自然言語処理 127－19
（1998. 9. 18）

文字の共起情報のみを利用した文字列抽出

延澤 志保、斉藤博昭、中西正和
{shiho,hxs,czl}@nak.ics.keio.ac.jp
慶應義塾大学計算機科学専攻

本稿では、共起関係という表面的な繋がり情報のみを利用して文の切り分けを行なう。本稿で示した実験では人名のみから得た文字間統計情報を利用する事で文字の人名としての繋がりやすさを基準に文中の人名の認識を行なう。人名中の文字の並びが他の語と比べて特徴的であり、その特徴を文字間共起情報として利用可能であることを示すため、候補文字列抽出等の前処理は行わず、テキスト全体から直接人名を認識する手法を採った。実験の結果新聞記事を用いた実験で 75% から 100% の人名の抽出に成功した。これにより人名は特有の文字の並びを持つ事が示され、文字間の共起情報のみを用いたシンプルな文字列抽出システムの構築が期待できる。

Japanese Linky String Extraction Using Statistic Information between Letters

Shiho Nobesawa, Hiroaki Saito, Masakazu Nakanishi
Dept. of Information and Computer Science, Keio University

In this paper we propose a method to extract strings using co-occurance information of letters only. In Japanese we have thousands of characters, and it is possible to find out strings using the co-occurance information of those letters. We used a Japanese person name list as the training corpus so that the system can extract name-like strings from test corpora. We had experiments on newspaper articles and the presision rates were 75% to 100% at max. No grammar or knowledge information was used to extract those names string.

はじめに

本稿では、文字列の意味的なかたまりとして捕らえず、共起関係という表面的な繋がり情報のみを利用して文を切り分ける。

本稿で示した実験では人名のみから得た文字間統計情報を利用する事で人名としての繋がりやすさを基準に文中の人名の認識を行なう。

実験の結果、珍しい名前を含んだコーパスからの抽出でも 75% 、一般的なもののみを含んだ理想的なコーパスでは 100% の人名の抽出に成功した。

1 有繋文字列の抽出

1.1 文字間統計情報による繋がりやすさの評価

複合語等の単語間の繋がりやすさ、文同士の繋がりやすさ等、連続した複数の事象の繋がりやすさはいろいろなレベルで議論されてよい。その中で、文字間の繋がりやすさは、言語を意味を持たない文字のレベルまで落として繋げ直す点で、他と大きく異なる。

1.2 有繋評価値を用いた文中の切れ目検索

事象を文中の文字とする。前述のように共起頻度の高い文字間の有繋性は高く、共起する可能性の低い文字間の有繋性は低くなるので、有繋性を示す評価値である有繋評価値をグラフに表してみると共起確率の高い部分には山が、低い部分には谷ができる事になる。二文字以上からなる文字列の場合その文字列全体で一つの山を形成する。従って谷の部分は意味のある文字列と文字列の切れ目、つまり文中の切れ目である可能性が高い。

2 有繋性の尺度

2.1 有繋性

有繋性とは、統計情報を基準にしてある二つの事象の繋がりやすさを示すものであり、その二事象が連続して出現する可能性が高い時有繋性が高いと呼ぶ [1]。また、 Nobesawa らは有繋性のみに基づいた意味のある文字列の抽出についての実験を行なったが、この手法により抽出されるような統計情報から観て有繋性が高いと判断された文字列を有繋文字列と呼ぶ [1]。

2.2 d-bigram 統計情報

本稿のシステムは、文字間統計情報を d-bigram モデルの形で取得、利用する。d-bigram とは連続でない二事象についても扱う bigram モデルの一種であり、その二事象間の距離も考慮する [2]。例えば、扱う事象を日本語文字とする場合、「こねこ (仔猫)」という 3 文字から成る文字列からは、(“こ”,“ね”;1) 、 (“こ”,“こ”;2) 、 (“ね”,“こ”;1) の 3 つの文字間 d-bigram 情報が得られる事になる。ここで、第三引数は二事象の距離であり、隣り合うものを距離 1 としている。d-bigram では距離を考慮するため、基本的に、 (“こ”,“こ”;0) と (“こ”,“こ”;2) 等距離が異なれば違う事象として数えられる。

