5U-03　新聞記事からの数値情報の抽出

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、新聞記事などに代表されるテキストも電子化されてきており、そこから重要な情報を抽出する必要性が高まってきている。本研究では、それら膨大な情報の中で数値を伴う情報に着目した。数値を伴う情報の特徴として、その情報の中核を構成することが多い。経済や株価などに加え、リストラなどの雇用人員の増減、イベントの日程など、数値がなければ情報としての価値がないことが多く無視できない。よって、電子化された新聞記事から数値を伴った情報の抽出を本研究の目的とする。

2．数値情報の抽出

2．1　抽出項目

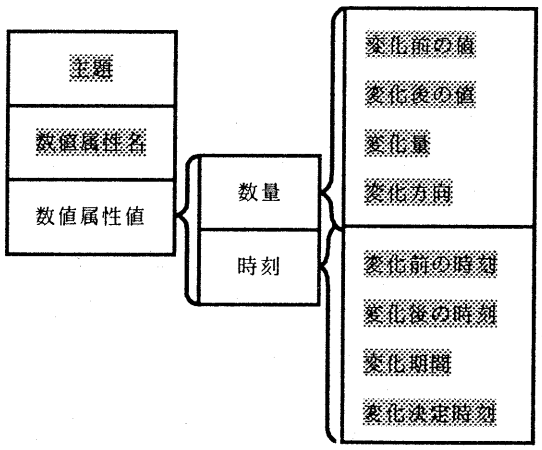

図１．抽出項目の細分化

数値情報として図１に示す項目を抽出する。

１．主題：１個以上の数値情報の意味する内容を表す語または句
２．数値属性名：その数値情報の属性値が表している内容を表す語または句
３．数値属性値：数値情報の中核をなし、数量、時刻の２種類の属性値を持つ
４．数量：時刻以外の属性値であり、変化の前後、量、方向に分けられる
５．時刻：日時を表す属性値であり、変化の前後、期間、決定時刻に分けられる

2．2　抽出方式

本研究が目指す数値情報の抽出方式は以下の通りである。

１．各記事に対して主題は一つまたは複数
２．各主題に対して数値属性名が一つ以上
３．数値属性名に対して数値属性値が一つ

また、図２の各手順について下記に記す。

① 主題の抽出方法は、直後の助詞が「は」「で」など特定のもので表現されている場合が多い。よって、それを手がかりにして抽出する方針である
② あらかじめ頻繁に出てくる数値単位をテーブルに入れておき文中の語と照らし合わせて一致したものを抽出するパターンマッチングを用いる
③ 数値属性名は数値属性値と同じ一文中にあることが多い。よって、それぞれにつ

Information Extraction about Numerical Values from Newspaper's Articles
Tsutomu Yamaguchi　Hiroshi Kinukawa
Graduate school, Tokyo Denki University

く助詞を手がかりに抽出パターンを作っておき、文と照らし合わせて抽出する方針である。また、主題と一致する可能性もあり、その場合は数値属性名を優先とする

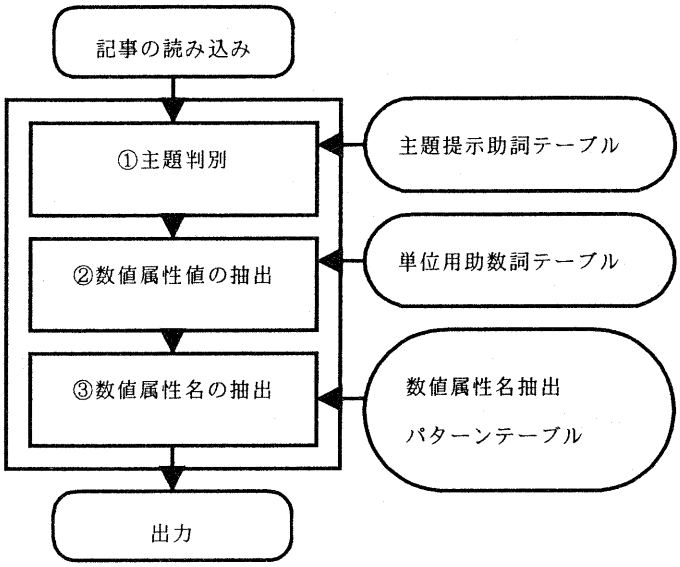

図２．数値情報の抽出方式

３．実験評価

現在は1998年度版日経新聞のCD-ROMより、あらかじめ一日単位もしくは一月単位で読み込んだ記事を用いて一文ごとに項目として区別せずに数値属性値を抽出した。このとき行った実験結果は以下の通りである。

・抽出条件

対象記事：1998年度版日経新聞
本紙地経面11月20日一日分

抽出方式：数値属性値部分の
パターンマッチング

助数詞テーブル内の単位数：76個

・抽出結果

抽出できた数値情報総数：957件

記事内にある数値情報総数：987件

正しく抽出できた数値情報数：895件

精度：93.5%　　再現率：90.7%

４．考察

数値属性値部分をパターンマッチングにより抽出した場合、「信二社長」「一人一人」等の表現を誤抽出してしまった。また、抽出できなかった数値属性値92件のうちテーブルに単位を入れてなかったものが80件で、残りの12件は数詞のみの記述だった。よってこれらを改善していく必要がある。

また、主題を判別する為には直後に続く助詞の出現頻度について検討する必要がある。数値属性名に関しても、文の構成をパターン化する為に、数値属性名と数値属性値の後に続く助詞について検討する必要がある。

５．おわりに

本研究では数値情報を３つに分けたが、そのうち数値属性値に関しては既存の辞書は使わない方針であるので、数値単位などの数を増やしていく必要がある。また、数値属性値の抽出精度の向上や項目として区別することや、数値属性名と主題の明確な区別についての検討および抽出が今後の課題である。

６．参考文献

[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・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

[3] 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