修 士 論 文

名詞と助数詞の呼応関係のコーパスからの自動獲得

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

矢野 修平

2009 年 3 月

修 士 論 文

名詞と助数詞の呼応関係のコーパスからの自動獲得

指導教官　白井 清昭 准教授

審査委員主査　白井 清昭 准教授
審査委員　島津 明 教授
審査委員　東条 敏 教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

0710072 矢野 修平

提出年月: 2009 年 2 月

概 要

日本語では名詞を数える際には一般的に助数詞を使い, その種類も豊富である. さらに, ある名詞を数える際には特定の助数詞のみが使われるという名詞と助数詞の呼応関係が存在する. 日本語の生成や解析において, 助数詞を適切に取り扱うためには, 呼応する名詞と助数詞の知識が必要となる. 例えば, この知識は自然言語処理の課題の一つである語義曖昧性解消の精度向上に利用できる. 本研究では, コーパスから呼応関係にある名詞と助数詞の組を大量に自動獲得し, 呼応する助数詞の情報を含む名詞の辞書を構築することを目指す.

関連する研究として, Bond らは, 個々の名詞の代わりに名詞の意味クラスに対して呼応する助数詞を割り当てている. また, Sornlertlamvanich は, タイ語を対象とし, 本研究と同様にコーパスから呼応関係にある名詞と助数詞の組を獲得する手法を提案している. その際, 人手で作成した抽出パターンを用いている. これに対し本研究では, 同じ意味クラスを持つ名詞は常に同じ助数詞と呼応関係にあるわけではないことから, コーパスから名詞と助数詞の呼応関係を網羅的に収集するというアプローチを取る. また, 名詞と助数詞は様々なパターンで出現し, 人手で抽出パターンを書き尽くすのは困難であるため, パターンマイニングにより抽出パターンの学習を行う.

はじめに, 予備調査として簡単なパターンマッチによってコーパスから (n, c) を抽出することを試みた. ここで, n は名詞, c は助数詞であり, (n, c) は呼応関係にある名詞と助数詞の組とする. ここでは, 「名詞 + 数字 + 助数詞」「数字 + 助数詞 + (の) + 名詞」「名詞 + (が) + 数字 + 助数詞」といった単語列にマッチしたときに名詞と助数詞を抽出するパターンを作成した. これら 3 つのパターンは, 呼応する名詞と助数詞が出現する典型的な単語の並びであると考えられる. しかし, 獲得された組の中には誤りが多く, 最も精度の高い抽出パターンでも, 獲得した名詞と助数詞が呼応関係にある割合は 54% であった.

この予備調査を踏まえ, 本研究では, 名詞と助数詞の呼応関係を正確に獲得するために, パターンマイニングより (n, c) を抽出するパターンを自動獲得する手法を提案する. 本研究で提案する手法は, 「例文検索」「抽出パターン獲得」「(n, c) の抽出」の 3 つのステップに分けられる. これらを反復することによって (n, c) を漸進的に獲得し, (n, c) のデータベース NC-DB に追加する. 初期の NC-DB にはシードを使用する. シードとは少量の正しい (n, c) の組であり, (n, c) の抽出パターンを学習する元となるデータである. シードの作成方法には様々な手法が考えられるが, ここでは人手で与えるものとする. また, 抽出パターンや (n, c) を獲得する際に用いるコーパスには, 予備調査の結果を踏まえ, (n, c) を正確に検出するために (1) 数字の連続はまとめてひとつの単語とする. (2) 助数詞は単位として使用されるものを除外する. (3) 名詞は抽象名詞等の数えられないものは除外し, 連続する名詞はひとつにまとめる. 以下の前処理の後, まずはじめに NC-DB に登録されている (n, c) について, 同一文中に名詞 n と助数詞 c が出現する例文をコーパスから検索する. 次に, 得られた例文に頻出する単語列をマイニングし, (n, c) を抽出するためのパターンを獲

得する. その際, 単語を表記のまま扱う手法と原型に直して扱う手法の 2 通りの手法が考えられる. 2 つの手法で作成されたパターンの候補をそれぞれ 3 つの評価基準に照らし合わせ, 獲得するパターンを選択する. 1 つ目は, そのパターンにマッチする例文の数. 2 つ目は, 獲得された (n, c) のうち, 正しい (n, c) の組が占める割合. このとき, 「正しい (n, c) の組」の定義に応じて 2 通りの手法が考えられる. シードのみを正しい (n, c) の組とみなす手法 (A) と, その時点で獲得された (n, c) を全て正しいとみなす手法 (B) である. 3 つ目の評価基準は, 獲得した (n, c) のうち, 最も頻出する (n, c) の割合である. 最後に, 得られた抽出パターンを用いて, コーパスから (n, c) の組を新たに獲得し, NC-DB に追加する.

提案手法を評価する実験を行った. 実験に使用するコーパスとして日経新聞の 2006 年の新聞記事データを用いた. また, 抽出パターンの作成方法として, 手法 (1) と手法 (2) の 2 つが, 抽出パターンの獲得方法として手法 (A) と手法 (B) の 2 つが存在するため, これらを組み合わせた 4 つの手法を用いて (n, c) の獲得を試みた. 手法 (1)(A) では, 「例文検索」「抽出パターン獲得」「(n, c) の抽出」の操作を 3 回反復した時点で新しい抽出パターンが獲得されなかったため, 処理を終了した. 獲得した (n, c) の数は 1,845 組あり, そのうち正しい (n, c) の推定数は 1,482 組, 正解率はおよそ 80%であった. 手法 (1)(B) では, 反復回数が 2 回の時点で処理を終了した. 獲得した (n, c) の数は 1,721 組あり, そのうち正しい (n, c) の推定数は 1,454 組, 正解率はおよそ 84%であった. 手法 (2)(A) では, 反復回数が 3 回の時点で処理を終了した. 獲得した (n, c) の数は 1,817 組あり, そのうち正しい (n, c) の推定数は 1,412 組, 正解率はおよそ 78%であった. 手法 (2)(B) では, 反復回数が 2 回の時点で処理を終了した. 獲得した (n, c) の数は 1,704 組あり, そのうち正しい (n, c) の推定数は 1,437 組, 正解率はおよそ 84%であった. 実験により獲得された (n, c) を見てみると, 新聞記事によく使われるような名詞に対して, それと呼応する助数詞が獲得されていることがわかった. このことから, 用いるコーパスを変えることにより, 専門用語に対しても名詞と助数詞の呼応関係の獲得が期待できる. また, 獲得されたパターンを見ると, 人手では作成の難しい精緻なパターンが学習されたことがわかった.

目 次

第1章　はじめに

1.1　研究の背景

日本語では名詞を数える際には一般的に助数詞を使い, その種類も豊富である. さらに, 例えば生徒は「人」では数えるが「個」では数えないといったように, ある名詞を数える際には特定の助数詞のみが使われるという名詞と助数詞の呼応関係が存在する. 名詞と助数詞の呼応関係を獲得することにより, 以下のようなメリットがある.

- 日本語文の生成
  機械翻訳等で日本語文を生成するときに利用される.

  (1) There are five buildings in the block.

  (2) このブロックにはビルが5棟ある.

  例えば, (1) の英文を (2) のように「ビルが5棟」と訳すためには, 「ビル」という名詞が「棟」という助数詞と呼応関係にあるという知識が必要となる.

- 日本語文の解析
  名詞と助数詞の呼応関係は, 日本語文の係り受け関係を解析する際にも利用できると考えられる. 係り受け解析器を用いて解析した結果と名詞と助数詞の呼応関係に矛盾が生じた場合, その解析結果は誤っている可能性があると判断することができる.

  (1) 3体の猿の彫刻.

  (2) (猿, 匹)

  例えば, (1) の文を解析したとき, 解析器が「3体」が「猿」に係ると判断したとする. しかし, この文では「3体」は「彫刻」に係るという解釈が正しい. このとき, (2) の知識, すなわち「猿」と呼応関係にある助数詞は「匹」であるという知識を持っていたなら, 「3体」と「猿」が対応しているという結果が誤りであると判断することができる.

- 語義曖昧性解消
  語義の曖昧性解消は自然言語処理において重要な課題の一つである. これは複数の意味を持つ単語に対して, その単語がどの意味で使われているかを判断する問題である. 名詞と助数詞の呼応関係は, 語義の曖昧性解消にも利用できると考えられる.

