岡山理科大学紀要 第38号 A pp. 141-147(2002)

新聞記事における数量表現認識のための読点分類法

小林 伸行*・木村 宏**・椎名 広光**

*岡山理科大学大学院理学研究科博士課程応用数学専攻

**岡山理科大学総合情報学部情報科学科

(2002年11月1日 受理)

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には、単位や度合を含んだ数値や範囲を表す数量表現が存在する。現在、テキストに対する検索手法としては、全文検索が一般的であり、テキストに含まれる単語から事前に索引を生成しておき、検索を行う際には索引を用いて検索を行う。しかし、この手法では金額などの数量表現は単なる文字列として扱われる。そのため、数量表現を活かした検索は行えない。例えば、「十万円以下のパソコン」といった大小関係を扱う検索を行うことができない。テキストに含まれる数量表現を利用した検索を実現するためには、テキストから数量表現を抽出し、四則演算可能な数値に変換する(これを「数値化」と呼ぶ)ことが考えられる。数値化を行うことで、数量表現を、数値や範囲に単位あるいは度合を付加した数値情報として扱うことができる。これによって「100、000円」、「十万円」など数字の表記に関係なく、数値の大小関係や範囲を指定した検索が可能になる。

本論文で取り上げた新聞記事には、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—一九〇五」というような、範囲などの表現もしばしば見受けられる。そのため正しく数値情報を抽出することが困難である。これまで数値情報の抽出法を研究したものとしては、斉藤ら[1)]、山口ら[2)]などがある。しかしながら、これらの研究は新聞の数量表現を抽出すること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。そこでわれわれは数値情報を表す構造を定義し、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の提案を行う。

通常、数量表現には読点「、」やダッシュ「—[1]」などを用いない単独の形式が多いが、読点やダッシュを用いた数量表現のほうが単独の数量表現に比べて、情報の重要度が高い。しかし、単独の数量表現に比べると件数が少ないので、正しく数量表現を認識するためには十分な計算機実験を行い、精度を評価することが必要不可欠である。本研究では、読点やダッシュなどを含む範囲を表す数量表現の抽出や分類についても述べる。特に、値表現に関する読点の分類に分類木を用いた新しい分類アルゴリズムを提案し、計算機による評価実験を行う。

本論文では、まず2章で数値情報の要素である基本数値情報の定義について述べる。次に3章では、数量表現を数値情報に変換する際に必要となる読点の分類を述べた後、読点の分類アルゴリズムを提案する。4章では、分類アルゴリズムを評価するための計算機実験について概要を述べ、5章では、計算機実験の結果を考察する。最後に、本論文のまとめと今後の課題を6章で述べる。

2. 数量表現の認識について

2－1 基本数値情報の定義

数値情報を扱うための要素として基本数値情報を定義する。基本数値情報は『上限値』、『下限値』、『度合』、『単位』の4項目からなり、新聞記事に存在する2種類の数量表現、すなわち範囲を表す表現と度合を含む表現の両方を表すことができる。

まず、範囲の表現は「1000—2000円」や「九時から十一時」のような数量表現であり、これを表すために『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「1」、「一万」と概数を表す単語「約」、「程度」を含む。また、『上限値』と『下限値』を同じ値にすると、基準となる数のような範囲を持たない数も表現することができる。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。例えば、「約一万五千円」と「六階から九階まで」は次のようになる。

[1]新聞記事は縦書きのため、チルダ「～」をあまり用いない。

・例1:「約一万五千円」の場合
上限値: 15000
下限値: 15000
単位: 円
度合: 約

・例2:「六階から九階まで」の場合
上限値: 9
下限値: 6
単位: 階
度合: なし

この基本数値情報の構造には2種類あり、構文定義をBNF記法で表すと次の通りである。ここで、数量表現の中の数字部分を値表現として別に定義する。例えば、「1、2、4位」や「一万五、六千円」などで、下線部が値表現である。

基本数値情報1 ::=
[前置度合] [前置単位] 値表現
[後置単位] [後置度合] (1)
基本数値情報2 ::=
基本数値情報1 [範囲表現]
[基本数値情報1] [範囲表現] (2)

