日本語形態素解析における付属語列の統計的性質について

伊藤良太　　宮崎正弘

新潟大学大学院自然科学研究科

1　はじめに

日本語文の形態素解析において、通常、付属語列は付属語の二項関係の連鎖によって処理される。従って、付属語列が長くなるにつれて、解析時における分割の曖昧性、品詞の曖昧性が発生し、その結果、途中状態数が多くなり解析の負担となっている。本稿では、このような問題の解決を目指して、日本語文コーパスを形態素解析することによって、付属語列の統計的性質を明らかにした。実験によれば出現頻度の高い約 1000 個の付属語列の累積頻度は 99%を占め、使用率の高い付属語列のデータベース化および 2 項関係の接続ルールの代わりに n 項関係の文法規則を用いることの有効性が示された。

2　形態素解析における曖昧性発生

日本語文の形態素解析では、様々な曖昧性が発生する。とくに、ひらがなで記述されることが多い付属語列部分は、曖昧性が発生しやすく、解析時の途中状態数が多くなり、解析の負担となっている。例えば、付属語列「たばかりである」を解析したときに発生する曖昧性を図 1 に示す。太字で表記された部分が正しい解析結果であるが、それ以外にも品詞の曖昧性が発生していることがわかる。

た:既定の助動詞「た」の終止形
└ばかり:限定、程度の副詞「ばかり」
├で:肯定の助動詞「だ」の連用形
│ ├ある:肯定の助動詞「ある」の終止形
│ └ある:本動詞「ある」の終止形
├で:本動詞「でる」の連用形
│ ├ある:形式動詞「ある」の終止形
│ └ある:本動詞「ある」の終止形
└で:連用格助詞「で」
　└ある:本動詞「ある」の終止形

図 1: 品詞の曖昧性

2.1　形態素解析における曖昧性発生に伴う問題点の解決策

形態素解析時における曖昧性発生を抑制する方法として、付属語の 2 項関係だけでなく、3 項以上の関係を準備することが考えられる。

しかし、文法規則に基づき、単純に付属語の接続を繰り返すことによって生成される 3 項以上の付属語列は実在性の低いものが多く含まれる。例えば、図 2 のように「ほど」に「まで」を接続し、さらに、「ほど」、「まで」、「です」と接続を繰り返して生成された「ほどまでほどまでです」という付属語列の実在性はかなり低いものであると考えられる。また、、図 3 のように「ほどまでほどまでほどまで...」という接続のループが発生しうるので生成数は無限となる。 表 1 に文法規則に基づいて接続することによって生成される付属語列の数を示す。形態素数 4 で 130 万を超え、そのうちの多くは実在性の低いものとなっている。

そこで、本稿では日本語文コーパスを形態素解析することによって実在する付属語列を抽出し、その統計的性質を示す。

図 2: 実在性の低い付属語列の例

図 3: ループする付属語列の例

表 1: 形態素数と付属語列数

形態素数	付属語列数
2	3137
3	65035
4	1310451

3　日本語文コーパスにおける付属語列の調査

3.1　調査対象

- 日本経済新聞 94 年度版記事全文データベース
- 機械翻訳プロジェクト日英対訳コーパス

調査対象の日本語文コーパスは以上の 2 つである。以下、それぞれを日経コーパス、機械翻訳 P コーパスとする。また、機械翻訳 P コーパスは、辞書の例文集であり、その日本語文を調査する。

3.2　調査方法

本研究の調査方法について述べる。まず、コーパスの文を日本語形態素解析システムを用いて、形態素に分割する。そして、分割結果から付属語が連続して出現する部分を付属語列として、その部分を抽出し、分析対象とした。本研究では、14 の助動詞と 118 種類の助詞を付属語とした。抽出された付属語列の例を以下に示す。

付属語数 2
「で/は」、「に/は」、「と/の」
付属語数 3
「だけ/で/なく」、「だっ/た/が」
付属語数 4
「べき/だ/と/の」、「だっ/た/だけ/に」
付属語数 5
「て/から/で/ない/と」
付属語数 6
「て/まし/た/と/ばかり/に」
付属語数 7
「だっ/た/から/で/あり/まし ょ/う」

3.3　日本語文コーパス内の付属語列数

それぞれのコーパスについて解析可能文数と抽出された付属語列の総数、付属語列の種類数についてまとめた結果を表 2 に示す。

表 2: コーパス内の付属語列数

	日経	機械翻訳 P
総文数	約 1500000	約 700000
解析可能文数	1376387	651385
付属語列総数	934587	121298
付属語列種類数	2564	1706