- マウス
  (1) 実験用に改良されたハツカネズミ
  (2) コンピュータの入力装置
- フィルム
  (1) 写真フィルム
  (2) 映画フィルム

例えば,「マウス」という名詞の場合, 呼応する助数詞が「匹」であるなら上記の (1) の意味で,「個」という助数詞と呼応関係にあるならば (2) の意味で使用されていると判断することができる. 同様に,「フィルム」という名詞の場合,「枚」と呼応関係にあるなら (1) の意味で使用されており,「本」と呼応関係にあるなら (2) の意味で使用されていると判断することができる.

このように, 名詞と助数詞の呼応関係は自然言語処理の分野において重要な知識であるといえる. 本研究では, 呼応する助数詞の情報を含む, 大規模な名詞辞書を構築することを目的としている. しかし, 名詞の種類は膨大である. 例えば, 様々な分野のコーパスでは多くの専門用語が使われる. また, 時が経つにつれて, 新しい名詞が作られることもある. そのため人手で辞書を作成するのは困難である. また, ある名詞に対して呼応する助数詞を人間が全て思いつくことができるとも限らない. そのため, コーパスから呼応関係にある名詞と助数詞の組を大量に自動獲得する.

1.2 研究の目的

本研究では, 日本語文の解析や生成, あるいは語義曖昧性解消等を支援することを目的とし, コーパスから呼応関係にある名詞と助数詞の組を漸進的に大量に自動獲得し, 呼応する助数詞の情報を含む名詞の辞書を構築することを目指す. その際, パターンマイニングにより抽出パターンの学習を行い, 呼応関係にある名詞と助数詞の組を正確に獲得できるような精度の高い抽出パターンを漸進的に獲得することを目指す.

1.3 本論文の構成

本論文の構成は以下の通りである。

第 2 章では, 助数詞の呼応関係を含む名詞辞書の整備に関する研究や, パターンマイニングに関する研究等を紹介する.

第 3 章では, コーパスから名詞と助数詞の呼応関係を獲得するための予備調査と, それに対する考察について述べる.

第 4 章では, 予備調査を踏まえた上で, 呼応する名詞と助数詞の組を獲得する手法を提案する.

第5章では, 第4章で提案した手法の評価, 実験について述べる. また実際に獲得できた名詞と助数詞の組を紹介する.

第6章では, 本研究のまとめと今後の課題を述べる.

第2章　関連研究

本章では, 本研究と関連のある研究について紹介する.

2.1　名詞の意味クラスと助数詞の呼応関係に関する研究

ここでは名詞の意味クラスと助数詞の呼応関係に関する研究を取り上げる. Bond らは, 日英機械翻訳システムの処理向上を目的とし, 助数詞の分析を行っている [1]. この研究では, 助数詞を UNIT, METRIC, GROUP, SPECIES の 4 つのタイプに分類し, さらに, UNIT は GENERAL, TYPICAL, SPECIAL に, METRIC は MEASURE と CONTAINER に分類し, それぞれのタイプに対して日本語と英語の特徴や違いに基づき分析を行っている. さらに, Bond らは, 名詞の意味クラスに対して対応する助数詞を人手で記述している. Bond らはまた, 個々の名詞の代わりにシソーラスにおける名詞の意味クラスと助数詞の呼応関係を利用し, 機械翻訳等で日本語文を生成する際の辞書を効率的に整備する研究を報告している [2]. 彼らは, 日本語語彙大系の意味クラスに対して呼応する助数詞を人手で割り当て, その意味クラスを持つ全ての名詞はその助数詞と呼応するとみなした. そして, そのような手続きで名詞と助数詞の呼応関係を決めたとき, 正しい呼応関係が得られた割合は 81%であったと報告している. また, Paik らによって韓国語を対象とした同様の試みが報告されている [3].

2.2　名詞と助数詞の呼応関係に関する研究

タイ語を対象とし, 本研究と同様にコーパスから呼応関係にある名詞と助数詞の組を獲得する手法が Sornlertlamvanich によって提案されている [4]. Sornlertlamvanich は, 抽出パターンを用い, 名詞と助数詞の組をパターンマッチで獲得した. その際, 使用したパターンの数は 12 あり, それらは全て人手で作成している. また, 呼応関係獲得の正解率についての報告はされていない.

日本語を対象とした名詞と助数詞の呼応関係に関する研究には白井らによるものがある [5]. 彼らは, 2 つの助数詞と, それらと呼応関係にある 2 つの名詞集合に包含関係があるかを調べ, 包含関係がある場合には助数詞の間に上位-下位関係があると推測した. さらに, これらの上位-下位関係を基に助数詞オントロジーを構築している. この研究では, 名詞と助数詞の組はコーパスから人手で収集している.

2.3 パターンマイニングに関する研究

ここではパターンマイニングに関する研究を取り上げる．Yang らは Last Position Induction Sequential Pattern Mining(LAPIN-SPAM) と呼ばれる新しい系列パターンマイニングのアルゴリズムを提案した [6]．LAPIN-SPAM は従来の SPAM[7] よりも最高で 3 倍の効率を示した．

- 系列パターン
  系列パターンとは, 文中に現れる連続または非連続の単語列である．連続している必要がないため, 例えば,

  (1) a *sheet* of paper ...

  (2) a *piece* of paper ...

  (3) a of paper

  (1) と (2) から共通する (3) の部分をパターンとして抽出する．

- SPAM と LAPIN-SPAM
  - SPAM とは, データベースのビットマップ表現を利用したアルゴリズムである．アイテムがあるならビットマップは 1 に, そうでないなら 0 となる．系列のビットマップは, それに含まれているアイテムのビットマップから作られる．
  - LAPIN-SPAM とは, SPAM において繰り返し行われていた AND 操作などを回避することにより, 大幅に効率を改善したアルゴリズムである．

2.4 本研究の特色

本研究では, 同じ意味クラスを持つ名詞は常に同じ助数詞と呼応関係にあるわけではないという考えから, コーパスから名詞と助数詞の呼応関係を網羅的に獲得するという手法を取り, この点で Bond らの研究とは異なる．また, マッチングに使用するパターンを人手で作成するのではなく, 学習によってコーパスから獲得することを試みるという点では, Sornlertlamvanich の研究と異なる．

本研究では, 名詞と助数詞の組を獲得するパターンをパターンマイニングで学習するが, このアルゴリズムについては, ここで紹介した SPAM や LAPIN-SPAM よりも単純なアルゴリズムを用いている．本研究では, 単語列の頻度をカウントし, 出現回数の多い単語列をパターンとして取り出す単純な手法を採用している．ただし, 系列パターンマイニングのようなより複雑なパターンマイニングアルゴリズムを利用して抽出パターンを獲得することは検討する必要がある．

第3章　予備調査

コーパスから名詞と助数詞の呼応関係を獲得するための予備調査として, 簡単なパターンマッチによって (n, c) を抽出することを試みた. ここで, n は名詞, c は助数詞であり, (n, c) は呼応関係にある名詞と助数詞の組とする.

3.1　調査手続き

予備調査では, 呼応する名詞と助数詞が出現する典型的な単語の並びと思われるものを抽出パターンとした. 使用したパターンを以下に示す.

$$
\begin{array}{ll}
P_1 & \text{名詞＋数字＋助数詞} \rightarrow (\text{名詞}, \text{助数詞}) \\
P_2 & \text{数字＋助数詞＋ (の) ＋名詞} \rightarrow (\text{名詞}, \text{助数詞}) \\
P_3 & \text{名詞＋ (が) ＋数字＋助数詞} \rightarrow (\text{名詞}, \text{助数詞})
\end{array}
$$

これらのパターンは, 左辺の単語の並びがあったとき, 「名詞」と「助数詞」に該当する単語を (n, c) として獲得する. 例えば, パターン P_1 なら「牛/3/匹」というような文から (牛, 匹) という組を抽出する. 同様に, パターン P_2 なら「2/本/の/鉛筆」というような, パターン P_3 なら「生徒/が/1/人」というような文からそれぞれ (鉛筆, 本), (生徒, 人) を抽出する. コーパスとして日経新聞の過去 16 年間 (1990 年ー1994 年, 1996 年ー2006 年) の新聞記事を用いた. 茶筌 [8] によって形態素解析を行い, 得られた品詞の情報を用いてパターンマッチを行った. その結果, パターン P_1 では, 66,708 組の (n, c) を, パターン P_2 では, 7,049 組の (n, c) を, パターン P_3 では, 367 組の (n, c) を獲得した. ところが, 獲得された (n, c) を調べたところ, 呼応関係にない組が誤って抽出された場合も多いことがわかった. 誤って獲得された (n, c) の例を表 3.1 に載せる.

