自然言語処理 114－3
（1996. 7. 18）

隣接文字情報を用いた n-gram 抽出文字列からの名詞句の自動抽出

下畑さより 杉尾俊之
{sayori,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

あらまし　本論文では、n-gram 統計により抽出された文字列の中から、名詞句だけを精度良く抽出する方法について述べる。本手法は、品詞および隣接文字情報を付与した n-gram 統計文字列を学習データとし、各文字の名詞に接続する確率を求める。そして、これを推定対象文字列の隣接文字に適用することにより、文字列の名詞となる確率を算出する。本手法を用いた名詞句推定実験では適合率 82.9 %、再現率 80.4 %を得ることができた。

和文キーワード　自動抽出, 名詞句, コーパス, n-gram

Extraction of Noun Phrases from Corpora Using n-gram Statistics

Sayori SHIMOHATA Toshiyuki SUGIO
{sayori,sugio}@kansai.oki.co.jp
Kansai Laboratory, Research & Development Group
Oki Electric Industry Co., Ltd.
1-2-27 Shiromi, Chuo-ku, Osaka 540, Japan

Abstract　This paper describes a method to extract noun phrases from strings derived through n-gram statistics from corpora. This method estimates the probability that a string belongs to noun by distribution of characters which precede or follow the string. The experimental results achieved 80.4 % recall and 82.9 % precision.

英文 **key words**　automatic extraction, noun phrase, corpus, n-gram

1 はじめに

近年、統計的な手法を用いて、コーパスから単語やイディオムなどの意味的にまとまりのある文字列を抽出する研究が行なわれている。

その中でも、n-gram 統計による文字列抽出の手法[長尾 93] は、未知語や分割位置について考慮する必要がなく、文字列の抽出に有効である。しかしながら、この手法では断片的文字列を含む様々な単位の文字列が混在して抽出されるため、実際に利用するには用途に応じてさらに文字列の選別を行なわなければならないという問題がある。

これに対し、重複する文字列を除去する手法[池原 95]、ヒューリスティックを用いて抽出対象を限定する手法[新納 95]、出現頻度を正規化する手法[中渡瀬 95] などが提案されている。我々も[下畑 95] において、隣接文字のエントロピーを基準に、表現の単位として有効な文字列を抽出する手法を提案した。

本論文では、自然言語処理で問題となる専門用語や未知語の多くは名詞であるという仮定から、n-gram 統計により抽出された文字列 (以下、n-gram 抽出文字列と呼ぶ) から名詞を抽出する方法を提案する。以下では、まず従来研究とその問題点について述べ、次に本手法の具体的方法を説明する。そして、本手法に基づく実験および評価を行ない、本方式の有効性を検証する。

2 従来研究

n-gram 抽出文字列の品詞推定に関する研究には、[森 95] [下畑 96] がある。これらの研究はいずれも、同じ品詞に属する文字列の隣接文字は類似しているという考えに基づき、各品詞の隣接文字の出現パターンと文字列の隣接文字の出現パターンとを比較することによって品詞推定を行なっている。

しかし、これらの手法には 3 つの問題がある。

第 1 に、これらの手法では全ての隣接文字を対象としているが、品詞推定に有効な隣接文字はある程度限られており、全ての文字を対象とすることで精度および処理効率が低下することになる。表 1は、EDR コーパス [EDR] 中の名詞の隣接文字の出現確率を集計したものである。EDR コーパスでは、名詞の前接文字として 33000 種類、また、後接文字として 25000 種類の文字が出現しているが、どちらも上位 5 文字で隣接文字全体のほぼ半分を占めている。品詞推定を行なう際には、全ての文字を対象とするのではなく、こうした特定の頻出する文字だけを対象とするほうが文字列による揺れも少なく、効率的である。

第 2 に、隣接文字の出現確率を比較しているが、出現確率は必ずしも品詞らしさの基準として適当であるとは限らない。例えば、表 1に示す隣接文字の出現確率は、EDR コーパス中のすべての名詞の隣接文字を合計したものであって、それぞれの名詞の隣接文字がこの割合で出現することを示すものではない。文字列の品詞を推定するには、隣接文字の出現確率の分布が品詞の平均的なパターンとどれだけ似ているかではなく、品詞に隣接する傾向の強い文字がどれだけ多く出現しているかを評価する必要がある。

