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。従来、辞書に登録する表現は人手で抽出されていたが、膨大なデータが必要となるため、計算機によって自動的に抽出する方法が考えられてきた。現在、N-gram 統計処理を応用して使用頻度の高い表現や定型的な言い回しを抽出する方法が提案されている。しかし大規模コーパスへの適用はコーパス自体が作られていなかったのでまだ行われていない。

そこで本研究では、重文複文を収集した CREST コーパス約 9 万文に対して、効率よくパターンを抽出するため、各品詞の置き換えを行い、連鎖共起 N-gram 抽出方法、離散共起 N-gram 抽出方法を用いて重文複文の構造パターンの抽出を試みた。

その結果、品詞の置き換えを行うことで効率よくパターンが抽出されることが確認できた。連鎖共起、離散共起 N-gram 抽出方法の両方で大規模コーパスから重文複文の構造パターンを抽出することができた。

目 次

図 目 次

表 目 次

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。従来、辞書に登録する表現は人手で抽出されていたが、膨大なデータを扱うために非常に困難である。

そこで計算機によって自動的に抽出する方法が考えられてきた。しかしその手法では膨大なデータが必要となり、不要な表現が多数混在してしまうという問題点があった。

従来の方法として、[1] 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。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 [1] で提案されている。[1] を応用した研究として単語単位に変換した原文を用いる方法 [2] や、重文複文に的を絞って適切な単語の置き換えを行う方法 [3] が行われている。

しかし大規模コーパスからの抽出はコーパスが無かったためにまだ行われていなかった。しかし最近になり電子辞書や新聞記事データ等が出版され、それを基に収集することが可能になった [4]。

そこで、本研究では各品詞の置き換えを行い、[1] の手法を用いて大規模コーパスから重文複文の構造パターンの抽出を試みる。

結果、品詞の置き換えを行うことで効率よくパターンが抽出されることが確認できた。連鎖共起、離散共起 N-gram 抽出方法の両方で大規模コーパスから重文複文の構造パターンを抽出することができた。

以下、2 章では N-gram 統計処理による表現抽出法について、3 章では本研究で行う抽出実験について、4 章では実験結果について、そして 5 章では考察について、最後に 6 章では結論と今後の課題について述べる。

2　N-gram 統計処理による表現抽出法

本研究で使用する、N-gram を使用した文字列抽出法として、[1] の方法を使用する。以下にその方法を要約して説明する。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。

2.1　連鎖共起 N-gram 抽出方法

連鎖共起 N-gram 抽出方法 [1] とは、図 1 で示す複数文中に 2 回以上の出現回数を持つ連続した文字列 α を抽出する方法である。一度抽出された文字列内部に含まれる部分文字列 β を抽出するか否かによって、次の抑制法がある。

- **無抑制型** 部分文字列でも抽出対象とする。図 1 に示す文を処理した結果を例 1 として示す。

  (例 1)「彼はいい」、「彼は」、「いい」等

- **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 1 に示す文を処理した結果を例 2 として示す。

  (例 2)「彼はいい」、「彼は」

- **強抑制型** 部分文字列は一切抽出しない。図 1 に示す文を処理した結果を例 3 として示す。

  (例 3)「彼はいい」

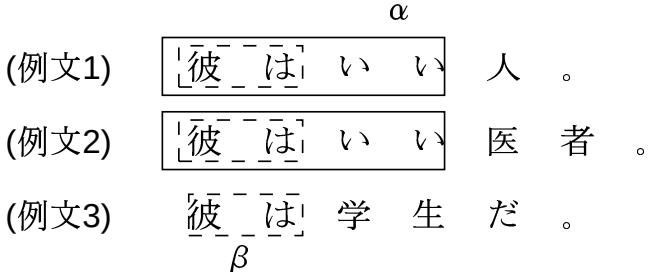

図 1: 連鎖共起表現の例

連鎖共起 N-gram 抽出方法はまとまった表現をとるのに適している。

次に連鎖共起 N-gram 抽出方法のアルゴリズムについて説明する。連鎖共起 N-gram 抽出方法のアルゴリズムは以下の手順で行われる。(参照論文 [1])