表 3.1: 誤って獲得された (n, c) の例

パターン	獲得された (n, c)	(n, c) を含む文
P_1	(夫婦, 人)	夫婦 二 人 で気軽に足を運んでもらう
P_1	(週, 便)	貨物便は同 6 便少ない 週48便 だった。
P_1	(初年度, セット)	初年度 百五十 セット の販売を目指す.
P_1	(残り, 件)	残り 六 件 は調査中。
P_2	(チョコ, 人)	資格を持つ八 人 の チョコ を詰め合わせた
P_2	(メーカー, 位)	液晶パネルで世界三 位 の メーカー。
P_2	(熱帯林, 平方キロ)	四十万 平方キロ の 熱帯林 を持続可能な管理下に置く
P_2	(国防予算, 年度)	二〇〇六 年度 の 国防予算 は前年度比一四・七%増。
P_3	(タイプ, 色)	起毛素材を採用した タイプ が 6 色、
P_3	(魚, 割)	青果が六割、肉・魚 が四 割。
P_3	(初日, 人)	乗客は 初日 が一万二千七百五十 人、
P_3	(首都圏, 割)	販売割合は 首都圏 が 5 割、

自動獲得された (n, c) をランダムにそれぞれ 100 個ずつ選択し, それらが呼応関係にあるか人手によるチェックを行った. 獲得された (n, c) が正しい呼応関係にある割合 (正解率) と, 正解率を用いて算出した正しい (n, c) の推定値を表 3.2 に示す.

表 3.2: 予備調査結果

	P_1	P_2	P_3
獲得された (n, c) の数	66,708	7,049	367
(n, c) の正解率	0.54	0.28	0.36
正しい (n, c) の数 (推定)	36,022	1,974	132

ここでは呼応関係にある名詞と助数詞が出現する典型的な単語の並びで, 高い精度で (n, c) を獲得できると思われるものを抽出パターンとして作成し, 実験を行った. しかし, P_1, P_2, P_3 いずれの抽出パターンで獲得された (n, c) の中にも誤りが多く含まれていた.

3.2 考察

この予備調査から, 人手で作成した単純なパターンマッチによる手法では, 呼応関係にない名詞と助数詞の組が多く抽出されるということがわかった. この結果にはいくつかの理由が考えられる.

(1) 抽象名詞
表 3.1 の 11 行目に示したように,「乗客は初日が一万二千七百五十人、」の文からパターン P_3 によって (初日, 人) の組が獲得された. しかし,「初日」という名詞は数えられない抽象名詞であるので, (初日, 人) の組は呼応関係にない. 表 3.1 の例では, 他に「週」や「残り」が抽象名詞にあたる.

- このケースの対策は簡単である. 獲得の対象となる名詞が抽象名詞であった場合, マッチングは不成立とすればよい. 抽象名詞かどうかの判定には, 日本語語彙大系から作成した抽象名詞の辞書を用いる. ただし, 抽象名詞としても具体名詞としても使われるものは抽象名詞の辞書には加えず, 獲得の対象とする.

(2) 単位として使用される助数詞
表 3.1 の 7 行目に示したように,「四十万平方キロの熱帯林を持続可能な管理下に置くためには、」の文からパターン P_2 によって (熱帯林, 平方キロ) の組が獲得された. しかし,「平方キロ」という助数詞は単位であり, 本研究で対象としている名詞と呼応関係を持つ助数詞ではない.

- 本研究では, 形態素解析を茶筌で行っている. 茶筌で「名詞-接尾-助数詞」となるものの中には単位も含まれるため, このようなことが起こる. このケースの対策も抽象名詞への対策と同様に, 単位の辞書を作成し, 獲得の対象となる助数詞が単位かどうかを判定すればよい. 単位かどうかの判定には, EDR 日本語単語辞書を用いる.

(3) 単位ではないが, 常に名詞と呼応しない助数詞
単位以外にも名詞と呼応しない助数詞が誤って検出されることが多い. 表 3.1 の 6 行目の「位」や 10 行目の「割」がそれにあたる. 例えば,「数字 + 位」は順位を表すためどんな名詞とも呼応しない. また,「数字 + 割」は歩合を表しどんな名詞とも呼応しない.

- このケースは, 3.1 節で述べた P_1, P_2, P_3 の 3 つのパターンによる予備調査で頻出する助数詞を人手でチェックし, 常に名詞と呼応しない助数詞のリストを作成することで対応する. 抜き出した助数詞のリストを先の (1), (2) のケースと同様にストップワードとして用いる. 作成したストップワードのリストを図3.1に示す.

(4) 抽象名詞ではないが, 常に助数詞と呼応しない名詞
日本語語彙大系では抽象名詞とタグ付けされていないものの, どんな助数詞とも呼応しない名詞も存在する. 表 3.1 の例では「初年度」がそれにあたる.

- このケースも (3) と同様に, 予備調査で頻出した名詞を人手でチェックし, 助数詞で数えることのできない名詞のリストを作成する. 助数詞と呼応しない名詞

カナダドル, シンガポールドル, マレーシアドル, スイスフラン, セント, 人民元
テラバイト, テラビット, デシベル, ミリワット, メガビット, メガワット, ギガワット
パーセント
カ日, カ年, 月来, 日来, 年来, 日間, 年間, 秒間, 分間, 年生, 年度, 年版, 周年
平方キロメートル, 立方キロメートル
位, 倍, 代, 割, 階, 種

図 3.1: 助数詞のストップワード

うち, ただ, 月, 史上, 初年度, 大小

図 3.2: 名詞のストップワード

のリストを用いて, (n, c) を獲得する際に, n がそのリストに含まれていればそれを抽出しない. 作成したリストを図 3.2 に示す.

(5) 誤検出
(1)ー(4) のいずれにも該当しないが, 獲得された名詞と助数詞が呼応関係にないもの. 表 3.1 では, 5 行目の「資格を持つ八 人 の チョコ を詰め合わせた」という文からパターン P_2 で獲得された (チョコ, 人) が該当する.

- 誤検出を減らすためには, パターン P_1, P_2, P_3 のような単純な抽出パターンではなく, 名詞や助数詞の前後に現れる単語等も考慮した条件の厳しい抽出パターンを使用する必要がある. ただし, 効果的なパターンを人手で作成するのは困難であるので, パターンマイニングにより (n, c) を抽出するパターンを自動的に獲得する.

本研究では, 名詞と助数詞の呼応関係を正確に獲得するために, 予備調査の結果を踏まえ, 上記のような対策を施した手法を提案する.

第4章　提案手法

本章では, コーパスから呼応関係にある名詞と助数詞の組を大量に自動獲得する手法について述べる.

4.1　提案手法の概要

提案手法における処理の流れを図 4.1 に示す. 図 4.1 における破線は一度だけ実施する処理の流れであり, 実線は反復する処理の流れである. また, 図中における NC-DB は呼応関係にある名詞と助数詞の組 (n, c) のデータベースであり, これをコーパスから自動獲得することが本研究の目標である.

まず, 正しい名詞と助数詞の組を少量用意する. 以下, これを「シード」と呼び, 初期の NC-DB とする. 本研究では, シードは人手で作成する. 当初は抽出パターンを用いたシードの自動作成を目指していた. しかし, 第 3 章で示したように, 最も正解率の高かったパターン「名詞 + 数字 + 助数詞」を用い獲得した (n, c) でさえ, その正解率は 0.54 であった. 一方, シードは抽出パターンを学習する基となるデータのため, 正しい (n, c) の組を集める必要がある. 正解率の低さから, パターン P_1 で獲得した (n, c) の集合は, シードとして使用するには不適切であると判断し, シードは人手で作成することにした. シードの人手による作成方法や量については様々なやり方が考えられる. ここでは, どのようなシードの作成方法が最適かについては議論せず, シードは完全に正しい少量の (n, c) の組であることだけを仮定する. 本研究の評価実験におけるシードの具体的な作成手順は 5.1 節で述べる. また, 抽出パターンや (n, c) を獲得する際に用いるコーパスには, 予備調査の結果を踏まえ, (n, c) を正確に検出するための前処理を加える. さらに, 以下の 3 つのステップを反復することによって (n, c) を漸進的に獲得し, NC-DB に追加する.

(1) 例文検索
NC-DB に登録されている (n, c) について, 同一文中に名詞 n と助数詞 c が出現する例文をコーパスから検索する.

