文字間統計情報に基づく口語文字列の自動抽出

延澤 志保[†] 斎藤 博昭[†] 中西 正和[†]

統計情報に基づく自然言語処理が盛んになる中で、訓練データとしてのコーパスの影響は非常に大きい。生コーパスをそのまま利用する場合には、コーパスの取得が容易であるため、目的に合ったドメインのコーパスを大量に入手できるという利点がある。しかし、生コーパスは人間の言語の性質上、未登録語や未知の言い回し、非文とされるような文の出現等を多く含むことがほとんどであり、これらが処理の精度の低下を招くという問題がある。特に、口語表現の処理は、電子メールでの利用等利用頻度の高いものであるにも関わらず、十分に研究されているとは言い難い。本稿では、生コーパスに含まれる未知の語句および言い回しに着目し、電子メール文書内に出現する意味のある文字列を自動的に抽出する実験を行なった結果について報告する。本システムは事前に与えられた電子メール文書中の各文字の共起確率を利用して、テストコーパスとして与えられた電子メール文書から意味のある文字列を抽出し出力する。本システムを利用することで、同じテストコーパスを既存の形態素解析ツールで解析した結果未登録語として処理された文字列の 69.06% を抽出することに成功した。

キーワード: 文字間共起情報, 文字列自動抽出, 口語表現

Automatic Extraction of Oral Expressions Based on Letter Cooccurrence Statistics

SHIHO NOBESAWA[†], HIROAKI SAITO[†] and MASAKAZU NAKANISHI[†]

Researches based on statistical information have been more significant in the field of natural language processing. The use of raw corpora is fascinating, as it is easy to obtain a certain amount of non-tagged texts. However raw corpora often contain unknown words and phrases, and this causes low accuracy of the experiments. Colloquialism has not been worked enough because of this problem, though the processing of colloquialism is strongly required for the emails and other tasks. In this paper we propose a simple method to obtain domain-specific sequences from unrestricted texts using statistical information only. Our method needs a non-tagged training corpus. We use the statistical information drawn from the training corpus to extract semantic character sequences automatically. We had experiments on sequence extraction on email texts, and succeeded in extracting significant semantic sequences in the test corpus. The sequences our system salvaged contain casual terms, proper nouns, and sequences with representation change such as pronunciation extension.

KeyWords: *Statistical information between letters, automatic string extraction*

[†] 慶應義塾大学 大学院理工学研究科, Department of Computer Science, Keio University

1 はじめに

統計情報に基づく自然言語処理では、訓練データとしてのコーパスの影響は非常に大きい。形態素情報や品詞情報等の情報を付加したコーパスを利用することで処理の精度の向上や処理の簡略化等が期待できるが、情報を付加する段階での労力が大きく、その精度に結果が大きく左右されるという問題がある。

生コーパスをそのまま利用する場合には、コーパスの取得が容易であるため、目的に合ったドメインのコーパスを大量に入手できるという利点がある。しかし、生コーパスは未登録語や未知の言い回し、非文とされるような文の出現等を多く含むことがほとんどであり、これらが処理の精度の低下を招くという問題がある。コーパスから得た情報を利用するようなシステムの場合、処理の基本は意味のある言語単位であるから、まずこれを正しく認識することが先の処理の精度の向上に必要である。日本語のように意味のある言語単位ごとの区切り目が明らかでない言語では、まずこれを認識することが処理の第一段階であると言っても過言ではない。そこで、本稿では、生コーパス中の意味のある文字列を推測し認識することで結果的にコーパス中の未登録語を推定するシステムを提案する。

本システムは、対象となるドメインの訓練用コーパスから取得した文字間共起情報を利用して、入力コーパス中の意味のある文字列を認識しこれを出力する。訓練用コーパス、テストコーパスともに事前のタグ付けは必要としない。

2 日本語における語句抽出

2.1 自然言語処理における処理単位

自然言語処理の処理単位としては、単語を基本とするものが一般的である。しかし、単語は多義性を持つものも多く、文脈の中では一意に意味を決められる場合でも単語ごとに分割した時点でその語の持つ意味を特定できなくなる場合があり、単語が最適な処理単位と言えるかには疑問が残る (Fung 1998)。処理の単位として意味的な塊としての語句をとった場合、このような多義性の問題等はある程度抑えることができる。従って、意味的な塊としての語句の認識は、自然言語処理の精度の向上の点で、重要な課題である。

現在研究されている語句抽出システムは、ほとんどが名詞句を対象としたものである。構文情報を基に名詞句を推定する方法および大規模コーパスからのドメイン固有の語句の抽出とが主に研究されている。Argamon らは、サブパターンの概念を利用して名詞句をパターンとして認識する、記憶ベースの手法を提案している (Argamon, Dagan, and Krymowski 1998)。また、Ananiadou は語句の組成についての形態論的ルールを利用して語句の認識を行なう手法を提案している (Ananiadou 1994)。

2.2 日本語における語句抽出

英語のように単語間に区切りを置く言語では、語句の認識は、区切りによって分けられた語を連結する作業となる。それに対し、日本語は単語間に区切りを置かない言語であるため、日本語においては、単語の抽出プロセスは、文の切り分け作業となる。

日本語は日常的に利用される文字が数千文字と非常に多い。また、数種類の字種を同時に利用するという点でも、日本語は特徴的である。

文字の共起情報を利用する手法としては、 n -gram の画期的な抽出法を提案した長尾らによるもの (Nagao and Mori 1994) や Oda らの n -gram を利用した手法 (Oda and Kita 1999) が挙げられる。