手順 1：「原文番地ファイルの作成」各レコードに原文番地をあたえる。原文番地は、言語データ上の文字番号から末尾までの部分文字列へのポインタを表す。

手順 2：「汎用ソートファイルの作成」原文番地ファイルの各レコードを、文字列の文字コード順にソートしたファイルをつくる。

手順 3：「一致文字数のカウント」汎用ソートファイルの各レコードの示す文字列を、その直後のレコードの文字列と先頭文字から比較し、一致した文字数を書き込む。

手順 4：「抽出文字数の記入」汎用ソートファイルの各レコードの文字列が、先頭から何文字抽出対象となるかを記述。

手順 5：「拡張原文番地ファイルの作成」汎用ソートファイルを原文番地順にソートし直す (拡張原文番地ファイル)。

手順 6：「有効無効判定処理」拡張原文番地ファイルの各レコードの抽出文字数を調べ、有効無効判定を行う。無効判定の方法は論文 [1] 参照。

手順 7：「再拡張汎用ソートファイルの作成」再度、汎用ソートファイルのレコード順にソートする (再拡張汎用ソートファイル)。

手順 8：「抽出文字列集計処理」再拡張汎用ソートファイルの採否表示、抽出文字数、一致文字数の関係を調べて抽出文字列を決定する。

次の例文 4 を以上で説明したアルゴリズムに適用させた例を次ページの図 2 に示す。

(例文 4)
さいたさくらがさいた。さくらのはながさいた。

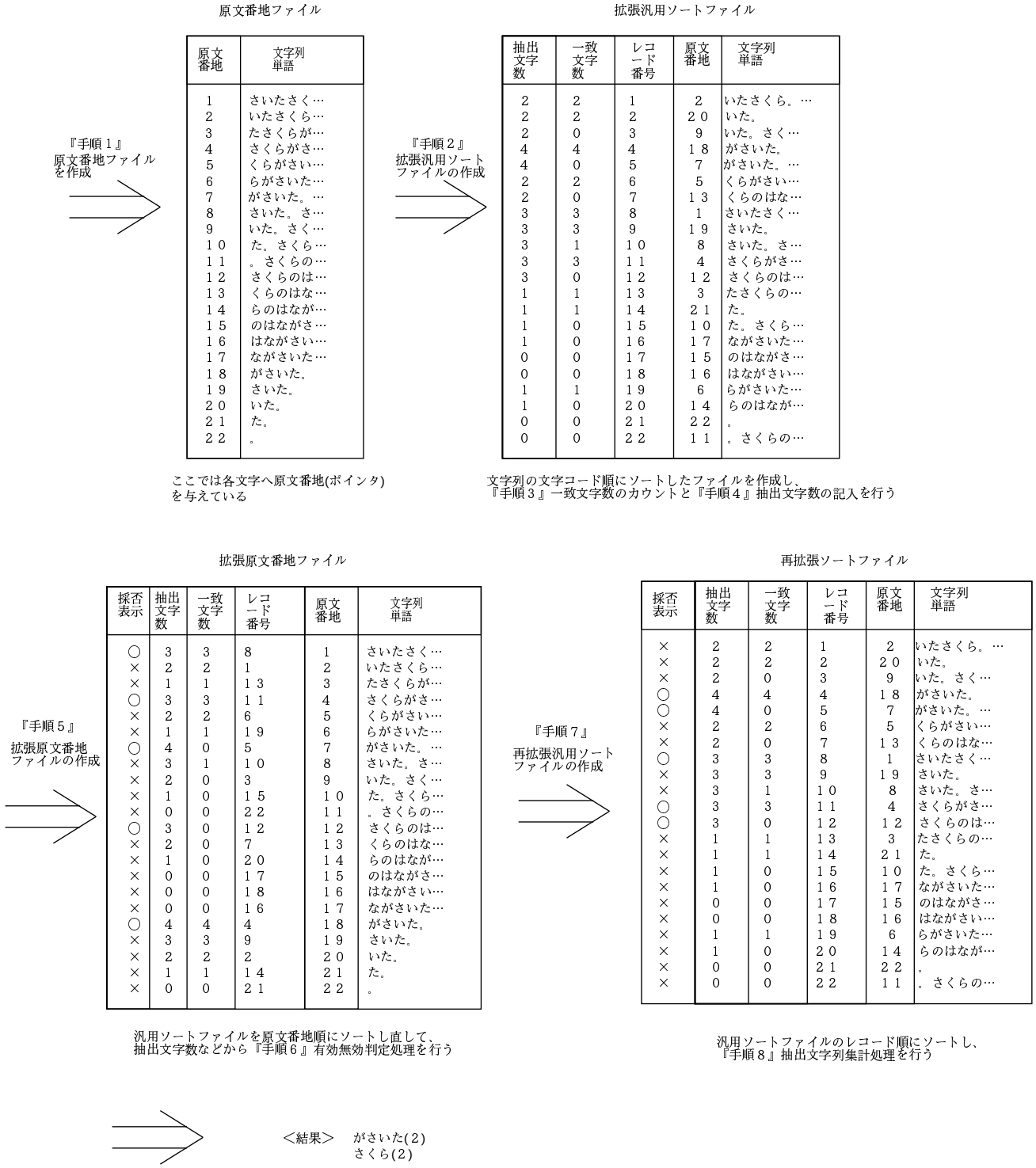

図 2: 連鎖共起表現抽出アルゴリズムの実施例

2.2 離散共起 N-gram 抽出方法

離散共起 N-gram 抽出方法 [1] とは、図 3 の γ のように原文データの離れた場所に現れ、共起する二つ以上の文字列の組合わせを抽出する方法である。組合わせ数とは離れた場所に現れた文字列の組合せの個数である。図 3 の γ は組み合わせ数 3 の文字列の抽出である。共起の仕方により、次の抑制法がある。

- **無抑制型** 共起する文字列が文中にあることだけを条件として抽出する。図 3 を処理した結果を例 4 として示す。

  (例 4)
  「彼は～彼は～選手」,「選手～有名～選手」等

- **弱抑制型** 上記の条件に加え、抽出する文字列は互いに異なるもののみ抽出する。図 3 を処理した結果を例 5 として示す。

  (例 5)
  「彼は～有名～選手」,「選手～彼は～有名」「有名～選手～彼は」,「彼は～選手～有名」「有名～彼は～選手」

- **強抑制型** 上記の二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、着目する文字列が二回以上現れないものを抽出する。図 3 を処理した結果を例 6 として示す。

