N-gramを用いた離散型共起表現収集の一手法

白井　諭[†]　池原　悟[†]　小見佳恵[‡]

[†]NTTコミュニケーション科学研究所　[‡]NTTアドバンステクノロジ

1 はじめに

自然言語処理において，大量のコーパスや用例の重要性が指摘され，それを分析する技術の必要性が増大している。例えば，機械翻訳では，単語単位ではうまく訳せないフレーズを集めフレーズ単位に翻訳する方法や，一定の構造を持つ表現を対訳パターン化して原言語と目的言語を対応づける方法などの実現が期待されている。これらの方法を実現するには，言語データの中から使用頻度の高い表現を効率よく抽出する手段が必要となる。

抽出の対象となる表現には，連語やフレーズのように連続した文字列を構成するもの（連鎖型共起表現）と，呼応や特定の名詞と特定の動詞の組み合わせのように2種類以上の文字列が文中の離れた位置に現れるもの（離散型共起表現）がある。

連鎖型共起表現に関しては，大量の言語データを対象に，任意のnに対するn-gram統計を高速に実行する方法が提案され[長尾94]，原文データ内にある任意の長さの文字列を自動的に抽出して集計することが可能となった。この方法を拡張し，抽出される文字列の相互関係を考慮して部分文字列の重複抽出を抑制することにより，有効最長文字列を網羅的に抽出することも可能となった[池原95]。

離散型共起表現は，連鎖型共起表現の文字列が文中で共起したものと考えられる。本稿では，連鎖型共起表現を網羅的に抽出する方法[池原95]を拡張し，離れた位置にある文字列の共起関係を自動的に抽出して集計する方法を提案する。

2 共起する文字列の抽出条件

文章表現上は，表現の統一や複数文にまたがる呼応などの共起表現も考えられるが，本稿では1文内に閉じた範囲を対象として文字列の共起を考える。本稿の前提となる連鎖型共起表現は，[池原95]の方法により複数文にまたがる文字列を除外して抽出されているものとする。また，文の区切りは句点を基準として扱うが，引用文を伴う文は引用部分を除外したものを1文，引用部分は句点を基準とする別の文としてそれぞれ扱うことにする。

離散型の文字列共起を考える場合，文字列相互の位置関係が問題となる。[池原95]の方法によれば2回以上現れる文字列は長い単位で抽出されるため，特定の文字列の連接は最大1回しか起こりえず，偶然の連接と見られるので抽出対象とする。また，文字列が重なり合う場合は抽出対象とはしない。従って，離散型共起は，重なり合わない2つ以上の文字列の前後関係を保存して抽出することが必要である。

3 抽出アルゴリズム

処理の手順は以下の通りである（図1参照）。

手順1：原文番地ファイルの作成
手順2：汎用ソートファイルの作成
手順3：一致文字数のカウント

手順1～3は[長尾94]と同じで処理ある。まず，先頭文字から順に，0,1,2,...,n-1文字を削除したn個の部分文字列（文字列単語と呼ぶ）に対して，各部分文字列の順に対応するn個のレコードを持った原文番地ファイルを作成し，文字列をコード順にソートした後，隣接する文字列の先頭からの一致文字数をカウントし，汎用ソートファイルを作成する。

手順4：抽出文字数の記入
手順5：拡張原文番地ファイルの作成
手順6：有効無効判定処理
手順7：再拡張汎用ソートファイルの作成

手順1～7は[池原95]と同じ処理である。汎用ソートファイルの各レコードに抽出対象の文字数（抽出文字数）を記入し，原文番地の順にソートし戻し

An extraction method of interrupted collocational expressionss by n-gram statistics
Satoshi SHIRAI[†], Satoru IKEHARA[†] and Yoshie OMI[‡]
[†] NTT Communication Science Laboratories and [‡]NTT Advanced Technology Corporation

た後，抽出文字数に着目して各レコードの採否を決定する。採否は，先頭レコードから「抽出文字数」の値を調べ，最初の0でない値nをもつi番目のレコードを「採用」とし，続くi+1, i+2, . . . , i+(n-1)番目のレコードの抽出文字数がn-1, n-2, . . . , 1より大きいものを「採用」，以下を「不採用」とする。これにより文字列単語の相互関係が考慮され，不要な部分文字列の抽出が抑制される。この後，「汎用ソートファイル」のレコード順にソートし直すことにより「再拡張汎用ソートファイル」を作成する。

手順8：文字列番号の付与

「再拡張汎用ソートファイル」の「採用」レコードの異なりごとに文字列番号を付与する。

手順9：再拡張汎用ソートファイルの再ソート

原文番地の値の順にソートし戻し，「拡張原文番地ファイル」の形式に戻す。

手順10：文番号の付与

手順9で得られたファイルに「文番号」の欄を設け，当レコードが所属する原文の文番号を記入する。

手順11：ファイルの圧縮

手順10で得たファイルから，「文番号」「文字列番号」「抽出文字数」「原文番地」の4つの欄以外を削除し，さらに，「文字列番号」の欄の値のないレコードを削除することにより「離散型共起圧縮ファイル」作成する。