また、Kashioka らは、文字の相互情報量を利用した文字クラスタリングシステムを提案し、これを利用した切り分け処理およびタグ付け処理を行なっている (Kashioka, Kawata, Kinjo, Finch, and Black 1998)。Kashioka らの手法では処理に必要な情報を得るため、事前に訓練コーパスの単語への切り分けが必要とされる。

単語の共起情報を利用する手法として、Borthwick は、最大エントロピ法に基づく固有名詞抽出の手法を提案している (Borthwick 1999)。Borthwick の手法ではあらかじめ Juman を用いて切り分けおよびタグ付けを行なっている。

3 システム概要

3.1 本システムの概要

本システムは入力として日本語文一文を採り、訓練コーパスから事前に抽出した文字共起情報を基に文中に含まれる文字列を調べ、意味のある文字列と認めたものを出力する。以下にシステムの処理の流れを示す (図 1)。本システムは入力されたテストコーパスを一文単位で処理する。入力文中の隣り合う文字ペアそれぞれについて、訓練コーパスから抽出された文字単位共起情報を基に、その二文字の繋がりのやすさを示す有繋評価値を算出し、これに基づいてこの二文字が同じ文字列に含まれるものであるかを推測する。この結果一繋がりと判断された文字列について、その有繋評価値が十分に高いものを統計情報に基づいて意味がある文字列と判断し、これを抽出結果として出力する。

3.2 コーパス

本システムの処理単位は文である。従って、訓練コーパスから文字単位共起情報を取得する際も、余分な情報を取得しないよう文単位で処理を行なう。

文末は、原則として、文末記号によって決める。文末記号としては「.」「!」「?」の他、「……」や「♪」、「★」等文末を示すと認められるものはすべて認める。また、複数の文末記号

図 1 システムの処理の流れ

が続く場合、その最後の文末記号までを一文とする。文末記号が現れても、それが引用中である等、文の途中であると認められるような場合にはそこで文を切らない。

e.g. 「こんにちは。」と言った。

また、文末記号がない場合でも、明らかに文の区切り目であると考えられる場合は、文末記号を置かないまま文末とした。

3.3 文字間統計情報

隣接 bigram だけでは、 $AXYB$ に現れる事象 XY と $CXYD$ に現れる XY とを区別することはできない。日本語のように形態素間の区切り目が明らかでない言語では、単純隣接 bigram を利用することで捨てられるこれらの文脈情報が重要な役割を果たすことがある。

本稿では、共起情報を取得する際の確率モデルとして d-bigram を採用した。d-bigram とは、事象間の距離を考慮した bigram モデルである (堤, 新田, 小野, 延澤 1993)。

d-bigram 確率モデル

d-bigram モデルは、2 つの事象および事象間の距離の 3 つのパラメータから成っている。

2 事象間の距離は、この 2 事象が隣接して並んでいる時 1 とする。例えば、 $ABBC$ という事象列からは 6 種類の d-bigram が取得できる。距離を考慮しない通常の bigram モデルでは (A, B) が 2 回カウントされるが、d-bigram モデルの場合、隣り合って出現している $(A, B; 1)$ と一事象間に置いて出現している $(A, B; 2)$ は別の事象として扱われる。

有繋評価値

Nobesawa らは文中のある文字の並びが意味のある塊を成すことの起こりやすさを算出するシステムを提案している (Nobesawa, Tsutsumi, Sun, Sano, Sato, and Nakanishi 1996). この手法は語彙についての知識を一切必要とせず, 文字間共起情報のみで文を意味のある塊に切り分けることが可能であることを示した. 実験は日本語について行なわれ, 日本語の各文字の共起関係の情報は文中での文字の繋がり方を示すのに十分な情報を持っていることを示した.

本稿ではこの点に着目し, 文字間の共起関係を利用して意味のある文字列を推測し自動抽出するシステムを提案することで, 辞書等の語彙を非常に小さい労力で補う手法を提案する. 本システムは, 有繋評価値と呼ばれる評価値 (Nobesawa et al. 1996) を導入する. 有繋評価値とは, 隣り合った二文字が一塊の文字列に属する事象の起こりやすさを示す値であり, この値が高いほど, 対象となっている隣接二文字ペアが同じ文字列に属する可能性が高い. 有繋評価値は統計情報のみに基づいて算出する値である. Nobesawa らは有繋評価値の計算に d-bigram を利用することで文脈情報を影響させている.

文中の i 番目の文字と $i+1$ 番目の文字の間の有繋評価値の算出式を式 (1) に示す. ただし, w_i はある事象 (本システムでは, 文 w の i 番目の文字), d は 2 事象間の距離 (d-bigram の定義による距離), d_{max} は有繋評価値の算出に利用される d-bigram の距離 d の最大値 (本稿で紹介する実験では $d_{max} = 5$ とした), $g(d)$ は距離の影響に対する重み付け関数¹とする.

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

また, 2 事象間の相互情報量の計算式を d-bigram に対応するよう拡張したものとして式 (2) を利用した. ただし, x, y は各事象, d は 2 事象間の距離, $P(x)$ は事象 x が起こる確率, $P(x, y; d)$ は d-bigram $(x, y; d)$ が起こる確率とする.

