# 大規模データにおける文節機能語列の調査と長単位機能語辞書

1N－9

村上裕　兵藤安昭　池田尚志
岐阜大学工学部

## 1　はじめに

日本語文解析の精度を向上させるためには、まず文節を正しく認識出来ることが前提となる。また、解析の対象となる文は'正しく'書かれた文だけとは限らないから、'誤りを含む入力文'に対しても適切に対処できなければならない。表現上の誤りを含む文を入力して誤り個所を指摘するといった応用は、文書校正、OCR後処理をはじめいくつもある。したがって、'正しい入力文'だけでなく、'正しくない入力文'についても文節を正しく解析できることが必要である。

形態素解析システムでは、可能な単語候補を切り出し、接続可能なあるいは接続確率の高い単語列を選び出すという手順が一般的である。しかしこの方法では、完全な接続規則や接続確率を定めることが困難であるために、一般に過剰生成になっている。つまりシステムが許容する表現は現実に現われる表現、あるいは正しい表現の範囲を大きく超えており、誤った表現も、他と特別に区別することなく、正しい文節として出力してしまう。高精度な解析、また誤りを指摘できる解析のためには、必要十分な範囲の文節を認識できるようにすることが必要になってくる。

我々は日本語文解析システムIBUKI[1]を開発しているが、そこでは比較的長い単位の機能語（約4,000語）を基本の機能語として辞書に登録して文節解析を行っている。もし機能語の単位をさらに長く取って、実際に現われる文節中の機能語列そのものを何らかの方法で全て(ほとんど)辞書に表現する事が出来るならば、文節解析の精度を高め、また'誤った表現'に対しても適切な指摘をすることが可能となろう。そこで我々は大量のテキストを解析して、日本語文中に実際に現われる文節中機能語列を数え上げることを試みた。

Statistics of Bunsetsu function words in large scale text data
Yutaka Murakami, Yasuaki Hyodo, Takashi Ikeda
Faculty of Engineering, Gifu University
Gifu-shi,501-11,Japan

## 2　文節機能語列の頻度統計

毎日新聞記事4年分(総文数3,981,448)をIBUKIで文節解析し、出現した機能語列の頻度を調べた(表1)。述べ数は26,462,869、異なり数は45,418であった。当初、新聞記事4年分という大量のデータをとれば、機能語列の異なり数はある値に収束すると予測したが、頻度1の機能語列が異なり数で21,413(47.15%)も存在し、この予測は成立しなかった。しかしカバー率で見れば上位1万語で99.75%を占め、実際に出現する機能語列は「ほとんど有限」と考える事が出来る。

また、出現した機能語列がどの品詞属性を持つ文節に含まれているかについて調べた(表4)。
名詞から続く機能語列については、上位5,000語で全体の99.98%をカバーしており、「有限性の度合い」は非常に高くなっている。

表1: 頻度順位毎の機能語列の統計

| 頻度の順位 | 延べ数 | カバー率[1](%) | 平均文字長[2] |
|---|---|---|---|
| 1 ～ 100 | 23,453,240 | 91.35 | 1.24 |
| 101 ～ 200 | 962,043 | 94.16 | 3.05 |
| 201 ～ 300 | 447,105 | 95.44 | 3.07 |
| 301 ～ 400 | 266,136 | 96.24 | 3.45 |
| 401 ～ 500 | 183,395 | 96.78 | 3.64 |
| 501 ～ 1,000 | 443,245 | 98.07 | 3.88 |
| 1,001 ～ 5,000 | 527,396 | 99.52 | 4.83 |
| 5,001 ～ 10,000 | 87,896 | 99.75 | 6.05 |
| 10,001 ～ 20,000 | 49,901 | 99.89 | 6.82 |
| 20,001 ～ 40,000 | 29,346 | 99.98 | 7.84 |
| total(45,418) | 26,462,869 | 100 | 1.53 |

表2: 頻度上位の機能語

| 機能語 | 頻度 |
|---|---|
| の | 4,787,766 |
| を | 2,979,964 |
| は | 2,208,996 |
| が | 2,187,504 |
| に | 2,090,847 |
| た | 1,593,872 |
| で | 1,275,852 |
| と | 1,257,585 |
| も | 503,091 |
| て | 472,261 |

表3: 頻度1の機能語列例

| 機能語列 |
|---|
| ていたからではないでしょうか |
| っぱなしだからだと |
| でいたからです |
| っぱなしだったのは |
| っぱなしだったのが |
| っぱなしだったからだ |
| っぱなしであることから |
| ていたからではないかとさえ |
| っぱなしでしたが |
| ていたからではないのか |