手順12：離散型共起表現の抽出

「離散型共起圧縮ファイル」先頭から「文番号」が同じレコードを，2文字列共起，3文字列共起などの目的に応じて組み合わせ，「文字列番号」の組をファイルに出力する。

手順13：離散型共起表現のカウント

手順12で作成したファイルをソートし，文字列番号の組の同じものをカウントする。なお，文字列番号を縦横に並べたマトリックスを用意して度数を書き込んでいくことにより手順12と手順13を同時に行なう方法も考えられるが，極めてスパースなマトリックスになるため効率的でない。

4 抽出実験

日経新聞記事3カ月分（892万字）を対象として，[池原95]による連鎖型共起文字列に引き続き，離散型共起文字列の抽出実験を行なった。

本稿では，連鎖型共起文字列として単独では10回以上出現した12,351件を対象として，このうちの2種類の文字列が1文内に離れて共起する場合を集計した。抽出された文字列の組の数を表1に，出現頻度の多い文字列の組を表2に，2回以上出現した文字列の組のうち合計文字数の多い文字列の組を表3に示す。対象とする連鎖型文字列の数にもよるが，本実験はXEROX ARGOSS 5270(SUN OS4.1.3)によりターンアラウンド時間十数分で結果が得られた。

表1　抽出される文字列の組の種類と延べ度数

抽出対象＼結果	文字列の組の種類数	延べ出現回数
2回以上	6，544	21，829
5回以上	941	9，057
10回以上	237	4，556
20回以上	61	2，291

（2種類の文字列の文内共起の場合）

表2　出現頻度の高い文字列の組の例

順位	前方文字列	後方文字列	度数
1位	価格	発売時期	257
2位	ゼネラル	モーターズ	117
3位	サミット	先進国首脳会議	86
4位	EC	欧州共同体	80
5位	イラン	ジャパン石油化学	80

表3　合計文字数の大きい文字列の例

順位	前方文字列	後方文字列	合計字数	度数
1	部会長、梅本純正武田薬品工業副社長	について協議した	25	2
2	本社夕張市、保全管理人山根喬氏	問題などについて	23	2
3	北炭夕張炭坑	本社夕張市、保全管理人山根喬氏	21	9
4	永田町のホテルニュージャパン	横井英樹社長	21	11
5	マルクは一ドル	フラン、英ポンドは一ポンド	20	5