2.3 有繋文字列の抽出

Nobesawa らは、有繋性の概念に基づいて文の切り分けを行なった [1]。この手法では、トレーニングコーパスから抽出した d-bigram 頻度情報を有繋性の評価に利用する。入力文の各文字間について、先の d-bigram 頻度情報を基にその有繋性を示す評価値を計算する。事象間の有繋評価値の高い隣接二事象は有繋性が高いと判断され、逆に有繋評価値の低いものは有繋性が低い、すなわち統計情報に照らした場合繋がりを持つ可能性が低いと認識する。この有繋性の「谷」の部分で文を切り分ける事で、有繋性の高い文字列の抽出を行なう。この時、有繋性の判断に用いる統計情報はその隣接二事象間だけでなく二事象の周りの事象についての情報も足し合わせるため、単純な bigram での推定と違い文字列に応じた切り分けが可能となっている。

図 1 は文中の各文字間の有繋評価値を表した評価値グラフの例である。図 1 は、文末記号も含めて 14 文字から成る一文について、各文字間の有繋評価値を縦軸で表している。図中の各アルファベットはそれぞれ文中の一文字を表し、最後の “!” は、句点等文末記号を示す。有繋評価値の高い部分は山型に、低い部分は谷になっている。谷は、前後の文字列との統計情報に照らして繋がる可能性が低いと判断される部分であり、文字列の切れ目と見做す事が可能である。図の例では、文字の並びの有繋性を調べた結果、この文は “AB”、 “CDEF”、 “G”、 “HIJK”、 “L”、 “M!” の 6 つの有繋文字列に切り分けるのが妥当と判断された事になる。

図 1: 有繋性による評価値グラフ

各文字間の有繋評価値 $UK(i)$ は以下の式 1 で算出される。

$$UK(i) = \sum_{d=1}^{d_{max}} \sum_{j=i-(d-1)}^{i} MI_d(w_j, w_(j+d); d) \times g(d) \quad (1)$$

ここで、ある二文字の d-bigram としての相互情報量 MI_d は以下の式 2 によって計算される。但

表 1: 日本語コーパスにおける各字種の割合 (%)

	bus.	ed.	人名
size(K chars)	42	275	6.7
ひらがな文字	30.2	58.0	1.3
漢字文字	47.5	34.6	98.7
カタカナ文字	19.3	4.8	0.1
数字・英文字	2.9	2.6	0.0

し、d はその二文字間の距離を示し、P はその事象の出現確率である。

$$MI_d(x,y;d) = log_2 \frac{P(x,y;d)}{P(x)P(y)} \quad (2)$$

3 日本語における固有名詞

3.1 固有名詞中の文字間の有繋性

一般に、固有名詞は漢字の割合が比較的高く、特に人名や地名ではカタカナやひらがなの割合が低くなっている。表 1 に日本語コーパス中の各字種の出現割合を示す。表 1 のうち、bus. コーパスと ed. コーパスについては Teller らの論文中の表から引用した [3]。bus. は読売新聞中の商業関係の記事から成るコーパス、ed. は朝日新聞中の教育関係の記事から成るコーパスである。比較対象として、本稿の実験に用いた人名コーパスでの字種割合を人名として並べて記載する。

表 1 で判るように、人名では漢字文字の割合が圧倒的に高く、一部ひらがなが用いられており、カタカナはほとんど使われていない。漢字以外の数字やアルファベットは一般の人名には利用されない。また、ひらがなやカタカナは、主に名前として用いられ、苗字に出現する事は稀である。

固有名詞と一般の語は、字種の比率だけでなく、文字の並び方の点でも大きく異なる。固有名詞では、頻出文字の種類が一般の語と異なり、またその並びもある程度決まったものが多い。文中の固有名詞は、文脈を調べなくてもある程度認識可能である。

3.2 未知語としての固有名詞

自然言語処理において、未知語は克服すべき課題の一つである。固有名詞はほぼ無限に存在するため、そのすべてを辞書に登録する事は不可能であり、従って処理の上で未知語となる事が多々ある。固有名詞をそれと認識する事が可能となれば、未知語の問題の解決に大きく貢献する事になる。

