概要

機械翻訳などの自然言語処理において、使用頻度の高い表現や定型的な言い回しなどを収集した日本語共起表現辞書が必要とされている。しかし、その辞書に登録する見出し語を、人手で収集することは困難である。そこで、使用頻度の高い表現や定型的な言い回しを自動的に抽出する方法が考えられているが、まだ確立されていない。従来、N-gram統計を応用した方法が提案されているが、抽出される表現には、まだ多数の不要な文字列が混在するため、その後の人手の作業は容易ではない。

そこで本研究では、原文の段階で抽出すべき、単文、重文、名詞句の文型の種類毎に適切な単語置き換えを行なうことによって、効率よく目的とする表現を抽出した。そして、N-gram統計の出力結果に対しても、品詞情報を用いて必要のない表現の削除も行なった。また、出力結果を調査しやすくするために相互情報量を用いて言い回しを絞り込んだ。

上記の手法を毎日新聞95年度の一面記事1万文に対して適用し、定型的な言い回しを抽出した。抽出された表現から、共起表現辞書の見出し語となりそう表現を人手で選択した結果、単文、重文、名詞句について、定型的な言い回しの含有率は100%, 50.5%, 87.2%であると分かった。従来に比べて断片的な表現が減少し、単文、重文、名詞句を構成する表現が効率よく抽出できることが確認された。

本手法により、大規模なコーパスから容易に日本語共起表現辞書の見出し語が収集でき、日本語共起表現辞書を作成できる見通しとなった。

目 次

1 まえがき 1

2 N-gram 統計の問題点 3

3 N-gram 統計処理による共起表現 4

3.1 連鎖型共起表現 4

3.2 離散型共起表現 5

4 日本語共起表現辞書の作成方法 7

4.1 共起表現抽出の考え方 7

4.1.1 表現種別に着目した置き換えと品詞情報の利用 7

4.1.2 相互情報量の利用 7

4.2 アルゴリズム 9

4.2.1 表現種別に着目した単語の置き換え 9

4.2.2 品詞情報の利用 10

5 実験結果と考察 11

5.1 実験方法 11

5.2 実験結果と考察 12

5.2.1 表現の種類数 12

5.2.2 相互情報量の効果 13

5.2.3 得られた定型的な言い回し 13

5.2.4 表現パターンの含有率 13

5.2.5 含有率からの考察 18

6 結論 20

図 目 次

1 連鎖型共起表現の例 4
2 連鎖型弱抑制共起表現の例 4
3 離散型共起表現の例 5
4 実験手順 11
5 重文・離散共起抽出 (2 組) の含有率の推移 16
6 単文・連鎖共起抽出のサンプルテスト結果 16
7 単文・離散共起抽出 (2 組) のサンプルテスト結果 16
8 名詞句・連鎖共起抽出のサンプルテスト結果 17
9 名詞句・離散共起抽出 (2 組) のサンプルテスト結果 17

表 目 次

1 離散型共起表現の組み合わせの例 6
2 相互情報量利用による順位変化の例 8
3 相互情報量の評価式 8
4 置き換えテーブル 9
5 置き換えの例 10
6 単文、重文、名詞句の品詞情報による抽出条件 10
7 重文の表現パターン抽出の場合の表現の種類数 12
8 重文の場合の相互情報量の効果 14
9 得られた定型的な言い回しの例 14
10 表現パターンの出力表現数 15
11 採用表現の含有率 15
12 相互情報量と頻度情報を共に用いた評価式 19

1　まえがき

機械翻訳などの自然言語処理において、使用頻度の高い表現や定型的な言い回しなどを収集した日本語共起表現辞書が必要とされている。しかし、その辞書に登録する見出し語を、人手で収集することは困難である。そこで、使用頻度の高い表現や定型的な言い回しを自動的に抽出する方法が考えられているが、まだ確立されていない。従来、N-gram 統計を応用する方法が提案されているが、抽出される表現には、まだ多数の不要な文字列が混在するため、その後の人手の作業は容易ではない。

N-gram 統計を用いた表現抽出の方法として、[1, 2] の研究がある。文献 [1] では大量の言語データを対象に、N-gram 統計を高速に実行する方法が提案されている。この方法を用いることで、言語データ内の任意の長さの文字列を自動的に抽出し、その出現回数をカウントできる。この結果、原文中に現われる文字列を、その長さの順かつ出現頻度の高い順に集計できることが報告されている。しかし、この方法では意味的にまとまりのない断片的な文字列が多数抽出されるという問題があった。

この問題を解決する方法として、[2] の方法が提案されている。この方法は大量の言語データから、使用頻度の高い表現および表現の組を自動的に発見し集計する方法である。具体的には、まず、任意の N-gram の計算法として提案された [1] のアルゴリズムを独立性の高い表現を抽出する観点から改良し、言語データ中に 2 回以上出現した文字列 (連鎖型共起表現) を「一度、抽出した文字列の部分文字列は、その後、抽出しない」という条件下で、もれなく自動的に抽出し集計する方法である。また、この方法で抽出された文字列を組み合わせて、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、頻度を求める方法も [2] で提案されている。

