圧縮統語森上での形態素解析候補の絞り込み - 品詞列統計情報の利用 -

Disambiguation of morphlogical ambiguity on packed forests using bigram of part of speech

20－1

伴光 昇　福田 譲　白井清昭　田中穂積

TOMOMITSU Noboru FUKUDA Yuzuru SHIRAI Kiyoaki TANAKA Hozumi

東京工業大学
Tokyo Institute of Technology

Abstract

This paper proposes a method to disambiguate morphological ambiguity on packed forest obtained from morphological and syntactic parsing of Japanese sentence using Generalized LR parsing algorithm. Disambiguation is to give preference to an element in a partial tree that has a highest probability based on statistical information of a sequence of part of speech. Because several elements that are analysed as the same syntactic category are packed, the syntactic constraint, that the elements belonging to the same category, is used for disambiguation of morphological ambiguity. So we can improve the precision of morphological analysis.

1 はじめに

日本語のように、単語と単語の間に空白などの切れ目を置かない言語の形態素解析では、数多くの解析候補が生成される。日本語の形態素解析の候補数を絞り込むヒューリスティクスとして、文節数最小法[2]が広く知られている。しかし、文節数最小法は、文節数の少ない候補に高い優先順位を与えることの根拠が明らかでない。

本稿では、一般化LR法に基づいて形態素解析と統語解析を同時に行うパーザ[3][6]により日本語文を解析し、その出力である圧縮統語森上で形態素解析の候補の絞り込みを行う。前述の文節数最小法は、文節を構成する文法規則に一定のコストを付与し、それらの規則の総適用数をその解析結果の文節数とすることで、実現している。しかし、文節数最小法では複数の候補が残るので、品詞列の統計情報より品詞列に関する bigram を抽出して、これにより候補の絞り込みを行う。すなわち、bigram により個々の候補の生成確率を計算し、より生成確率の高い候補に優先度を与える。圧縮統語森上で候補の絞り込みを行うことは、統語カテゴリにより圧縮されているという統語解析の制約を、形態素解析の候補の絞り込みに利用していることを意味する。

2 文節数最小法

本研究で使用しているパーザシステムでは、一般化LR法に基づいて形態素解析と統語解析を同時に行っている。このため、文節数最小法の文節の定義のために、統語解析の文法を利用することができる。例えば、以下の文法において、

文 → 後置詞句, 述語.
後置詞句 → 名詞, 助詞.
述語 → 動詞.

2,3番目の規則を文節を構成する規則とし、それぞれにコスト1を付与する。これにより、解析結果の総コストを求めることでその解析結果の文節数を求めることができる。また、

複合名詞 → 名詞, 名詞.

連絡先： 伴光 昇 東京工業大学工学部情報工学科田中研究室 〒152 目黒区大岡山 2-12-1
Tel:03(3726)1111 ex.4188 Fax.03(5734)2915 E-mail:tomomitu@cs.titech.ac.jp

のような規則にコスト 2 を付与することで、自立語数最小法にも対応することができる。

文節数最小法は有効であるが、複数の候補が残ってしまう。また、文節数の少ない候補に高い優先順位を与えることの根拠が明らかでない。次の章以降では、品詞列統計情報を用いた候補の絞り込みについて説明する。

3 品詞 bigram

シンボル連鎖 $X = x_1, x_2, \ldots, x_n$ の生成確率は、確率モデルを用いて以下のように表される。

$$P(X) = \prod_{i=1}^{n} P(x_i|x_1, x_2, \ldots, x_{i-1})$$

以後、x は形態素解析で得られた品詞を表すと考える。

一般に、形態素解析では、すべての解析結果 X は有限個の x の系列で構成されているが、解析結果の数が非常に大きいため、すべての解析結果 $X = x_1, x_2, \ldots, x_n$ について、シンボルの生成確率 $P(x_i|x_1, x_2, \ldots, x_{i-1})(i < n)$ を求めるのは現実的ではない。本稿では、N 個のシンボル連鎖の出現確率 (N-gram) を用いて生成確率の近似を行う。N-gram は言語情報のモデル化に優れており、従来から広く用いられている ([1])。

N-gram モデルによる X の出現確率は、以下の近似式で求められる。

$$P(X) = \prod_{i=1}^{n} P(x_i|x_{i-N+1}, x_{i-N+2}, \ldots, x_{i-1})$$

N-gram によるモデル化の利点として、言語の持つ連鎖の局所依存性の表現に適合すること、トレーニングセットの性質をコンテキストの相対頻度で抽出するだけなので、モデル生成の計算量がわずかで、トレーニングの結果が安定して再現できることがある。