4 人名の抽出

4.1 人名の文字間統計情報

本稿では、Nobesawa らの有繋文字列抽出アルゴリズム [1] を基に、日本語文中の人名抽出を試みた。本実験で利用する統計情報は、人名のみからなるコーパスから得た d-bigram 頻度情報である。本稿の実験で利用した人名コーパスは以下の形式で手作業で作成した。

- 日本語の人名のみ
- 男女共含む
- 一行一名とし、一名分として苗字と名前を一組とする
- 苗字と名前の間に区切りを置かない

本稿では、固有名詞の文字間統計情報を利用する事でどの程度固有名詞の認識が可能かを検証するため、比較的小さな人名コーパスをトレーニングコーパスとした。これにより、人名コーパスに登録されていない未知の人名の抽出について考察を行なう。

なお本稿では、「苗字」、「名前」、「氏名」、「人名」の各語を区別して用いる。「氏名」はフルネームの事を指し、「苗字」、「名前」、「氏名」の全てを括って「人名」と呼ぶ事にする[1]。

4.2 人名抽出システム

本稿で提案する人名抽出システムは、人名文字間 d-bigram 統計情報を基準として、入力文中の各文字列間の有繋評価値を計算する。人名での文字間の繋がりのみを利用して有繋性を判別するため、人名としての繋がりを持たない文字間の評価値は非常に低くなり、逆に人名コーパスに出現した文字間の評価値は十分に高くなる。このそれぞれの評価値を利

[1]氏名という場合には「延澤志保」等フルネームを指し、他に「延澤」のみや「志保」のみの場合も入れて議論する場合には「人名」と表現する。

用して、評価値の十分に高い部分を人名候補として抽出する。先の有繋文字列抽出システムと違い、人名抽出システムでは人名と認識された部分のみを出力すればよい。

人名以外の文字列についてはその有繋性についての統計情報が無いため、文中の文字列は全て人名らしさの点だけについてその有繋性を問われる事になる。人名における文字の共起が一般の文字列と異なる場合、人名以外の文字列の有繋評価値は非常に低くなるため、人名部分の評価値は他と区別できるだけ十分高くなるはずである。そこで本システムでは人名らしさを判断するための閾値を導入し、この閾値を越えた評価値を持つ文字列を人名候補として出力する。本稿の実験ではこの閾値の値として -27 を用いた。この値については 6.3 で説明する。

5 実験

5.1 実験条件

表 2 に実験の条件を示す。

表 2: 実験環境

トレーニングコーパス
(日本語人名)

	総計	男性名	女性名
人名数	1,714	1,294	420
文字数	6,749	4,982	1,767
一名前当たりの文字数	3.94	3.85	4.21

テストコーパス
(新聞記事)

テストコーパス	A	B
文数	7,502	86
文字数	339,491	7,238
コーパス中の人名数	344	42

本実験では、登録された人名とそうでないものとの認識状況を比較するため、総数 1,714 という比較的小さな人名コーパスを用いた。また、テストコーパスとして新聞社説約一年分から成るコーパス (A) を利用した。比較評価のため、これとは別に新聞の文化面等の記事 3 編から成るコーパス (B) についても実験を行なった。表 3 に人名抽出実験の結果を示す。表中の $+\alpha$ とは、余分な文字列が付随した形で抽出された事を示す。

テストコーパス A では、343 の人名のうち 256 個の抽出に成功した。これは全体の 74.65% に当たる。抽出された人名のうち 92.58% はその人名全体が欠ける事無く認識されていた。テストコーパス B では、抽出成功率は 88.10% と A よりも高くなっている。ここで示した数値には、人名としても使われるがテストコーパス中では人名以外として出現した例、例えば「千葉」等の地名は含まれていない。

表 3: 実験結果

テストコーパス	A	B
人名	196	21
人名 $+\alpha$	41	6
人名の一部	10	9
人名の一部 $+\alpha$	9	1
計	256	37
	74.65%	88.10%