ただし、[]は省略可能を表す。

式(1)の基本数値情報1は、ひとつの値表現で、数値や範囲、複数の数値を表す構造である。ここで、値表現には、数値、範囲、複数の数値を含む。基本数値情報1の例としては、「約六百億円」、「3万—5万円」などがあり、下線部が式(1)の値表現である。

一方、式(2)の基本数値情報2は、複数の値表現で数値、範囲などを表す定義である。基本数値情報2の例としては、「六階から九階まで」、「24日から4日間」などがあり、下線部が式(2)の値表現を表す。

2－2 基本数値情報の抽出

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 基本数値情報の定義に従い、数量表現の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。
⑤ 数量表現に度合用辞書を適用し、度合を決定する。

3. 読点の分類

3－1 読点の種類

数量表現を数値情報に変換するときは、式(1)を用いる。このとき、値表現に読点を含むと、読点は読点前後の数量によって複数の意味を持つために、正しい数値の判定が困難である。例えば、「340、245ミリリットル」は「340ミリリットルと245ミリリットル」を表し、「一、七一一メートル」は「1711メートル」を表すが、読点の種類が正しく判断されないと、「340245ミリリットル」や「1メートル、711メートル」として変換される。他の例として「五万二、三千人」、「一八六〇、七〇年代」は正しく認識されると「52000人、53000人」、「1860年代、1870年代」だが、「50002人、3000人」、「1860年代、70年代」と誤認識される恐れがある。ここでは数量表現を数値情報に変換する際の読点の違いを読点の意味から、『桁区切り』、『列挙』、『置換』、『前後組合せ』、

表 1 読点の分類と数量表現の例

読点の分類	意味	数量表現の例	実際の数値
桁区切り	「,」の代わり、ひとつの数値を表す	一三、五四二トン 一五、六七〇平方メートル	13542 トン 15670 平方メートル
列挙	前後の数量がそれぞれの数値を表す形式	五百、千トン 23、30日	500, 1000 メートル 23, 30 日
置換	前数量の下2桁を後数量に置き換えた形式	1991、92年度 一九六〇、七〇年代	1991, 1992 年度 1960, 1970 年代
前後組合せ	前後の数量を組み合わせた形式	一万七、八千円 六、七十人	17000, 18000 円 60, 70 人
その他	名前や前後の数量に関連がないもの	1、2－ジクロロメタン 1992・12・23、234回	なし なし

『その他』の5種類に分類する(表1)。それぞれの分類項目を示す。

桁区切り：『桁区切り』は、縦書きの新聞特有の表現と考えられ、桁を表すカンマの代わりで、「一、〇〇〇、〇〇〇円」のように3桁置きに用いる。つまり、読点前後の数量を全てまとめてひとつの数値を表している。したがって、読点を含む値表現を数値に変換する場合に、値表現をひとつの値として変換する必要がある。

列挙：『列挙』は、同じ単位を持った数量を、読点で区切って複数列挙しているにすぎない。例えば、「二、三、四、八ヶ月」などである。したがって、読点前後二つの数量は、別々に数値に変換され、複数の数値を表す。

置換：『置換』は、主に「年」を表すとき、「一九八九、九一年」のように用いられる。読点の後ろの数量は、読点の前の下2桁を置き換えた数値を表す。したがって、読点の前後二つの数量を各々数値に変換した後、読点の後ろの後数量は、読点の前の前数量の百の位以上の値を付加した数値に置き換える。すなわち、「1968、69年」は「1968年」と「1969年」として数値化する。

前後組合せ：『前後組合せ』は、「五万二、三千円」のような値表現を「52000円」と「53000円」として数値化する際の分類である。このような数値を認識させるためには、値表現を三つの部分に分割する必要がある(図1)。すなわち、読点の前後1文字までを「判断部」と呼び、判断部の前を接頭部、後を「接尾部」と呼ぶことにする。そして、判断部の読点の前後の数字それぞれに対して、「接頭部」と「接尾部」の数字を組み合わせることで正しく数値に変換できる。

その他：上述の4種類に該当しないものを『その他』とした。『その他』には、例えば化学物質の名称内に読点が出現するものや、「12・29、34回」が「12月29日」と「34回」を表すような前後の数量の単位が異なる数量表現などが含まれる。