(2) 抽出パターン獲得
(1) で得られた例文に頻出する単語列をマイニングし, (n, c) を抽出するためのパターンを獲得する.

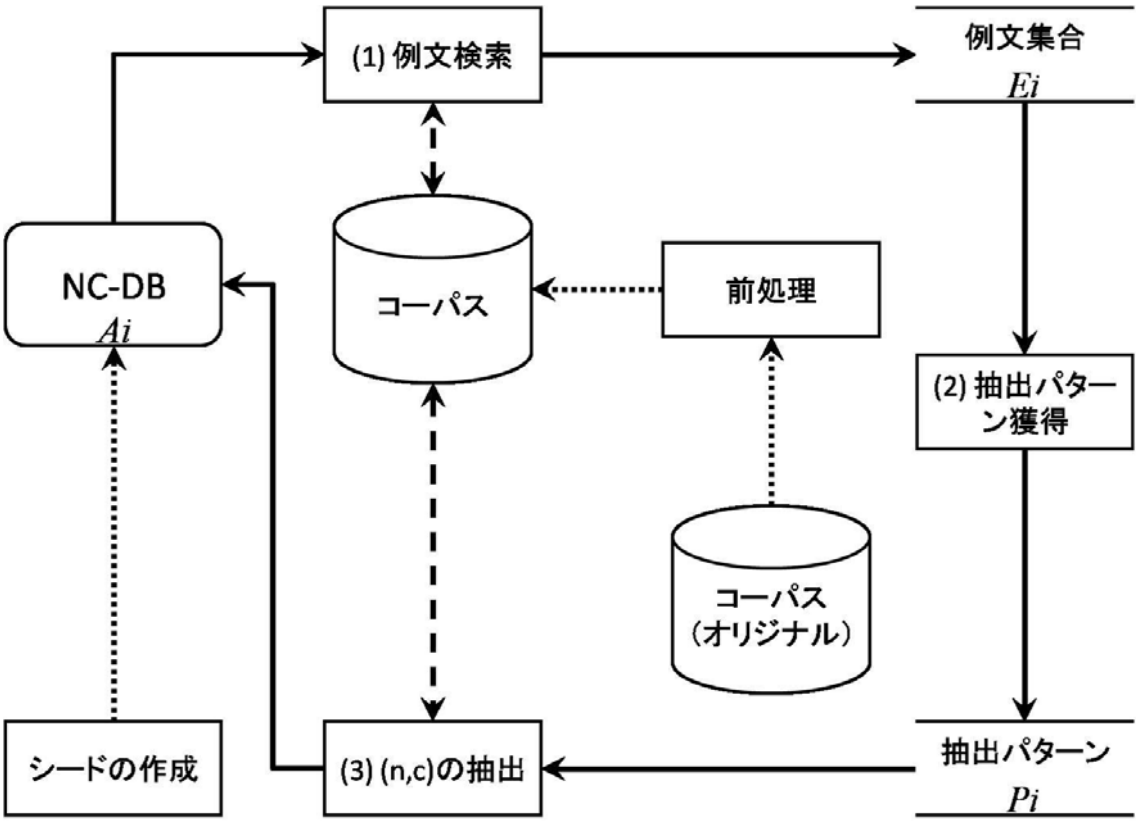

図 4.1: 提案手法の概要

- ここでは, 抽出パターンを式 (4.1) のフォーマットで記述する.

$$\text{条件} \rightarrow (N, C) \tag{4.1}$$

  条件は単語ならびにシンボル N, C, M の並びである. ここで N, C, M は前処理によって検出された名詞, 助数詞, 数字である. 式 (4.1) のパターンは, 条件にマッチする文があるとき, 名詞 N と助数詞 C を呼応関係にあるとみなして抽出することを表す. パターンの具体例は 4.4 節で紹介する.

(3) (n, c) の抽出
(2) で得られた抽出パターンを用いて, コーパスから (n, c) の組を新たに獲得し, NC-DB に追加する.

以下, 4.2 節で前処理について述べた後, 4.3, 4.4, 4.5 節で上記 (1), (2), (3) の処理の詳細について述べる. なお, 以降の説明では, i 回目の反復処理の時点で獲得されている例文の集合を E_i, 抽出パターンの集合を P_i, (n, c) の集合を A_i と記述する. また, シードは A_0 で表す.

4.2 前処理

(n, c) を抽出するためにコーパスに対する前処理を行う. まず, コーパスを茶筌で形態素解析する. 次に, 数字, 助数詞, 名詞を以下の手続きで検出する.

数字の検出

品詞が「名詞-数」である単語を数字として検出する. 数字の連続はまとめてひとつの単語とする.

- 「名詞＋が＋3＋助数詞＋ある」
- 「名詞＋が＋2＋0＋助数詞＋ある」

上の 2 つを同じパターンとして扱うために必要な処理である. 以下, 数字として検出された単語は「M」で表す. これにより, 上記の 2 つの単語列は以下のように同じものとして取り扱われる.

- 「名詞＋が＋M＋助数詞＋ある」

助数詞の検出

品詞が「名詞-接尾-助数詞」である単語を助数詞として検出する. ただし, これらの単語には「ミリ」のような単位も含まれる. しかし, そのような単位は, 本研究で対象としている名詞と呼応関係にある助数詞ではないので, 除外する必要がある. そこで, EDR 日本語単語辞書の品詞が「JUN(単位)」である 746 個の単語は助数詞から除外する. 除外した単語には「アール」「オーム」「カロリー」「ガウス」「ギガ」「センチ」「デシベル」「ドル」「ヘルツ」「ヤード」等がある. また, 3.2 節で作成した助数詞のストップワードに含まれている語も除外する. 以下, 助数詞として検出された単語は「C」で表す.

名詞の検出

以下の条件を満たす 1 つ以上の単語の並びをひとつにまとめて名詞として検出する.

- 品詞が「名詞-一般」「名詞-サ変接続」「名詞-接尾-一般」のいずれかである.
  茶筌で名詞に該当する品詞タグとしては, その他にも「名詞-固有名詞」「名詞-代名詞」等があるが, 数えられないものが大半であるため, 本研究では上の 3 つのみを名詞として扱う.
- 先頭の単語の品詞は「名詞-接尾-一般」ではない.
  「名詞-接尾-一般」は接尾語を表し, 本来なら語の先頭に現れることはないためである.
- 末尾の単語の品詞は「名詞-サ変接続」ではないとする.
  「経験」「募集」「関係」等のサ変名詞は, 何らかの行為を表す場合が多く, 助数詞を使って数えることができない場合が多い. そこで, サ変名詞が名詞列の最

後にある場合は名詞として検出しない. ただし, 名詞列の先頭, あるいは途中に現れる場合には「経験/者」「人材/募集/用/ＰＲ/ビデオ」「関係/省庁」等, 名詞列全体としてみれば数えることができるため, このような複合名詞は 1 つの名詞として検出する.

- 末尾の単語が抽象名詞でない.
  抽象名詞も数えることができないので除外する必要がある. 抽象名詞かどうかの判定は, 日本語語彙大系を用いて行った. 具体的には, 日本語語彙大系の意味クラスが「1000(抽象)」の下位ならストップワードとしてリストに加えた. 意味クラスが 2 つ以上ある名詞は, その全てが「1000」の下位であるときだけリストに加え, 1 つでもそれ以外の意味クラスが存在するならば, 具体名詞として使われる可能性があるため, リストには加えない.
- 末尾の単語が「数」ではない.
  「死者数」「退職者数」のような複合名詞を名詞として検出しないようにするための処理である. これらの複合名詞は助数詞を用いて数えることができない. また, 「死者数」「退職者数」の例では, 「数」という単語を除外することにより, 代わりに「死者」「退職者」を名詞として検出する.
- ストップワードに含まれていない.
  3.2 節で作成した, 助数詞で数えることのできない名詞のリストに載っている名詞は除く.

以下, 名詞として検出された単語は「N」で表す.

上記の処理によって検出された N または C を呼応関係を獲得する名詞または助数詞の候補とする.

4.3 例文検索

それまでの処理で獲得された (n, c) の集合 A_{i-1} の各要素について, n と c を同時に含む文をコーパスから検索する. コーパスにおける文書を句点, 読点, 空白, 「▽」のいずれかで分割したものを一文とする. 文境界として読点も含めているので, ここで抽出するのは文よりも短い節である. ただし, 本論文では説明のわかりやすさのため, 上記の処理で分割された節を「例文」と呼ぶ. 一文中に呼応関係にある名詞 n と助数詞 c が含まれているものを抽出し, 例文集合 E_i を得る.