6 考察

6.1 人名文字間の有繋性を用いた人名抽出

表 3 で示したように、本システムは条件が悪い場合でも 74.65% 、そうでない場合にはそれ以上の精度でコーパス中の人名を抽出可能である事が判る。この精度は、表 6 にあるように、 100% に達する。本稿の実験で利用した人名コーパスは非常に小さいものであるため日本の人名を十分に含んでいるとは言えないが、この人名コーパスを使ってこれだけの正解率を挙げられる事は、固有名詞が特有の文字の並びを有する事、この特有の並びの情報を統計情報のみで取得可能である事を示している。

6.2 名前の多様性

テストコーパス A では、人名の認識率はテストコーパス B に比べて低くなった。この原因の一つに、名前の種類の数の違いがあげられる。

表 4 は、各テストコーパス中に含まれる人名の種類の数を示したものである。テストコーパス A はテストコーパス B よりも大きいため、それに含まれる人名の数も 8.2 倍と多い。しかし、文中に含まれる人名の種類の数は、 B の 2.6 倍に過ぎない。人名数の違いを考えると、 A は明らかに出現

人名が偏っている事が判る。表 4 中の「異なり人物数」とは、実際に文中に現れた人物の人数である。同じ人物が「宮沢」と苗字だけで呼ばれたり「宮沢喜一」と氏名合わせて出現したりする場合があるため、異なり人物数は異なり人名数より少なくなる。

表 4: 各テストコーパス中の異なり人名数

テストコーパス	A	B
人名数	344	42
異なり人名数	82	32
異なり人物数	65	27

テストコーパス A は社説であり、特定の人物が高い頻度で出現する。例えば当時の首相である「海部」は 74 回出現したが、これはこのコーパスの全人名数の 20% を越えている。このようにテストコーパス A で頻繁に出現する「海部」は、本実験で使用した人名コーパスには登録されていない。そのため、これが人名と認識されない例も多く、認識率を下げる大きな要因となった。テストコーパス A では同じ人名が複数回出現する事例が多かったため、人名コーパスに無く認識し辛い人名の出現が認識率低下をもたらした。テストコーパス A での各人名の平均出現回数は 4.2 回であり、各人物の平均出現回数にすると 5.3 回となる。

それに対し、テストコーパス B では各人名の平均出現回数は 1.3 回である。これから、テストコーパス中に珍しい人名が出現しても、全体の認識率に大きな影響を与えなかった事が判る。

6.3 有繋評価値の閾値

表 5 は、実験で算出された有繋評価値を示したものである。

表 5: 有繋評価値

有繋評価値	平均値	最大値	最小値
全文字間	−40.30	1.04	−45.67
人名中の文字間	−15.73	1.04	−26.79

表 5 から、人名を構成する文字間の有繋評価値は平均に比べてかなり高く、その最小値も全文字間の有繋評価値の平均に比べて十分大きい事が判る。本実験での有繋評価値は人名コーパスから得られた共起情報に基づいて算出されたものである。表 5 は、人名を構成する文字の共起関係は一般の語の共起関係と明らかに異なる事、すなわち、人名コーパスから得た文字間統計情報を基にある文字列の人名らしさを評価する事は可能である事を示している。

本システムではこの有繋評価値の差を利用して人名の認識を行なっている。人名中の文字間の有繋評価値の最小値は −26.79 であるので、これを閾値として用い、この閾値を越えた部分を人名として抽出すれば、有繋性の上で人名らしいと判断された文字列を取り出す事が可能である。この評価値は文字間の繋がりを表すものなので、ある部分の評価値が閾値を越えた場合には、その評価値を示した連続した二文字を人名として抽出する。複数の評価値が連続して閾値を越えた場合、その複数の評価値にまたがる文字列全体を一つの人名として抽出する。本実験で用いた人名コーパスは苗字と名前の区切りが無いため、苗字と名前が繋がって一つの人名として抽出される事がある。実際の実験では、 −27 を閾値として設定した。

6.4 人名認識例

表 6 にテストコーパス B を構成する記事の一つである B1 についての実験結果を挙げる。B1 は 1,801 文字から成る 45 文の記事で、13 の人名を含んでいる。人名中に含まれる文字の数は 47 である。