確率	前接		後接	確率
18	の		の	16
13	、		を	11
11	(文頭)	《名詞》	に	11
4	は		が	8
3	た		は	8
51	(その他)		(その他)	46
				(出現確率は%)

表 1: EDR コーパスにおける名詞の隣接文字

第 3 に、一般の品詞タグつきコーパスを学習データに使っているが、こうしたコーパスでは n-gram 抽出文字列の特徴を反映した隣接文字の情報を得ることができない。表 2は、[下畑 95] による抽出文字列 (新聞記事 1 カ月分) の上位 100 件を単位ごとに分類したものである。この表からも分かるように、n-gram 抽出文字列には複数の要素 (単語) からなる文字列や断片的文字列を含む様々な単位の文字列が混在している。これらの文字列の品詞を正しく推定するためには、n-gram 抽出文字列の特徴を考慮した学習データを用意する必要がある。

出現数	単位	文字列の例
25	名詞	買い, 東京
21	格助詞	から, には
15	名詞+付属語	企業の, 株を
15	動詞+付属語	した, している
8	付属語	である, める
7	副詞	しかし, また
5	断片的文字列	
4	動詞	する

表 2: n-gram 抽出文字列の分類

3 隣接文字情報を利用した品詞推定

以上の議論から、本論文では隣接文字の品詞接続性に基づく品詞推定方式を提案する。

品詞接続性とは、文字がある品詞の文字列に接続す

る傾向の強さを表す。ある文字が特定の品詞pos_iの隣接文字として頻出し、それ以外の品詞の隣接文字にはあまり出現しない時、その文字はpos_iに対する品詞接続性が高いという。

このような文字は品詞推定を行なう上で有効な手がかりとなる。例えば、推定対象の文字列の隣接文字にpos_iへの品詞接続性の高い文字が多く出現している場合、その文字列の品詞はpos_iである可能性が高いと考えられる。本方式では、学習データを用いて文字の品詞接続性を獲得し、これを文字列の隣接文字に適用することにより、品詞推定を行なう。

隣接文字の出現傾向は文字列の前方と後方で異なるため、品詞接続性は前接文字と後接文字とでそれぞれ求める必要がある。以下では、片方向の品詞性の推定方法について説明するが、実際には同様の方法で両方向の品詞性を求めたうえで、文字列の品詞性を推定することになる。

3.1 学習データの作成

学習データとして必要な情報は、文字列の品詞、隣接の種類、各隣接文字の出現回数である。隣接文字情報は、前接、後接それぞれに対して求める。2の議論から、学習データにはn-gram抽出文字列を用いることが望ましい。この場合、文字列および隣接文字情報は、文字列抽出処理の過程で自動的に獲得することが可能である。

3.2 品詞接続性の計算

品詞接続性$P(c, pos_i)$は、文字cがpos_iの文字列に隣接する確率で表す。本論文では、これを以下のように定義する。

$$P(c, pos_i) = \frac{f\_kind(c, pos_i)}{\sum_{k=1}^{n} f\_kind(c, pos_k)} \quad (1)$$

この時、$f_kind(c, pos_i)$は、文字cが隣接するpos_iの文字列strの種類数を示す。式(1)で文字列の出現回数でなく種類数を対象とするのは、特定の文字列にだけ隣接する文字の影響を抑制し、品詞pos_iの文字列に偏りなく出現する文字を計数するためである。

