# 日本語文章における句読点自動最適配置*

4N－2

鈴木英二　島田静雄　近藤邦雄　佐藤尚†
埼玉大学工学部情報工学科‡

## 1　はじめに

本研究の目的は句読点、特に読点の最適な配置ルールを提案し、それを計算機上で実現することである。わかち書きをしない日本語文章にとって、句読点の役割は大きい。しかし、句読点の用法、特に読点の用法は著者に一任されているのが現状である。熊野ら [1] は、英日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基準を提案した。しかし、この方法では英文中の単語数を利用するため、基本となる英文が必要である。

本研究では、従来の句読点の用法を改めて調査した。そして、読点と読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する。なお研究対象を漢字仮名混じりの科学技術論文に限定した。また、なるべく簡単な形態素解析で処理を進めていくことを目標としている。

## 2　従来の読点の用法

従来の読点の用法は主に次の4つが挙げられる [2]。

(1) 節の読点としての用法

(2) 語句の読点としての用法

(3) 論理の読点としての用法

(4) 並列の読点としての用法

(1) は文と文をつなげるときの用法である。この用法の読点の前は、連用中止法や接続助詞が用いられている。この用法が読点挿入の原則とされている。(2) は格要素が長くなったときの用法である。(3) は平仮名が続いてどこで区切られるかが明確でない、または修飾語と被修飾語が離れていて、どれに係るかわからないときの用法である。(4) は並列句や同格を表すときの用法である。ナカテン「・」で代用する場合もある。

*Automatic Recognition of Optimal Punctuation in Japanese Documents

†Eiji SUZUKI,Shizuo SHIMADA,Kunio KONDO,Hisashi SATO

‡Department of Information & Computer Sciences,SAITAMA University

## 3　論文文章の調査

計算機で判読可能な複数の科学技術論文から抽出した500の文をサンプルとして、以下の調査を行った。

### 3.1　読点の直前にある語

局所的情報として読点の前にある語句が何であるかを調査した。その結果を表1に示す。

表1：読点の前にある語句の出現率

| | |
|---|---|
| とりたて詞「は」 | 30.7% |
| 格助詞 | 14.3% |
| 連用中止形 | 12.8% |
| 接続助詞 | 11.5% |
| 文頭の接続詞 | 11.1% |
| 語句の並列 | 10.5% |
| 文頭の慣用句 | 6.8% |
| その他 | 2.3% |

文頭の慣用句とは「一方・次に・近年・今日」などの文頭の語である。

また特にとりたて詞「は」が多い。これは主語を明示するためだと考えられる。

### 3.2　連用中止形・接続助詞

調査対象から連用中止形・接続助詞を抽出し、その直後に読点があるかどうかを調査した。その結果、全体の89.4%の連用中止形・接続助詞の直後に読点が存在した。

### 3.3　読点によって区切られる文字列

次に読点によって区切られる文字列の長さを調査した。その結果を表2に示す。

15ないし20字ごとに読点を打つのが読み易いとされている [2]。平均18.6文字ごとに読点は打たれていた。しかし25文字まではほぼ一様に分布している。そこで、短い文字列(15文字以下)と長い文字列(16文字以上)とに分けて調査した。その結果、短い文字列には、文頭の接続詞・とりたて詞「は」・文頭の慣用句が含まれている頻度が高い。長い文字列は、「に」「が」「を」などの格助詞、連用中止形で終っている頻度が高い。

また、句点側の文字列だけを調査すると平均の長さは、23.5 文字になり、他の文字列よりも長くなる傾向がある。

表 2. 読点によって区切られる文字列の長さ

| | |
|---|---|
| 1 ～ 5 文字 | 16.9% |
| 6 ～ 10 文字 | 14.2% |
| 11 ～ 15 文字 | 16.5% |
| 16 ～ 20 文字 | 18.3% |
| 21 ～ 25 文字 | 14.0% |
| 26 ～ 30 文字 | 8.9% |
| 31 文字以上 | 11.0% |

## 4 読点挿入箇所の決定

以上の調査結果から、読点を挿入するべき箇所を決定する規則を下のように提案した。なおここで述べる「文字列」とは、読点と読点で区切られる文字の並びのことである。

(1) 一文中の連用中止法・接続助詞の直後に読点を挿入する。これは、読点挿入の原則である「節の読点としての用法」を最優先させるためである。

(2) この時点で、基準の長さに満たない文字列 (本研究では 15 文字、ただし句点の直前の文字列は 25 文字) は、処理の対象から外す。

(3) 残った文字列は依然として長いままである。その中で文字列が文頭のものであるなら、文頭の接続詞・文頭の慣用句の直後に読点を挿入する。文頭の接続詞・文頭の慣用句からなる文字列は、十分短いので処理の対象から外す。また、文頭の接続詞・文頭の慣用句を外したことによって、短くなった文字列も外す。

(4) 残った文字列の中でとりたて詞「は」が含まれているのなら、その直後に読点を挿入する。それによって、基準の長さに満たなくなった文字列は処理の対象から外す。このとき、その長さの基準は (2) の基準よりも長いものである (本研究では 25 文字、ただし句点の直前の文字列は 35 文字)。

(5) 残った文字列の中で格助詞が含まれているのなら、その直後に読点を挿入する。それによって、基準の長さに満たなくなった文字列は処理の対象から外す。このとき、その長さの基準は (4) の基準よりも長いものである (本研究では 30 文字、ただし句点の直前の文字列は 40 文字)。

(6) 残った文字列は読点が挿入できないものとして処理を終える。

読点を挿入するべき箇所を決定する際に局所的情報を基に優先度を与えた。その優先度は、連用中止形・接続助詞→文頭の接続詞・文頭の慣用句→とりたて詞「は」→格助詞の順とした。さらに、優先度が高い語を含む文字列ほど長さが短くなるようにした。

## 5 処理例

上記の規則に従って読点を挿入した例を示す。

(1) 本研究の目的は、句読点特に読点の最適な配置ルールを提案し、それを計算機上で実現することである。

(2) さらに、今日の生成文法規則は、初期の規則中心の文法から大きくその姿を変え、句構造規則というものは完全になくなった。

(3) その際、いろいろな手法があるが、その中でも視覚に訴える方法が聴衆に理解を促し、説得する上でもっとも有効な方法である。そのため、現在のプレゼンテーションは、黒板や、スライドＯＨＰ・ＶＴＲ・パーソナルコンピュータなどの、視覚効果のあるものを用いて行なわれる。

## 6 おわりに

実際の文章から読点の用法を調査し、読点の直前にくる頻度が高い語がわかった。また、読点と読点で区切られる文字列の長さに注目して、読点を挿入するべき箇所を決定する規則を提案した。そして、それに従って文に読点を挿入できた。

本実験では格助詞を一括してしまったが、格助詞の間でも優先度の違いがあると考えられる。また、節の読点の用法と語句の読点としての用法の一部を扱ったのみで、論理の読点としての用法、並列の読点としての用法及び「…結果、…」などの用法は、検証していない。読点で区切られる文字列の長さにもまだ調整の余地が残っている。今後はこれらの点を検証して研究を進めていく。

## 参考文献

[1] 熊野他:「自然な日本語生成のための指針」情処第 41 回全国大会,4S-8,1990

[2] 小泉 保:「日本語の正書法」大修館書店,1978

[3] 本多勝一:「日本語の作文技術」朝日新聞社,1982

[4] 「TRIE 構造辞書とその形態素分類体系の概要」(財)新世代コンピュータ技術開発機構,1992