表 6: 実験結果 (記事 B1)

文字列	抽出数	
人名	11	
人名 $+\alpha$	2	13 (100%)

記事 B1 では、文章中の全ての人名が認識された。13 の人名のうち 84.6% に当たる 11 個は人名そのままの形で抽出されている。他の 2 個は余分な一文字が付いた形で抽出された。一つは人名の後に助詞の「は」が付いたものであり、もう一つは人名の後に「夫」が付いた。これは「加藤周一」が「加藤周一夫」になったものだが、この「夫」は「夫妻」の頭の一文字である。「加藤周一夫」という文字列の最後の二文字だけ取り出すと「一夫」となるが、これは人名として使われる文字列であり、実際に本実験で用いた人名コーパスにも名前として含ま

表 7: 「加藤周一夫」中の各文字ペアの共起頻度

	藤	周	一	夫
加	11	0	0	0
藤		0	6	0
周			0	0
一				1

れている。そのためシステムは人名らしい繋がりを要すると判断し、「夫」を付けた形で出力したのである。本システムは隣接 bigram だけでなくその周囲も評価値に影響を与える。この例の場合、「周一夫」という文字列が不自然である事から「一夫」が登録人名であるにも関わらずその評価値は多少低くなっている。これは、「周一夫」という文字列が人名コーパス中に無い事よりも、「周」の字そのものが「加」「藤」「一」「夫」の 4 文字と共起した例が無いため「周」の存在が全体の評価値を下げたとみるべきである (表 7)。表中の頻度は、その「加藤周一夫」という文字列中の距離での共起頻度であり、各文字ペアの bigram 頻度ではない。表 7 から、文字列「加藤周一夫」が人名候補として抽出されたのは主に (“加”,“藤”;1) および (“藤”,“一”;2) の 2 つの d-bigram 情報に依っており、一例登録されていた (“一”,“夫”;1) のために「夫」が繋がってしまった事が判る。

本システムでは文字列の「人名らしさ」についてのチェックのみを行ない、それ以外の文字列については考えない。「夫妻」等人名以外の語の存在を考えて判断する事でこの問題の解決を図る事が可能であると考えられる。

6.5 人名以外の文字列の誤抽出

表 8 に、人名でないのに人名として認識された文字列の数を示す。

表 8: 人名以外の文字列の抽出結果 (記事 B1)

文字列	抽出数	
人名以外		
(文頭)	15	
(文頭以外)	28	43

これらには大きく分けて、地名など他の固有名詞である場合と、文字列中に人名に含まれうる文字列が発見された場合とがある。他の固有名詞である場合については 6.7 節で述べ、ここではそれ以外の、明らかに間違いである場合について検討する。

人名の一部が他の語に含まれる事は珍しくない。例えば、「あい」というひらがな文字列はそれそのもので女性名として使われるが、それ以外の語に含まれる例も多い。また、「かり」や「すか」等も「ばかり」や疑問形の「ですか」等人名以外の文字列に含まれる事が頻繁にあるが、それぞれ「ゆかり」「あすか」等人名中にも含まれる。本稿で提案しているシステムでは有繋評価値の算出に際して d-bigram を用いるため、「か」と「り」等の間の共起情報だけでなくその周りの情報も影響を与える。人名以外の文字列中ではたいていの場合周りの文字列の情報が負の影響として評価され、これらの二文字の間の有繋評価値を下げ、人名として抽出される可能性を低くしている。しかし、その二文字の人名としての出現頻度が非常に高い場合や、周りの文字列が人名である可能性を十分下げられなかった場合などには、人名として抽出されてしまう事がある。

また、表 8 を見ると、人名以外の文字列が人名として抽出された事例は、文頭で多く起こっている事が判る。これは、文頭の文字列では周りとして扱う情報が減るため、評価値を下げにくくなるためであると考えられる。

6.6 人名文字列の欠落

表 9 にテストコーパス B の一部である記事 B2 の実験結果を示す。記事 B2 は 688 文字から成る 16 文のコーパスであり、11 の人名を含んでいる。