3－2　読点の分類方法

前節で述べた読点の分類は、数量表現を数値情報に変換する際に必要である。読点を分類するために利用した新聞記事は、既にCD-ROM化されている「毎日新聞 CD-ROM '94 データ集」、「毎日新聞 CD-ROM '95 データ集」である。この中から1ヶ月分の新聞記事データに対して、「茶筌」を用いて形態素解析を行い、文を単語に分割し品詞付けを行う。その後、「数詞、読点、数詞」の順番に並ぶ語句をすべて取り出す。次に、取り出した語句とその前後の文脈などを判断して人手によって、表1の5種類に分類する。すなわち、現在、読点の前数量と後数量の分類のためのアルゴリズムが存在しないため、膨大なデータの分類を人手に頼らざるおえない。文脈判断せずに値表現だけで自動的に判定を行い分類ができるならば、高速で高能率なアルゴリズムが提案できる。

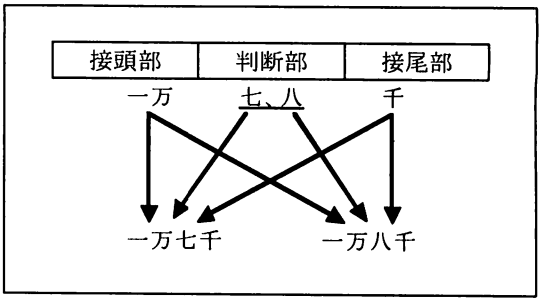

図 1　前後組合せの認識方法

読点の種類を判別する分類木を生成するための分析項目は、次に示す7項目である。この項目はすべて値表現のみから得られる。

① 前数量の文字数
② 後数量の文字数
③ 位の有無
④ 中点の有無
⑤ 前数量と後数量の差が9未満
⑥ 前数量と後数量が等しい
⑦ 前数量の最後の文字と後数量の最初の文字が連番

これらの分析項目のうち「位の有無」は、前数量と後数量のどちらかに「十、百、千、万、億、兆、京」などの位を表す漢字を含むかどうかの判断である。また、「中点の有無」は、小数点を表す中点が、前数量と後数量のどちらかに含まれるかを示す。なお、中点を小数点として用いるのは、新聞特有の表現である。さらに、「前数量と後数量の差」の計算は、読点前後の数量を個別に数値に変換し、それらの数値を比較する。例えば「五万二、三千」の前数量と後数量は、それぞれ「50002」と「3000」になる。最後の項目の「前数量の最後の文字と後数量の最初の文字が連番」では、「0、1」と「九、十」の組み合わせは、連番とみなさないことにする。これは「10、11人」や「九、十六日」など「前後組合せ」にはならないため、読点を分類するためには利用できないからである。

本論文では、図2に示す分類木で読点の種類を判

定する。

① 後数量の文字数は3である。
Yes→②、No→⑦

② 位がある。Yes→③、No→④

③ 前数量と後数量の差は9未満である。
Yes→『列挙』、No→『前後組合せ』

④ 前数量の文字数は3以下である。
Yes→⑤、No→⑥

⑤ 前数量と後数量は等しい。
Yes→『その他』、No→『桁区切り』

⑥ 中点がある。
Yes→『その他』、No→『列挙』

⑦ 位がある。Yes→⑧、No→⑨

⑧ 前数量の最後の文字と後数量の最初の文字が連番である。
Yes→『前後組合せ』、No→『列挙』

⑨ 前数量と後数量の差は9未満である。
Yes→『列挙』、No→⑩

⑩ 前数量の最後の文字と後数量の最初の文字が連番である。
Yes→『前後組合せ』、No→⑪

⑪ 前数量の文字数は4である。
Yes→『置換』、No→『列挙』

4．読点分類の計算機実験

4－1　実験の概要

前章で提案したアルゴリズムを用いて、実際に読点の分類を行う。実験データは分類木の生成に利用した「94年1月」および「94年10月」、「95年4月」、「95年12月」の記事4ヶ月分をデータとする。記事から「茶筌」によって「数詞、読点、数詞」の順に出現する部分を取り出す。取り出した部分を実験データとし、図2の分類木を用い分類を行い、実際の記事上の語句と対応しているかを確認する。ただし、スポーツ面の記事は特殊な数量表現が多いため、今回は除外する。スポーツ面は他の新聞記事と異なり、テニスの試合の結果「2—1(6—4、3—6、6－1)」のようなスポーツ面特有の表記やマラソンの順位と記録を表形式でを含むからである。