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

図 2 に文脈情報の影響のイメージを示す. ある隣接 2 文字 w_i と w_{i+1} の間の有繋評価値の算出には, この 2 文字ペアの共起情報だけでなく, その周りに現れる文字ペアの共起情報 (例えば $(w_{i-1}, w_{i+2}; 3)$ 等) が影響する.

3.4 文字列抽出

本稿で用いるシステムは文字間共起情報を利用して算出する有繋評価値を基に抽出すべき文字列の選択を行なう. 図 3 に有繋評価値のグラフの例を挙げる. 「ABCDEFGHIJK!」という

¹ 距離が遠いほど事象の影響が小さくなるとする主張 (Church and Hanks 1989) を実装したものであり, 本稿で紹介する実験では $g(d) = d^{-2}$ (佐野, 堤, 孫, 延澤, 佐藤, 大森, 中西 1996) としている.

図 2 d-bigram を用いた有繋評価値の算出 (Nobesawa et al. 1996)

12 文字から成る文字列を入力文としたとし、グラフ中の X 軸上のアルファベットはそれぞれ文中の各文字を示す。末尾の「!」は文末記号を表す。

図 3 有繋評価値スコアグラフ (Nobesawa et al. 1996)

有繋評価値は隣り合う文字ペアそれぞれについて算出される。グラフ中では、各文字ペア間の有繋評価値を Y 軸で表している。有繋評価値は文字間共起情報を基にしており、共起する確率が高い場合ほど値が高くなる。スコアグラフが山状になっている部分は文字間の繋がり強い文字の並びであり、この部分は一塊の意味を成すとみなすことが可能である。有繋評価値の算出には該当文字ペアだけでなくその周りの文字との共起情報も利用されるため、長い文字列では、その文字列に含まれる文字それぞれの共起関係の相乗効果により、文字列の有繋評価値が高くなる傾向がある。偶然隣り合って並んだ文字ペア²の場合、有繋評価値は相対的に低くなり、スコアグラフ上では谷を成す。

本システムでは、スコアグラフで言うところの山状の部分に相当する文字列を有繋文字列 (有繋評価値に基づいて一塊と判断された文字列 (Nobesawa et al. 1996)) として抽出する。図 3 の例では、AB, CDEF および HIJK がそれぞれ一塊の文字列として出力される。

この手法では、どこまでを山とするかの基準が必要になる。抽出に際しての基準として、有

² 対象が言語であるため、「偶然」という表現は実際には適切でない。ここでは、必ずしも隣り合う可能性が高いとは言えない 2 単語の境界にある 2 文字ペアを指して偶然の並びと言っている。

繋評価値に閾値を設けこの値を越えたものを山とみなす方法と、山の部分の勾配について閾値を設けて前後の文字列との区切り目が明らかなものを山とみなす方法が挙げられる。この閾値を操作することで、抽出する文字列の種類や精度をある程度調節することが可能である。確実に一塊となる文字列のみを抽出したい場合には、抽出対象を選出する際の基準を高く設定すればよい。閾値が高い場合、有繋文字列として出力される文字列の長さが短くなる傾向がある。これは、例えば複合語等が単語に切り分かれる³等、確実に一塊となる部分のみを残そうとする作用が強くなるためである。

4 実験

本稿では、与えられた訓練コーパスから得た文字間共起情報のみを利用した文字列抽出実験の結果を報告する。本稿では、口語文を実験対象とし、これに含まれる頻出文字列を抽出することで、口語文に多く含まれその処理を困難にする要因となっている辞書未登録語の自動認識および自動抽出を行なうことを目的とする。

本稿で報告する実験では、その結果を評価するため、同じ入力コーパスを日本語形態素解析ツール茶筌 ver. 1.51(松本, 北内, 山下, 平野, 今, 今村 1997) で処理した結果と比較し、茶筌が未登録語とした文字列および茶筌が解析に失敗した文字列について、これを本システムで抽出できたかについて調査する。

4.1 コーパス

本実験では、コーパスとして電子メール文書を利用した。口語の記述文としては対話コーパス等他にも利用可能な文書が存在するが、これらはほとんどの場合話し言葉を他者が記述したものであり、その正確性や発話者の意図の反映等の点では電子メール文書が優ると考える。また、電子メール文書の場合、特に口語で記述されるものは一般の文書に比べて未登録語が多いためかな漢字変換等の失敗が多く、他の文書を対象にしたものに比べてユーザの不満が多いと推測される。従って、一般の辞書では未登録語とされるような頻出文字列が自動的にユーザ辞書等の形で組み込まれるシステムへのニーズは大きい。

さらに、電子メール文書はそのまま機械処理可能な状態で存在しかつある程度の量を収集することが少なくとも物理的には簡単である⁴ことも、電子メール文書を実験コーパスとして選択したことの大きな理由に挙げられる。統計情報を利用した処理システムの場合、十分な量のコーパスの確保が問題になることを考えると、ほとんど労力を掛けずに大量のコーパスを取得できる電子メールは非常に有用である。

³ 複合語は、一般的に、複数の山が連なって一つの大きな山を構成するような形状となる。閾値が高くなると、それぞれの山(大抵の場合、単語)の切れ目の谷の部分で切り分けられる場合が増え、結果的に複合語を構成する単語がそれぞれに抽出される。