表 9: 実験結果 (記事 B2)

文字列	抽出数	
人名	5	
人名 $+\alpha$	1	
人名の一部	3	9 (81.81%)

記事 B2 では 3 個の人名が一部分欠落した状態で抽出された。一つは氏名のうち苗字部分が欠如、一つは氏名のうち名前部分が欠如、そしてもう一つは氏名のうち苗字の最初の一文字が落ちた形で抽出された。これは、これらの人名中の文字が十分な情報を持っていなかったために起こったものである。名

前が落とされた人名では、漢字二文字の名前(「峻也」)の最初の一文字(「峻」)が人名コーパス中に出現しない未知語であった。そのためこの文字の前後で統計情報の空白が生まれ、名前が欠落した。このような場合でも二文字目の「也」が人名中に頻出するためこれと苗字との間の有繋性が全体を繋げる場合が多いのだが、この人名の場合苗字も珍しいものであったため、十分な有繋性を示すことができなかったと推察される。他の二例の場合は、未知語ではないが、未知の繋がりであったため切れてしまっている。「民谷」という苗字では「民」という文字が人名コーパス中で苗字に使われた例が無かったため苗字としての有繋性を持たずにこの一文字のみ欠落してしまっている。また、「赤尾」という苗字では「尾」という文字が苗字の二文字目に来た例が無く、また「赤」という字も人名コーパス中にほとんど出現しなかったため、苗字部分が欠落してしまった。

表 10 は、テストコーパス B 中の記事 B3 の実験結果を示したものである。記事 B3 は 1,044 文字から成る 26 文のコーパスで、 18 の人名を含んでいる。

表 10: 実験結果 (記事 B3)

文字列	抽出数	
人名	5	
人名 $+\alpha$	3	
人名の一部	6	
人名の一部 $+\alpha$	1	15 (83%)
人名以外		
(文頭)	6	
(文頭以外)	11	
(地名)	10	27

記事 B3 では、苗字の一部が欠落した例が 3 例あった。そのうち二つは同じで「笹川」の「笹」が欠落したものであり、もう一つは「佐久間」の「間」が欠落したものである。この 3 例は同じ原因に依る。これらの苗字に出現した各文字は人名コーパス中でも苗字を構成する文字として出現していたが、「笹」と「川」という組み合わせや「佐久」と「間」という組み合わせが無かったのである。これにより、「笹川」では出現頻度も低い「笹」が欠落し、また「佐久間」では「佐久」だけが人名として抽出された。「佐久間」という苗字は、苗字のみで出現した場合には「佐久」の部分のみ抽出されたが、「佐久間久子」と氏名合わせて出現した箇所では「佐久」「間」「久子」それぞれ人名としての有繋性を持っていたため相互に有繋評価値を高め合い「佐久間久子」と完全な形で抽出されている。

表 11 はテストコーパス B 中の一部欠落した人名の数を表にしたものである。

表 11: 一部欠落した人名

		B1	B2	B3
苗字	欠落	0	1	0
	一部欠落	0	1	3
名前	欠落	0	1	2
	一部欠落	0	0	2
	計	0	3	7

本実験で用いた人名コーパスは小さなものであり、その知識は日本の人名を覆うには小さ過ぎるものである。これらの失敗例はこの事に依るものであり、質量共に十分な大きさの人名コーパスを利用する事で十分な認識率を挙げる事が可能である。

ここで、どの程度認識できれば十分か、という問題が出てくる。日本人の人名は、新しい名前の創作が容易であるため、全ての人名を完全に認識する事は非常に難しい。特に、名前は時代に依って傾向が異なり、全ての世代に対応する事は容易ではない。また僧侶名や芸名等一般の名前とは異なった文字の並びを有するものもあるため、その全てを認識する事は困難である。人名の自動認識システムでは、認識対象の範囲をきちんと決める事も認識率の向上において重要な点である。

6.7 地名

表 10 では、人名でないのに人名として抽出された 27 の文字列のうち 10 個が地名またはその一部であった事を示している。そのうちの 6 個は二文字から成る文字列で、「千葉」が 3 回、「市川」が 2 回、「志野」が 1 回抽出されている。但し、「志野」は「習志野」の一部である。これらの 3 種類の文字列はそれぞれ人名として人名コーパスに登録されており、抽出されるのが当然であった[2]。