4－2　実験の評価法

実験結果を表2に示す。評価は次式で求められる適合率と再現率で行う。

$$\text{適合率} = \frac{\text{正しく分類された数}}{\text{分類木で分類された数}} \tag{3}$$

$$\text{再現率} = \frac{\text{正しく分類された数}}{\text{分類項目に属している数}} \tag{4}$$

表2から各月のデータ数を見ると、どの月も550件から700件ほど取り出すことができた。極端に出現数が少ないものはなく、分類項目の出現数についても、どの月もほぼ同じくらいの割合になっている。

なお、表2の『正しく認識』は分類木を用いて正しく

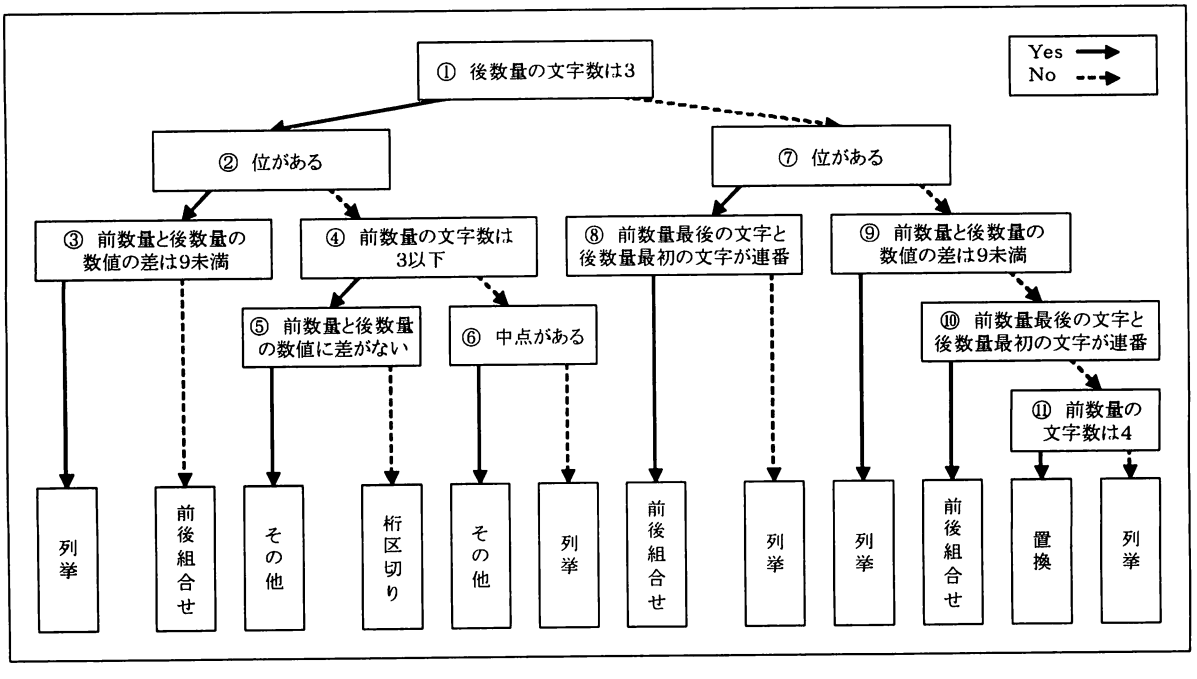

図 2　読点の分類木

表 2　分類された読点の数と精度の評価

	データ数						列挙					前後組合せ				
	列挙	前後組合せ	桁区切り	置換	その他	計	正しく認識	誤認識	未認識	適合率	再現率	正しく認識	誤認識	未認識	適合率	再現率
1994年1月	451	58	24	10	23	566	449	16	2	96.6%	99.6%	58	1		98.3%	100.0%
1994年10月	547	76	54	8	11	696	520	9	27	98.3%	95.1%	76	4		95.0%	100.0%
1995年4月	429	50	33	7	37	556	426	38	3	91.8%	99.3%	50	3		94.3%	100.0%
1995年12月	504	45	34	6	12	601	497	12	7	97.6%	98.6%	45	5		90.0%	100.0%
計	1931	229	145	31	83	2419	1892	75	39	96.2%	98.0%	229	13		94.6%	100.0%