一方、長距離依存性や構文的制約の階層性を獲得できない欠点や、カバー率の低下による信頼性の低下などの問題がある。

形態素解析においては、品詞 (構文カテゴリ) の並びに関して、長距離依存性よりも局所的な依存性の方が重要であると考えられ、また、情報の抽出・反映が容易であることから、本稿では $N = 2$ とした bigram を用いる。bigram の式は

$$P(X) = \prod_{i=1}^{n} P(x_i|x_{i-1})$$

である。

4 品詞列での順位づけ

ここでは、パーザの解析結果である圧縮統語森を展開しフラットな品詞列としたものに対しての bigram による生成確率の付与について、具体的に説明する。

「足で移動をする」

を解析した結果、

「足」「で」「移動」「を」「する」

と単語に区切られ、それぞれの品詞の並びに関して以下の 4 つの解析結果 X_i $(i = 1, \cdots, 4)$ が得られたとする（正解は 3 である)。

1. X_1 : n, jd, n, jo, v
2. X_2 : n, jd, v, jo, v
3. X_3 : n, jo, n, jo, v
4. X_4 : n, jo, v, jo, v

ここで、n は名詞、jd は助動詞、jo は助詞、v は動詞をそれぞれ表している。これらに対し、品詞 (n、v、jo、jd) に関する bigram によりそれぞれの生成確率を計算する（bigram の個々の値についてはここでは省略する）と、

$$\begin{aligned}
1.\ P(X_1) &= P(jd|n) \times P(n|jd) \\
&\quad \times P(jo|n) \times P(v|jo) \\
&= 1.15 \times 10^3 \\
2.\ P(X_2) &= P(jd|n) \times P(v|jd) \\
&\quad \times P(jo|v) \times P(v|jo) \\
&= 3.94 \times 10^5 \\
3.\ P(X_3) &= P(jo|n) \times P(n|jo) \\
&\quad \times P(jo|n) \times P(v|jo) \\
&= 6.90 \times 10^2 \\
4.\ P(X_4) &= P(jo|n) \times P(v|jo) \\
&\quad \times P(jo|v) \times P(v|jo) \\
&= 8.86 \times 10^3
\end{aligned}$$

となり、正解である X_3 の生成確率が最も高くなっている。

このように、品詞列に対して bigram により生成確率を付与することは有効であると考えられるが、圧縮統語森を一度展開しなければならないため、統語構造を (圧縮統語森のまま) 統語解析後の処理に渡すことができない。次の章では、圧縮統語森を展開することなく bigram を利用する方法を説明する。

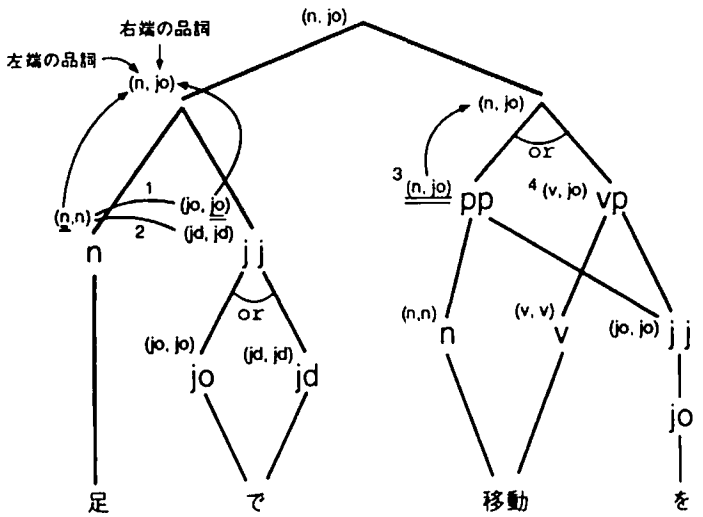

図 1: 「足で移動を」の確率計算による順位付け

5 圧縮統語森上での順位づけ

形態素解析と統語解析の解析結果を後の処理 (意味処理等) において利用するためには、曖昧性をできるだけ圧縮した形すなわち圧縮統語森としておくことが有効である。ここでは、圧縮統語森を展開することなく、bigram 情報を利用する手法について述べる。

