n-gram 統計による抽出文字列の品詞推定

下畑さより 介弘達哉 杉尾俊之
{sayori,sukehiro,sugio}@kansai.oki.co.jp
沖電気工業（株）研究開発本部 関西総合研究所
〒540 大阪市中央区城見 1-2-27

1 はじめに

自然言語処理において、コーパスから未知語や専門用語、定型表現などの文字列を自動的に抽出する研究が行なわれている。代表的な研究に、長尾、森によるn-gram統計を用いた文字列抽出方式 [1] がある。これは、コーパスに対してn-gram統計処理を行ない、文字数および頻度の順に文字列を抽出する方法である。この方法では、形態素解析などの前処理を行なうことなく文字列を抽出できるという利点がある反面、断片的な文字列がかなりの割合で混在するという問題点がある。

これに対し我々は、前後に出現する文字（隣接文字）の分散の度合を基準として、有効な文字列を抽出する方法を提案した [5]。この他にも、相互に重複する部分文字列を除去する方法 [2]、ヒューリスティックを用いて抽出対象を限定する方法 [3]、出現頻度を正規化する方法 [4] などが提案されている。これらの手法で抽出した文字列を言語処理に利用する方法は、様々なものが考えられるが、いずれの場合にも、抽出文字列の性質を正しく認識することが必要となる。

本論文では、隣接文字の情報を使って、n-gram統計により抽出された文字列の品詞を推定する方法について述べる。以下の章では、まず、隣接文字に関する考察を行ない、品詞推定処理の概要と具体的な実現方法について説明する。また、実験の結果と考察を述べる。

2 基本となる考え

本論文では、「同一の性質を持つ文字列の隣接文字の出現パターンは類似している」という考えに基づき、抽出文字列の性質を推定する方法を提案する。「性質」とは、品詞や意味分類のように語の振舞いを決定する言語情報を指す。また、隣接文字の出現パターンとは、隣接文字の種類と出現確率を示す[1]。品詞を例にとると、ある文字列の品詞が「名詞」である時には後接文字に格助詞となる文字が多く出現し、「動詞」である時には活用語尾となる文字が多く出現する傾向がある。

こうした出現パターンの特徴を手がかりに、抽出文字列の性質が何であるかを推定する。すなわち、性質ごとの隣接文字の出現パターンから性質を決定する要素を規則として獲得し、それを抽出文字列の出現パターンに適用することにより、抽出文字列の性質を推定する。ここでは、品詞を対象として、抽出文字列の性質を推定する方法を説明する。

本論文で提案する品詞推定方法の基本的な流れは以下の通りである。

1. 隣接文字の出現パターンの獲得
2. 品詞推定規則の獲得
3. 品詞推定規則の抽出文字列への適用

第1に、各品詞に属する単語の隣接文字の出現パターンを求める。次に、これを学習データとして品詞推定の規則を獲得する。規則の獲得には、帰納学習アルゴリズムを利用する。これは、属性、属性値、クラスの形式のデータから、クラスを決定するのに有効な属性とその値を統計的に計算した決定木を学習するものである。本手法では、隣接文字を属性、各隣接文字の出現確率を属性値として、クラスである品詞を分類する決定木を学習し、それを品詞推定規則とする。最後に、得られた品詞推定規則を抽出文字列の隣接文字の出現パターンに適用し、文字列の品詞を推定する。

同様の研究に、森、長尾による研究 [7] がある。これは、同一品詞の隣接文字を集計して品詞ごとの出現パターンを求め、文字列の出現パターンと比較することにより品詞を推定する手法である。この手法では、出現パターンがその品詞に属する単語の平均的な値となり、個々の特徴あるパターンが埋没してしまう可能性が高い。これに対し、我々の手法では、個々の単語の出現パターンから帰納的に品詞推定規則を生成するため、品詞分類に有効な値を規則に反映することができる。また、前者は対象となる文字列をすべての品詞の出現パターンと比較する必要があるが、我々の手法では品詞決定に有効なものから順に配列された規則を適用するため、効率良く品詞推定処理を行うことができるという特徴もある。