4.4 抽出パターンの獲得

前節で得られた E_i は呼応関係にある名詞と助数詞を含む例文の集合である. したがって, E_i に頻出する単語列は, 名詞と助数詞の呼応関係を抽出するための手がかりとなる.

そこで, 以下の手続きで抽出パターンを獲得する.

まず, E_i から, 呼応関係にある n と c の間にある単語, n の直前の単語, c の直後の単語の列を抽出パターンの候補として抽出する. ただし, n に対応する単語は N, c に対応する単語は C, 数字として検出された単語は M というシンボルに置き換える. このとき, 単語を表記のまま扱うか, 原型に戻してから扱うかによって 2 通りの手法が考えられる. 前者を手法 (1), 後者を手法 (2) とする. 以下に例文からの抽出パターンの候補の作成の例を示す.

例文: 描いた*作品*約 58#点#を展示する。
パターン: た+N+約+M+C+を → (N,C)

また, c の後に n が出現する例文からも同様にパターンの候補を作成する. この場合の例文からの抽出パターンの候補の作成例を以下に挙げる.

例文: 当初三#人#の*遺体*が発見され、
パターン: 当初+M+C+の+N+が → (N,C)

また, 手法 (1) と手法 (2) で異なるパターンが作成される例を以下に挙げる.

例文: 特に「良い」とする*企業*が昨年より八#社#増えた半面、
手法 (1) のパターン: する+N+が+昨年+より+M+C+増え → (N,C)
手法 (2) のパターン: する+N+が+昨年+より+M+C+増える → (N,C)

ここでは「増えた」は「増え/た」と形態素解析されている. 手法 (1) は, 表記である「増え」をそのままパターンに用いるが, 手法 (2) は「増え」の原型である「増える」をパターンに利用する. そのためこのような差異が生まれる.

作成されたパターンの候補のうち, 以下の式 (4.2) の条件を満たさないものは, 信頼性が低いと判断し除外する.

$$\text{抽出パターンの左辺の単語列の } E_i \text{における出現頻度が 5 以上} \tag{4.2}$$

次に, 得られた抽出パターンの候補の評価を行う. パターンの候補を p とするとき, 以下の 3 つの条件を全て満たすものを選別し, 抽出パターンの集合 P_i とする.

$\langle 1 \rangle \quad m(p) \geq T_m$

$\langle 2 \rangle \quad r(p) \stackrel{def}{=} \dfrac{|A_p \cap A_c|}{|A_p|} \geq T_r$

$\langle 3 \rangle \quad i(p) \stackrel{def}{=} \dfrac{A_p\text{で最も頻出する } (n,c) \text{ の数}}{|A_p|} \leq T_i$

⟨1⟩ は p にマッチする例文の数 $m(p)$ が T_m 以上であるという条件である. すなわち, マッチングにあまり成功しない p は有効でないとみなす. 本研究では $T_m = 50$ とした. この条件は, コーパス全体に抽出パターンを用いてマッチングを行った結果パターンにマッチする例文数に対する条件であり, 式 (4.2) に示した例文の集合 E_i における出現頻度に対する条件とは異なる.

⟨2⟩ はパターンの信頼度 $r(p)$ が T_r 以上であるという条件である. ここで A_p は p を適用して獲得される (n, c) の集合, A_c は正しい (n, c) の集合である. すなわち, 正しい (n, c) をある程度の割合で抽出できるパターンは信頼度が高いとみなす. ここでは A_c の定義に応じて以下の 2 通りの手法を考える.

手法 (A)
$A_c = A_0$, すなわちシードのみを正しい (n, c) の組とみなす手法

手法 (B)
$A_c = A_{i-1}$, すなわちその時点で獲得された (n, c) を全て正しいとみなす手法

T_r の値は, 手法 (A) のときは 0.1 とした. 手法 (B) のとき, 1 回目の処理のときは 0.1, 2 回目以降では 0.55 とした. 手法 (B) で 1 回目の閾値を低く設定しているのは, 初期段階では正しい (n, c) の組はシードのみであり, 量が十分にないことを考慮したためである.

条件 ⟨3⟩ は, 同じ (n, c) しか抽出できないような p は, 抽出した n と c に呼応関係がない可能性が高いため, 有効ではないという考えに基づいている. 例えば,

$$\text{警視庁} + N + M + C + \text{は} \rightarrow (N, C)$$

というパターンから抽出されるのは (捜査, 課) がほとんどである. しかし, 「警視庁捜査一課」はいわば定型表現に近く, この中に出現する名詞と助数詞の間には呼応関係がないため, 誤った組が抽出されている. ⟨3⟩ はこのような誤抽出を避けるための条件である. 本研究では $T_i = 0.7$ とした.

4.5 名詞と助数詞の呼応関係の獲得

獲得された抽出パターンの集合 P_i をコーパスに適用し, (n, c) の組を獲得し, A_i とする. また, 抽出回数が T_e 未満の組は信頼度が低いとみなして除去する. ただし, 本研究では $T_e = 1$, すなわち抽出パターンによって獲得された (n, c) は全て正しいとみなして A_i に加えた.

4.6 処理の流れ

これまで説明してきた処理の流れを, 5 章で述べる実験で実際に得られた A_i, E_i, P_i を用い紹介する. ここでは例として (作品, 点) をシードとしたとき, 新しい抽出パターンや呼応関係にある名詞と助数詞の組が得られる過程を説明する.

(1) シードとして (作品, 点) がある.

(2) (作品, 点) を含む例文をコーパスから検索する.
(作品, 点) を含む例文 486 個の一部を図 4.2 に示す. ただし, n は*に挟まれた単語で, c は#に挟まれた単語である.

S_1	二十一カ国・地域の*作品*約二千#点#を所蔵。
S_2	任性珍さんと大石祐子さん２人の日用食器などの*作品*約８０#点#を展示。
S_3	八十#点#の*作品*から伝わってくる孤高の極み。
S_4	食器などの*作品*約７０#点#を展示。
S_5	おもちゃなど木の*作品*５０#点#を展示。
S_6	鈴木春信といった江戸期を代表する浮世絵師の*作品*約８０#点#を公開。
S_7	動物などの*作品*約２５#点#を展示。
S_8	千五百#点#以上の京焼*作品*を確認した。
S_9	東日本を拠点に活動する若手作家から人間国宝まで二十八人の*作品*約八十#点#を伝統、
S_{10}	*作品*は茶わんなど約百#点#。

図 4.2: (作品, 点) を含む例文の例

(3) 得られた例文からパターンを作成する
S_1, S_2, S_4, S_6, S_7, S_9 の文からは「の + N + 約 + M + C + を → (N, C)」という抽出パターンの候補が作成される. 他の文からも同じパターンが獲得され, 式 (4.2) の条件を満たす. また, ⟨1⟩, ⟨2⟩, ⟨3⟩ の条件も全て満たしている. したがって, このパターンを採用し, P_i に加える.

(4) パターンから新たな (n, c) が獲得される.
「の + N + 約 + M + C + を → (N, C)」という抽出パターンからは, 267 組の (n, c) が獲得された. その一部を図 4.3 に示す.

(n, c)	パターンの条件部にマッチした例文
写真, 点	の写真約三十点を
短歌, 首	の短歌約三万首を
裁判官, 人	の裁判官約七百七十人を
自動車教習所, 校	の自動車教習所約四百三十校を
実包, 発	の実包約二百発を
苗木, 本	の苗木約六千二百五十本を
民間企業, 社	の民間企業約一万社を
出張, 件	の出張約三千件を
錠剤, 錠	の錠剤約五万六千錠を
仏像, 体	の仏像約八十体を

図 4.3: パターン「の + N + 約 + M + C + を → (N, C)」から獲得された (n, c) の例

第5章　実験

本章では, コーパスより呼応関係にある名詞と助数詞の組を獲得する実験について述べる. 5.1 節では実験を行う際の条件について, 5.2 節では実験結果について述べる.