  (例 6)
  「彼は～有名～選手」、「選手～彼は～有名」、「有名～選手～彼は」

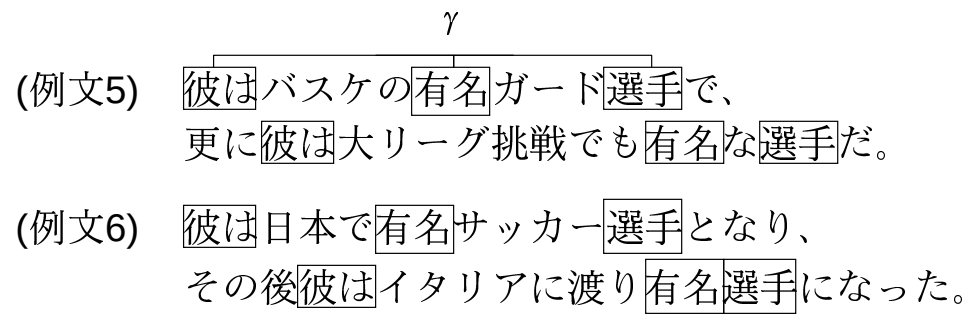

図 3: 離散共起表現の例

次に各抑制において、図 3 を処理した結果を表 1 にまとめて示す。

表 1: 離散共起表現の組み合わせ例

抽出表現	無抑制	弱抑制	強抑制
彼は ～ 彼は ～ 有名	○		
彼は ～ 彼は ～ 選手	○		
彼は ～ 有名 ～ 彼は	○		
彼は ～ 有名 ～ 有名	○		
彼は ～ 有名 ～ 選手	○	○	○
彼は ～ 選手 ～ 彼は	○		
彼は ～ 選手 ～ 有名	○	○	
彼は ～ 選手 ～ 選手	○		
有名 ～ 彼は ～ 有名	○		
有名 ～ 彼は ～ 選手	○	○	
有名 ～ 有名 ～ 選手	○		
有名 ～ 選手 ～ 彼は	○	○	○
有名 ～ 選手 ～ 有名	○		
有名 ～ 選手 ～ 選手	○		
選手 ～ 彼は ～ 有名	○	○	○
選手 ～ 彼は ～ 選手	○		
選手 ～ 有名 ～ 選手	○		

このようにして、離散共起抽出方法は長い文における離れた場所にある表現を抽出することができるため、重文・複文における表現も発見可能である。

次に離散共起 N-gram 抽出方法のアルゴリズムについて説明する。離散共起 N-gram 抽出方法は連鎖共起 N-gram 抽出方法のアルゴリズムで述べた手順 7 まで同じ手順で行う。そして前準備として連鎖共起 N-gram 抽出方法で抽出された文字列に文字列番号を付与する。次に以下の手順で行う。

手順 9：「再拡張汎用ソートファイルの再ソート」 再拡張汎用ソートファイルを原文番地順にソートし、拡張原文番地ファイルのレコード順に戻す。

手順 10：「文番号の付与」 各レコードに文番号を付与する。

手順 11：「ファイルの圧縮」 文番号、文字列番号、抽出文字数、原文番地の 4 つの欄以外は削除する。そして文字列番号欄の値がないレコードを削除する (離散共起型圧縮ファイル) 。

