抽出パターンの学習に基づく名詞と助数詞の呼応関係の自動獲得

矢野　修平　　白井　清昭
北陸先端科学技術大学院大学 情報科学研究科
{shuhei, kshirai}@jaist.ac.jp

1 はじめに

日本語では名詞を数える際には一般に助数詞を使い，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．日本語の解析や生成において，助数詞を適切に取り扱うためには，呼応する助数詞の情報を含む名詞の辞書が必要となる．本研究では，そのような大規模な名詞辞書を構築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大量に自動獲得する手法について述べる [5].

助数詞の呼応関係を含む日本語の名詞辞書の整備に関する研究として，Bond らは，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に対して呼応する助数詞を割り当てることにより，機械翻訳における文生成のための辞書を効率良く整備する方法を示している [1]. また，韓国語を対象とした同様の試みが Paik らによって報告されている [3]．これに対し，同じ意味クラスを持つ名詞は常に同じ助数詞と呼応するわけではないことから，本研究では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．一方，タイ語を対象とし，本研究と同様にコーパスから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]．この研究では人手で作成した抽出パターンを用いるのに対し，本研究では抽出パターンの学習も試みる点が異なる．また，Sornlertlamvanich は獲得された名詞と助数詞の組の評価については報告していない．

2 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．予備調査では，呼応関係獲得のためのパターンとして以下を用いた．

$$\text{名詞} + \text{数字} + \text{助数詞} \quad \rightarrow \quad (\text{名詞}, \text{助数詞}) \qquad (1)$$

このパターンは，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，「牛/3/匹」という文から (牛, 匹) という組を抽出する．コーパスとして日経新聞の 15 年分の新聞記事を用いた．茶筌 [1] 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，84,307 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) の例を以下に挙げる．

(夫婦, 人)　⇐ 夫婦 二 人 で気軽に足を運んでもらう
(週, 便)　⇐ 貨物便は同 6 便少ない 週 4 8 便 だった。
(全国, カ所) ⇐ ○四年には 全国 七十 カ所 に拠点を持つ

自動獲得された (n, c) をランダムに 100 個選択し，それらが呼応関係にあるかを人手でチェックしたところ，正しい呼応関係にある (n, c) の割合は 64%程度であった．式 (1) のパターンの他にも，「数字+助数詞+(の)+名詞」(ex. 2/本/の/鉛筆) や「名詞+(が)+数字+助数詞」(ex. 生徒/が/1/人) など，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が，やはり獲得された (n, c) の中には誤りが多く含まれた．

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，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ことを試みる．

3 提案手法

提案手法における処理の流れを図 1 に示す．図 1 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構築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．さら

[1] http://chasen.naist.jp/hiki/ChaSen/

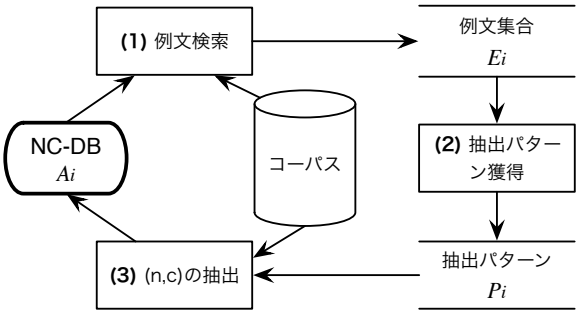

図 1: 提案手法の概要

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．

(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，3.1 項で前処理について述べた後，3.2, 3.3, 3.4 項で上記 (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 で表わす．

3.1 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．以下，数字として検出された単語は「M」で表わす．

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．そこで，EDR 日本語単語辞書の品詞が「JUN(単位)」である単語は助数詞から除外する．以下，助数詞として検出された単語は「C」で表わす．

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．ただし，先頭の単語の品詞は「名詞-接尾-一般」ではなく，末尾の単語の品詞は「名詞-サ変接続」ではないとする．
- 末尾の単語が抽象名詞ではない．抽象名詞かどうかの判定は，日本語語彙体系を用いて行った．具体的には,「1000(抽象)」以下の意味クラスに属する名詞を抽象名詞とみなした．
- 末尾の単語が「数」ではない．これは「死者数」「退職者数」のような複合名詞を除外するための条件である．

以下，名詞として検出された単語は「N」で表わす．

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．

3.2 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，呼応関係にある名詞 n と助数詞 c が含まれる文または句を抽出し，例文集合 E_i を得る．

3.3 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．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n 約５８ 点c を展示する。
パターン: た+N+約+M+C+を → (N,C)

また，上記の例では n の後に c が出現しているが，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．作成された抽出パターンの候補のうち，E_i における出現頻度が 5 未満の単語列を左辺とするパターンは除外する．

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．

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

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

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

$\langle 1\rangle$ 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．

$\langle 2\rangle$ 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 $\langle 3\rangle$ 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，

警視庁+N+M+C+は → (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〜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．$\langle 3\rangle$ 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．

3.4 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 実験

4.1 実験条件

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』[2] 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，単位を表わす助数詞や,「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．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，なるべく少量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．

コーパスは日経新聞の 2006 年の新聞記事データを用いた．また，3.3 節で述べたように，抽出パターンの獲得条件 $\langle 2\rangle$ 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 2 つの手法を用いて (n,c) の獲得を試みた．

4.2 実験結果

表 1 は手法 (A) による実験結果を表わす．

表 1: 実験結果 (手法 A)

i	1	2	3
$\lvert A_i\rvert$	964	1,808	1,845
$\lvert P_i\rvert$	17	33	34
$A_i \setminus A_{i-1}$ の正解率*	0.88	0.71	0.92
正しい (n,c) の数*	848	1,448	1,482
C の再現率	4.48%	5.06%	5.07%

手法 (A) では，図 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 $\langle 1\rangle$,$\langle 2\rangle$,$\langle 3\rangle$ 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 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わす．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．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 1 における「C の再現率」は，4.1 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，コーパスに出現する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．

次に，手法 (B) によって (n, c) を獲得した．手法 (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 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 (A) とあまり変わらなかった．

表 2: 実験結果 (手法 B)

i	1	2
$\|A_i\|$	964	1,721
$\|P_i\|$	17	34
$A_i \setminus A_{i-1}$ の正解率*	0.87	0.80
正しい (n, c) の数*	848	1,454
C の再現率	4.48%	5.26%

4.3 例

手法 (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 2 に示す．

(水彩画, 点) (観光客, 人) (逮捕者, 人)
(ミンククジラ, 頭) (和牛, 頭) (原発, 基)
(ATM, 台) (爆弾テロ, 件) (DRAM, 個)

図 2: 抽出された (n, c) の例

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 3 は獲得された抽出パターンの例である．2 節で，式 (1) のパターン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．パターン p_2 や p_3 も基本的に

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)

図 3: 獲得された抽出パターンの例

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たことがわかった．

5 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．

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．

参考文献

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

[2] 飯田朝子. 数え方の辞典. 小学館, 2004.

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

[5] 矢野修平. 名詞と助数詞の呼応関係のコーパスからの自動獲得. Master's thesis, 北陸先端科学技術大学院大学, 3 2009.