⁴ 電子メールの利用はプライバシーの問題等がある。

なお、電子メールでは他のメールの引用が含まれることが多いが、同じ文を複数回記録することを避けるため、本実験で用いたコーパスでは引用部分はすべて削除した。

訓練コーパス

有繋評価値の算出に利用する共起情報を得るための訓練コーパスとして、本実験では、1998年から1999年にかけて友人宛てに書かれた口語調の電子メールを利用した。送信者は17人(全員10代から30代の女性)で、送られたメールはすべて同一の受信者(女性)に宛てたものである。本コーパスは351の電子メール中の7,865文から成っており、含まれる文字数は176,380(一文当たりの平均文字数22.4)である。

テストコーパス

テストコーパスは、1999年に友人宛てに書かれた口語調のメール文書を利用した。送信者は3人(訓練コーパスの送信者の一部)であり、同一の受信者(訓練コーパスの受信者と同じ)に宛ててメールを書いている。テストコーパスは訓練コーパスの一部ではなく、独立したものである。テストコーパスは1,118文24,160文字から成っており、一文当たりの平均文字数は21.6である。

4.2 実験結果

有繋評価値の分布

本実験における有繋評価値の分布を図4に示す。有繋評価値の平均値は0.34であった。この値を受けて、文字列抽出実験では、抽出の閾値を0.50に設定した。

文字列抽出結果

本システムが抽出した文字列を分類すると、表1のようになった。ここで、「抽出成功」とは単語や熟語のような文字列が過不足なく抽出された数、「過接合」とは複数の単語等が一つの文字列として抽出されたものおよび意味のある文字列にその前後の文字列の要素である文字が付着した形で抽出されたもの、「過分割」とは文字列が途中で分割されたもののうち元の文字列が推測できるもの、「抽出失敗」とは過分割・過接合のため意味をなさなくなったものを指す。文字列の例のうち、過分割、抽出失敗として挙げたものは、抜け落ちた部分を括弧内に示した。

定義にあるように、過接合とは複数の文字列が接合した形であり、頻出文字列の抽出を行う本システムの手法では多く抽出されるものである。過接合文字列は、複数の文字列が繋がる場合と、文字列の前後に他の文字列が付着する場合とに大きく二つに分類できる。

- 複数の文字列の有意味な繋がり

図 4 有繋評価値の分布

表 1 本システムの抽出文字列

	文字列数	割合	文字列の例
抽出成功	1920	42.76%	「やっぱり」「北海道」「html 形式」「24 時間」「トモダチ」
過接合	1876	41.78%	「メールを」「にお手紙書く」「いいねー!」「私は」
過分割	564	12.56%	「メーラ (ー)」「(O)utlook」「ゴールデンウイ (ーク)」
抽出失敗	130	2.90%	「を遣 (わせる)」「が起こ (った)」「(つか) めてないっ(す)」
計	4490		

- 頻出文
- 頻出言い回し
- 文字列の前後に他の文字列の全部または一部分が付着したもの
 - 助詞、文末記号の付着
 - 前後の文字列が過分割され付着

これらのうち、頻出文、頻出言い回しが一塊として抽出されることは、本システムの手法を考えると自然であり、失敗とは言えない。また、文字列の前後に助詞や文末記号が付着することは、例えば「と思う」の助詞「と」の付着のように、その文字列の主たる利用法によるものであり、頻出言い回しの一種と考えることができる。

表 2 に、過接合とされた文字列の内訳を示す。文の形をしたもの等意味のある文字列が過接合とされたのは 531 例であり、これらを含めた有意味文字列の抽出文字列全体に対する割合は 54.59% となる。この割合は、本システムの適合率と考えてよい。さらに、文末記号・助詞付着

表 2 抽出文字列中の過接合文字列

	文字列数	割合	文字列の例
文	339	18.08%	「今年もよろしくね。」「あははははは。」 「メール遅くなってごめんね。」
有意味文字列	192	10.24%	「関西方面」「新バージョン」「いい人」「ちょっと不安」 「でもやっぱり」「メール送らないで」
助詞付着 & 記号付着	198	10.56%	「で大変です。」「といます。」
助詞のみ付着	337	17.97%	「それは」「と一緒に」「メールを」
記号のみ付着	523	27.89%	「だけど,」「だよ.」「です.」
その他	287	15.30%	「って感じ」「かったらハッキリ」
計	1876		

を原因とする過接合をも加えると、その割合は 84.70% に上がる。

表 1 にある「過分割」は、そのほとんどが動詞である。これは、動詞についての判定を、助動詞等まで含めた一塊を抽出した場合にのみ抽出成功としたためである。本稿では、複合動詞「書き直す」が二つの動詞に切り分けられた例や「言っちゃった」のように活用し助動詞等が付着した文字列で「っちゃった」等付着部分が切り分けられた例等もすべて過分割に分類している。

頻出文字列

表 3 に、高い有繋評価値を示した隣接文字ペアの一部を示す。ここに挙げられた隣接文字ペアは有繋評価値の値が 10.00 を越えたものであり、表中の数字はそのペアが評価値 10.00 以上で出現した頻度を示す。表 3 に挙げたような特に評価値の高い隣接文字ペアは、評価値の高い

表 3 有繋評価値の高い隣接文字ペア

文字ペア	頻度	文字ペア	頻度
ワタ	19	0 0	7
ット	12	友達	6
ネッ	11	返事	4
タシ	11	登録	4
ポスト	10	絶対	4
遊び	9	示板	4
関西	8	揭示	4
誕生	7	H P	4

文字列の一部であることが多い(表 4)。

表 4 に本システムが抽出した文字列の一部を示す。ここでは、文字列に含まれるすべての隣接文字間の評価値が 5.00 以上の文字列のみを示し、この場合の頻度のみを掲載した。

