# 日本語形態素解析用文法規則の使用状況

6M－4

荻野　紫穂
日本アイ・ビー・エム株式会社　東京基礎研究所

## 1. 序説

最近の自然言語処理においては、文法の巨大化傾向や、それによって生じる無駄がよく指摘される。こういった問題を解決するために、例えば、規則のコストを調整する方法(文献[6])等が研究されている。しかし、文法の縮小と精度の関係や、解析結果分析の際の、《安定しているデータ》のための必要量の基準、更には「あるドメインは解析しやすい／しにくい」といった基準などは、非常に重要であるにも関わらず、あまり明らかではない。本稿では、あるドメインの処理における日本語形態素解析の文法規則の使用頻度と出現時点、および、各ドメインに共通に使われる規則などについて調査すると同時に、解析しやすさ／しにくさなどの観点について考察する。

データとして、読売新聞社説記事を約4,500文、IBMのメインフレーム関係のマニュアルを約4,000文、UNIX系のマニュアルを約4,800文使用した。これらのデータを形態素解析する際に使用した規則を調べ、その頻度や共通部分について調査した。使用した形態素解析システムは、3型文法で書かれた規則を、約5,000持っている(文献[9])。

## 2. 規則の使用状況と処理語数

### 2.1. 新しい規則の出現状況

まず、使用規則の異なり数が、どのように変化するかについて述べる。図1は、処理語数と使用規則の異なり数のグラフである。3本のグラフはそれぞれ、新聞記事、UNIXマニュアル、メインフレームのマニュアルでの結果を示す。どのドメインも、20,000語を処理するまでは、処理中に初めて使われる規則が急激に増えるが、その後、徐々に伸びが緩やかになる。

UNIXマニュアルや、メインフレームのマニュアルの規則数の伸び率は、新聞記事に比べて緩やかである。新聞記事での規則の異なり数は、130,000語を処理した時点でも、ほぼ一定に伸びている。これは、新聞記事処理の文法は、同じ量の他のドメインをカバーする文法に比べて、適用範囲が広くなければならないということであり、他の分野に比べて処理が難しい、と言われる一面を示していると言える。

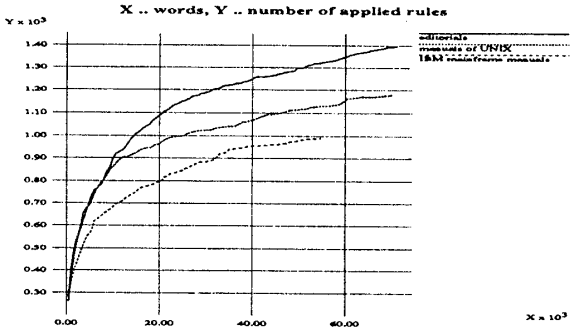


図1: 新規則の出現状況

新聞記事130,000語を処理した時点で使われた規則数は、約1,800であり、使用した形態素解析システムが持つ規則の3分の1弱が使われたことになる。

### 2.2. 出現規則の使用頻度

次に、処理中のある時点で現れる規則が、全処理中に何回使われたかについて述べる。図2は、500語ずつに処理時点を区切り、その範囲の処理中に初めて使われた規則の、全処理中での使用頻度の平均を示している。

どのドメインにおいても、処理開始直後に現れた規則の使用頻度が一番高い。新聞記事は、始めの500語を処理する際に現れた規則の平均使用頻度が約1,200回、UNIXマニュアルが900回、メインフレームのマニュアルが600回である。その後、平均使用頻度は急激に減少し、20,000語処理した後は、メインフレームのマニュアルの所々の増加を除いて、ほぼ1回に近くなり、40,000語を過ぎると、その傾向は更に強まる。

これは例えば、あるドメインにおいて、約20,000語までに現れるある現象を処理するための規則を新たに追加すると、その規則は後に現れる他の部分の処理に役立つ場合が多いが、20,000語を越えた時点で新たに現れた現象を処理するために規則を追加しても、その規則は、後の処理で、それほど効果的には働かないかもしれない、という可能性を示している。

しかし、図1から分かるように、使用規則の異なり数はある一定の割合で伸び続ける。従って、もし規則だけ

---

The Frequency of Applied Rules for Morphological Analysis in Analyzing a Set of Sentences
Shiho OGINO
IBM Research, Tokyo Research Laboratory