手順 12：離散型共起表現の抽出とカウント　文字列番号の全ての組合せを書き出し、同一の組数をカウントする。

以上で説明したアルゴリズムを連鎖共起抽出法アルゴリズムで用いた例文 4 で適用させた例を図 4 に示す。

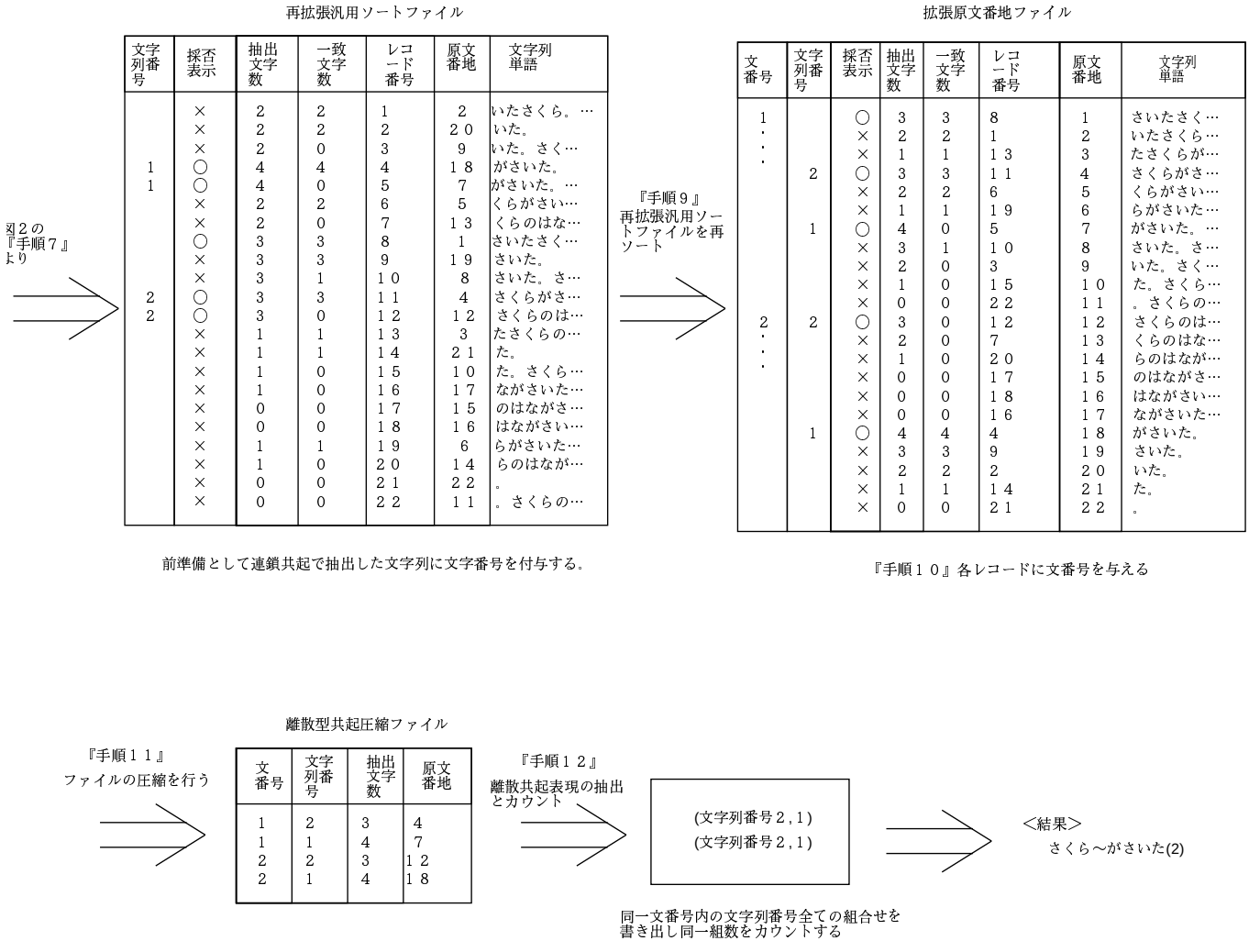

図 4: 離散共起表現抽出アルゴリズムの実施例

2.3 単語に着目した抽出

N-gram 抽出法では原文データに対して文字単位もしくは単語単位による抽出を行うことができる。本研究では断片的な文字列の抽出を少なくするために、原文データに対してあらかじめ形態素解析を行い、単語単位による抽出を行なう。本研究で行う単語単位による抽出は、ある 1 文字の全角文字を単語に割与える方法で行う。単語単位におけるパターン抽出までの流れを以下の図 5 に示す。図 5 内の単語テーブルはある単語に対して与える文字を対応させたテーブルを意味する。図 5 内の下線部は計算機による処理を表している。図 5 内の二重枠内は入出力データを意味する。