圧縮統語森内での bigram による候補の順位付けは、

> 1 つの統語カテゴリにパックされた複数の候補のそれぞれの生成確率を求め、最大となる候補、その値、左右端の品詞を記憶しておく (マークをつける)。上位のカテゴリ (構成規則 $A \rightarrow A_1, A_2, \cdots$) では、記憶してある値、品詞により、
>
> A の確率値
> $= A_1$の確率値
> $\times P(A_2$の左端品詞 $|A_1$の右端品詞)
> $\times A_2$の確率値 $\times \cdots$
>
> のように確率を求める。

に基づいて行う。そして、後の処理ではマークを利用して最適な候補を取り出す。

前出の「足で移動を」の例を図 1 に示す。まず、「付属語列 (jj)」の「で」に関して「助詞 (jo)」と「助動詞 (jd)」のいずれかという曖昧性がある。しかし、ここでは両者に確率の優劣を付けられないため、上位の統語カテゴリ「後置詞句 (pp)」に両方の情報を伝達する。「足」より「名詞 (n)」、「で」より「助詞 (jo)」、「助動詞 (jd)」を受けとった「後置詞句 (pp)」は、1.「名詞 (n)、助詞 (jo)」、2.「名詞 (n)、助動詞 (jd)」の 2 通りの生成確率を計算し、より値の大きい 1. の情報を上位の統語カテゴリへ伝達する。渡す情報は、その後の生成確率の計算に必要な左右端の品詞、その部分品詞列の生成確率である。また「付属語列 (jj)」では、「助動詞 (jd)」より「助詞 (jo)」が優先されたことを記憶 (マーク) しておく。これによ

り、後の処理(意味処理等)で「助詞(jo)」が優先して利用されることになる。

同じように「移動を」に関して「後置詞句(pp)」と「動詞句(vp)」の曖昧性があるが、より値の大きい「後置詞句(pp)」が持つ情報(3)を上位の統語カテゴリに伝達する。

以上のように、圧縮された統語構造に対して、bi-gramによる生成確率を計算し、それが最大の統語構造の品詞情報を、上位の統語カテゴリへ伝達していく。

ここで述べた方法は、パックされた部分の順位づけをその他の部分より優先して行う点で、前章で述べた品詞列の場合と異なっている。パックされた部分は統語的に結び付きが強いと考えられるため、それらを優先することにより統語的制約を反映させることができる。

6 実験

ここでは、これまで述べた方法に基づいた実験とその結果について述べる。

実験では、品詞数33、規則数86の文法を用いて平均長22.9文字の300文を解析した。文節数最小法を利用して候補を絞り込んだ場合（候補数平均2.9）と、文節数最小法を利用しない場合（候補数平均76.4）のそれぞれについて、解析結果を品詞列に展開してからの生成確率による順位づけと、圧縮統語森上での生成確率による順位づけを行い、1位の候補の解析正解率を調べた。ここでは、単語の分割と品詞の割り当てが正しく行われているものを正解としている。結果は以下の通りである。

	文節数最小法	
	なし	あり
品詞列	88.3%	88.5%
圧縮統語森	89.6%	89.6%

この結果から、文節数最小法の利用の有無は、解析精度にほとんど影響ないことがわかり、生成確率による順位づけがより少ない単語に分割された候補を優先するという性質が、文節数最小法の効果を包含していると結論できる。また、展開した品詞列に対しての場合よりも、圧縮統語森上での場合の方が正解率が高いことから、統語解析により得られる統語的制約が形態素解析の候補の順位づけに貢献していると結論できる。

7 おわりに

本稿では、一般化LR法を基にして形態素解析と統語解析を同時に行った出力である圧縮統語森上で、形態素解析結果の曖昧性解消を行う手法を提案した。実験により、生成確率による順位づけは有効であることが確かめられた。

今後の課題として、形態的情報により品詞を細分化すること、trigram等依存距離の長い統計情報の利用、未知語処理への応用が挙げられる。

参考文献

[1] Bahl, L.R. et.al.: Recognition of a Continuously Read Natural Corpus
IEEE ICASSPproc, pp29-32, 1978

[2] 吉村賢治, 日高 達, 吉田 将: 文節数最小法を用いたべた書き日本語文の形態素解析
情報処理 Vol.24, No.1, pp.40-46, 1983

[3] 相澤道雄, 徳永健伸, 田中穂積: 一般化LR法を用いた形態素解析と統語解析の統合
信学技報 NLC93-2,1993

[4] EDR 日本語単語辞書
日本電子化辞書, 1993

[5] EDR 日本語コーパス
日本電子化辞書, 1993

[6] 伴光 昇: 形態素解析と統語解析の統合処理システムに関する研究
東京工業大学修士論文, 1994