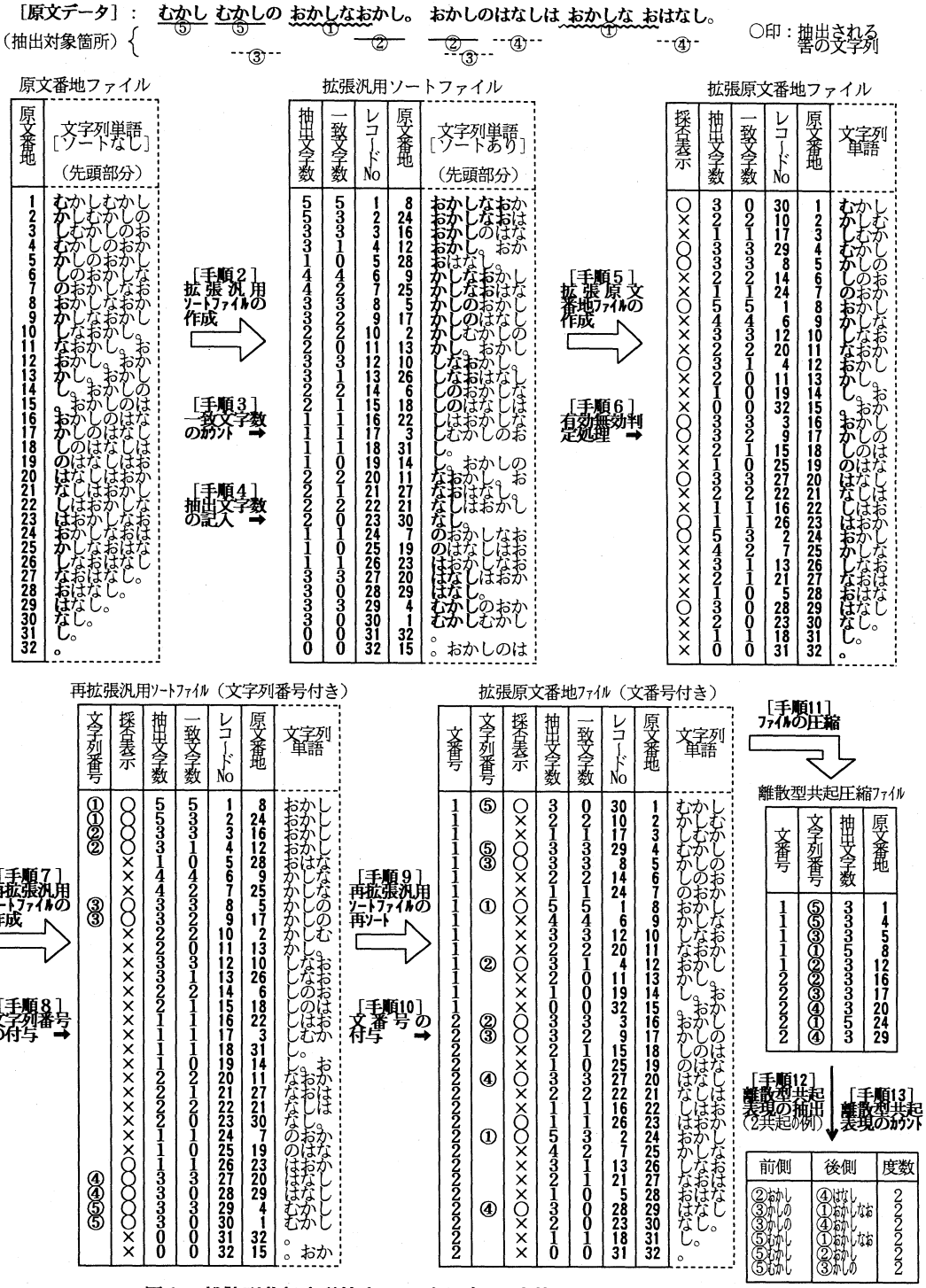

図1　離散型共起表現抽出アルゴリズムの実施例

これらの表から，出現頻度の高い離散型共起は名詞同士の共起が圧倒的に多いことがわかる。

抽出された表現の組は２度数以上が6,544件であるので，人手により目的に応じた表現の組を取り出すのが可能であると思われる。しかし，抽出対象の表現の範囲を予め指定すれば，離散型共起表現は能率的に抽出される。例えば，表現上の言い回しに着目して共起表現を抽出するには，手順８で得られた連鎖型共起表現の中で不要と見られる表現を削除した後，手順９以降の処理を行なえばよい。そこで，固有名詞や数詞を含む文字列を除去した残りの文字列に対して離散型共起を求めてみた。その一部を表４に示す。表から新聞記事の文型に相当するような離散型共起表現が抽出されることがわかる。

表４　抽出された離散型共起表現の例

番号	離散型共起文字列の例		度数
１	としながらも	と述べた	９
２	の質問に答え	と述べた	９
３	その内容は	というもの	６
４	われわれは	と語った	６
５	さらに首相は	と述べた	５
６	その内容は	など	５
７	とし、	と述べた	５
８	についても	としている	４
９	いかにも	らしい	４
１０	つまり	である	４
１１	にしろ	にしろ	４
１２	このほか	などと語った	４
１３	これに対し	と答えた	４
１４	この中には	も含まれている	３
１５	これに対し	と反論した	３
１６	これに対して	と答えた	３
１７	するつもりだ	と語った	３
１８	これからは	という	３
１９	その骨子は	というもの	３
２０	にせよ、	にせよ	３
２１	その内容は	などとなっている	３
２２	などから	とみている	３
２３	にするか	にするか	３
２４	と述べるとともに	と語った	３
２５	なり	なり	３

なお，本方式では，原文データは文字列として扱い，２文字以上の文字列複数個が文内で共起する場合を対象としたが，共起する文字列が１文字の場合にも文法的に重要な意味を持つ場合が考えられる。そのような場合は文字列に対する処理では限界が考えられるので，原文データを形態素解析した後，形態素列に対して本稿の方法を適用するなどの改良が有効であると考えられる。

５ おわりに

言語コーパスなどの膨大な言語データの中から，２種類以上の文字列が文中の離れた位置に現れるものを自動的に発見し集計する方法を提案した。具体的には，任意のnに対するｎ-gramの計算法として提案されたアルゴリズム[長尾94]を改良し，言語データの中から，相互に部分重複せず，かつ，２回以上出現した文字列（連鎖型共起表現）を網羅的に収集する方法[池原95]を発展させることにより，連鎖型共起表現を組み合わせて，文中の離れた位置に共起する文字列の組（離散型共起表現）を抽出し，その頻度を求める方法を示した。

新聞記事約890万字を用いた適用実験によれば，２度数以上の文字列の組が6,544件抽出され，さらに固有名詞や数詞などを除去することにより，新聞特有の表現形式が抽出されることがわかった。

本稿では日本語文字列への適用を考えたが，この方法は任意の記号列に適用できるため，形態素解析により得られた単語列や文法的要素列などへの応用が考えられる。また，文字列への適用では断片的文字列の混入が少なからず見られるが，単語列や文法的要素列への適用ではより良い結果が期待される。これらの点を考えて，今後は種々の適用実験により改良を加えていく予定である。

＜謝辞＞
本検討にご協力くださったスバルインターナショナルの渡辺亮嗣氏ならびに河津勝己氏に感謝する。

＜参考文献＞
[池原95] 池原，白井，河岡：N-gramを用いた連鎖型共起表現の自動抽出法，言語処理第1回年次大会C4-1
[長尾94] Nagao, M. and Mori, S.: 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, The Proc. of the 15th International Conference on Computational Linguistics, Kyoto