# 読点挿入のための確率モデルの提案とそれを用いた読点自動挿入システム

情報科学科　坂 祥太朗　　指導教員：山村 毅

## 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 [?].

本研究では，文に読点を挿入するシステムの開発を行う．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，

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

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．

表 1　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 2　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 |

$P_l(j)$ 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

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 文を用い再現率と適合率を求めた．その結果を表 1-2 に示す.

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．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入できるということである．

## 4　おわりに

本システムは文の長さと 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，しかし再現率はやや低かった．

また，今回の実験で固有名詞に対しての形態素解析の精度が特によくなかった．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] 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

[2] 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