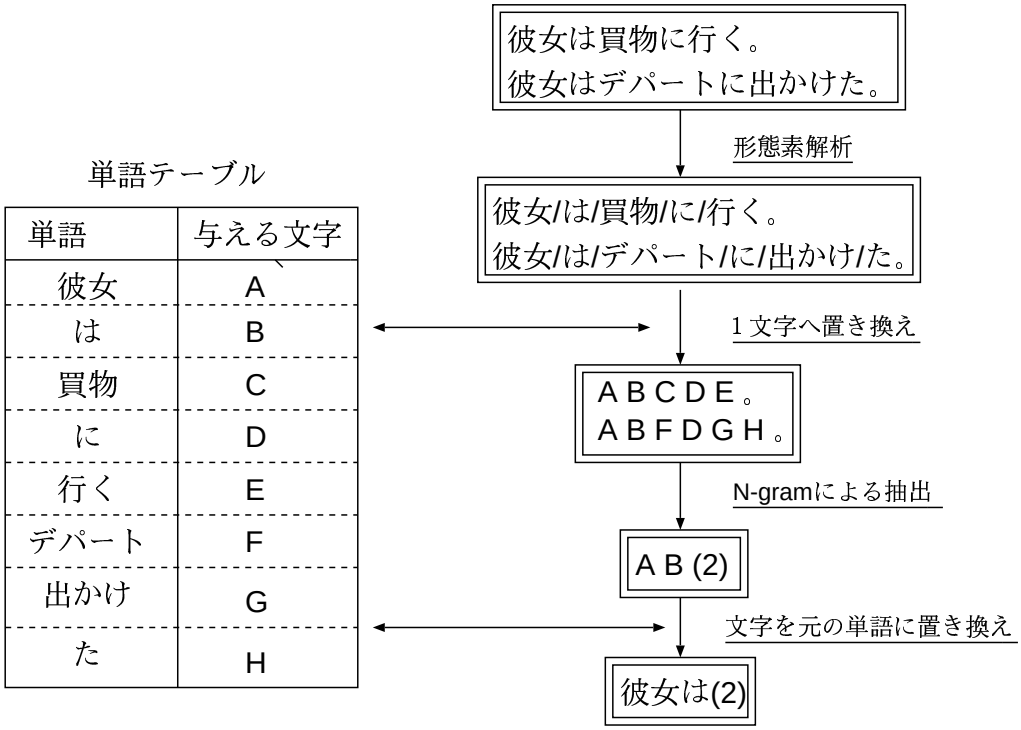

図 5: 単語単位によるパターン抽出の流れ

図 5 より、まず原文データに対し形態素解析を行う。そして単語テーブルより参照して単語に対応した一文字に置き換える。そして置き換えたデータで N-gram 抽出を行う。そして得られた結果は、一文字に置き換え後のデータが出力される。そこでもう一度単語テーブルを参照して、一文字から対応した単語に置き換える。以上の過程で、単語単位による抽出を行う。

3　抽出実験

3.1　連鎖共起 N-gram 抽出法による実験

本実験では連鎖共起 N-gram 抽出法では弱抑制型の抽出を行う。無抑制型では全ての部分文字列を抽出してしまい、強抑制では部分文字列が抽出できない問題があるため行わない。本実験では調査対象を単語数が 5 以上で構成される単語列で行い、品詞の置き換えを行った場合 (3.3 節参照) についても調査をする。また単語列を構成する単語数を増やすことでパターン抽出精度の向上が期待できると考え、単語列を構成する単語数が 7 以上、9 以上についても調査を行う。5 以下で構成される単語列は文型パターンを成す抽出が少ないため本実験では行わないこととする。

ここでいう単語数とは、次の例 7, 例 8 に示す様な単語列を構成する単語の数を意味する。例中の'/' は単語境界線を意味する。

(例 7)　彼/は/外/に/行く　　単語数 5

(例 8)　彼女/は/今日/は/町/に/買物/に/行く　　単語数 9

3.2　離散共起 N-gram 抽出法による実験

離散共起 N-gram 抽出法による実験は以下の 3 つの種類で行い、品詞の置き換えを行った場合 (3.3 節参照) においても調査する。

実験条件 1) 弱抑制型で単語列の組み合わせ数 3

実験条件 2) 弱抑制型で単語列の組み合わせ数 4

実験条件 3) 強抑制型で単語列の組み合わせ数 3

ここでいう組み合わせ数とは、次の例 9,10 に示す様な単語列の組み合わせの数である。例中の'/' は単語境界線を意味する。

(例 9)　彼/は 〜 を/して 〜 を/する　　組み合わせ数 3

(例 10)　彼/は 〜 で/あり 〜 彼女/は 〜 で/ある　　組み合わせ数 4

3.3 品詞の置き換え

本研究では重文複文のパターン抽出を効率良く行うために、連鎖共起 N-gram 抽出法による実験、離散共起 N-gram 抽出法による実験それぞれに品詞の置き換えを以下の 5 つの種類で行う。

置き換え A) 単語単位

置き換え B) 名詞→ N

置き換え C) 名詞→ N、動詞→ V

置き換え D) 名詞→ N、「N の N」→ N、「NN…」→ N、「形容詞+N」→ N

置き換え E) (置き換え D) かつ動詞→ V

本研究で行う品詞置き換えについて名詞や連続名詞などは重文複文の構造に関係の無い品詞であると考えられるため置き換えている。動詞については文の核をなす品詞であるため、置き換えは不適であると考えられるが、文の構造を効率よく抽出するために置き換えている。

次に以下に示す例文を上の各品詞置き換えを行った場合の例を示す。例中の'/' は単語境界を意味する。

> (例文 7) 私の研究はどうやら見通しが明るくなりましたので逐次研究結果をお知らせします。

置き換え A) 私/の/研究/は/どうやら/見通し/が/明る/くなり/まし/た/ので/逐次/研究/結果/を/お知らせし/ます/。

置き換え B) N/の/N/は/どうやら/N/が/明る/くなり/まし/た/ので/N/N/N/を/お知らせし/ます/。

置き換え C) N/の/N/は/どうやら/N/が/明る/V/まし/た/ので/N/N/N/を/V/ます/。

置き換え D) N/は/どうやら/N/が/明る/くなり/まし/た/ので/N/を/お知らせし/ます/。

置き換え E) N/は/どうやら/N/が/明る/V/まし/た/ので/N/を/V/ます/。

3.4 精度調査方法

本研究では重文複文を中心に収集した CREST 例文約 9 万文 [4] に対して N-gram を応用した文字列抽出法 [1] を行う。CREST 例文は長短様々な重文複文を収集している。更に CREST 例文には英語の対訳も掲載されている。CREST 例文において、日本文一文を構成する平均単語数は約 10.6 単語である。CREST 例文の一部を以下に示す。

(例文 8) 私は慶応義塾大学で教育を受けたので、ビジネスに関する理解、とくに財務、マーケティング及び戦略計画の分野における理解が深まりました。

対訳>> My education at Keio has enhanced my understanding of business, especially in the areas of finance, marketing and strategic planning.

(例文 9) 私はそう言っている人を知っている。

対訳>> I know a man who says so.

(例文 10) 彼は詩人でもあり外交官でもある。

対訳>> He is a poet and diplomat.