5.1　実験条件

5.1.1　シードの作成

まず, シードとして少量の (n, c) の組を用意する. 本研究では,『数え方の辞典』[9] を参照してシードを用意した.『数え方の辞典』は様々な名詞とそれらを数える際に用いられる助数詞を網羅的に記載した辞典である. ただし, 単位を表す助数詞や,「パック」「山」など個体を数えずに集合を数えるような助数詞は人手であらかじめ除去した. また,「つ」という助数詞は一般的すぎるために除外した. コーパスに対して行った前処理と合わせるため, 日本語語彙大系の意味クラスによって抽象名詞と判断された名詞も除外した. 最終的に 7,135 組の (n, c) を得た. 以下, このようにして得られた正しい (n, c) の集合を C とする.

今回の実験では, なるべく少数のシードから新しい (n, c) を獲得できるかを調べたかったため, C よりも小さい集合をシードとすることにした. 具体的には, C に含まれる名詞のうち, コーパスにおける出現頻度の上位 100 個の名詞を選定し, それらの名詞ならびにそれと呼応する助数詞の組の集合をシード A_0 とした. A_0 の要素数は 213 であった. シードとして用意された (n, c) の例を図 5.1 に示す. また, シードの一覧を付録 A に載せる.

(企業, 社)
(写真, カット)(写真, 齣)(写真, 葉)(写真, ポーズ)(写真, 点)(写真, 枚)
(図, 図)(図, 点)(図, 枚)
(人, 名)(人, 方)(人, 体)(人, 個)(人, 人)(人, 口)(人, 頭)(人, 氏)
(情報, 報)(情報, 件)(情報, 本)(工場, 棟)
(工場, 軒)(工場, 箇所)
(商品, 個)(商品, 点)
(地域, 郭)
(事件, 件)
(声, 声)
(国, 国)(国, か国)
(証券, 通)(証券, 枚)
(株, 株)(株, 枚)(株, 本)
(業績, 点)(業績, 本)
(銀行, 軒)(銀行, 行)
(グループ, 班)
(店舗, 店)(店舗, 軒)(店舗, 店舗)
(法人, 法人)
(支店, 支店)
(核, 発)(核, 個)
(病院, 軒)(病院, 院)
(客, 名)(客, 人)(客, 家族)
(機関, 機関)
(課題, 課題)
(自動車, 台)
(ファンド, 口)(ファンド, 本)
(大学, 大学)(大学, 校)
(ホテル, 棟)(ホテル, 軒)
(チーム, チーム)
(ネット, 枚)
(銘柄, 銘柄)
(業者, 社)(業者, 軒)(業者, 人)
(住宅, 棟)(住宅, 戸)(住宅, 軒)(住宅, 邸)
(時代, 時代)

図 5.1: シード (抜粋)

5.1.2 実験の概要

提案手法によって (n, c) ならびに抽出パターンを自動獲得するプログラムを実装した. 計算速度向上のため, それぞれの反復ステップでは本研究では A_i, P_i, E_i は差分を計算し, 新規に獲得された (n, c) の組, パターン, 例文を求め, これをそれぞれ A_{i-1}, P_{i-1}, E_{i-1} とマージすることにより A_i, P_i, E_i を求めている.

コーパスは日経新聞の新聞記事データを用いた. 2006 年の新聞記事データのみを使用する実験と 2002 年ー2006 年の 5 年分の新聞記事データを使用する実験の 2 通りの実験を行った. また, 4.4 節で述べたように, 抽出パターンを作成する際, 表記を用いる手法 (1) と, 原型を用いる手法 (2) の 2 つがある. また, 抽出パターンの獲得条件 ⟨2⟩ の設定方法として, シードのみを正しい (n, c) とする手法 (A) と, 自動獲得された (n, c) も全て正しいとする手法 (B) の 2 つがある. これらを組み合わせた手法 (1-A), (1-B), (2-A), (2-B) の 4 つを用いて (n, c) の獲得を試みた.

5.2 実験結果

実験結果を手法 (1-A), 手法 (1-B), 手法 (2-A), 手法 (2-B) の順に述べる.

5.2.1 手法 (1-A) の結果と考察

表 5.1 は手法 (1-A) による実験結果を表す. コーパスには 1 年分の新聞記事データを用いた.

表 5.1: 実験結果 (手法 1-A)

i	1	2	3	4
$\|A_i\|$	964	1,808	1,845	1,845
$\|P_i\|$	17	33	34	34
$\|E_i\|$	7,299	39,875	52,320	52,733
$A_i \setminus A_{i-1}$ の正解率*	0.88	0.71	0.92	-
A_i の正解率*	0.88	0.80	0.80	0.80
正しい (n, c) の数*	848	1,448	1,482	1,482
C の再現率	4.48%	5.06%	5.07%	5.07%

手法 (1-A) では, 図 4.1 に示した一連の操作を 3 回反復した. これは, 4 回目の反復操作で条件 ⟨1⟩, ⟨2⟩, ⟨3⟩ を満たす抽出パターンを新たに獲得することができなかったためである. 表 5.1 において, $|A_i|$ はそれぞれの段階で獲得された (n, c) の数 (213 個のシードは除

く), $|P_i|$ は獲得された抽出パターンの数である. 「$A_i \setminus A_{i-1}$ の正解率」は, i 番目の反復操作で新たに獲得された (n, c) のうち, 正しい組の割合を表す. ただし, 表に掲載した正解率は, 最大でランダムに 100 個サンプリングした (n, c) を人手でチェックして算出した近似値である. また, 「正しい (n, c) の数」は, A_i のうち, 上記の正解率を用いて算出した正しい (n, c) の数の見積もりである.
この算出式を式 (5.1) に示す. 式 (5.1) において, a_i は i 回目のステップにおける正しい (n, c) の推測値, r_i 集合 $A_i \setminus A_{i-1}$ からランダムサンプリングされた (n, c) を人手でチェックして求めた正解率である.

$$a_i = a_{i-1} + (|A_i| - |A_{i-1}|) \times r_i \tag{5.1}$$

この実験の結果から, 自動獲得した抽出パターンを用いると, 0.7 から 0.9 の正解率で (n, c) を抽出できることがわかった.

一方, 表 5.1 における「C の再現率」は, 5.1.1 項で作成した 7,135 組の正しい (n, c) の集合 C のうち, 提案手法で獲得することのできた組の割合である. C の再現率は 5%程度と低く, 提案手法では C とは異なる (n, c) の組が得られていることがわかる. これは, コーパスには出現するが必ずしも一般的ではない名詞に対して, 呼応する助数詞が新たに獲得されたためと考えられる.

5.2.2 手法 (1-B) の結果と考察

次に, 手法 (1-B) によって (n, c) を獲得した. コーパスには 1 年分の新聞記事データを用いた. 手法 (1-B) では, 3 回目の反復操作において, 条件を満たす新たな抽出パターンが得られなかったため, 反復回数は 2 回となった. 結果を表 5.2 に示す. 獲得できた正しい (n, c) の組, 抽出パターンの数, C の再現率などは手法 (1-A) とあまり変わらなかった.

表 5.2: 実験結果 (手法 1-B)

i	1	2	3
$\|A_i\|$	964	1,721	1,721
$\|P_i\|$	17	34	34
$\|E_i\|$	7,299	39,875	51,962
$A_i \setminus A_{i-1}$ の正解率*	0.88	0.80	-
A_i の正解率*	0.88	0.84	0.84
正しい (n, c) の数*	848	1,454	1,454
C の再現率	4.48%	5.26%	5.26%

5.2.3 手法 (2-A) の結果と考察

次に, 手法 (2-A) によって (n, c) を獲得した. コーパスには 1 年分の新聞記事データを用いた. 手法 (2-A) では, 一連の操作を 3 回反復し, 4 回目の反復操作において, 条件を満たす抽出パターンを新たに得ることができなくなった. 結果を表 5.3 に示す. 獲得できた正しい (n, c) の組, 抽出パターンの数, C の再現率などは手法 (1-A) とあまり変わらなかった.

表 5.3: 実験結果 (手法 2-A)

i	1	2	3	4
$\|A_i\|$	954	1,780	1,817	1,817
$\|P_i\|$	18	33	34	34
$\|E_i\|$	7,299	39,499	51,854	52,267
$A_i \setminus A_{i-1}$ の正解率*	0.83	0.71	0.92	-
A_i の正解率*	0.83	0.77	0.78	0.78
正しい (n, c) の数*	792	1378	1412	1412
C の再現率	4.43%	4.99%	5.00%	5.00%

5.2.4 手法 (2-B) の結果と考察

最後に, 手法 (2-B) によって (n, c) を獲得した. コーパスには 1 年分の新聞記事データを用いた. 手法 (2-B) では, 3 回目の反復操作において, 条件を満たす抽出パターンを新たに獲得することができなかったため, 反復回数は 2 回となった. 結果を表 5.4 に示す. 獲得できた正しい (n, c) の組, 抽出パターンの数, C の再現率などは手法 (1-A) とあまり変わらなかった.