4.3 茶釜における未登録語

未登録語として扱われる文字列のカテゴリとしては以下のものが挙げられる。

表 4 共起関係の強い文字列

文字列	頻度	文字列	頻度	文字列	頻度
…	72	(笑)	36	ネット	20
そう	52	ワタシ	29	!!	16
けど	48	それ	26	リンク	15
から	43	…	25	友達	13
メール	39	自分	20		

- 新しい語句
 - － 新語
 - － 俗語
 - － 固有名詞
- 既存の語句の異表記
 - － カタカナ表記 (「私」を「ワタシ」と表記する等)
 - － アルファベット表記 (「グッド」を「GOOD」と表記する等)
 - － 発音変化に伴う表記変化
- 擬音語, 擬態語等
- 解析エラー

比較に用いた日本語形態素解析ツール茶釜は辞書ベースのシステムであり, 未登録語は「未定義語」のタグを付加して出力される. 表 5 は, テストコーパスを茶釜にかけた結果未定義語タグを付加されて出力された文字列を出現頻度毎にまとめたものである. 本稿で利用するシステムでは品詞タグを利用しないため, 茶釜の品詞タグ付けエラー (文字列の切り出し方は正しいが付加された品詞タグが不適切なもの) については無視し, 未定義語タグが付加された文字列および切り分けに失敗した文字列のみを茶釜における抽出失敗とみなす. 表 5 では, 茶釜が未定義語タグを付加して出力した文字列の本システムでの抽出状況についても示している. 「抽出成功」とは本システムが抽出に成功した文字列の数, 「抽出失敗」は本システムが認識できなかった文字列の数, 「部分抽出」とした欄は文字列の一部が認識されなかった文字列の数を示す.

茶釜は, テストコーパスの形態素解析処理の結果 627 の未定義語タグ付き文字列を出力した. テストコーパスは 1,118 文から成っているため, 単純に平均すると 56.08% の文に未登録語が含まれていたことになる. 表 5 に示すように, 本稿で提案する手法により, 茶釜が未定義語タグを付加した文字列の 69.06% を回復することが可能である. この割合は, 本システムの再現率に相当する値である.

本システムは, 頻度の高い文字列についてさらにより結果を示す (表 5). 茶釜が未定義語タグを付加した文字列のうち, テストコーパスでの出現頻度が 10 回を越す文字列では, 81.85% が正しく抽出されている. 本実験で用いたテストコーパスは訓練コーパスの一部ではないが, 条件に近いものであるため, テストコーパスでの出現頻度は訓練コーパスとある程度似ている

表 5 茶筌が未定義語タグを付加した文字列の扱い

未定義語タグ 出現頻度	総出現数	本システムでの抽出結果		
		抽出成功	部分抽出	抽出失敗
10 以上	281	230	7	44
3 ~ 9	143	100	13	30
2	56	43	4	9
1	147	60	44	43
計	627	433	68	126
		69.06%	10.85%	20.10%

ものと考えることができる。テストコーパス中に二度以上出現した未登録語に限った場合、本システムでの抽出成功の割合は 77.71% となる。テストコーパスに一度しか現れていない未登録語も、本システムを用いることで 40.82% を抽出することができた。

本システムが抽出した意味のある文字列中、茶筌が未定義語タグを付加した文字列の割合は 11.22% であった。抽出の閾値を上げると、出現頻度の低い文字列の抽出に失敗する可能性が高くなるが、この割合の値は大きくなる。茶筌が未定義語タグを付加して出力した文字列のうち、本システムが抽出に成功したものの有繋評価値の平均は 6.11 であり、実験の閾値 0.50 を大きく上回る。有繋評価値の分布 (図 4) から見ても、この数値は十分高いものである。

表 6 は、茶筌が未登録語とした文字列を分類しそれぞれについての本システムでの抽出結果を示したものである。表 6 が示すように、未登録語の最大の原因は表記の変化である。口語調

表 6 茶筌が未定義語タグを付加した文字列の分類

未定義語タグ カテゴリ	本システムでの抽出結果			
	計	抽出成功	一部抽出	抽出失敗
固有名詞	60	39	17	4
新語 (固有名詞を除く)	70	48	12	10
文字挿入	119	89	4	26
表記変化	276	194	28	54
文末記号	58	43	0	15
スマイリー	15	9	6	0
その他	29	12	1	16
計	627	433	68	126

の文章では、文字列を強調する等の場合にその表記方法を変えることがあり、これによって、辞書に載っている語でも見出し語と異なる表記のために辞書とマッチせず未登録語とされる。文字挿入も、強調等の目的で行なわれることが多く、その意味では表記変化の一部と見ることができる。この 2 種類を足し合わせると、茶筌が未登録語として出力した 627 文字列のうち 356 個、56.78% が表記の変化によるものということになる。

文末記号が一部未登録語とされたのは、(1) 記号が文末記号の代わりに用いられた事例と (2) 複数の文末記号が並んで使われた事例の 2 種類の場合であった。文末記号として頻繁に利

用された記号には「♪」や「…」等がある．本システムは辞書を利用せず訓練コーパスから得た共起情報のみを用いるため，これらの記号も他の通常の文末記号と同様に文末に来る確率が高い文字列として抽出している．また(2)では「!!」や「??」等，同じ記号を並べて利用することが多い．並ぶ個数には規則はないが，記号同士の並び方(「!?」や「…!」は頻繁に起こるが「!.」は起こらない，等)にはある程度規則があり，d-bigram 共起情報を利用することで習得可能である．