そして、従来、単語の置き換えを行なう方法 [3] や、単語単位に変換した原文データに対して、N-gram を適用する方法 [4] 等が [2] の方法を改良した方法として提案されている。

置換えを用いた N-gram による言語表現の抽出 [3] では、対象となる原文データに対して、(1) 数字の置き換え、(2) 企業名の置き換え、及び、企業名が連続している部分をひとまとまりにしての変換、(3) 括弧内の省略、(4) 引用文全体の置き換えを行っている。このような原文データに変換を施すと、N-gram 出力時の表現パターンを減らせる結果となる。しかし、言い回しを抽出するまでには至っていない。

また、N-gram 統計を応用した文型パターンの自動抽出法の研究 [4] では、原文データに対して、単語境界を付加し対象データを文字列データとしてでなく、単語列データと

して扱っている。また、名詞の中の数詞、固有名詞を縮退させた。単語境界付加により断片的な文字列の抑制ができ、1文字単語の抽出が可能となった。名詞の縮退により同一表現がまとめられたことが報告されている。

しかし、上記[3, 4]の手法では、まだ多数の不要な文字列が混在するため、その後の人手の作業は容易ではない。

そこで本研究では、原文の段階で抽出すべき、単文、重文、名詞句の文型の種類毎に適切な置き換えを行なうことによって、効率よく目的とする表現を抽出する。そして、N-gram統計の出力結果に対しても品詞情報を用いて、必要のない表現の削除も行う。また、出力結果を調査しやすくするために相互情報量を用いて言い回しを絞り込んだ。

上記の手法をプログラム化し、毎日新聞95年度の一面記事1万文に対して適用して、定型的な言い回しの抽出を行なった。

以下、2章ではN-gram統計の問題点について、3章ではN-gram統計処理による共起表現について、4章で日本語共起表現辞書の作成方法について、そして、5章では実験結果と考察について、最後に7章で結論を述べる。

2 N-gram 統計の問題点

本研究では、日本語共起表現辞書の作成を主眼としている。そのためには定型的な言い回しを自動的に抽出が必要である。しかし、完全に自動な抽出は現実には不可能であるため、最終段階では人手によって各言い回しの正当性を検証する。

従来、[2] の方法を用いた以下のような N-gram 統計を応用した言い回しの抽出法が提案されている。単語の置き換えを行なう方法 [3] や、単語単位に変換した原文データに対して、N-gram を適用する方法 [4] 等が存在する。

置換えを用いた N-gram による言語表現の抽出 [3] では、対象となる原文データに対して、(1) 数字の置き換え、(2) 企業名の置き換え、及び、企業名が連続している部分をひとまとまりにしての変換、(3) 括弧内の省略、(4) 引用文全体の置き換えを行なっている。原文データに変換を施すと、N-gram 出力時の表現パターンを減らせる結果となることが報告されている。しかし、言い回しを抽出するまでには至っていない。

また、N-gram 統計を応用した文型パターンの自動抽出法の研究 [4] では、原文データに対して、単語境界を付加し対象データを文字列データとしてでなく、単語列データとして扱う。また、名詞の中の数詞、固有名詞を縮退させた。単語境界付加により断片的な文字列の抑制ができ、1 文字単語の抽出が可能となった。名詞の縮退により同一表現がまとめられたことが報告されている。

しかし、上記 [3, 4] の手法では、まだ多数の不要な文字列が混在するため、その後の人手の作業は容易ではない。

そこで本研究では、原文の段階で抽出すべき、単文、重文、名詞句の文型の種類毎に適切な置き換えを行なうことによって、効率よく目的とする表現を抽出する。そして、N-gram 統計の出力結果に対しても品詞情報を用いて、必要のない表現の削除も行う。また、出力結果を調査しやすくするために相互情報量を用いて言い回しを絞り込んだ。

3　N-gram 統計処理による共起表現

本研究では、N-gram を使用した文字列抽出法として、[2] の方法を使用する。以下にその方法を要約し説明する。共起表現には、連鎖型共起表現と離散型共起表現の 2 種類がある。

3.1　連鎖型共起表現

原文データに現れる文字列をすべて抽出する方法である。図 1 のαのように原文中から一度抽出された文字列の内部に含まれる部分文字列を抽出するか否かによって次の 3 つの抑制法がある。部分文字列とは、図中のβのような文字列を指し、γのような文字列は含まない。(実験では、2 文字以上、頻度 2 以上の文字列を抽出対象とした)