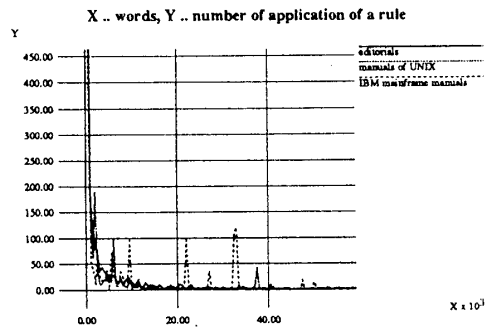


図 2: 規則の使用頻度

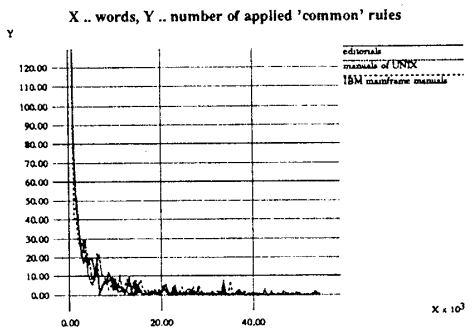


図 3: 共通規則の出現状況

で処理を行なうとすれば、20,000 語以降に現れる現象に関する規則を持っているかどうか（言い替えれば、使用頻度が少ない規則を含んでいるかどうか）が、特に処理データが増えるにつれて、文法全体の適用範囲を大きく左右することになる。また、メインフレームのマニュアルのように、章ごとにスタイルが大きく違う文書は、その章でしか使われない規則の使用頻度が、部分的に著しく高くなる場合もあるので、20,000 語までに現れた現象をカバーする文法で、ほとんどの現象が処理できるという訳ではない。

### 2.3. 共通に使用される規則

では、各ドメインで共通する使用された規則はどのくらいあり、どの処理時点で現れたのかについて述べる

それぞれのドメイン 54,500 語を処理する際に、全てのドメインに共通して使われた規則は約 700 で、システムが持つ規則の約 7 分の 1 だった。新聞記事とメインフレームのマニュアルに共通する規則は 800 弱、新聞記事と UNIX マニュアルは約 860、UNIX マニュアルとメインフレームのマニュアルは約 840 である。

図 3は、全てに共通して使われた約 700 の規則が、各ドメイン処理の、どの時点で初めて現れたかを示している。どのドメインにでも、処理開始直後から約 20,000 語までの処理で現れる規則が多く、40,000 を過ぎると、ほとんど現れなくなる。これは、どのドメインでも、20,000 から 30,000 語をカバーする文法は、各ドメイン共通の文法規則のほとんどを含んでいる可能性を示すと言える。ただし、それはあくまで《共通の》規則を含む可能性であり、あるドメイン 30,000 語に現れる現象全てをカバーする文法が、別のドメイン 30,000 語を高精度でカバーするということではない。

## 3. まとめ

形態素解析において、異なったドメインの文書処理に使われる規則の使用状況を調査し、使用規則の異なり数の違いによる処理の難易度の目安と、共通規則の出現時点の可能性とを示した。

今後の課題として、データ数の増加によって、規則の異なり数の伸び率がどう変わるか、別なドメインを加えた共通規則の出現状況も変わらないかどうかなどの調査が挙げられる。

## 文献

[1] 水谷静夫 (1983) 計量語彙論から見た文章展開,『朝倉日本語新講座 2 語彙』朝倉書店, pp. 155 - 164.

[2] 田中章夫 (1983) 抄録のための言語処理,『朝倉日本語新講座 6 運用 II』朝倉書店, pp. 1 - 41.

[3] 北村博 (1985) Zipf の法則と text の情報量,『情報処理学会研究報告』85-NL-52-2.

[4] 北村博 (1986) 機械翻訳の辞書量と未知語比率,『情報処理学会第 33 会全国大会講演論文集』II, pp. 1777 - 1778.

[5] 坂野俊哉, 森本逞 (1989) 音声認識における正規文法活用の有効性,『電子情報通信学会技術研究報告』SP89-95.

[6] 小松英二, 安原宏 (1991) コスト最小法形態素解析のコストルールの作成法,『情報処理学会研究報告』91-NL-85-1.

[7] 丸山宏, 荻野紫穂, 渡辺日出雄 (1991) 確率的形態素解析,『ソフトウェア科学会第 8 回全国大会論文集』, pp. 177 - 180.

[8] 北野宏明 (1992) 情報量と計算量で人工知能はどう変わるか,『合同研究会 "AI シンポジウム '92" 講演資料』人工知能学会研究会資料 SIG-F/H/K/S/I-9202-3.

[9] Hiroshi MARUYAMA, Shiho OGINO, Masaru HIDANO (1993) The Mega-Word Tagged-Corpus Project, "Proceedings of the 5th International Conference on Theoretical and Methodological Issues in Machine Translation," pp. 15 - 22.