表 5.4: 実験結果 (手法 2-B)

i	1	2	3
$\|A_i\|$	954	1,704	1,704
$\|P_i\|$	18	36	36
$\|E_i\|$	7,299	39,499	51.985
$A_i \setminus A_{i-1}$ の正解率*	0.83	0.86	-
A_i の正解率*	0.83	0.84	0.84
正しい (n, c) の数*	792	1,437	1,437
C の再現率	4.43%	5.20%	5.20%

5.2.5　5 年分のコーパスを使用して行った実験

次に, コーパスの量を増やして名詞と助数詞の呼応関係を獲得する実験を行った. ここでは日経新聞の 2002 年から 2006 年の 5 年分のコーパスを用いた. 5.2.1ー5.2.4 の実験結果から, 提案する 4 つの手法には大きな差がないことがわかったため, ここでは手法 (1-A) のみを用いて実験を行った. 5 年分のコーパスを用いた実験でも, 1 年分のコーパスを用いたときと同じように, 4 回目の反復操作において, 条件を満たす抽出パターンを新たに獲得することができなかったため, 反復回数は 3 回となった. 結果を表 5.5 に示す.

表 5.5: 実験結果 (5 年分のコーパス)

i	1	2	3	4
$\|A_i\|$	7,482	9,289	9,412	9,412
$\|P_i\|$	119	169	172	172
$\|E_i\|$	36,106	428,910	468,915	470,153
$A_i \setminus A_{i-1}$ の正解率*	0.71	0.64	0.78	-
A_i の正解率*	0.71	0.70	0.70	0.70
正しい (n, c) の数*	5,312	6,468	6,564	6,564
C の再現率	10.36%	10.97%	11.07%	11.07%

この実験結果から, コーパスの量を 5 倍にすると獲得される (n, c) の数も約 5 倍になることがわかった. 一方, 反復回数は 1 年分のコーパスを用いたときと同じく 3 回であった. また, 正解率は 0.70 程度と 1 年分のときと比べて下がることがわかった. これは, 式 (4.2) における出現頻度の閾値や条件 $\langle 1 \rangle$, $\langle 2 \rangle$, $\langle 3 \rangle$ の閾値 T_m, T_r, T_i をコーパス 1 年分を用いた実験と同じものにしてしまったためと考えられる. これらの閾値を厳しく設定すれば, 獲得される名詞と助数詞の組の正解率は向上するが, 獲得数は少なくなる. したがって, これらの閾値は正解率と獲得数のバランスを考慮して最適化する必要があるが, この際コーパスの規模の大きさも考慮に入れる必要がある.

5.3　例

手法 (1-A) によって実際に抽出された (n, c) のうち, 正解集合 C に含まれていない組の例を図 5.2 に示す. 新聞記事によく使われるような名詞に対して, それと呼応する助数詞が獲得されていることがわかる. このことから, ドメイン固有のコーパスに提案手法を適用することにより, 専門用語に対しても名詞と助数詞の呼応関係が獲得できるのではないかと考えている.

図 5.3 は獲得された抽出パターンの例である. 3 章で, 「名詞+数字+助数詞 → (N,C)」というパターン P_1 では正しくない呼応関係も数多く抽出されることは既に述べた. 図 5.3

のパターン p_2 や p_3 も P_1 と同じように基本的には「名詞 + 数字 + 助数詞」という並びにマッチするが, p_2 のように M(数字) の前に「約」という単語があったり, p_3 のように「N+M+C」の後に「当たり」という単語があると, パターンにマッチした文では名詞 N の数を数えている可能性が高いと考えられる. このように, 呼応関係にある名詞と助数詞の組を抽出する精緻なパターンが学習されていることがわかった.

(死者, 人)(ＤＲＡＭ, 個)(代表作, 点)
(油彩画, 点)(求職者, 人)(爆弾テロ, 件)
(農家, 戸)(上場企業, 社)(個人情報, 件)
(リンパ節, 個)(トラブル, 件)(チラシ, 枚)
(不払い, 件)(案件, 件)(自動車教習所, 校)
(就業者, 人)(光ファイバー, 本)(原発, 基)
(八ミリ作品, 本)(取締役, 人)(自販機, 台)
(細管, 本)(公認会計士, 人)(金メダル, 個)
(ゲンジボタル, 匹)(ヒマワリ, 本)(缶, 本)
(ケース, 件)(カップル, 組)(ＡＴＭ, 台)
(薬きょう, 発)(無人レジ端末, 台)(実包, 発)
(ミンククジラ, 頭)(申し込み, 件)(基板, 枚)
(架空請求書, 通)(レンタカー, 台)(ユリ, 本)
(ユーザー, 人)(プロジェクト, 件)(ＣＭ, 本)
(洋蘭, 鉢)(パターン, 通り)(バイヤー, 人)
(ＲＶ車, 台)(デモカー, 台)(チョウ, 羽)
(ダイヤ, 個)(ゼネコン, 社)(シミ, カ所)
(クロマツ, 本)(かっぱ, 着)(コアラ, 匹)
(Ｂ４判, 枚)(注射器, 本)(トレーナー, 人)
(洋楽, 曲)(アナリスト, 人)(どじょう, 匹)
(閣僚, 人)(ゴマ, 粒)(コンサルタント, 人)
(部品, 点)(油彩作品, 点)(ロケット, 発)
(製品, 個)(顧客情報, 件)(記念写真, 点)
(無所属議員, 人)(物件, 件)(男児, 人)
(入院ベッド, 床)(夫婦, 組)(若者, 人)
(台風画像, 枚)(水彩画, 点)(加盟店, 店)
(申請データ, 件)(失業者, 人)(和牛, 頭)
(戸建て住宅, 戸)(冷蔵車, 台)(例, 件)
(客室, 室)(プレーヤー, 人)(猟師, 人)
(コレクション, 点)(壁, 面)(苗木, 本)

図 5.2: 抽出された (n, c) の例

p_1	:	M+C+以上+の+N+を → (N,C)
p_2	:	た+N+約+M+C+を → (N,C)
p_3	:	N+M+C+当たり → (N,C)
p_4	:	から+M+C+の+N+が → (N,C)
p_5	:	た+N+は+M+C+と → (N,C)
p_6	:	の+N+約+M+C+を → (N,C)
p_7	:	た+N+は+M+C+。 → (N,C)
p_8	:	の+N+数+は+M+C+。 → (N,C)
p_9	:	が+N+M+C+に → (N,C)
p_{10}	:	の+N+M+C+から → (N,C)

図 5.3: 獲得された抽出パターンの例

第6章　おわりに

6.1　おわりに

本研究では, パターンマイニングにより抽出パターンを学習することにより, コーパスから呼応関係にある名詞と助数詞の組を抽出する手法について述べた. 以下に本研究についてまとめる.

6.1.1　本研究のまとめ

予備調査

抽出パターンの作成に際して予備調査を行った. 呼応する名詞と助数詞が出現する典型的な単語の並びと思われるものを3つ抽出パターンとしたが, 獲得された (n, c) の中には誤りが多く含まれていた. この調査により, 人手で作成した単純なパターンマッチによる手法では, 呼応関係にない名詞と助数詞の組が抽出されたり, 抽象名詞のように数えられない名詞が抽出されることが多いことがわかった. この結果を踏まえ, 本研究ではパターンマイニングにより (n, c) を抽出するパターンを自動獲得することを試みた.

提案手法

- 前処理
  (n, c) を抽出するためにコーパスに対する前処理を行った. 始めに, コーパスを茶筌で形態素解析し, 次に, 数字, 助数詞, 名詞を所定の手続きに従い, それぞれ「M」「C」「N」で表し, N または C を呼応関係を獲得する名詞または助数詞の候補とした.
- 例文検索
  それまでの処理で獲得された (n, c) の集合の各要素について, n と c を同時に含む文をコーパスから検索した.
- 抽出パターンの獲得
  例文集合から, 呼応関係にある n と c の間にある単語, n の直前の単語, c の直後の単語列を抽出パターンの候補とした. また, c の後に n が出現する例文から

も同様にパターンの候補を作成した. 作成された抽出パターンの候補を評価基準に従って評価をし, 全ての基準を満たしたものを抽出パターンとした.

