新聞記事からの数値情報の抽出と数値化*

3Z－02

小林伸行†　北川文夫　木村宏　椎名広光‡
岡山理科大学大学院 理学研究科§　岡山理科大学 総合情報学部¶

1　はじめに

近年，新聞記事やWWWなど大量のテキストが電子化されている．その大量のテキストから必要な情報を取り出すことは困難であり，効率的な検索が必要とされている．検索には，テキストに含まれる単語から事前に索引を生成し，索引を用いて検索を行う手法がしばしば用いられている．この手法では金額などの数値情報は単なる文字列として扱われるため，「十万円以下のパソコン」といった大小関係を扱う検索は行えない．テキストに含まれる数値情報を利用した検索を実現するためには，テキストから数値情報を抽出し，四則演算可能な数値に変換する（これを数値化と呼ぶ）ことが必要である．数値化を行うことで，「100,000円」「十万円」など数字の表記に関係なく，意味上の数値情報の検索が可能になる．

ここでは，新聞記事の数字を含んだ文字列を数量表現とするが，新聞記事には様々な数量表現が存在する．例をあげると「２００１年」「二〇%」「千四百六十一円」などである．これに加え，「一万五、六千円」「１、２、４回」「一八五六――九〇五」というような，範囲などの表現もしばしば見受けられるため，抽出がより困難である．数値情報の抽出法を研究したものとしては，斉藤ら[1]，や山口ら[2]などがある．これらの研究は新聞の数値情報を文字として抽出はできているものの，四則演算可能な数値への変換法は提案されていない．

そこで本研究では，単なる数だけではなく，パターンを利用して範囲などを含む数量表現を，四則演算可能な数値に変換する方法を提案する．この研究の有効性は，数字の表記に依存せず数値の検索が可能になる点である．

*Numeral Information Extract and Translation into Values from Newspaper Articles
†Nobuyuki KOBAYASHI
‡Fumio KITAGAWA, Hiroshi KIMURA, and Hiromitsu SHIINA
§Graduate School of Science, Okayama University of Science
¶Faculty of Informatics, Okayama University of Science

2　数値情報

新聞記事に存在する数量表現は主に二種類あり，それは範囲を表わす表現と度合を含む表現である．この二種類の数量表現の数値情報を『上限値，下限値，度合，単位』の4項目からなる基本数値情報を定義することで，表わすことにする．

まず，範囲を表わす表現の場合は「1000―2000円」や「九時から十一時」のような数量表現であり，これを表現するために『上限値』と『下限値』を用いる．次に，度合を含む表現の場合は「約1キロメートル」「一万人程度」のような数量表現であり，基準となる数と概数を表す単語を含む．『上限値』と『下限値』を同じ値とすることで，基準となる数のような範囲をもたない数を表現する．概数を表す単語を『度合』とする．これに『単位』を加えて，基本数値情報とする．

今後，基本数値情報を組み合わせることで，より複雑な情報を的確に検索できるようにする．複数の基本数値情報を組み合わせることで，時間と時刻，郵便番号，電話番号，住所の一部などの数値情報が表現可能になる．

3　抽出方法

基本数値情報を抽出する手順を以下に示す．

1. 茶筌[3]を用いて新聞記事の形態素解析を行い，文章を単語単位に分割し品詞付けを行う．
2. 数詞とその前後の文字列を数量表現とする．
3. 数量表現を範囲抽出パターンでマッチングし，上限値と下限値を決定し，数値化を行う．
4. 数量表現に単位用辞書を適用し，単位を決定する．該当する単位が存在しない場合は数字部分の直後の単語を単位候補語として登録する．ただし，直後の単語が助詞，助動詞，括弧などの場合は候補なしとする．

表 1: 抽出および数値化の適用結果

正しく抽出できた数字の数	抽出された数字の数	記事に含まれる数字の数	適合率	再現率
1561	1875	2088	83.2%	74.8%

5. 数量表現に度合用辞書を適用し，度合を決定する．

4 抽出および数値化の実験結果

「毎日新聞 CD-ROM '94 データ集」の1月の記事から無作為に200件を実験対象とした．ただし，抽出対象にスポーツ面は含めない．野球のスコアや順位表など，表を用いた表現が多く特殊なパターンを必要とするためである．

このデータに対して数値化を適用した結果が表1である．なお，適合率と再現率は以下の式を用いる．

$$\text{適合率} = \frac{\text{正しく抽出できた数字の数}}{\text{抽出された数字の数}}$$

$$\text{再現率} = \frac{\text{正しく抽出できた数字の数}}{\text{記事に含まれる数字の数}}$$

正しく数値化を行った例を図1に示す．

元の記事：(約六千六百億円) とみられ	
上限値：	6.60E+11
下限値：	6.60E+11
度合：	約
単位：	円
元の記事：調査二百五十―三百匹。	
上限値：	250.0
下限値：	300.0
度合：	(なし)
単位：	匹

図 1: 正しく抽出できた例

適合率低下の原因を述べるため，誤認識した314件の内訳を図2に示す．

単位の誤認識・未認識	150件
抽出すべきでない数値	73件
単語の誤認識	49件
パターン不足による誤認識	40件
文字化けによる誤認識	2件

図 2: 誤認識した数字の原因

単位の誤認識・未認識では「山田恵さん（４０）」のように年齢を抽出するパターンを追加することで対応できると考える．また，抽出すべきでない数値というのは「その１」「その２」や「(1)」「(2)」などの項目番号である．単語の誤認識は固有名詞に数字を含む場合に生じ，形態素解析のシステムに依存するため，解決が困難である．パターン不足による誤認識は，電話番号や郵便番号が正しく認識できていない．これは新たにパターンを登録することで解決できると考える．

再現率低下の原因として最も多かったのは，「１月」から「１２月」など，品詞が「数詞」ではない場合だった．これは新たにパターンを登録することで解決できると考える．

5 おわりに

本研究では，数値情報を利用した検索を実現するため，新聞記事から数値情報を抽出して，四則演算可能な数値に変換する数値化を提案した．実際の適用では，適合率74.8%，再現率83.2%と優れた結果を得た．

今後，さらに本研究を拡張して，数値化されたDBを利用した検索システムの開発や，さらに数値化法の精度向上，数値情報の単位を利用した単位間の変換機能を追加し，より完成度の高いシステムを開発する予定である．

参考文献

[1] 斉藤公一，迫田昭人，中江富人，岩井禎広，田村直良："数値情報をキーとした新聞記事からの情報抽出"，情報処理学会研究報告, NL125-14, pp.63-64, 1998.

[2] 山口 努，絹川博之："新聞記事からの数値情報の抽出と判別"，第63回情報処理学会全国大会, 1L-6, 2001.

[3] 形態素解析システム【茶筌】: http://chasen.aist-nara.ac.jp/