[1]カバー率：累積延べ数の総延べ数に対する割合
[2]頻度で重み付けした平均文字長

表 4: 文節自立語の品詞別に分類した統計

| 頻度の順位 | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 ～ 100 | 17,863,560 | 417,496 | 3,828,220 | 48,461 | 16,876 | 146,143 | 0 |
| 101 ～ 200 | 218,166 | 117,923 | 484,689 | 23,878 | 24,419 | 0 | 0 |
| 201 ～ 300 | 104,914 | 55,140 | 230,139 | 10,643 | 8,869 | 0 | 0 |
| 301 ～ 400 | 53,636 | 29,670 | 151,911 | 15,312 | 0 | 0 | 0 |
| 401 ～ 500 | 46,051 | 20,819 | 98,469 | 5,341 | 3,986 | 0 | 0 |
| 501 ～ 1,000 | 63,804 | 53,957 | 258,366 | 31,463 | 14,951 | 0 | 0 |
| 1,001 ～ 5,000 | 48,959 | 60,430 | 340,434 | 36,911 | 16,179 | 1,197 | 981 |
| 5,001 ～ 10,000 | 2,649 | 11,270 | 60,614 | 5,998 | 3,412 | 146 | 212 |
| 10,001 ～ 20,000 | 873 | 6,269 | 36,153 | 3,105 | 1,931 | 60 | 62 |
| 20,001 ～ 40,000 | 365 | 3,540 | 22,061 | 1,615 | 957 | 7 | 26 |
| total(45,418) | 18,403,227 | 777,955 | 5,520,540 | 183,712 | 92,159 | 147,568 | 1,305 |

A:名詞 B:名詞述語 C:動詞 D:形容詞 E:形容動詞 F:副詞, 連体詞 G:引用文、から続く機能語列

## 3 機能語列の分割による統計

表 1 に見るように、低頻度語の文字列長は大きい。そこで機能語列を意味的な切れ目と考えられる個所で分割すれば当初の予測が成立するかと考え（実際には終止形ないし連体形の個所で分割した）、同様の統計をとってみたが (表 5)(述べ数は 29,267,893、異なり数は 2,212)、やはり頻度 1 の (短) 機能語列が 312(14.1%) も存在した。しかし、カバー率では、上位 1,000 で 99.96%とさらに高い「有限性」がみられた。また、延べ数に対する異なり数の割合は、分割しない状態では 0.17%であったものが、0.008%と約 20 分の一と集約度が高まっていることが解った。

表 5: 分割した機能語列に対する統計

| 頻度数の順位 | 延べ数 | カバー率 (%) | 平均文字長 |
|---|---|---|---|
| 1 ～ 100 | 28,038,370 | 95.80 | 1.30 |
| 101 ～ 200 | 679,128 | 98.12 | 3.10 |
| 201 ～ 300 | 256,218 | 98.99 | 3.76 |
| 301 ～ 400 | 122,386 | 99.41 | 3.63 |
| 401 ～ 500 | 66,759 | 99.64 | 3.89 |
| 501 ～ 1,000 | 94,676 | 99.96 | 4.21 |
| 1,001 ～ 2,000 | 10,144 | 99.99 | 5.17 |
| total(2,212) | 29,267,893 | 100 | 1.39 |

## 4 機能語列切り出しの確信度

頻度が 10 以上の機能語列は異なり数 8,251、延べ数 26,383,102 で全体の 99.69%をカバーしている。もし、文節解析におけるこれらの切り出しに関して何らかの確信度のようなものを表現することが出来るならば、いろいろな応用システムで有効に利用できる。その指標の一つとして、機能語列の途中から引き続く自立語の存在について調べてみた。(図 1)

結果は、機能語列の語尾部分で自立語 (EDR 辞書を用いた) の語頭となり得るものが異なり数で 316、そのような語尾部分を持つ機能語列が 8,251 中 6,285 であった。このように、かなりの機能語列が自立語と交差するという結果となり、このまま単純に切り出しの確信度と結びつけることは難しい。今後、①現実のテキスト中での切り出しの曖昧さの出現具合を調べてみること、②考慮すべき機能語列は約 8,000 と限定できるのだから、個々の機能語列に関してテキストの分野を考慮した実際上の曖昧さ解消の方策の可能性を追求してみることなど、取り組んでいく予定である。

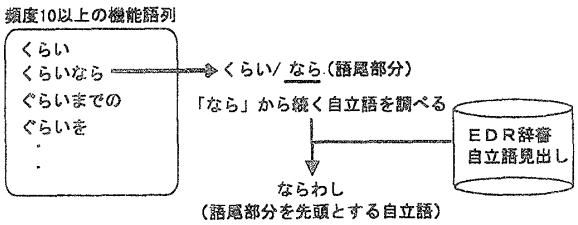


図 1: 機能語列語尾部分の調査

## 5 おわりに

新聞記事 4 年分を文節解析して、実際に出現する機能語の分布について調べた。実際に出現する機能語は、「ほとんど有限」であることが分かった。接続条件だけでなく、これらの統計情報を活用することで、文節解析の精度向上、また「誤りを含む入力文」の誤り個所の検出/訂正に活用していく予定である。

## 参考文献

[1] 文節単位のコストに基づく日本語文節解析システム、兵藤安昭、池田尚志 (1999)
(言語処理学会第 5 回年次大会発表論文集)