付属語列総数に対し、付属語列種類数が少ないということから日本語文で用いられる付属語列の異なり数は少ないということがわかる。

また、解析可能文数に対する付属語列総数の割合に着目すると、機械翻訳 P コーパスが日経コーパスより低くなっている。その理由は、機械翻訳 P コーパスの文は辞書の例文であり、長い付属語列を含まないような単純な文が多いためであるということが考えられる。

表 3: 形態素数別の種類数、存在数

	形態素数	2	3	4	5	6	7	総計
日経	種類数	912	1020	503	115	12	2	2564
	存在数	877424	52564	4301	278	18	2	934587
	割合 (%)	93.8836	5.6243	0.4602	0.0298	0.0019	0.0002	100.0000
機械翻訳 P	種類数	802	653	214	34	3	0	1706
	存在数	109421	11058	757	58	4	0	121298
	割合 (%)	90.2084	9.1164	0.6241	0.0478	0.0033	0.0000	100.0000

表 4: 付属語列数と付属語列総数に占める割合

	付属語列数	100	200	300	400	500	1000
付属語列総数に	日経	92.6	96.6	97.9	98.6	99.0	99.7
占める割合 (%)	機械翻訳 P	85.2	92.5	95.3	96.7	97.6	99.3

3.4 形態素数別の種類数、存在数

コーパスから抽出した付属語列を形態素数別に種類数、存在数をまとめた結果を表 3 に示す。

付属語列の割合に着目すると、両コーパスとも形態素数 2 の付属語列が 90 %以上を占め、形態素数 3 の付属語列を加えると 99 %以上を占めている。このことから、実際に使用される付属語列は形態素数 2、3 程度が多数を占め、形態素数が 4 を越えるものの割合は少ないということがわかる。

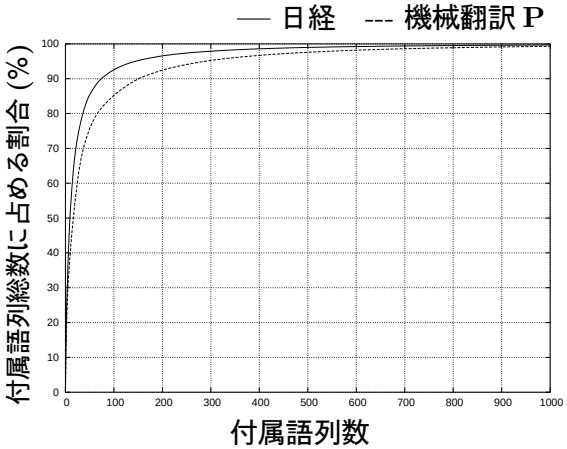

図 4: 付属語列の累積頻度

3.5 付属語列の累積頻度

コーパスから抽出した付属語列の累積頻度をまとめた結果を図 4、表 4 に示す。

日経コーパスにおいては 100 個で 90 %以上、500 個で 99 %以上となっている。また、機械翻訳 P コーパスにおいては 200 個で 90 %以上、1000 個で 99 %以上となっている。このことから、少数の付属語列で実際に使用される付属語列の多くをカバーができるということがわかる。

4 形態素解析への応用

本研究で抽出した付属語列を既存の形態素解析に組み込み、付属語列部分の解析の負担を軽減することによって処理の高速化が見込める。

また、漢字かな混じり文よりも曖昧性が発生しやすいかなべた文の形態素解析では、付属語列部分の曖昧性発生を抑え、付属語部分と自立語部分の推定を可能にすることが考えられる。

5　まとめ

本稿では、日本語文形態素解析における付属語列の統計的性質について論じた。日本語文における付属語列の異なり数は少ないということを示し、また、分野によっては 1000 個の付属語列を準備することにより、99 %以上をカバーできるということを示した。従って、使用率の高い付属語列のデータベース化および 2 項関係の接続ルールの代わりに n 項関係の文法規則を用いることの有効性が示された。

従来の接続表に代わり、CFG 形式の文法規則を用いた拡張型チャートパーザによる日本語形態素解析システム jampar[2] では、n 項関係の文法規則を用いることを可能としており、存在数の多い付属語列をデータベース化し、形態素解析に組み込むことが容易に行える。

参考文献

[1] 宮崎正弘, 白井諭, 池原悟; 言語過程説に基づく日本語品詞の体系化とその効用, 自然言語処理,Vol.2,No.3,pp.3-25(1995)

[2] 宮崎正弘, 川辺諭, 武本裕; 構造化チャート法に基づく日本語形態素解析器 Jampar, 言語処理学会第 13 回年次大会,PA1-3(2007)