構造的類似文の自動検出と縮退方法

尾高 智大　村上 仁一　徳久 雅人　池原 悟
鳥取大学工学部知能情報工学科
{odaka,murakami,tokuhisa,ikehara}@ike.tottori-u.ac.jp

1　はじめに

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、辞書に登録する表現を人手で抽出することは、膨大なデータを扱うため困難である。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ている。従来の方法として、N-gram を応用した連鎖共起表現や離散共起表現を抽出する手法 [1] が提案されている。[1] を応用した研究として単語単位に変換した原文データを用いる方法 [2] や、重文複文に的を絞って適切な単語の置き換えを行い定型的な言い回しを抽出する方法 [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。

2　構造パターンの抽出

2.1　N-gram によるパターンの抽出

本研究では、パターンを抽出する方法として、N-gram を応用した文字列抽出法 [1] を使用する。[1] の方法を以下に説明する。

2.1.1　連鎖共起 N-gram 抽出法

連鎖共起 N-gram 抽出法 [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** すべての部分文字列を抽出対象する。図 1 中の例文 1〜3 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

  無抑制型は膨大な候補を出力するため通常使用されない。
- **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 中の例文 1〜3 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 中の例文 1〜3 を処理した結果を例 3 として示す。

  (例 3)「彼はいい」

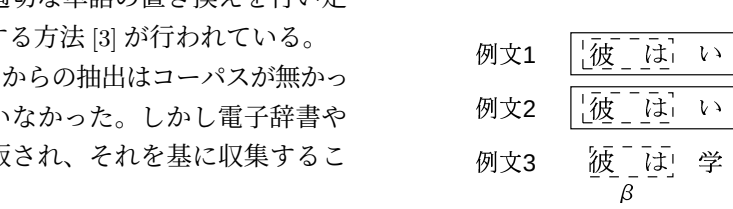

図 1: 連鎖共起表現の例

2.1.2　離散共起 N-gram 抽出法

離散共起 N-gram 抽出法 [1] とは、図 2 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 2 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 2 中の例文 4、5 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 2 中の例文 4、5 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 2 中の例文 4、5 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」

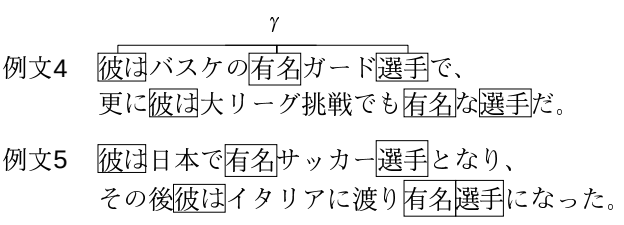

図 2: 離散共起表現の例

2.1.3 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。

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。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。

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

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 かつ動詞→ V
置き換え D) 名詞→ N かつ「N の N」→ N かつ「NN …」→ N かつ「形容詞+N」→ N
置き換え E) (4) かつ動詞→ V

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 6) 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。

(例文 7) 私は書くペンがない。

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 7、例 8 の様なデータを抽出したとする。

(例 7)　を聞いて喜んだ
(例 8)　彼の言うことは

例 7 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 8 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。

表 2: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: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 %

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 1 に示す。表 1 の抽出パターン中の'/' は単語境界を意味する。表 1 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 1 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 1 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 1: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	○
N/は/N/の/ある/N/だ	7	B	○
V/て/N/を/V/	5	C	○
に/気/が/着い/た	5	A	×
N/は/N/の/N/を/N	7	B	×
N/の/N/を/V	5	C	×

次に表 2 に連鎖共起 N-gram 抽出法による全抽出種類数を示す。表 2 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。

表 3 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 3 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 3 より、「置き換え B」を行うと抽出精度が悪くなったが、「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」に連続した名詞や「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 4 に示す。表 4 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 4 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 4 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 4: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	○
彼に払うべき〜を払った〜上に礼を	3	A	○
N は N よりたとい〜としても〜はない	3	B	○
の N で〜V れる〜N 週〜で V	3	C	○
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	×

次ページに離散共起 N-gram 抽出法による全抽出種類数を表 5 に示す。表 5 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。

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 6 に示す。表 6 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。

表 5: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: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)

5　考察

5.1　名詞の置き換えについての考察

表 2、表 4 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 2、表 4 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 9、例 10 に示す。

(例 9)　V/て/V/V/た

(例 10)　N/を/V/て/N/を/V た

例 9、例 10 は重文複文の構造であることは判るが、意味を判断することは困難であると考えられる。

5.3　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 5 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。

5.4　離散共起 N-gram 抽出法の強抑制についての考察

表 4 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。しかし抽出種類数が少ないものの抽出精度は高かった。

6　おわりに

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、品詞の置き換えは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。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。

今後の課題として新たな単語単位での抽出法の考案と、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。

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)