(例文 11) 私は書くペンがない。

対訳>> I have no pen to write with.

次に各実験を行い、出力されたデータから抽出頻度数の多い上位 100 データを対象に、手作業により重文複文のパターンであるかを調査する。調査の例として次の例 11、例 12 の様なデータを抽出したとする。

(例 11)　を聞いて喜んだ

(例 12)　彼の言うことは

例 11 の場合、「〜を聞いて、〜喜んだ」といった並列した文に分解できるため、重文のパターンであると判断する。例 12 の場合、「言うこと」で埋め込み文となっているため、複文のパターンであると判断する。

4 実験結果

4.1 連鎖共起 N-gram 抽出法による実験結果

連鎖共起 N-gram 抽出法による実験を行った。連鎖共起 N-gram 抽出法によって抽出したパターン例を表 2 に示す。表 2 の抽出パターン中の'/' は単語境界を意味する。表 2 の単語数は抽出したパターンが構成する単語数を意味する。表 2 の置き換え内の「A,B」等は品詞の置き換え (3.3 節参照) 内の「置き換え A, 置き換え B」に対応している。表 2 の「評価」は重文複文のパターンを構成していると思われる表現を◯とし、そうでないものを×としている。

表 2: 連鎖共起 N-gram 抽出法による抽出パターン例

抽出パターン	単語数	置き換え	評価
に/なる/と/思い/ます	5	A	◯
役/に/立つ/もの/と/思い/ます	7	A	◯
手/に/汗/を/握っ/て/勝負/を/眺め	9	A	◯
と/いう/N/が/ある	5	B	◯
N/は/N/の/ある/N/だ	7	B	◯
N/の/N/に/当たっ/て/N/が/ある	9	B	◯
V/て/N/を/V/	5	C	◯
N/に/V/て/V/V/た	7	C	◯
N/を/V/N/に/N/を/V/た	9	C	◯
N/と/いう/N/だ	5	D	◯
N/は/N/に/対する/N/が	7	D	◯
N/を/V/て/V	5	E	◯
N/を/V/て/N/を/V	7	E	◯
に/気/が/着い/た	5	A	×
御/迷惑/を/お/かけ/し/て	7	A	×
大家/は/彼/に/部屋/を/出て/行く/よう	9	A	×
N/N/の/N/を	5	B	×
N/は/N/の/N/を/N	7	B	×
N/の/N/に/は/N/の/N/が	9	B	×
N/の/N/を/V	5	C	×
N/の/N/が/V/V/た	7	C	×

次に表 3 に連鎖共起 N-gram 抽出法による全抽出種類数を示す。

表 3: 連鎖共起データの全抽出種類数

単語数	置き換え A	置き換え B	置き換え C	置き換え D	置き換え E
5 以上	4,698	40,259	77,562	33,792	64,562
7 以上	1,031	11,784	38,374	9,194	30,523
9 以上	252	2,401	10,900	2,165	8,403

表 3 より、単語数を増やすことで全抽出種類数が下がることが示された。
次ページの表 4 に連鎖共起 N-gram 抽出法による抽出パターンの調査結果を示す。

表 4: 連鎖共起データの調査結果

上位 100 データ対象

単語数	置き換え A	置き換え B	置き換え C	置き換え D	置き換え E
5 以上	21 %	2 %	57 %	14 %	63 %
7 以上	45 %	9 %	89 %	49 %	93 %
9 以上	57 %	46 %	100 %	69 %	100 %

表 4 より、抽出する単語数が多い程、パターンの抽出精度が向上することが示された。表 4 より、名詞を置き換える「置き換え B」を行うと抽出精度が悪くなったが、名詞と動詞を置き換える「置き換え C」では抽出精度が良くなった。更に「置き換え B」、「置き換え C」よりも、連続した名詞や「N の N」等の置き換えを行った「置き換え D」、「置き換え E」で精度が向上することが示された。

4.2 離散共起 N-gram 抽出法による実験結果

離散共起 N-gram 抽出法による実験を行った。離散共起 N-gram 抽出法による抽出パターン例を表 5 に示す。表 5 の条件内の「1,2」等は離散共起 N-gram 抽出法による実験 (3.2) 内の「実験条件 1, 実験条件 2」に対応している。表 5 の置き換え内の「A,B」等は品詞の置き換え (3.3) 内の「置き換え A, 置き換え B」に対応している。表 5 の「評価」は重文複文のパターンを構成していると思われる表現を○とし、そうでなかったものを×としている。

表 5: 離散共起 N-gram 抽出法による抽出パターン例

抽出パターン	条件	置き換え	評価
夏休みに〜へなり〜たいと思う	1	A	○
には自分の〜に対する権利がある〜のは当然のことだ	1	A	○
を変えて〜て何度〜何度も〜繰り返した	2	A	○
彼に払うべき〜を払った〜上に礼を	3	A	○
N に向け〜N 隻〜N に襲われた	1	B	○
の NNNN〜N 数を N〜へと NN し〜と NN した	2	B	○
N は N よりたとい〜としても〜はない	3	B	○
N の N ダースは〜NN 日までに、N の N〜に V れなければ V ません	1	C	○
N は N つの N に V〜、N 番〜の N には〜N が VV	2	C	○
の N で〜V れる〜N 週〜で V	3	C	○
N で発行さ〜N 週〜である	1	D	○
N は、N 年 N 日に〜N 日に N で、N 日〜、N 日に N〜N で開催される N	2	D	○
VN によって N〜N 一 N が V れ〜N が V れ V ます	1	E	○
N は「N」N〜の N で V れる〜N で〜N 週〜で V	2	E	○
部と〜の部分〜構成されている	1	A	×
の解決策〜一つの戦略〜の戦略へ	1	A	×
を始めたの〜たのが〜が５〜５時	2	A	×
N は N 上で〜N 数を N〜へと N し	1	D	×