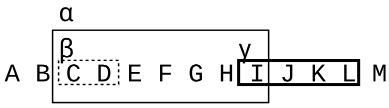

図 1: 連鎖型共起表現の例

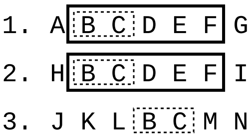

図 2: 連鎖型弱抑制共起表現の例

- 無抑制型　部分文字列でも抽出対象とする。
- 弱抑制型　部分文字列でも、他の場所で独立して出現していれば、抽出の対象とする。
  図 2 を例として説明する。1、2 の文があった場合、BCDEF は 2 回出現しているので抽出される。ここで、1、2 の文しか無い場合、BC は BCDEF の文字列の内部に含まれている部分文字列であるため、抽出対象とならない。しかし、3 の文がある場合は、独立して出現しているので、抽出対象となる。
- 強抑制型　部分文字列は一切、抽出の対象としない。

3.2　離散型共起表現

原文データの離れた場所に現れ、共起する2つ以上の文字列の組を抽出する方法で、共起の仕方により、次の3つのタイプがある。図3を例とした場合の抽出の組合わせを表1に示す。(実験では、頻度2以上の共起する組を抽出対象とした)

- 無抑制型　共起する文字列が同一文中にあることだけを条件として、抽出をする。
- 弱抑制型　上記の条件に加えて、抽出する文字列は互いに異なるもののみ抽出する。
- 強抑制型　上記の2つの条件に加えてさらに、抽出する表現の先頭の文字列と末尾の文字列との間には、着目する文字列が2回以上現れないものを抽出する。

文頭　文末

～(1)～(2)～(3)、～(1)～(2)～(3)、～(1)～(2)～(3)～。

図 3: 離散型共起表現の例

図3から離散共起表現を抽出した場合、抑制の仕方を表1に示す。図中の「○」は、左の表現が抽出されることを示している。

例えば、(1)...(2)...(3)の共起表現は、どの抑制法においても抽出される。(1)...(3)...(2)の共起表現は、無抑制、弱抑制の場合にのみ抽出され、強抑制の場合には抽出されない。

弱抑制では、「抽出する文字列は互いに異なるもののみ抽出する」という条件のため、(1)...(2)...(3)のような共起表現は抽出するが、(1)...(2)...(2)のように同じ文字列を含む共起表現は抽出されない。

また、強抑制では、弱抑制に加えて「抽出する表現の先頭の文字列と末尾の文字列との間には、着目する文字列が2回以上現れないものを抽出する」という条件のため、(1)...(2)...(3)のような共起表現は抽出するが、(1)...(3)...(2)のように抽出する表現の先頭の文字列と末尾の文字列との間に、(3)のような着目する文字列が現われている共起表現は抽出されない。

表 1: 離散型共起表現の組み合わせの例

抑制型	無抑制	弱抑制	強抑制
(1) ... (1) ... (1)	○		
(1) ... (1) ... (2)	○		
(1) ... (1) ... (3)	○		
(1) ... (2) ... (1)	○		
(1) ... (2) ... (2)	○		
(1) ... (2) ... (3)	○	○	○
(1) ... (3) ... (1)	○		
(1) ... (3) ... (2)	○	○	
(1) ... (3) ... (3)	○		
(2) ... (1) ... (1)	○		
(2) ... (1) ... (2)	○		
(2) ... (1) ... (3)	○	○	
(2) ... (2) ... (1)	○		
(2) ... (2) ... (2)	○		
(2) ... (2) ... (3)	○		
(2) ... (3) ... (1)	○	○	○
(2) ... (3) ... (2)	○		
(2) ... (3) ... (3)	○		
(3) ... (1) ... (1)	○		
(3) ... (1) ... (2)	○	○	○
(3) ... (1) ... (3)	○		
(3) ... (2) ... (1)	○	○	
(3) ... (2) ... (2)	○		
(3) ... (2) ... (3)	○		
(3) ... (3) ... (1)	○		
(3) ... (3) ... (2)	○		
(3) ... (3) ... (3)	○		

4 日本語共起表現辞書の作成方法

4.1 共起表現抽出の考え方

4.1.1 表現種別に着目した置き換えと品詞情報の利用

従来の方法では、置き換えを行なってはいるが、置き換えが名詞のみに限られていた。そのため、名詞については文字が違うことによって、別の表現になっているものがまとめられ、表現の抽出の種類は大幅に抑えられる結果となっている。しかし、名詞のみの置き換えでは、抽出される表現には、まだ多くの不要な文字列が混在するため、その後の人手の作業は容易ではない。

そこで本研究では、原文の段階で抽出すべき、単文、重文、名詞句の文型の種類毎に適切な置き換えを行なうことによって、効率よく目的とする表現を抽出する。そして、N-gram 統計の出力結果に対しても品詞情報を用いて、必要のない表現の削除も行う。