表 6 のカテゴリ中未登録語とされるべきものは新語および固有名詞だが，これは未登録語全体の 20.73% であった．本実験ではこれらのうち 66.92% を正しく抽出することに成功した．

未登録語に含まれる字種

表 7 に，茶筌が「未登録語」とした文字列に含まれる文字の字種を示す．数字は各字種の文字の数であり，文字種とあるのはその字種に該当する文字の異なり数である．茶筌が未登録語として出力した 627 語(表 5)は，合計 1,493 の文字から成っており，平均文字数は 2.38 であった．未定義語タグを付加された文字列に含まれる文字の 70.40% はカタカナであった．

表 7 茶筌が未定義語タグを付加した文字の字種

字種	未定義語タグ		本システムでの抽出結果		
	文字種	計	抽出成功	一部抽出	抽出失敗
漢字	1	19	19	0	0
ひらがな	12	200	155	7	38
カタカナ	73	1051	712	188	151
アラビア数字	1	1	0	1	0
アルファベット	23	122	43	72	7
記号	22	100	39	37	24
	計	1493	968	305	220
			64.84%	20.43%	14.74%

漢字および数字については，それぞれ 1 つずつしか未登録語とされていない．数字は一文字一文字を独立した数字として扱うことが可能であり，このため，未登録語となる可能性が非常に低い．また，漢字は表意文字であり，一文字でもなんらかの意味を持つ．そのため，誤分割によって一文字だけ独立して切り分けられた場合でも，その一文字で一つの文字列と扱われることがあり，その結果，未登録語として出力される可能性が低くなっている．例えば，「この世界」を「この世」と「界」とに切り分けた例があったが，この場合，「界」は「文学界」等に見られるような名詞接尾語とされていた．このように，漢字文字列の未登録語は，未登録語とされずに無理な切り分けをされ，タグ付けの誤りを引き起こす原因となっている．

異表記に起因する未登録語

表 8 に、茶釜が未登録語としたもののうち、表記変化が原因となっているものを示す。ほとんどの辞書は、基本的に、それぞれの語について見出し語を一つしか持たないため、例えば同じ語をカタカナで表記された場合これを同じと判定することは困難である。しかし、口語表現等では語の強調や表記の簡素化等のため、基本的な表記を用いずかなで表記することも多い。

表 8 表記変化に起因する未登録語

未定義語タグ サブカテゴリ	本システムでの抽出結果			
	計	抽出成功	一部抽出	抽出失敗
語形変化	40	33	3	4
カタカナ表記	137	102	12	23
語形変化 & カタカナ表記	55	34	10	11
その他	44	25	3	16
計	276	194	28	54

発音延長に起因する未登録語

日本語では「ん」以外の音がすべて母音を伴うため、ある音を伸ばす場合にはその音の含む母音をさらに付加する。表 9 に、発音の延長のために添付された文字が未登録語となった 119 例について、本システムによるその抽出結果を示す。未登録語とされた添付文字は 7 種類の小字および「ン」の 8 種類である。表 9 の「抽出失敗」は、添付文字は文字列に付加されたが、文字列自体の抽出に失敗したものを示す。

表 9 添付文字が未登録語とされた例

文字	あ	い	う	え	お	っ	ッ	ン	計
付加成功	39	2	5	32	7	3	1	0	89
抽出失敗	0	0	0	0	0	4	0	0	4
付加失敗	5	1	4	2	1	7	5	1	26
計	44	3	9	34	8	14	6	1	119

表 9 に示すように、本システムを利用することで、茶釜で未登録語とされた添付文字の 74.79% が元の文字列に添付された形で抽出された。これにより、添付文字に起因する未登録語のおよそ 75% が、未登録語としてでなく、文字列の一部として抽出することが可能になった。

この場合、抽出される文字列は添付文字を付けた形であり辞書の見出し語と異なるが、これは見出し語が変形したものであり、添付文字まで含めて一塊の文字列であることは否めない。従って、語形変化した文字列と見出し語とを結びつけることができれば、語形変化した文字列を辞書に登録することが可能である。未登録語とされた文字と元の文字列との関係が明らかに

なり元の文字列の語形変化であることが示されれば、元の文字列の属性を継承することで語形変化した文字列に十分な情報を与えることが可能である。例えば、「すごく」を強調するため添付文字「っ」を挿入して「すっこく」とした場合、茶筌のような辞書ベースのツールでは、「す」「っ」「ごく」等分割されて出力される。この時、挿入された「っ」は未登録語とされる。本システムでは辞書とは関係なく文字単位の共起情報を基に文字列を抽出するため、頻出語である「すっこく」は一塊の文字列として抽出される。ここで、「っ」が茶筌では未登録語とされていること、この「っ」を抜かした「すごく」が辞書に登録されていること、周りの語句との共起情報等から、「すっこく」が「すごく」の語形変化であることを推測することが可能となる。

4.4 茶筌の解析誤り

表 10 に、茶筌が解析に失敗した文字列の抽出結果を示す。本システムは茶筌が解析に失敗した文字列のうち 70.88% の認識に成功した。カテゴリ B は数字を含む文字列であり、このカ