3 出現パターン

帰納学習アルゴリズムの利用において、データの形式、すなわち出現パターンの形式をどうするかは重要な

[1] 文字列の前方と後方では出現する文字の役割が異なっているので、前接文字と後接文字の出現パターンは分けて考える必要がある。

問題となる。ここでは、出現パターンを構成する個々の要素の設定方法を具体的に説明する。

3.1 隣接文字

隣接文字は文字列の前後に出現する文字で、コーパス中に出現するすべての種類の文字が隣接文字となり得る。しかし、隣接文字の種類は文字列ごとにかなりばらつきがあり、すべての文字を属性として設定しても、実際にはデータとしてかなり疎な状態となる。また、個々の文字が品詞を決定する要素とならない場合は、細かく分けても意味がない。そこで、文字の持つ機能的な働きに着目し、以下の観点から、合計77種類の属性を設定する。

- ひらがなは、助詞、助動詞、活用語尾のように機能語として用いられることが多いため、各文字を1つの属性とし、77の属性を設定する。
- 句読点は、文の切れ目を表すため、まとめて1つの属性とする。文頭、文末も便宜上この属性に含める。
- 漢字、カタカナといったその他の文字は、主に自立語の構成要素として用いられ、機能的な意味は持たないため、まとめて1つの属性とする。

隣接文字は前接文字と後接文字に分け、各文字に対して出現確率を計算する。出現確率は、文字列の前後に各隣接文字が出現する回数を文字列の出現回数で除することにより求める。

3.2 品詞分類

品詞は、前接文字から導き出される品詞（前接品詞）と、後接文字から導き出される品詞（後接品詞）を分けて定義する。品詞体系は、一般文法によるものと若干違っている。これは、片方向の隣接文字の特徴から品詞を特定することが困難であったり、同じ品詞でも隣接文字の特徴が違う場合があったりするためである。前接品詞の分類を表1に、後接品詞の分類を表2に示す。

- 名詞、形容詞などの自立語は、先行する文字と独立した関係であり、これらの品詞の間に前接文字の差はほとんどないと考えられる。よって、これらの品詞は同じ前接品詞 (jr) に分類する。
- 接続助詞、助動詞のように用言と接続する付属語（表では「助詞」と記す）は、品詞ではなく接続可能な用言の活用形によって前接文字が決まる。よって、接続する用言の活用形ごとに別の前接品詞に分類する。
- 助詞や副詞はともに連用修飾の働きを持ち、直後の文字とは独立した関係であり、これらの品詞の間には後接文字の差はほとんどないと考えられる。よって、これらの品詞は同じ後接品詞 (ry) に分類する。
- 動詞は活用型により、後続する語尾の種類が決まるので[2]、活用型ごとに別の後接品詞に分類する。
- 助動詞の活用は、動詞、形容詞、形容動詞に準拠しているため、これらの後接品詞に含める。

前接品詞	品詞
jr	名詞、形容詞、形容動詞 動詞、副詞、連体詞
gb	語尾
kj	格助詞
js_mz	助詞（未然形に接続）
js_ry	助詞（連用形に接続）
js_rt	助詞（連体形に接続）
	:
js_tai	助詞（体言に接続）

表 1: 前接品詞の分類

後接品詞	品詞
ms	名詞
ks	形容詞
kd	形容動詞
rt	連体詞
ry	副詞、格助詞、接続助詞
gb	語尾
1d	動詞（上1段活用、下1段活用）
5k	動詞（か行5段活用）
	:
sh	動詞（さ行変格活用）
kh	動詞（か行変格活用）

表 2: 後接品詞の分類

4 品詞推定

4.1 品詞推定規則の獲得

前接文字および後接文字による品詞推定規則を各々求める。品詞推定規則の獲得には帰納学習アルゴリズムC4.5[6]を利用する。C4.5は、与えられた出現パターンから、各隣接文字が品詞の決定にどれだけ有効かをエントロピーを用いて統計的に計算し、有効な隣接文字から順に配列した決定木を生成する帰納学習プログラムである。