文型の種類毎に適切な置き換えを行なうことによって、効率よく目的とする表現を抽出する。そのため、各文型の種類に閉じた言い回しの抽出が出来るため出力結果も抑えられる形となる。かつ、辞書に見出し語を登録する際、文型ごとに登録することができ利用しやすい。

品詞情報は、出力される結果から必要な文の構成をしているものだけを抽出するために用いる。置き換えを行なったことで、表現パターンの種類数は減少しているがまだ多い。そのため、品詞情報を用いて、表現が求める構成と異なるものは削除し、求める構成の表現のみを選択する。

4.1.2 相互情報量の利用

上記の方法では、頻度の高い表現が上位に現われることになるが、離散共起表現においては頻度が高いからといって、定型的な言い回しとは限らないので、相互情報量を適用し表現を絞り込む。相互情報量により、表現を表現の結び付きの強さを評価する。相互情報量の値により順位づけを行なうことによって、出現頻度に関係なく、結び付きの強い言い回しが得られるようになる。

表 2 に、相互情報量利用の例を示す。

表中の離散共起表現の結果では、「もし...である」のほうが頻度は高いが、相互情報量を計算すると、「もし...なら」のほうが高くなり上位にくる。これは、「もし...である」の「である」は 5,000 回出現していて、そのうち 90 回が「もし」と共起しているが、「も

し...なら」の「なら」は200回出現して、そのうち80回が「もし」と共起しているためである。つまり、「なら」のほうが「である」に比べてより「もし」と強く結び付いているためである。

表 2: 相互情報量利用による順位変化の例

連鎖共起表現

順位	表現	頻度
1	である	5,000
2	なら	200
3	もし	100

離散共起表現

順位	表現	頻度
1	もし ... である	90
2	もし ... なら	80

⇒

相互情報量利用

順位	表現	相互情報量
1	もし ... なら	9.36
2	もし ... である	4.89

相互情報量は表3の評価式を用いて評価する。

表 3: 相互情報量の評価式

$$
\begin{aligned}
MI(x_1,\cdots,x_n) &= log_2\left(\frac{P(x_1,\cdots,x_n)}{P(x_1)\cdot\cdots\cdot P(x_n)}\right) \\
&= log_2\left(\frac{\dfrac{f(x_1,\cdots,x_n)}{N_c}}{\dfrac{f(x_1)}{N_f}\cdot\cdots\cdot\dfrac{f(x_n)}{N_f}}\right)
\end{aligned}
$$

$f(x_1,\cdots,x_m)$: $x_1,\cdots,x_m$ が共起する離散共起頻度
$f(x_m)$: x_m の連鎖共起頻度
N_c : 離散共起の総頻度
N_f : 連鎖共起の総頻度

4.2　アルゴリズム

4.2.1　表現種別に着目した単語の置き換え

原文の段階で抽出すべき、単文、重文、名詞句の文型の種類毎に適切な置き換えを行なうことによって、効率よく目的とする表現を抽出した。置き換えはn対1の連続的な置き換えであり、表4に示す置き換える品詞を一単語以上の連鎖を表現する一単語の「γ」に置き換える。各文型の鍵となる品詞及び、それに付随する品詞は残して、それら以外の単語は置き換えを行なうように、表4の置き換えテーブルを作成した。

表 4: 置き換えテーブル

	単文	重文	名詞句
記号	−	−	−
名詞	−	−	○
動詞	○	○	−
形容詞・形容動詞	−	−	−
副詞	−	○	−
連体詞	−	−	−
接続詞	−	○	−
感動詞	−	−	−
接辞	−	−	−
助動詞	−	○	−
格助詞 (に、より、を、をば)	○	○	○
格助詞 (その他)	○	−	○
副助詞	○	○	○
接続助詞	○	○	○
終助詞	−	−	−
準体助詞	−	−	−

[○：残す品詞、 −：「γ」に置き換える品詞]

以下に、単文、重文、名詞句の各文型の鍵となる品詞を示す。

- 単文： 動詞、助詞
- 重文： 動詞、接続助詞
- 名詞句： 名詞、助詞

複文の置き換えの例を表5に示す。「／」は単語境界であり、「γ」は置き換え後の単語である。

表5: 置き換えの例

置き換え前：	毎日新聞が入手した寄付リストによれば、業態別には別表のようになる。
	↓
単語境界付加：	／毎日／新聞／が／入手し／た／寄付／リスト／に／よれば／、 ／業態／別／に／は／別表／の／ように／なる／。／
	↓
置き換え後：	／γ／入手し／た／γ／によれば／γ／に／は／γ／ように／なる／γ／

4.2.2 品詞情報の利用

まず、品詞情報を原文データに埋め込む。その際には置き換えを施さないものにのみ品詞情報を付加した。N-gram出力結果から単文、重文、名詞句の文型の種類毎の抽出条件を満たす表現のみを選択する。表現の抽出条件を表6示す。