また、$P(c, pos_i)$は、出現回数の低いcに対して大きくなる性質がある。そこで、以下の閾値を設定し、条件を満たす場合だけを計算の対象とする。

$$f(str, c) \geq f_{min} \quad (2)$$

$$\sum_{k=1}^{n} f\_kind(c, pos_k) \geq fk_{min} \quad (3)$$

$f(x)$：xの出現回数

3.3 文字列の品詞推定

文字列strが品詞pos_pである確率$Pr(str, pos_p)$は、隣接文字の出現回数に品詞接続性を乗じたものを累算することで求める。本論文では、これを以下のように定義する。

$$Pr(str, pos_p) = \frac{\sum_{i=1}^{m} f(str, c_i) \cdot P(c_i, pos_p)}{\sum_{i=1}^{m} f(str, c_i)} \quad (4)$$

$$C = \{c | c_1, c_2 ... c_n\}$$

隣接文字c_iが学習データ中に出現しなかったり、式(2)(3)の閾値に達しなかった場合には、$P(c_i, pos_p)$が未知となるが、その場合は計算の対象としない。

$Pr(str, pos_p)$は、隣接文字の品詞接続性が高いほど、また、品詞接続性の高い隣接文字の出現回数が多いほど大きくなる。

4 名詞句の抽出実験

4.1 実験条件

学習データは、日本経済新聞[日経94]の株式関連記事1カ月分(約4500文)から[下畑95]の方法で抽出したn-gram抽出文字列のうち、エントロピー値1以上、出現回数10以上の文字列1777件である。これに、人手で品詞情報を付与した[1]。

品詞情報は2通りの方法で与える。第1の方法では、文字列を1つのまとまりと考え、文字列全体の働きを示す品詞を与える。また、第2の方法では、文字列がいくつかの要素(単語)からなる場合を考え、文字列の前接部分、後接部分に着目した品詞を与える。品詞付与の例を表3に示す。この結果、第1の方法では553件が、第2の方法では前接で848件、後接で663件が名詞となった。

文字列	第1の方法	第2の方法	
		前接	後接
株式	ms	ms	ms
の株式	not	not	ms
株式が	not	ms	not
の株式を	not	not	not

ms は名詞
not はその他の文字列

表 3: 品詞付与の例

次に、式(1)を使ってそれぞれの学習データから各文字の名詞接続性を求める(第1、第2の品詞付与方

[1]この場合、名詞句の抽出が目的であるため、品詞情報は名詞か否かを与える。

法による名詞接続性の情報をテーブル1、テーブル2と呼ぶ)。本実験では、f_{min} を5、fk_{min} を10に設定した。この結果、テーブル1、2ともに前接で67文字、後接で56文字の名詞接続性が獲得された。

式(4)を使って、n-gram抽出文字列の名詞性を計算する。実験では、以下の3通りの方法でn-gram抽出文字列の名詞性を計算し、閾値ごとに再現率と適合率を求めた[2]。

実験1　テーブル1を用いて学習データと同じn-gram抽出文字列の名詞性を計算する。

実験2　テーブル1を用いて学習データと同じn-gram抽出文字列の名詞性を計算する。

実験3　テーブル1を用いて学習データとは異なる株式関連記事1カ月分のn-gram抽出文字列(全1810件、うち名詞560件)の名詞性を計算する。

4.2 実験結果

実験1～3の結果を表4、表5、表6に示す。表7は、実験3の結果名詞性が高いと推定された文字列の一部である。

品詞による名詞性の値　図1は、実験3における名詞とその他の文字列の名詞性の値の分布を示している。全体的に名詞は名詞性の値が高く、その他の文字列は値が低くなっており、本手法による名詞句の推定が妥当であることが分かる。

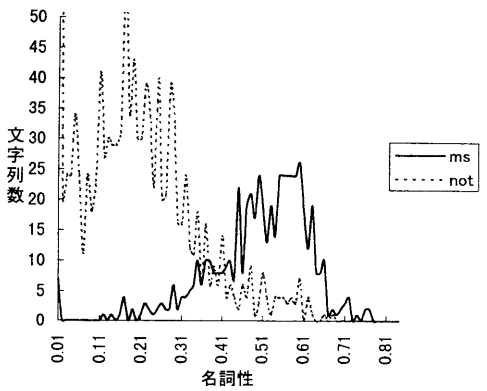

図 1: 名詞性と文字列数の相関

[2]実験では前接文字、後接文字の名詞性のうち小さい方を文字列の名詞性とした。

名詞性	正解数	抽出数	適合率	再現率
0.7	5	6	83.3	0.9
0.6	92	101	91.0	16.6
0.5	315	352	89.4	57.0
0.4	444	535	82.9	80.4
0.3	505	724	69.7	91.4
0.2	534	1043	51.1	96.7
0.1	546	1407	38.8	98.9
0.0	552	1777	31.0	100