表 10 茶筌による解析失敗

解析失敗		本システムでの抽出結果		
カテゴリ	計	抽出成功	一部抽出	抽出失敗
A 英文字を含むもの	42	41	1	0
B 数字を含むもの	60	35	10	15
C 固有名詞	92	81	5	6
D 新語 (固有名詞以外)	11	2	5	4
E 言い回し	8	4	3	1
F 表記変化	176	106	37	33
G 字種変化	19	10	5	4
H 強調表現	257	154	73	30
I 文末記号	253	233	6	14
J 解析誤り	115	82	19	14
計	941	667	159	115
		70.88%	16.90%	12.22%

テゴリに含まれる文字列はすべて数字に助数詞が付加された形になっている。本システムは 25 の文字列の抽出に失敗しているが、「30日」が「30」と「日」に切り分けられたもの等、25 のうち 12 は助数詞が切り分けられてしまったものである。

カテゴリ F と G と H が、文字列の語形変化に起因する誤りである。

カテゴリ F は、口語での利用等のための発音変化に起因する表記変化のため茶筌が認識に失敗した文字列である。「やはり」の口語化である「やっぱ」等がこのカテゴリに入る。

カテゴリ G は強調等のため他の字種で書かれた文字列である。

カテゴリ J は解析誤りによる失敗である。本システムでは解析誤りのため茶筌が切り分けに失敗した文字列のうち 71.30% の認識に成功した。

カテゴリ H で、本システムで抽出に失敗した 103 の文字列のうちの 53 個は、長音化のための附属文字が欠落したものであり、抽出された文字列は意味的に正しいものであった。

口語等で強調のために添付する文字は、複数個になることもある。例えば、「まさか」を強調するために「まさかー」等と長音記号を複数添付することも可能だが、この場合、添付する文字の個数には特に制限がない。本システムでは、例えば長音記号は複数回並ぶ可能性がある、という情報を d-bigram の形で保有しているため、このような複数個の文字の添付に対応可能である。これは添付文字に限ったことではなく、例えば笑い声を示す「ははははは」や文末の「!!!」等の抽出も可能である。

4.5 茶釜への語句登録

辞書ベースのツールでは一般に語句の登録を許しているが、その登録は主に人手によるものである。本システムは、辞書ベースの解析ツールのための辞書作成支援システムとして利用することが可能である。本手法は文字間の統計情報のみを利用して自動的に文字列の抽出を行なうため過分割・過接合の問題があるが、文字列抽出の閾値を上げることでかなり抽出文字列を絞ることが可能である。

例えば茶釜では、新規ファイルを用意し図 5 のフォーマットで各語句を記述することで、語句の登録を行なうことができる(松本他 1997)。茶釜の辞書の見出し語に付加されている数値は

(品詞 (名詞 固有名詞 人名 一般))((見出し語 (竹取の翁 2000)))(読み タケトリノオキナ))

図 5 茶釜の辞書項目の記述

単語コストである。この値の決定には、本システムが文字列抽出に際して計算した有繋評価値を加工して利用することが可能である。

本システムでは読みがなや品詞の決定はできないが、読みがな付加に関しては KAKASI⁵等のツールを利用も可能である。未登録語とされる文字列のうち、カタカナ表記に起因するものについては、カタカナとひらがなは一対一に対応するので、読みがな付加は容易である。また品詞情報については、既存の辞書の情報との差別化が必要な場合、本システムの出力した文字列を対象とした新たな品詞名を設定すればよい。

品詞情報を必要としないシステムでは、本システムの出力結果をそのまま、あるいは人手による選別を得て、利用することが可能である。特に、かな漢字変換システム等では、本システ

5 KAKASI 漢字-かな (ローマ字) 変換プログラム。 <http://kakasi.namazu.org/>

ムの利用により頻出言い回しの登録が容易となることで精度の向上が期待できる。

4.6 他の手法との比較

日本語における語句抽出の研究は、主に、名詞句の抽出、固有名詞の抽出、および語句切り分けに関するものである。統計情報を利用した語句抽出の手法は、主として品詞情報を利用するもの、単語の共起情報を利用するもの、文字の共起情報を利用するものに大別できる。

Nagao らは n -gram 頻度を用いた文字列抽出を提案した (Nagao and Mori 1994)。 n -gram は n の値が大きくなるにしたがって出現頻度が低下するが、有意味文字列は他の文字列に比べて高い頻度で出現する。Nagao らの手法はこの性質を利用したものであり、本稿と同様品詞等の情報を利用せずに文字列の抽出を行なっている。

Kitani らは固有名詞全般を対象に、固有名詞の前後に出現しやすい語を接辞として扱うことで固有名詞の抽出を行なっている (Kitani and Mitamura 1994)。また、久光らは対象を人名に絞り、辞書と共起情報を利用した手法を提案している (久光, 丹羽 1997)。久光らの手法でも、人名の直後に現れる接辞を手がかりとして抽出を行なっており、さらに、人名接辞の獲得支援や、姓と名との分割・判別についても提案している。また、福本らは、接辞の他、固有名詞の特性に基づいたヒューリスティクスを導入し精度の向上を図っている (福本, 下畑, 梶井 1998)。

また、Cha らは韓国語を対象として、構文解析中に出現した未登録語の抽出を行なっている (Cha, Lee, and Lee 1998)。Cha らは、未登録語発見のための形態素パターン辞書を利用して、未登録語に対しても他の語と同様にタグ付けを行なうという手法を提案している。

単語間共起情報や品詞情報を利用する手法では、前処理として切り分けおよびタグ付けが必要となる。辞書を利用した手法では未登録語の問題は避けられないが、未登録語に起因する解析エラーがその後の処理の精度を下げるという問題が生じる。本稿で提案する手法では、辞書を利用せず、前処理にあたる訓練フェーズも文字単位の共起情報の取得だけであるため完全に自動的に行なうことが可能である。