表4の置き換えテーブルと、表6の抽出条件はセットになっている。表4の置き換えを行ない、表6の抽出条件により、言い回しを抽出した。

品詞情報による抽出条件(表6)を、重文の場合を例に説明する。重文とは、「動詞~接続助詞~動詞」の文型をしているので、重文のみを抽出する場合は抽出条件を「動詞+接続助詞+動詞」とするよう定めた。

表6: 単文、重文、名詞句の品詞情報による抽出条件

	抽出条件
単文	助詞 + 動詞
重文	動詞 + 接続助詞 + 動詞
名詞句	助詞 + 名詞

5　実験結果と考察

5.1　実験方法

対象とする原文データを毎日新聞95年度の一面記事1万文とし、図4の手順で実験を行なった。

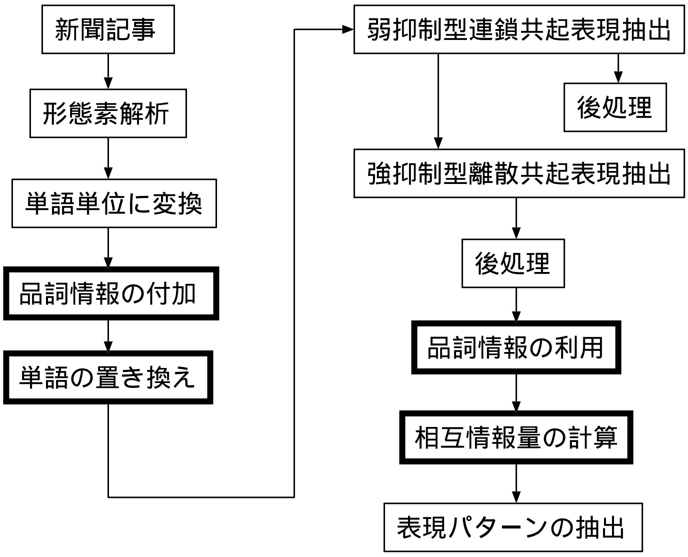

図 4: 実験手順

図4中の太枠の処理は、本研究で提案し、プログラムを作成した処理である。また、図中の『後処理』とは、単語単位の構成をしていない断片的な表現を削除する処理であり、プログラムを作成し不要な表現の削除をした。

連鎖共起表現抽出においては、弱抑制型で抽出をした。これは、まとまりのある表現のみを抽出するためである。無抑制では、まとまりのない断片的な表現が多数抽出され、強抑制では、部分文字列が一切抽出できないという問題があるため、弱抑制型連鎖共起表現抽出を使用した。

離散共起表現抽出においては、強抑制型で抽出をした。これは、定型的な言い回しのみを抽出するためである。無抑制、弱抑制では、あまりに表現の種類が膨大になり言い回しではない表現までも抽出し、定型的な言い回しの抽出に向かないため、強抑制型離散共起表現抽出を使用した。

5.2 実験結果と考察

5.2.1 表現の種類数

重文の表現パターン抽出の場合の表現の種類数を表 7 に示す。表中の「ｙ」は置き換えた単語、「／」は単語境界を表している。

表 7: 重文の表現パターン抽出の場合の表現の種類数

	連鎖共起	離散共起 (2 組)	離散共起 (3 組)
N-gram 出力時	36,895	668,385	出力不能
単語境界付加	21,914	386,322	出力不能
置き換え処理 N-gram 出力時	11,035	125,449	690,601
置き換え処理 単語境界付加	5,615	31,185	78,436
置き換え処理 単語境界付加 及び ／ｙ／のみと共起 している表現の削除	5,614	23,067 (81.7%減少)	43,178 (93.8%減少)

表 7 の表現の種類数は、重文の表現パターン抽出の場合であり、動詞、接続助詞、副助詞、格助詞 (に、より、を、をば)、副詞、接続詞、助動詞以外を一単語「ｙ」に置き換えた場合の表現の種類数である。

- 置き換えを行なわない場合、N-gram 出力時、単語境界付加の離散共起 (3 組) は表現が増えすぎて、計算機では出力不能となった。
- 置き換え処理後の N-gram 出力時と比べて、置き換え処理後の単語境界付加及び、／ｙ／のみと共起している表現の削除を行なった場合では、表現の種類が、離散共起 (2 組) で、81.7%、離散共起 (3 組) で、93.8%減少した。

実験の結果、形態素の情報を用いて単語境界を明示し利用することで、大幅に断片的な文字列が減少した。また、置き換え文字列である／ｙ／のみと共起している表現を削除すると表現が減少した。

5.2.2 相互情報量の効果

重文の場合の相互情報量の効果について説明する。相互情報量により表現と表現の結び付きの強さを評価した結果を、表8に出力結果から上位10表現を示す。表中の「...」は置き換えた単語列を表している。