また、「大和田」という地名も人名として抽出されているが、「大和田」は人名コーパス中には出現

[2]「千葉」「市川」は苗字として、「志野」は女性名として人名コーパスに登録されている。

しない。しかし、「大和」、「大根田」、「和田」という文字列がそれぞれ苗字として人名コーパス中に出現しており、これらから抽出された ("大","和";1)、("和","田";1)、("大","田";2) という d-bigram 情報を利用する事で「大和田」は人名として認識されたと考えられる。このように、d-bigram 情報を用いる事で人名コーパスに無い文字列についてもその人名らしさの評価が可能となっている。n-gram では「大」と「田」のような離れた文字間の関係を利用する事が難しい。

地名や組織名は人名として用いられるものも多く、これらの区別を行なうには文字列の並びの情報だけでなくその前後の文脈や語等の解析が必要である。本稿では文字の並びのみでの固有名詞の認識の可能性について考察するものであり、人名地名等の区別については考えない。本稿の実験結果では、人名の抽出という観点から、人名としてテストコーパス中に出現した文字列についてのみ認識率を示したが、本稿の立場では人名以外の固有名詞でも人名として利用されうるものについては認識されるべきであり、これを正解率に含めると本システムの認識率はさらに高くなる。

7 結論

本稿では文字間の有繋性を基にコーパス中の人名を抽出する手法を提案し、その実験結果について考察を行なった。本稿で提案したシステムは、人名コーパスから得た人名特有の文字間共起情報を利用して入力コーパス中の各文字間の有繋性を計算するものである。実験の結果、小さな人名コーパスで得た統計情報を利用した場合でも、コーパス中に出現する人名の種類に依って 100% から 75% 程度の人名の抽出に成功した。抽出された人名には、人名コーパスに登録されたものだけでなく、人名コーパス中の d-bigram 統計データから推測された未登録の人名も多く含まれた。本稿の実験結果は、人名が特有の文字間共起関係を有する事を示し、またこの特有の文字間共起関係のみから人名らしい文字列の自動抽出が可能である事を示した。

本システムでは人名らしい繋がりを有するかを調べるだけのシンプルなものであったため、人名でないものを抽出してしまう事象を抑える事ができなかった。人名のみを抽出するためには、人名らしさの推測と同時に、人名でない可能性の推測が必要になる。統計情報のみに依る自動人名抽出システムの確立は、さまざまな自然言語処理アプリケーションの前処理として有用である。

本稿では人名を抽出対象としたが、トレーニングデータを変えることで、同じシステムを他の文字列の抽出にも利用可能である。

参考文献

[1] Shiho Nobesawa, Junya Tsutsumi, Da Jiang Sun, Tomohisa Sano, Kengo Sato, and Masakazu Nakanishi. Segmenting Sentences into Linky Strings Using D-bigram Statistics. *Coling-96*, 1996.

[2] 堤 純也, 新田 朋晃, 小野 孝太郎, 延澤 志保. 統計情報を用いた多言語間機械翻訳システム. **人工知能学会研究会**, pages 7–12, 1993.

[3] Virginia Teller and Eleanor O. Batchelder. A Probabilisitic Algorithm for Segmenting Non-Kanji Japanese Strings. *AAAI*, 1994.

[4] 久光 徹, 丹羽 芳樹. 辞書と共起情報を用いた新聞記事からの人名獲得. **情報処理学会 自然言語処理研究会報告** *NL-97-118 No.1*, pages 1–6, 3 1997.

[5] Pascale Fung, Min yen Kan, and Yurie Horita. Extracting japanese domain and technical terms is relatively easy. pages 148–159, 1996.

[6] Pascale Fung. Extracting key terms from chinese and japanese texts. 1998.

[7] 森 大毅, 阿曽 弘具, 牧野 正三. 再現性を考慮した文字列に基づく統計的言語モデル. **信学技報** *NLC97-47, SP97-80*, pages 29–34, 1997.