適合率、再現率は%

表 4: 実験1の結果

名詞性	正解数	抽出数	適合率	再現率
0.9	2	2	100	0.3
0.8	17	18	94.4	30.0
0.7	183	196	93.3	33.1
0.6	349	385	90.6	63.2
0.5	446	532	83.8	80.7
0.4	495	668	74.1	89.6
0.3	527	879	59.9	95.4
0.2	542	1216	44.5	98.1
0.1	546	1454	37.5	98.9
0.0	552	1777	31.0	100

適合率、再現率は%

表 5: 実験2の結果

名詞性	正解数	抽出数	適合率	再現率
0.7	12	12	100	2.1
0.6	93	100	93.0	16.7
0.5	304	353	86.1	54.6
0.4	449	551	81.4	80.7
0.3	520	746	69.7	93.5
0.2	544	1064	51.1	97.8
0.1	553	1424	38.8	99.4
0.0	556	1810	30.7	100

適合率、再現率は%

表 6: 実験3の結果

名詞性	文字列		名詞性	文字列	
0.77	日興	(41)	0.71	日米	(36)
0.76	欧州	(83)	0.71	米証券	(16)
0.75	大手証券	(49)	0.70	野村	(75)
0.75	ＢＡｅ	(15)	0.69	シカゴ	(13)
0.73	大和	(33)	0.68	地域	(14)
0.72	山一	(49)	0.68	国の＊	(63)
0.71	フランクフルト	(46)	0.67	日銀	(20)
0.71	外為	(11)	0.67	上値	(24)
0.71	所得税減税	(13)	0.67	ニューヨーク	(115)
0.71	シンガポール株	(10)	0.66	分野	(12)

() 内は出現回数
＊は名詞でない文字列

表 7: 名詞性の高い文字列の例

名詞接続性の高い文字の特徴　テーブル 1、テーブル 2 において、名詞接続性が高くなった文字を表 8、表 9に示す。

前接		後接	
接続性	文字	接続性	文字
1.00(18)	「	0.91(11)	市
0.90(20)	(	0.88(17)	(
0.79(19)	米	0.84(19)	)
0.66(147)	(文頭)	0.82(34)	株
0.66(211)	の	0.82(11)	金

( ) 内は出現回数

表 8: テーブル 1 の名詞接続性

前接		後接	
接続性	文字	接続性	文字
1.00(20)	(	1.00(10)	物
1.00(19)	米	1.00(11)	市
1.00(18)	「	1.00(11)	金
0.95(21)	・	1.00(19)	)
0.95(19)	式	0.89(94)	を

( ) 内は出現回数

表 9: テーブル 2 の名詞接続性

名詞接続性の高い文字は 2 種類ある。1 つは一般的に名詞に接続しやすい文字 (「の」「を」など) で、比較的出現回数が多く、分野を選ばずに名詞性の推定に適用可能である。もう 1 つは学習データの分野に依存した文字 (「米」「株」「金」など) で、は出現回数が少なくても高値となる場合が多く、適用分野によっては前者より適用率が高い。これらの文字をうまく分類して使い分けることにより、推定率の改善が可能であると考えられる。

品詞付与方法による名詞句抽出の精度　実験 1 と実験 2 では、品詞付与方法の異なる名詞接続性のテーブルを参照し、名詞句の推定を行なった。その結果、表 4、表 5に示すように、名詞性の値そのものは実験 2 の方が高くなるものの、再現率と適合率の相関はほぼ同じであることが分かった。

非学習データでの精度　実験 3 ではテーブル 1 を用いて、非学習データの名詞性を推定した。その結果、表 6のように実験 1 とほぼ同等の精度が得られ、本手法が学習データ以外の n-gram 抽出文字列に対しても有効であることが分かった。ただし、この実験は同分野から抽出した文字列を対象としており、異なる分野に対しては、精度は低下すると考えられる。