- 離散共起のそのままの結果では、上位には、ほとんど助詞同士の共起表現が現われた。
- 相互情報量を用いた方が結び付きの強い表現が上位に現われた。
- また、相互情報量を用いた場合、頻度2の表現でも結び付きが強ければ、上位に現われた。

5.2.3 得られた定型的な言い回し

得られた言い回しの例を表9に示す。

5.2.4 表現パターンの含有率

出力された表現パターンの数を表10に示す。単文、名詞句では、連鎖共起、離散共起(2組)の表現パターンを、重文では、連鎖共起、離散共起(2組)、離散共起(3組)の表現パターンを出力した。

重文の出力表現数はすべてを合計しても、380表現ですべての表現を人手で調査可能な数なので、すべての表現を人手で検討した。実際に人手で検討した重文の表現は参考資料、出力表現 p1-4に示す。しかし、単文、名詞句の出力表現数は膨大であり、とても人手ですべてを調査することが可能な表現数ではない。そこで、単文、名詞句ではサンプルテストを行なうことにした。

サンプルテストは、出力表現を約5分割する位置の先頭から100表現を抜き出し、検討の必要があると思われる表現を採用することにした。例えば、単文の連鎖共起の場合は、サンプルデータをグループA,B,C,D,Eとした。それぞれは表現番号1,601,1201,1801,2401から各先頭100表現である。そして、各グループの採否を決定した。サンプルテストの結果を表6 ~ 9に示す。実際に人手で検討した単文、名詞句の表現は参考資料、出力表現 p5-24に示す。連鎖型共起の結果は、頻度の高い順に表現が並んでいて、離散型共起抽出の表現は、相互情報量の値が高い順に表現が並んでいる。

表 8: 重文の場合の相互情報量の効果

頻度により並び替え

順位	離散共起	頻度	相互情報量
1.	...は...に...を	490	2.794
2.	...は...を...た	449	2.891
3.	...に...を...た	312	4.242
4.	...は...に...た	268	4.329
5.	...は...を...に	235	3.844
6.	...は...を...を	193	3.768
7.	...は...を...ている	188	3.862
8.	...は...は...を	180	1.734
9.	...に...は...を	150	1.869
10.	...は...て...を	131	1.180

相互情報量により並べ替え

順位	相互情報量利用	頻度	相互情報量
1.	...に備える...を...ほど...にする...を	2	27.47
2.	...である...にも...も...分からない	2	26.89
3.	しかし...に入りきれず...を...なども...追いついていない	2	26.67
4.	...に備える...を...ほど...を伴う	2	25.89
5.	しかし...に入りきれず...を...が...に...追いついていない	2	25.57
6.	...である...にも...だ...分からない	2	25.30
7.	...に...を...に...を決める...に...たい...を示し	2	25.22
8.	...たとえ...か...か...だろう	2	25.08
9.	...は...を...も...を始め...たい...だが	2	24.89
10.	しかし...に入りきれず...を...も...は...追いついていない	2	24.83

表 9: 得られた定型的な言い回しの例

単文	...との...を示し ...の...に対する ...で...が発生する ...は...を...から...に推薦する
重文	...やっても...どんどん...なる ...おびえながら...を過ごした ...を乗り継げば...まで...移動できる ...に向けて...問われている
名詞句	...に...見込み ...からの参加 ...までの余震は ...を対象に

表 10: 表現パターンの出力表現数

	連鎖共起	離散共起 (2 組)	離散共起 (3 組)
単文	2,888	8,728	-
重文	60	199	121
名詞句	6,765	48,284	-

表 6 ~ 9 から以下のことがわかる。

- 単文はどちらの結果も常に 100%の含有率である。
- 名詞句の含有率にはばらつきがある。

単文、重文、名詞句の含有率を表 11 に示す。重文の N-gram の出力表現数は人手で調査可能な表現数であるため、出力されたすべての表現を人手で選択した結果である。単文、名詞句の含有率はサンプルテストの含有率の平均値である。