- 名詞と助数詞の呼応関係の獲得
  獲得された抽出パターンの集合をコーパスに適用し, (n, c) の組を獲得した.

実験

シードとして 213 個の (n, c) の組を用意し, 実験を行った. コーパスには日経新聞の新聞記事データを用いた. 2006 年の新聞記事データのみを使用する実験と 2002 年ー2006 年の 5 年分の新聞記事データを使用する実験を行った. また, 4 通りの提案手法を用いて (n, c) の獲得を試みた. その結果, 1 年分のコーパスを使用した実験では, いずれの手法においても, 自動獲得した抽出パターンを用いると, 0.7 から 0.9 の正解率で (n, c) を抽出できることがわかった. 5 年分のコーパスを使用した実験では, 0.7 程度の正解率で (n, c) を抽出できることがわかった. 実際に獲得された (n, c) のうち, 正解集合に含まれていない組を見てみると, 新聞記事によく使われるような名詞に対して, それと呼応する助数詞が獲得されていることがわかった. また, 獲得された抽出パターンを見ると, 呼応関係にある名詞と助数詞の組を抽出する精緻なパターンが学習されたことがわかった.

6.2 今後の課題

最後に今後の課題について述べる. 提案手法では, 獲得された名詞と助数詞の呼応関係の正解率は比較的高かったものの, 反復操作は 2,3 回で終了したことから, 大量の名詞と助数詞の組をコーパスから網羅的に獲得できたとは言い難い. そこで, 今回の抽出パターンは単語の並びにマッチするものであったが, 品詞にマッチしたり任意の単語にマッチする要素を抽出パターンの条件部に許すことを検討している. これらの抽出パターンは例文にマッチする回数が増えるため, より多くの (n, c) が獲得されると期待できる. より大規模なコーパスに対して提案手法を適用することも試みる必要があろう. また, 本論文では提案手法における閾値 (T_m, T_r, T_i, T_e) をアドホックに決定していたが, これらを最適化する方法を検討する必要がある.

謝辞

本研究を進めるにあたり, 北陸先端科学技術大学院大学・自然言語処理学講座の島津明教授, 白井清昭准教授, 中村誠助教には様々な御指導して頂き, 大変お世話になりました. また, 自然言語処理学講座の諸先輩方, 同期生, 後輩の皆様方には研究に関する助言をして頂きました. お世話になった皆様に心より御礼を申し上げます.

参考文献

[1] Francis Bond and Kentaro Ogura and Satoru Ikehara. Classifiers in Japanese-to-English Machine Translation. In *Proceedings of the COLING*, pp. 125-130, 1996.

[2] Francis Bond and Kyonghee Paik. Reusing an ontology to generate numeral classifiers. In *Proceedings of the COLING*, pp. 90-96, 2000.

[3] Kyonghee Paik and Francis Bond. Multilingual generation of numeral classifiers using a common ontology. In *Proceedings of the ICCPOL*, pp. 141-147, 2001.

[4] Virech Sornlertlamvanich, Wantanee Pantachat, and Surapant Meknavin. Classifier assignment by corpus-based approach. In *Proceedings of the COLING*, pp. 556-561, 1994.

[5] 白井清昭, 徳永健伸. 呼応する名詞の包含関係に着目した助数詞オントロジーの自動構築と評価. 情報処理学会自然言語処理研究会, Vol. 2007, No. 94, SIGNL-181-20, pp. 127-134, 2007.

[6] Zhenglu Yang and Masaru Kitsuregawa. LAPIN-SPAM: An Improved Algorithm for Mining Sequential Pattern. In *Proceedings of the ICDE*, pp. 1222, 2005.

[7] Jay Ayres, Johannes Gehrke, Tomi Yiu, and Jason Flannick. Sequential PAttern Mining using A Bitmap Representation. In *ACM SIGKDD Conference*, pp. 429-435, 2002.

[8] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸. 日本語形態素解析システム 茶筌 version2.3.3. 奈良先端科学技術大学院大学 情報科学研究科 自然言語処理学講座, 2003.

[9] 飯田朝子. 数え方の辞典. 小学館, 2004.

付 録A　シードの一覧

5章の評価実験に用いたシードは213個の正しい名詞と助数詞の組から構成されている. シードに含まれる (n, c) の一覧を以下に示す.

(企業, 社)
(写真, カット)(写真, 齣)(写真, 葉)(写真, ポーズ)(写真, 点)(写真, 枚)
(図, 図)(図, 点)(図, 枚)
(人, 名)(人, 方)(人, 体)(人, 個)(人, 人)(人, 口)(人, 頭)(人, 氏)
(情報, 報)(情報, 件)(情報, 本)(工場, 棟)
(工場, 軒)(工場, 箇所)
(商品, 個)(商品, 点)
(地域, 郭)
(事件, 件)
(声, 声)
(国, 国)(国, か国)
(証券, 通)(証券, 枚)
(株, 株)(株, 枚)(株, 本)
(業績, 点)(業績, 本)
(銀行, 軒)(銀行, 行)
(グループ, 班)
(店舗, 店)(店舗, 軒)(店舗, 店舗)
(法人, 法人)
(支店, 支店)
(核, 発)(核, 個)
(病院, 軒)(病院, 院)
(客, 名)(客, 人)(客, 家族)
(機関, 機関)
(課題, 課題)
(自動車, 台)
(ファンド, 口)(ファンド, 本)
(大学, 大学)(大学, 校)
(ホテル, 棟)(ホテル, 軒)

(チーム, チーム)
(ネット, 枚)
(銘柄, 銘柄)
(業者, 社)(業者, 軒)(業者, 人)
(住宅, 棟)(住宅, 戸)(住宅, 軒)(住宅, 邸)
(時代, 時代)
(作品, 作)(作品, 編)(作品, 点)(作品, 本)(作品, 作品)
(手, 人)(手, 手)(手, 枚)(手, 本)
(目, 個)
(マンション, 棟)(マンション, 戸)(マンション, 邸)
(大会, 回)
(円, 個)
(道, 条)(道, 筋)(道, 本)
(方法, 手)(方法, 方法)
(場所, 所)(場所, 箇所)
(メーカー, 社)(メーカー, 人)
(本, 部)(本, 冊)(本, 帙)(本, 点)
(水, 滴)(水, 雫)(水, 筋)(水, 本)
(車, 輪)(車, 乗)(車, 台)
(機械, 台)(機械, 基)
(パソコン, 台)
(都市, 都市)
(テーマ, テーマ)
(データ, アイテム)(データ, 個)(データ, 件)(データ, 点)
(土地, 区画)(土地, 面)
(協会, 協会)
(ブランド, 銘柄)(ブランド, ブランド)
(空港, 空港)(空港, 箇所)
(言葉, 言)
(ビル, 棟)(ビル, 軒)(ビル, 本)
(家族, 世帯)(家族, 家族)
(学校, 校)
(テレビ, 台)
(流れ, 筋)(流れ, 本)
(保険, 件)(保険, 口)
(機器, 台)
(子供, 男)(子供, 児)(子供, 人)(子供, 女)
(法案, 法案)(法案, 本)

(道路, 条)(道路, 筋)(道路, 本)
(映画, 巻き)(映画, 作)(映画, 本)(映画, 作品)
(例, 例)
(方式, 方式)
(事務所, 軒)(事務所, 箇所)
(柱, 茎)(柱, 本)
(世代, 世代)
(建物, 棟)(建物, 軒)
(原則, 原則)
(顔, 面)
(頭, 個)
(路線, 路線)(路線, 本)
(カード, 枚)
(記事, 項目)(記事, 点)(記事, 本)
(舞台, 幕)
(百貨店, 店)(百貨店, 軒)
(債権, 通)(債権, 枚)
(店, 店)(店, 軒)(店, 店舗)
(音楽, 曲)
(人間, 人)
(項目, 項目)(項目, 点)
(会員, 員)(会員, 人)
(足, 脚)(足, 本)
(乗用車, 台)
(バス, 便)(バス, 台)(バス, 本)
(高校, 校)
(家, 棟)(家, 戸)(家, 軒)(家, 宇)(家, 邸)
(遺体, 体)(遺体, 人)
(人員, 員)(人員, 人)
(ポイント, 点)
(野菜, 個)(野菜, 株)(野菜, 本)
(場面, カット)(場面, 幕)(場面, シーン)(場面, 場面)
(花, 片)(花, 輪)(花, 個)(花, 枝)(花, 本)
(ミサイル, 発)(ミサイル, 機)