不適切な推定結果の分析　名詞性が低くなった名詞の例を表 11に、名詞性が高くなったその他の文字列の例を表 10に示す。

名詞性	文字列	
0.68	国の	(63)
0.66	国内の	(11)
0.63	テク	(14)
0.62	面で	(21)
0.62	国で	(27)

表 10: 名詞性の高い文字列 (名詞以外)

推定結果が不適切であった文字列は、大きく以下の 3 種類に分類される。

1. 先頭あるいは末尾にひらがなを含む文字列
2. 意味のあるまとまりの部分文字列となる文字列
   “テク” … ハイテク, 財テク, テクノロジーなど
   “スター” … スタート, ミスター, スター TV など

名詞性	文字列	
0.13	気味	(11)
0.11	体制	(12)
0	理事	(25)
0	住友	(15)
0	スター	(16)

表 11: 名詞性の低い文字列（名詞）

3. 学習データに出現しない文字が隣接する文字列

先頭や末尾がひらがなとなる文字列は、名詞や助詞と接続するため、名詞との区別がつきにくい。この問題を解消する方法としては、ひらがな（特に助詞や語尾になりうるひらがな）を含む文字列を特別に扱うヒューリスティック規則の導入が考えられる。実際に、先頭や末尾がひらがなとなる名詞は限られているため[3]、それ以外の文字列の名詞性を低くすることにより、推定率の向上が期待できる。

2のような文字列は、本手法ではうまく扱うことができない。「スター」のように単独で名詞として使われることもある文字列は、名詞として出現した時の隣接文字の情報が、部分文字列の一部として出現した時の隣接文字情報に埋もれて、名詞性が低く算出されてしまう。この問題に対処するには、出現回数の情報を加味するなどの方法が考えられる。また、「テク」のようにそれ自体が単独で出現することはなく、常に意味のある文字列の部分文字列として出現している場合には、n-gram 抽出文字列を抽出する際に抑制できるよう文字列抽出処理を改善する必要がある。

また、データ数の不足による精度の低下は、統計情報を用いる手法において回避することのできない問題である。今回の実験では、名詞接続性の計算において、$P(c_i, pos_p)$ が未知の場合は計算の対象としなかったため、有効なデータ数が極端に少ない文字列が存在した。このような場合には、単に推定結果を出力するのではなく、十分なデータから得られた場合とそうでない場合とを区別するような仕組みを設けることが必要であろう。

5 まとめ

本論文では、品詞により隣接する文字の傾向は類似しているという考えに基づき、隣接文字の品詞接続性を利用して n-gram 抽出文字列中の名詞を抽出する手法を提案した。本手法は、品詞推定に有効な文字だけを対象に品詞への接続性を求めるため、重要性の少ない文字の影響が少ないという特徴がある。また、学習データに n-gram 抽出文字列を利用することにより、n-gram 抽出文字列における名詞の出現傾向を反映させることが可能となった。実験の結果、再現率 80 %で 80 %以上の適合率が得られ、本手法の有効性が確認できた。

今後は、前節で述べた文字種によるヒューリスティックスの導入や、頻度情報の利用などを検討し、さらに推定率の向上を計りたいと考えている。また、名詞以外の品詞についても拡張を行なう予定である。

謝辞 本論文で使用したテキストデータは、日本経済新聞記事データ CD-ROM94 版の記事を使用しました。使用を許可して下さった日経総合販売（株）およびデータの研究利用に尽力して下さった皆様に深く感謝致します。

参考文献

[長尾 93] 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)

[池原 95] 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)

[新納 95] 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)

[中渡瀬 95] 中渡瀬, 木元: 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)

[下畑 95] 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)

[森 95] 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)

[下畑 96] 下畑, 介弘, 杉尾:n-gram 統計による抽出文字列の品詞推定, 言語処理学会,(1996)

[EDR] 日本電子化辞書研究所 (1994)

[日経 94] 日本経済新聞 CD-ROM 版 (94 年版), 日本経済新聞社 (1994)

[3]動詞の連用形 (“取り引き”,“見込み” など) や形容詞、形容動詞に接尾辞がつくもの (“強さ”,“速さ”) がほとんどである。