表 11: 採用表現の含有率

	合計出力表現数	含有率
単文	11,616	100%
重文	380	50.5%
名詞句	55,049	87.2%

重文の離散共起抽出 (2 組) の含有率の推移を図 5 に示す。

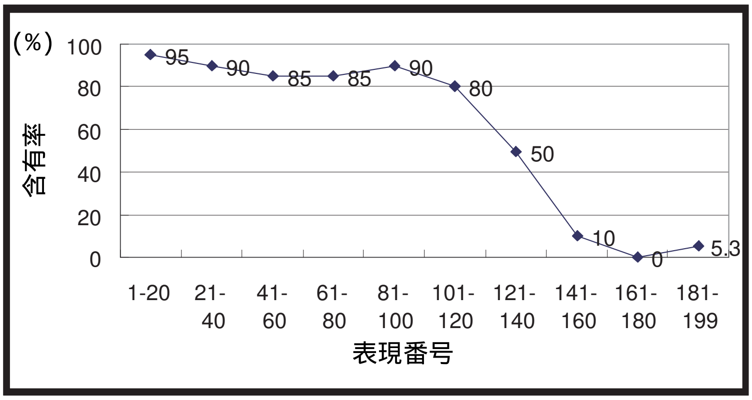

図 5: 重文・離散共起抽出 (2 組) の含有率の推移

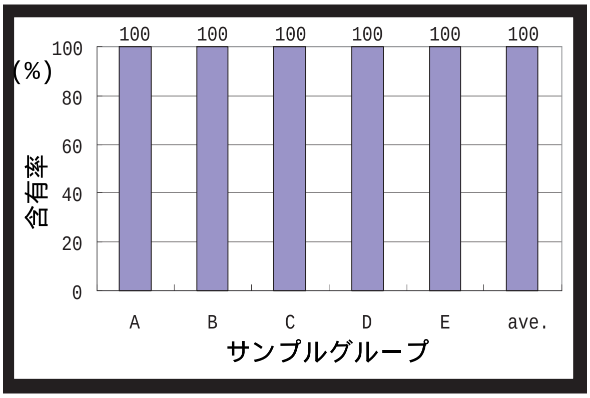

図 6: 単文・連鎖共起抽出のサンプルテスト結果

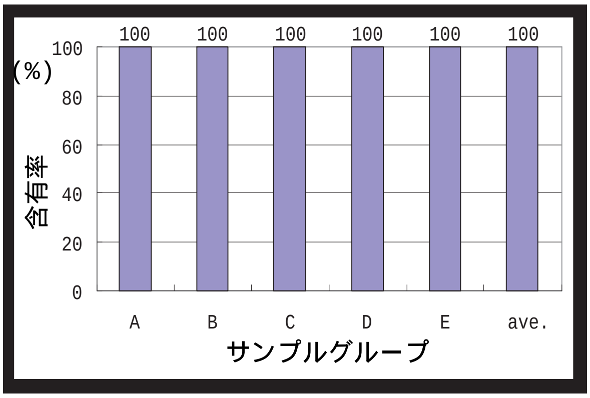

図 7: 単文・離散共起抽出 (2 組) のサンプルテスト結果

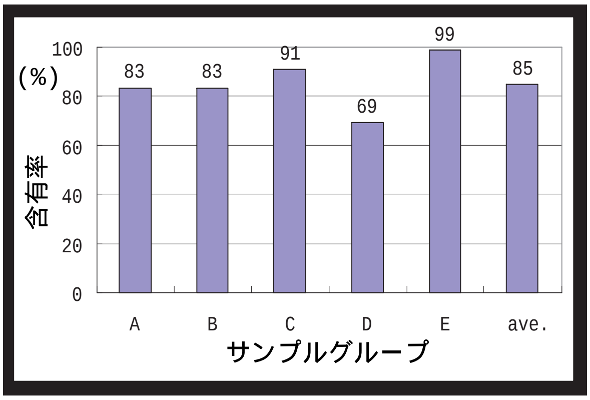

図 8: 名詞句・連鎖共起抽出のサンプルテスト結果

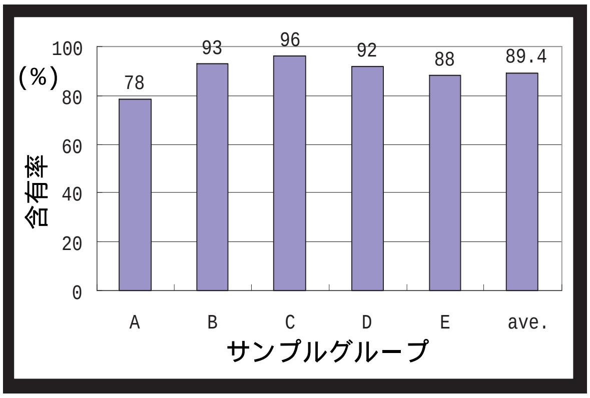

図 9: 名詞句・離散共起抽出 (2 組) のサンプルテスト結果

5.2.5 含有率からの考察

重文・離散共起抽出 (2 組) の含有率の推移 (図 5) からの考察

表現番号が大きくなるにつれて含有率が低下する傾向にあることが分かる。減少の傾向も一定ではなく、表現番号の最後のあたりで急に落ち込んでいるが、これは相互情報量の効果が現われているからである。相互情報量を利用することにより人手の労力を軽減することができた。また、相互情報量の値を足切り値をして使うことで、出力される表現を抑制することも可能である。

