新聞記事からの数値情報の抽出と判別

1L－6

山口　努　　絹川博之
東京電機大学　大学院　工学研究科　情報通信工学専攻

1．はじめに

近年、コンピュータの発展により、情報量は増加の一途を辿ってきた。それに伴い、必要な情報の入手が困難になってきており、情報抽出技術の必要性が高まってきている。本研究ではそれらの膨大な情報の中から数値を伴う情報に着目し、電子化された新聞記事から数値情報を抽出することを研究目的とした。

2．数値情報の抽出

2．1　抽出項目

本研究では、数値情報を下記の５項目へと分類し、数値属性値に関してはさらに数量・時刻の２つの中項目、７つの小項目に分類した。

・大項目

①主題：数値情報の内容を表す語句
②数値属性名：属性値の内容を表す語句
③数値属性値：情報の中核となる数値部分
④変化基準：数量を修飾する語句
⑤変化方向：数量の増減を表す語

・小項目－数量－

①変化前数量：その事柄の前回の値
②変化後数量：その事柄の（今回の）値
③変化量：変化前数量と変化後数量の差

・小項目－時刻－

①変化前時刻：その事柄の前回の時刻
②変化後時刻：その事柄が起こった時刻
③変化期間：変化前時刻と変化後時刻の間
④変化決定時刻：その事柄が決定もしくは発表された時刻

Information Extraction and Discrimination of Numerical Values from Newspaper's Articles

Tsutomu Yamaguchi,　Hiroshi Kinukawa
Graduate school of Engineering, Tokyo Denki University

2．2　抽出方式

全体的な流れとしては、指定した新聞記事を読み込み、一文ごとに大項目に対応する５つの定義パターンテーブルとの文字列照合を用いて該当する語句を抽出する。

数値属性名：数値属性値直前の平仮名以外の文字列語句や「は」「が」「では」文節などを抜き出した。

主題：数値属性名が決定時に、その前の数値属性名候補を主題とみなし抽出した。

数値属性値：単位のついた数値のみを抽出し、単位のないものは対象外とした。

変化基準・変化方向：マッチングしたもののうち、数値の前後にあり直接修飾しているものを抽出した。

また、数値属性値に関しては抽出した語句の前後の連接パターンや前後にある項目の配置具合から７つの小項目へと判別した後、抽出した。数値情報の抽出方式を図１に示す。

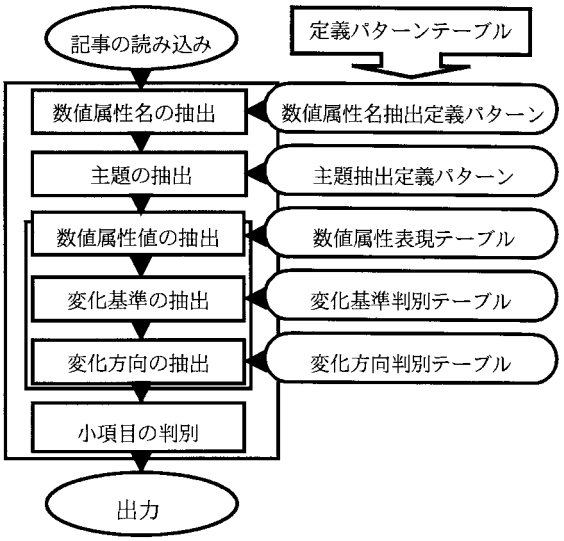

図１．数値情報の抽出方式

3．抽出実験の結果

3．1　抽出条件

抽出対象：1998 年度版日本経済新聞本紙地形面
10 月分中の数値情報記事 500 件

定義パターンテーブル内の語数

数量単位：57 個　　時刻単位：10 個

期間単位：26 個　　変化基準：112 個

変化方向：75 個

3．2　抽出結果

表１．抽出結果（件）と精度・再現率（%）

抽出項目名	正解抽出数（部分）	誤抽出	抽出漏れ	精度（部分含）	再現率（部分含）
主題	166(66)	99	89	50.2(70.1)	51.7(72.3)
数値属性名	877(128)	152	208	75.8(86.9)	72.3(82.9)
数値属性値	2514(6)	108	56	95.7(95.9)	97.6(97.8)
変化基準	276(0)	9	233	96.8(96.8)	92.3(92.3)
変化方向	799(5)	50	31	93.6(94.1)	95.7(96.3)

表２．数値属性値の判別精度

判別した項目名	正解数（件）	はずれ数(件)	精度（%）
変化前数量	15	33	34.1
変化後数量	827	49	94.4
変化量	759	56	93.1
変化前時刻	59	73	44.7
変化後時刻	159	32	83.7
変化期間	246	78	75.9
変化決定時刻	117	17	87.3

4．考察

表１、表２より精度、再現率共に主題・数値属性名が低く、特に主題は数値属性名と比べてもかなり値が低かった。これは数値属性名が数値属性値直前にあることが多いため判別しやすいのに対し、主題は位置が決まっておらず候補が多く存在することや連接パターンが数値属性名に比べて数が多く曖昧な点などが挙げられる。改善策としては、主題の連接パターンに抽出優先順位を置くなどが考えられる。この方法を確立できれば誤抽出 99 件中抽出する語句を間違えた 30 件が正解となり、精度が 79.2%、再現率が 81.6%まで上がることになる。

また、表３では数量･時刻共に変化前の精度・再現率が低いが、ほとんどの場合、その原因として下記の例のように、同じ連接パターンでも文脈から変化の前か後かが変わることが挙げられ、別の判別基準を設ける必要がある。

・**変化後時刻の例（10 月時の調査）**：市が発表した 10 月までの予算は五千万円。

・**変化前時刻の例（10 月時の調査）**：10 月までの予算は 3 月までの税収の 50%に相当。

5．おわりに

今回の結果から、現在の抽出方式では数値情報全体の抽出では主題や数値属性名の抽出、数値属性値の項目の判別では変化前の抽出が不十分であることが分かった。

本研究では、構文解析や意味解析など深いレベルではなく、表記パターンなどの表層部分からの解析による方式を重視しているが、主題や数値属性値の変化前時刻・変化前数量部分の抽出に関しては表記パターンからの判別は 100%には程遠く、特に主題抽出には構文解析や意味解析を用いる必要があることが分かった。

ただ、変化前数量は変化前時刻の後に続く場合もかなりあった為、数値を読み取るなど他の判断基準を設けることにより、現在の方式でも改良すれば精度を上げる余地は残っている。

また、数値属性名に関しては、数値の直前にあることが多い為、数値直前の連接パターンを増やすことで精度や再現率の向上に対応できると考えられる。抽出範囲の拡大を含め、以上のことが今後の課題である。

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・中川裕志：数値譲歩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 98-NL-125：情報処理学会：P63～70

[2] 井出裕二・永井秀利・中村貞吾・野村浩郷：単一項目テンプレートによる新聞記事からの製品情報抽出：情報処理学会研究報告 97-NL-122：情報処理学会：P63～70

[3] 井出裕二・藤吉　誠･永井秀利・中村貞吾・野村浩郷：構造化テンプレートを用いた新聞記事からの製品情報抽出：情報処理学会研究報告 97-NL-118：情報処理学会：P7～14