名詞と助数詞の呼応関係のコーパスからの自動獲得

矢野 修平 (0710072)

北陸先端科学技術大学院大学 情報科学研究科

2009 年 2 月 5 日

キーワード: 名詞, 助数詞, パターンマイニング, 知識獲得, 日本語.

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．