サンプルテスト結果 (図 6 ~ 9) からの考察

単文のサンプルテストでは、含有率がどの場合も100%となっている。単文では表現とその表現の頻度の両方が重要となる。

名詞句のサンプルテストでは、抜き出す表現によって含有率にばらつきがある。その原因は、名詞の種類については一切置き換えを行なっていないので、数詞が断片的な文字列として出力されたからである。数詞を置き換える案もあるが、それでは名詞句特有の「2月3日の節分」といった数詞を含む表現が失われてしまうため、本研究では置き換えないことにした。

名詞句・連鎖型共起抽出のサンプルテスト結果 (図 8) からの考察

サンプル E の頻度は低い値だが、含有率は99%と高い値になっている。この結果からも分かるように、頻度が低いからといって定型的な言い回しが含まれていないわけではなく、むしろ頻度が低い値の表現には定型的な言い回しが多く含まれている可能性もある。また、逆に頻度が高いからといって、含有率が高いわけでは無いこともこの表から読み取れる結果となった。

名詞句・離散型共起抽出 (2 組) のサンプルテスト結果 (図 9) からの考察

相互情報量の値が大きいからといって、定型的な言い回しであるとは限らないことが分かる。サンプル A の相互情報量の値は他の表現に比べて大きい値となっているが、定型的な言い回しの含有率は78%と、他のサンプルに比べて一番低い値となっている。これは、2回しか現われず、表現が常に決まった表現とのみ共起している表現がサンプル A に多く含まれているからである。2回しか現われない表現が即、定型的な言い回しではないとは限らないが、本研究では相互情報量の値のみしか用いておらず、頻度の情報が欠

落しているために、このような結果になってしまったのではないかと考えられる。この結果をふまえて、相互情報量を利用する場合には、頻度情報を利用することが考えられる。しかし、頻度を利用する場合には、利用するバランスが重要になってくる。語と語の結び付きの強さ、つまり相互情報量と頻度がちょうど釣り合うように評価式を定めなければならない。相互情報量と頻度情報を用いた評価式として12の式が考えれてるが、この評価式では頻度の割合が高くなり過ぎる。そのため、頻度の高い表現がやはり上位にくるようになり、相互情報量を用いる利点が薄くなってしまう。このように、相互情報量と頻度情報を共に用いる場合には、そのバランスが鍵となってくる。

表 12: 相互情報量と頻度情報を共に用いた評価式

$$
\begin{aligned}
MI(x_1,\cdots,x_n) &= P(x_1,\cdots,x_n)\cdot log_2(\frac{P(x_1,\cdots,x_n)}{P(x_1)\cdot\cdots\cdot P(x_n)}) \\
&= \frac{f(x_1,\cdots,x_n)}{N_c}\cdot log_2(\frac{\frac{f(x_1,\cdots,x_n)}{N_c}}{\frac{f(x_1)}{N_f}\cdot\cdots\cdot\frac{f(x_n)}{N_f}})
\end{aligned}
$$

6　結論

本研究では、原文の段階で抽出すべき、単文、重文、名詞句の文型の種類毎に適切な単語置き換えを行なうことによって、効率よく目的とする表現を抽出した。そして、N-gram 統計の出力結果に対しても、品詞情報を用いて必要のない表現の削除も行なった。また、出力結果を調査しやすくするために相互情報量を用いて言い回しを絞り込んだ。

上記の手法を毎日新聞 95 年度の一面記事 1 万文に対して適用し、定型的な言い回しを抽出した。抽出された表現から、共起表現辞書の見出し語となりそう表現を人手で選択した結果、単文、重文、名詞句について、定型的な言い回しの含有率は 100%, 50.5%, 87.2%であると分かった。従来に比べて断片的な表現が減少し、単文、重文、名詞句を構成する表現が効率よく抽出できることが確認された。しかし、例文を参照した検討により、本当の意味での言い回しの含有率は現在の値より下がる可能性がある。

本手法により、大規模なコーパスから容易に日本語共起表現辞書の見出し語が収集でき、日本語共起表現辞書を作成できる見通しとなった。

今後の課題としては、本手法を大規模なコーパスに対して適用することで、日本語共起表現辞書の見出し語を増やしてことがあげられる。また、他分野のコーパスに適用することで、分野に適した表現辞書の開発も可能である。

本手法では、N-gram 統計の後処理という形で単語境界を利用し、断片的な文字列の削除を行なったが、N-gram 統計の内部処理として行なうことで、処理時間の大幅な向上と処理の効率化が見込まれる。

参考文献

[1] 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, *Proceedings of the 15th International Conference on Computational Linguistics* (1994)

[2] 池原、白井、河岡：大規模コーパスからの連鎖型および離散型の共起表現の自動抽出法, 情報処理学会論文誌 Vol.36 (1995)

[3] 内野、白井、池原、新田見：置換えを用いた n-gram による言語表現の抽出, 情報処理学会自然言語処理学会第 2 回年次大会 (1996)

[4] 山田：N-gram 統計を応用した文型パターンの自動抽出法の研究, 鳥取大学卒業論文 (1998)