読点数の割合

		読点数					
		0	1	2	3	4	5
形態素数	10	0.704	0.272	0.021	0.003	0.001	0.000
	15	0.440	0.478	0.070	0.008	0.002	0.001
	20	0.238	0.568	0.168	0.022	0.004	0.001
	25	0.122	0.534	0.280	0.053	0.008	0.002
	30	0.066	0.442	0.369	0.098	0.018	0.004
	35	0.039	0.341	0.417	0.159	0.032	0.008
	40	0.021	0.243	0.431	0.223	0.060	0.014
	45	0.017	0.178	0.404	0.279	0.091	0.022
	50	0.08	0.129	0.367	0.304	0.138	0.040

表 2: 文字数による読点数の割合

		読点数					
		0	1	2	3	4	5
文字数	10	0.904	0.090	0.005	0.001	0.000	0.000
	20	0.626	0.343	0.026	0.003	0.002	0.000
	30	0.315	0.558	0.111	0.012	0.003	0.000
	40	0.143	0.549	0.256	0.042	0.006	0.002
	50	0.066	0.443	0.372	0.096	0.018	0.004
	60	0.039	0.313	0.431	0.172	0.036	0.007
	70	0.020	0.214	0.421	0.246	0.078	0.017
	80	0.012	0.142	0.360	0.315	0.114	0.040
	90	0.008	0.109	0.283	0.341	0.171	0.064

表 3: 文節数による読点数の割合

		読点数					
		0	1	2	3	4	5
文節数	5	0.577	0.391	0.028	0.000	0.000	0.000
	10	0.096	0.526	0.309	0.055	0.010	0.002
	15	0.018	0.217	0.447	0.237	0.061	0.014
	20	0.009	0.082	0.296	0.338	0.186	0.057
	25	0.008	0.043	0.161	0.279	0.268	0.153
	30	0.004	0.012	0.134	0.178	0.253	0.221

この表から分かるように，今回の実験では再現率に関しては半分以下の値であったが，適合率に関しては，形態素数，文字数，文節数それぞれにおいて高い結果を得ることができた．これは読点の挿入数は少ないが，挿入された読点は高い精度で正しい位置に挿入されていることを意味する．

5 考察

5.1 形態素数, 文字数, 文節数による違い

挿入された読点の数は，文長を文字数で定義した場合が一番多く，文節数で定義した場合が一番少ないという結果になった．換言すれば，文を小さい単位に分割し，文の長さを測ったほうが，より多くの読点が挿

表 4: 実験結果 (単位%)

	再現率	適合率
形態素数	45.6	74.0
文字数	46.0	73.0
文節数	44.9	73.4

表 5: 直前の品詞別結果 (単位%)

	形態素数		文字数		文節数	
	再現率	適合率	再現率	適合率	再現率	適合率
名詞	47.5	82.7	47.2	81.0	46.4	82.0
助詞	38.7	56.6	39.6	55.8	37.1	55.3
動詞	54.7	90.7	55.1	90.8	55.1	90.8
接続詞	56.7	89.5	63.3	95.0	66.7	95.2
助動詞	50.0	75.0	46.7	70.0	43.3	68.4
副詞	14.3	50.0	14.3	50.0	14.3	50.0
形容詞	16.7	100	16.7	100	25.0	100
その他	33.3	100	33.3	100	33.3	100

入できるということである．

一方，適合率に関しては，異なる傾向を示した．形態素数で定義した場合のものがもっともよく，次いで文節数の場合であった．

ただし，これらの差はわずかであるため，実際にはどの方法でも大差ないと言える．

5.2 品詞別評価

表 5 に読点の直前の品詞別の結果を示す．名詞，動詞，接続詞，形容詞に続く読点の適合率は極めて高いが，助詞は低い．再現率はどの品詞の場合も低い．

今回の実験に使用した原文では名詞，助詞，動詞の直後に打たれていた読点が全体の 92.0%であり, 挿入された読点も文の長さを形態素数で定義した場合の 93.0%，文字数で定義した場合の 93.0%，文節数で定義した場合の 92.6%が名詞，助詞，動詞の直後であることから，この 3 つの品詞について考察する．

名詞，動詞の直後については高い適合率を得ることができたが，助詞の直後については再現率，適合率いずれも低い結果になってしまった．以下に助詞の挿入ミスが起こった例を示す．

原文: 政府は次期通常国会に，０９年度第２次補正予算案と１０年度予算案を提出する．

出力文: 政府は，次期通常国会に０９年度第２次補正予算案と１０年度予算案を提出する．

原文では「に」の後に，出力文では「は」の後に読点が入っている．このような文では読点の挿入は人によって異なる．原文とは違うが，出力された文も不自然ではない．実際，助詞の直後に挿入される読点はこのような読点が多く，他の品詞より再現率，適合率が低くなってしまってのではないかと考えられる．

以下に, その他の読点の挿入がうまくいかなかった例を示し，その原因を考察する．

文の長さへの依存

本システムでは，特に短い文に対して挿入する読点の数がその長さに依存しすぎており，再現率を下げている一つの原因であると考えられる．以下に読点が挿入できなかった短い文を示す.