5 まとめ

辞書ベースの自然言語処理では未登録語が大きな問題の一つである。本稿では、処理対象となるドメインの生コーパスを訓練コーパスとして取得した文字共起データのみを利用して対象ドメインの頻出文字列の自動抽出を行なう手法を提案し、口語を多く含む電子メール文書に対して適用しその有効性を示した。

本稿で利用したシステムでは、純粹に二文字間の共起情報のみを利用し、与えられた入力コーパス内の各隣接文字間の関係を推測することで、文中の意味のある文字列の認識を行なっている。本稿では、口語を多く含む電子メール文書をコーパスとし、コーパス中に頻出する口語

表現および異表記表現等の抽出を行なった。これらの口語表現、異表記表現は一般的な辞書に登録されていないものが多く、辞書ベースの解析の際にノイズとなるものである。本稿で示した実験では、辞書ベースの形態素解析ツールである茶筌が未登録語と判断した文字列の 69.06% を正しく認識した。また、茶筌がなんらかの解析結果を出力した文字列についても、解析誤りのため正しく切り分けが行なわれなかった文字列のうち 70.88% について、本システムは正しい切り分けを行なった。本システムは品詞タグを利用しないため、この数字は切り分け誤りに関するものであり、本システムを利用することによって正しく認識された文字列を辞書に組み込むことで、切り分け誤りだけでなく、品詞タグ付けの誤りについても、減少を図ることができる。と期待する。

参考文献

- Ananiadou, S. (1994). "A Methodology for Automatic Term Recognition." *Coling-94*, 1034–1038.
- Argamon, S., Dagan, I., and Krymolowski, Y. (1998). "A Memory-Based Approach to Learning Shallow Natural Language Patterns." *Coling-ACL'98*, 67–73.
- Borthwick, A. (1999). "A Japanese Named Entity Recognizer Constructed by a Non-Speaker of Japanese." *IREX Workshop*.
- Cha, J., Lee, G., and Lee, J.-H. (1998). "Generalized Unknown Morpheme Guessing for Hybrid POS Tagging of Korean." *The 6th Annual Workshop on Very Large Corpora*, 85–93.
- Church, K. W. and Hanks, P. (1989). "Word Association Norms, Mutual Information, and Lexicography." *The 27th Annual Conference of the Association of Computational Linguistics*.
- 福本淳一, 下畑光夫, 梶井文人 (1998). "固有名詞抽出における日本語と英語の比較." 情報処理学会 自然言語処理研究会報告 NL-98-126 No.15, 107–114.
- Fung, P. (1998). "Extracting Key Terms from Chinese and Japanese texts." *The International Journal on Computer Processing of Oriental Language, Special Issue on Information Retrieval on Oriental Languages*.
- 久光徹, 丹羽芳樹 (1997). "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, 1–6.
- Kashioka, H., Kawata, Y., Kinjo, Y., Finch, A., and Black, E. W. (1998). "Use of Mutual Information Based Character Clusters in Dictionary-less Morphological Analysis of Japanese." *Coling-ACL'98*, 658–662.
- Kitani, T. and Mitamura, T. (1994). "An Accurate Morphological Analysis and Proper Name Identification for Japanese Text Processing." *In Transactions of IPSJ Vol.35, No.3*, 404–

413.

松本裕治, 北内啓, 山下達雄, 平野善隆, 今修一, 今村友明 (1997). “日本語形態素解析システム『茶釜』 version 1.5 使用説明書.” テクニカル・レポート NAIST-IS-TR97007, 奈良先端科学技術大学院大学. <http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html>.

Nagao, M. and Mori, S. (1994). “A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese.” *Coling-94*, 611–615.

Nobesawa, S., Saito, H., and Nakanishi, M. (2000). “Automatic Semantic Sequence Extraction from Unrestricted Non-Tagged Texts.” *Coling 2000*, 579–585.

Nobesawa, S., Tsutsumi, J., Sun, D. J., Sano, T., Sato, K., and Nakanishi, M. (1996). “Segmenting Sentences into Linky Strings Using D-bigram Statistics.” *Coling-96*, 586–591.

Oda, H. and Kita, K. (1999). “A Character-Based Japanese Word Segmenter Using a PPM*-Based Language Model.” *ICCPOL'99*, 527–532.

佐野智久, 堤純也, 孫大江, 延澤志保, 佐藤健吾, 大森久美子, 中西正和 (1996). “自然言語文評価における d-bigram 情報の活用方法に関する実験.” 言語処理学会 第 2 回年次大会, 185–188.

堤純也, 新田朋晃, 小野孝太郎, 延澤志保 (1993). “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, 7–12.

略歴

延澤 志保: 1994 年慶應義塾大学理工学部数理科学科卒業。1996 年同大学院理工学研究科計算機科学専攻修士課程修了。同年、慶應義塾大学理工学研究科博士課程進学、現在に至る。自然言語処理の研究に従事。

斎藤 博昭: 1983 年慶應義塾大学理工学部数理科学科卒業。工学博士。慶應義塾大学理工学部専任講師。

中西 正和: 1966 年慶應義塾大学工学部管理工学科卒業。工学博士。慶應義塾大学理工学部教授。

(2000 年 10 月 10 日 受付)

(2001 年 4 月 13 日 採録)