表3は、名詞、形容動詞、形容詞の後接文字の出現パターンの例である。表3のデータにプログラムを適用すると、図1のような決定木が学習される。

[2]活用語は語幹を単位とし、語尾は別の品詞と考える

品詞ごとの出現パターンの獲得には、品詞タグつきコーパスを利用する。すなわち、各品詞に属する単語を対象として、コーパス中の単語の出現回数と隣接文字の出現回数を計数し、各隣接文字の出現回数を単語の出現回数で除することで求める。

クラス	あ	い	…	が	…	な	…
ms	0	0		7		0	
ms	0	0		13		0	
kd	0	0		0		26	
kd	0	0		0		24	
ks	0	57		0		0	
ks	0	73		0		0	

(単位は%)

表 3: 出現パターンの例

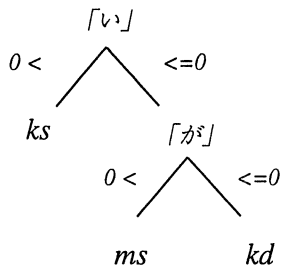

図 1: 決定木の例

4.2 抽出文字列への適用

抽出文字列の品詞推定は、以下の手順で行なう。まず、抽出文字列の隣接文字の出現パターンを計算する。次に、この出現パターンに 4.1で述べた品詞推定規則に適用し、得られた前接品詞と後接品詞の組合せから、文字列の品詞を推定する。

抽出文字列の出現パターンは抽出文字列の出現回数と抽出文字列より 1 文字多い文字列の出現回数から求める。例えば、文字列 "ab" の出現回数 f(ab) が 5、文字列 "abc" の出現回数 f(abc) が 3 である時、 "ab" の後接文字 "c" の出現確率 P(ab,c) は以下の式で求められる。

$$P(ab, c) = \frac{f(abc)}{f(ab)} = 0.6$$

これを "ab" を含む 3 文字の文字列すべてに対して行なうことにより、文字列 "ab" の出現パターンを求めることができる。

次に、 4.1により得られた品詞推定規則に従って文字列の前接品詞および後接品詞を推定する。ここで得られる推定結果は片方向での品詞的特徴を示すものであるから、これらを組み合わせて最終的な文字列の品詞を推定する。表 1、表 2の品詞分類に基づき、前接品詞と後接品詞が一致すれば、その文字列の品詞は一致した品詞である可能性が非常に高いと考える。前接・後接品詞とそこから導き出される文字列の品詞の対応例を表 4に示す。

- 前接品詞が「jr」で後接品詞が「ms」であれば、文字列の品詞は「名詞」とする。
- 前接品詞が「kj」で後接品詞が「ry」であれば、文字列の品詞は「格助詞」とする。
- 前接品詞が「js_mz」で後接品詞が「ks」であれば、文字列の品詞は「助動詞（未然形に接続、形容詞型活用）」とする。

文字列	前接	後接	品詞
ファイル	jr	ms	名詞
として	kj	ry	格助詞
な (い)	js_mz	ks	助動詞

表 4: 前接品詞と後接品詞の組合せ

5 実験

5.1 実験の条件

以上に述べた手順に従い、 n-gram 統計による抽出文字列の品詞を推定する実験を行なった。学習データは、表 1, 表 2に示した各品詞に属する単語 100 件ずつを EDR コーパス [8] から取り出し、隣接文字の出現パターンを計算したものである[3]。また、実験データは、コンピュータマニュアル 39,429 文 (1,034,933 文字) から文献 [5] の方法で抽出した文字列のうち、数字・記号列を除いた上位 100 件である。帰納学習プログラムは、 C4.5(オプション指定なし、枝刈なし) を使用した。

5.2 実験結果

まず学習データから品詞推定規則を獲得し、その精度を評価した。評価方法は、 5 分割によるクロスバリデーションである。結果を表 5に示す。次に、この規則を実験データに適用して前接品詞と後接品詞を求めた。結果を表 6に示す。

[3]ただし、出現回数が少ないと信頼性が低下するため、出現回数が 100 回以上の単語のみを対象とした。そのため、品詞によっては、学習データ数が 100 に満たないものもある。