次に離散共起 N-gram 抽出法による全抽出種類数を表 6 に示す。

表 6: 離散共起データの全抽出種類数

実験条件	置き換え A	置き換え B	置き換え C	置き換え D	置き換え E
弱抑制型, 組み合わせ数 3	913	1,805	1,935	1,105	1,291
弱抑制型, 組み合わせ数 4	1,259	4,281	4,077	1,702	2,268
強抑制型, 組み合わせ数 3	13	5	3	9	4

表 6 より、強抑制による全抽出種類数が少なくなることが示された。次に離散共起 N-gram 抽出法による抽出パターンの調査結果を次ページの表 7 に示す。

表 7: 離散共起データの調査結果

上位 100 データ対象

実験条件	置き換え A	置き換え B	置き換え C	置き換え D	置き換え E
弱抑制型, 組み合わせ数 3	29 %	10 %	30 %	19 %	41 %
弱抑制型, 組み合わせ数 4	34 %	14 %	37 %	29 %	46 %
強抑制型, 組み合わせ数 3	35 % (5)	60 % (3)	100 % (3)	66 % (6)	100 % (4)

表 7 より、品詞の置き換えによる抽出精度の傾向は連鎖共起表現による実験結果と類似していた。単語列の組み合わせ数を増やすことで、精度が多少あがることもわかった。強抑制に注目すると、抽出種類数は少ないが抽出精度は高いことがわかる。

5　考察

5.1　名詞の置き換えについての考察

表 4、表 7 より名詞を置き換える「置き換え B」でパターン抽出精度が悪くなった。置き換え後の抽出データを調べると大半が「N は N の NN に」等の名詞句の抽出であった。そのため名詞や連続名詞を置き換える「置き換え D」を行うことで精度が向上したと考えられる。

5.2　動詞の置き換えについての考察

表 4、表 7 より動詞を置き換えることによりパターン抽出精度が向上した。しかし抽出したパターンから意味まで理解するのは、動詞は文の中心となる品詞であるため困難な場合が多い。例を例 13、例 14 に示し、各例の原文も示す。

(例 13)　V/て/V/V/た

(原文)
はね/て/飛ん/でいっ/た

し/て/逃げ/てしまっ/た

当たっ/て/ 輝い/てい/た

連れ/て/行っ/てくれ/た

つかっ/て/立っ/てい/た

(例 14)　N/を/V/て/N/を/V/た

(原文)
知らせ/を/聞い/て/肝/を/つぶし/た

私/を/だまし/て/金/を/奪っ/た

顔/を/見/て/昔/を/思い出し/た

茶わん/を/落とし/て/ひび/を/入れ/た

猛勉強/を/し/て/健康/を/害し/た

例 13、例 14 は重文複文の構造であり、原文を見ると構造的にも類似しているが、意味的に類似していないものが多いと考えられる。

5.3 離散共起 N-gram 抽出法の強抑制についての考察

表 6 より強抑制による全抽出種類数が少なかった。原因は本研究で使用したコーパスが一般的に用いられる重文複文を収集したため、1 文を構成する単語数が少ないと考えられる。そのため実験を行う際に、「離散共起 N-gram 抽出法における強抑制型組み合わせ数 3」の抽出条件を満たせなかったと考えている。抽出条件を満たせなかった例として以下の図 6 中の文を示す。

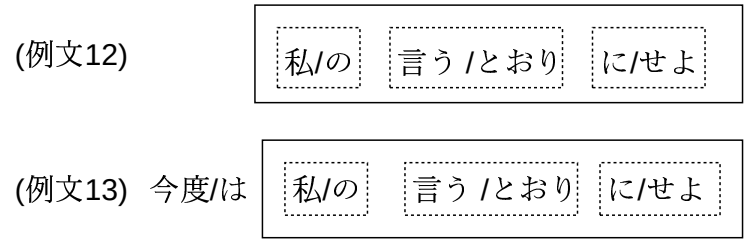

図 6: 離散共起 N-gram 抽出法の強抑制による抽出例 1

図 6 中の文は単語数 6 からなる文と単語数 8 からなる文である。図 6 中の文の場合、離散共起 N-gram 抽出を行ってもパターンは抽出されず、連鎖共起 N-gram 抽出として「私の言うとおりにせよ」が抽出される。次に以下の図 7 中の文を示す。