原文: 政権交代を機に，日本の統治は転換期に入った．

出力文: 政権交代を機に日本の統治は転換期に入った．

文の長さが十分でないため，読点が入らない典型的な例である．この長さだと人間でも読点を入れない人がいるのではないだろうか.

同じように文が短い場合でも，読点が入る場合がある．

原文: 大人９００円，大高生７００円，中学生以下無料.

出力文: 大人９００円 { 人名 } 生７００円，中学生以下無料.[1]

これは,「円」と「中学生」の間に読点の入る確率が 84.4 %と非常に高いからである (この文では「大高生」の「大高」が人名として間違った解析がされている)．このように，形態素間に読点の入る確率 (式 (3) の $P(q_i|w_i, w_{i+1})$) が十分大きい場合は，文が短い場合でも読点が挿入されることがある．

しかしこのような短い文では 2 つ目の読点が入ることはなかった．

人名に対しての読点挿入

以下のような人名の並ぶ文に読点がうまく挿入されなかった.

[1] 2 で述べたように，本システムでは固有表現は特別な記号 (この場合 { 人名 }) に置き換えている．

原文: ４９年選挙では池田勇人，佐藤栄作，前尾繁三郎らが初当選した．

出力文: ４９年選挙では，{ 人名 }{ 人名 }{ 人名 }{ 人名 }{ 人名 }{ 人名 }{ 固有地域 } らが初当選した．

こういった文は，読点が挿入されていないと，漢字の羅列になってしまい，読むことが困難になってしまうため，確実に読点を挿入しなければならない．

この問題は，$P(q_i|w_i, w_{i+1})$ を計算する際の形態素解析の結果，人名が姓と名の 2 つに分けられてしまうことが原因である．実際に { 人名 } と { 人名 } の間に読点の入る確率は 3.8 %ととても低い数値になってしまっている (この文では「三郎」が固有地域として間違った解析がされている)．

不自然な読点挿入

以下のような文には不自然な読点が挿入された．

原文: 前回覇者の広島皆実を地区大会決勝で破った広島観音は山形中央を降して１６強．

出力文: 前回覇者の { 固有地域 }，{ 固有地域 } を地区大会決勝で破った { 固有地域 } 観音は { 固有地域 } 中央を降して１６強．

この文では広島皆実という学校名が「広島」と「皆実」に分かれて解析され，それぞれ固有名詞の地域に分類された．その結果この 2 つの形態素の間に読点が挿入されてしまった．

これ以外の文にも，地名の入った名前，学校名は，正しく形態素解析することができなかった．

6　おわりに

本研究では，文の長さに依存して決まる読点の個数の確率と形態素と形態素の間に読点が入る確率とを用いて読点を挿入する方法を提案した．文の長さと n-gram を用いる極めて簡単化したモデルであるが，高い適合率を実現することができたが，再現率はやや低かった．特に，短い文に対して読点が挿入されないという問題があった．これについては，式 3 で示した確率が一定以上の場合は文の長さに関係なく読点を挿入していくなどの工夫の余地がある．

また，今回の実験では地名の入った学校名や，人物名，カタカナや漢字の羅列に対しての形態素解析の精度が特によくなかった[2]．このことは，再現率を下げる一つの原因と言える．

本システムは，形態素解析の精度や形態素間に読点の入る確率の質を高めていくこと (本研究では 3-gram に近似したが例えばこれを 4-gram にするなど) によって，より実用的なシステムとすることができると考える．

参考文献

[1] 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

[2] 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”, 情報処理学会第 50 回全国大会講演論文集, No.3, pp.185-186, 1995

[3] 村田 匡輝, 大野 誠寛, 松原 茂樹: “日本語テキストにおける読点位置の検出”, 言語処理学会第 16 回年次大会発表論文集, pp.812-815, 2010

[4] 村田 匡輝, 大野 誠寛, 松原 茂樹: “読点の用法分類に基づく自動読点挿入”, 情報処理学会研究報告, Vol.2010-NL-196, No.8, pp.1-8, 2010

[5] 村田 匡輝, 大野 誠寛, 松原 茂樹: “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

[6] 石黒 圭: “論文・レポートの基本”, 日本実業出版社, 2012

[2]固有表現に対して強い MeCab の NEologd という辞書を使用しようとしたが，固有名詞を過剰に検出してしまいうまくいかなかった．