	(1)	(2)
前接	94.5	86.3
後接	95.0	87.0

単位は%

(1) 学習データの正解率
(2) テストデータの正解率

表 5: 品詞推定規則の評価

	(1)	(2)
前接	98	94
後接	84	71

単位は%

(1) 候補中に正解が含まれていたもの
(2) 第1候補が正解であったもの

表 6: 実験データの品詞推定結果

5.3 考察

n-gram 抽出文字列には、単語だけでなく文・句単位の文字列や一般文法の品詞に当てはまらない文字列も多く含まれているが、本手法ではこうした文字列も柔軟に扱うことができる。表7の「これによって」「として」は、複数の単語から構成される表現であるが、本手法ではこれらの働きを正しく認識し、連用修飾句、格助詞と推定されている。

また、コーパスでの出現傾向が反映されるので、複数の品詞の解釈が可能な場合でも、対象コーパス内で最適な品詞に絞り込むことができる。例えば、「適切な」は形容動詞の連体形とも考えられるが、本手法では連体詞と推定されている。実際にコーパスを調べたところ、出現する「適切な」はすべて名詞にかかっていた。

反対に失敗したものは、3.2の品詞分類において、特徴が類似している品詞どうしを取り違えたものがほとんどであった。例えば、前接品詞での誤りは、すべて「kj」と「gb」を取り違えたものであった。「kj」の前接文字には名詞が、「gb」の前接文字には動詞や形容動詞などの語幹がくることが多いが、これらは句読点がこない、漢字が多くひらがなが少ないという点で類似しており、差異をうまく数値化することができなかったものと考えられる。後接品詞では、名詞と連用修飾句、連体詞と連用修飾句の取り違えが多かった。

本論文では、文法的機能の類似性に着目して品詞分類を行ったが、隣接文字の出現パターンを分析し、従来の品詞体系とは違う観点から分類を考える必要がある。また、今回の実験では学習データと実験データのコーパスは違う分野のものを使用したが、同一分野のコーパスを用いることで、正解率が向上することが予想される。

文字列	前接	後接	推定品詞
します	kj*	ry*	格助詞*
から	gb*	ry*	助詞*
これによって	jr	ry	副詞
として	kj	ry	副詞
マウント	jr	sh	動詞(サ変)
適切な	jr	rt	連体詞

*は誤り

表 7: 品詞推定結果の例

6 まとめ

本論文では、n-gram による抽出文字列の品詞を推定する方法として、隣接文字の出現パターンを利用する方法を提案した。また、コンピュータマニュアルに対して品詞推定の実験を行ない、前接で98 %、後接で84 %の正解率を得た。

今後は、品詞分類、学習データの分野や量、学習データ中の各品詞の配分といった条件を変えて実験を行ない、推定結果との関係を調べていく。また、品詞以外の言語情報についても本方式の有効性を検証したいと考えている。

参考文献

[1] 長尾, 森: 大規模日本語テキストのnグラム統計の作り方と語句の自動抽出, 情報処理学会自然言語処理研究会報告 96-1, pp1-8(1993)

[2] 池原, 白井, 河岡: 大規模日本語コーパスからの連鎖型および離散型共起表現の自動抽出, 電子情報通信学会技術研究報告 (NLC95-3), Vol.95, No.29, pp17-24(1995)

[3] 新納, 井佐原, 疑似 N グラムを用いた助詞的定型表現の自動抽出, 情報処理学会論文誌, Vol.36, No.1, pp32-40(1995)

[4] 中渡瀬, 木本: 統計的手法によるテキストからの重要語抽出メカニズム, 情報処理学会情報学基礎研究会報告 39-6, pp41-48(1995)

[5] 下畑, 杉尾, 永田: 隣接文字の分散値を用いた定型表現の自動抽出, 情報処理学会自然言語処理研究会報告 110-11, pp71-78(1995)

[6] Quinlan,J.R:C4.5 Program for Machine Learning, Morgan Kaufmann(1993)

[7] 森, 長尾:nグラム統計によるコーパスからの未知語抽出, 情報処理学会自然言語処理研究会報告 108-2, pp7-12(1995)

[8] EDR コーパス: 日本電子化辞書研究所 (1994)