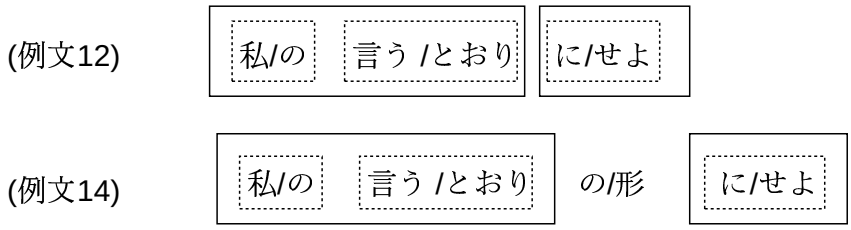

図 7: 離散共起 N-gram 抽出法の強抑制による抽出例 2

この図 7 中の文も単語数 6 からなる文と単語数 8 からなる文である。この場合でも離散共起 N-gram 抽出を行ってもパターンは抽出されず、連鎖共起 N-gram 抽出として「私の言うとおり」と「にせよ」が抽出される。

次に以下の図 8 中の文を示す。

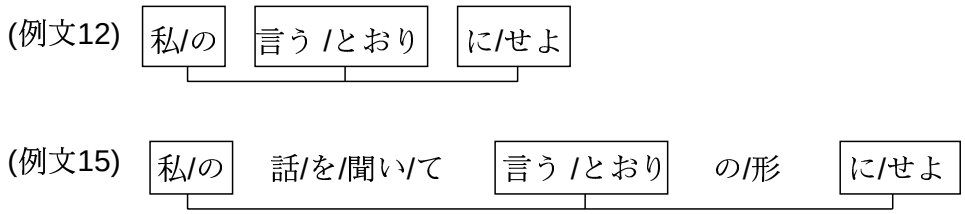

図 8: 離散共起 N-gram 抽出法の強抑制による抽出例 3

この図 8 中の文は単語数 6 からなる文と単語数 11 からなる文である。この場合で離散共起 N-gram 抽出として「私の〜言うとおり〜にせよ」が抽出される。以上で示すように、本実験においては離散共起 N-gram 抽出の強抑制による抽出は文を構成する単語数、構造等により制限を受けて、抽出数が少なくなったと考えられる。

しかし離散共起 N-gram 抽出の強抑制による抽出は抽出種類数が少ないものの抽出精度は高かった。

5.4 離散共起 N-gram 抽出法の弱抑制についての考察

離散共起表現 N-gram 抽出法における弱抑制では、重なった表現が多数抽出された。表 6 より、「組み合わせ数 4」による全抽出数は「組み合わせ数 3」による全抽出数の、約 2 倍抽出されたことがわかる。しかし「組み合わせ数 4」による実験で抽出されたデータを調査したところ、重なった表現の抽出が「組み合わせ数 3」による抽出データよりも更に多く見られた。したがって重なった表現の抽出により全抽出数の増加に影響したと考えている。本実験において「組み合わせ数 4」による抽出は効果があまり見られなかったと思われる。

5.5　抽出頻度についての考察

本研究で抽出精度の調査を行った際に、抽出頻度上位に単文等のパターンが多く現れ、頻度下位の方には重文複文パターンが多く現れた様に感じられた。この傾向の理由として、一般的に重文複文のパターンは複数の単文パターンを重ねたものであるため、単文パターンが頻度上位にどうしても集まってしまい、複数の単文パターンが重なった重文複文パターンが下位に集まってしまったと考えられる。そのため本研究で行った、抽出頻度上位 100 データに対する調査だけでは、頻度下位に現れた重要なパターンが抽出できていないことが考えられる。この問題を解決するためには、データ中のどの頻度にどれほど重文複文パターンが出現したのかといった分布密度等の調査を行い、ターゲットを絞った抽出が必要である。

6　おわりに

6.1　結論

本研究では、各品詞の置き換えを行い N-gram 抽出法 [1] を用いて、大規模コーパスから重文複文の構造パターンの抽出を行った。

実験の結果、品詞の置き換えにおいて最もパターンの抽出精度が良好だったのは名詞や連続名詞や動詞を置き換える「置き換え E」であった。しかし抽出したパターンから文の意味等を理解するのは困難であると思われる。そこで文の意味を残すことを考慮すると、本研究では名詞や連続した名詞を置き換える「置き換え D」が最も効率良くパターンを抽出できたと考えている。精度としては連鎖共起 N-gram 抽出法において、単語のままで行う「置き換え A」より約 10 %向上した。

離散共起 N-gram 抽出法の強抑制による実験ではパターンの抽出数が少なくなり、良好な結果を得られなかった。しかし抽出数は少ないものの、抽出したパターンは整った構造であった。

6.2　今後の課題

今後の課題として、コーパスをさらに拡大化し抽出パターン種類の充実化が必要である。そして離散共起 N-gram 抽出法の強抑制の抽出数の現象を改善するための新たな単語単位での抽出法の考案が必要である。

また、本手法を英語に適用していくことで、英語と対応づけた重文複文パターンを抽出する必要がある。

参考文献

[1] 池原、白井、河岡：大規模コーパスから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)

[2] 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)

[3] 斎藤:大規模コーパスからの重文復文の統語構造の自動抽出, 鳥取大学卒業論文 (2000)

[4] 村上、池原、徳久：日本語英語の文対応の対訳データベースの作成,「言語、認識、表現」第 7 回年次研究会,(2002.11)