	桁区切り					置換					その他				
	正しく認識	誤認識	未認識	適合率	再現率	正しく認識	誤認識	未認識	適合率	再現率	正しく認識	誤認識	未認識	適合率	再現率
1994年1月	24	1		96.0%	100.0%	10	1		90.9%	100.0%	6		17	100.0%	26.1%
1994年10月	54	23		70.1%	100.0%	8			100.0%	100.0%		2	11	0.0%	0.0%
1995年4月	31		2	100.0%	93.9%	6	1	1	85.7%	85.7%	1		36	100.0%	2.7%
1995年12月	30	3	4	90.9%	88.2%	6	2		75.0%	100.0%	1		11	100.0%	8.3%
計	139	27	6	83.7%	95.9%	30	4	1	88.2%	96.8%	8	2	75	80.0%	9.6%

分類できた数である。また、『誤認識』は分類木を用いて分類した分類項目が誤っていた数であり、『未認識』は正しい分類項目の中で認識できなかった数である。例えば、正しい分類項目は『列挙』であるが、分類木のアルゴリズムを用いた分類項目が『置換』と誤っている場合、『置換』の分類項目は『誤認識』になり、『列挙』の分類項目は『未認識』となる。

5．実験結果

5－1　列挙の分類

『列挙』の分類は、95年4月の適合率が低くなっている。これは「0120・899901、 899802」のような電話番号の省略をうまく判別できないためである。しかし、電話番号を認識するためには、前述の通り前後の文脈を判断する必要があるので、今回の評価の範囲外である。それにもかかわらず、91.8%と良い結果を示している。また、94年10月の再現率が、他の月に比べて低くなっている。この原因は、特集面の「94広島アジア大会のゴルフの記録」というスポーツ記事や3面の「JT株の当選番号発表」の記事である。特集面の記事では、「日本（横尾、小島、尾家） 864（218、215、216、215）」のように、ゴルフの団体戦の合計スコアの後に個人の記録を列挙している。このようにスポーツ記事は、実験の概要でも示したが、日常表記と異なりスポーツ特有の表記を用いるため認識率が低下する。また、大きなスポーツイベントの記事は、スポーツ面以外の1面や特集面に掲載されることがあり、記事の内容を分類することも必要と考えられる。一方、3面の記事は「【下三ケタ】195、293、332、366、695、803、887」のように、3桁の数字が連続で続く場合である。このように3桁の数字が連続で続く場合は、『桁区切り』と『列挙』を判別するための他の分析項目が必要になる。『列挙』全体の判定は適合率、再現率ともに95%を超え、かなり優れた結果となっている。

5－2　前後組合せの分類

『前後組合せ』の誤認識は、「31、2日7時」、「一九四三、四四年の二年間」のように位がない場合で、偶然連番になる場合である。そのため、文脈を判断する必要があり、今回の提案アルゴリズムからは範疇外となる。再現率については100%と今回のテストデータで最も良い結果を得た。

5－3　桁区切りの分類

『桁区切り』の分類で94年10月の誤認識は、『列挙』の分類で述べた。それ以外の誤認識は「プルトニ

ウム239、240などを含む」、「このところ100、101円台の狭い幅」、「小売価格は340、245ミリリットル缶とも」の3件だけである。これらの値表現を『桁区切り』と『列挙』に正しく分類するためには、文脈を判断して分類する必要があり、値表現だけで分類することは、かなり困難だと考えられる。一方、再現率低下の原因である未認識は「前日終値四、一六八・四一ドル」などであり、小数点を含む場合のデータが認識できていないことがわかった。これは分類木に小数点を含む場合も考慮することで、改善可能である。『桁区切り』全体の適合率は83.7%、再現率も95%を超え、非常に良い結果を示している。

5－4 置換の分類

『置換』の分類は、データ数が少ないので1件認識できないだけで、85.7%と再現率の低下が大きい。94年1月の誤認識は、記事中の表であり、「1938・1・5～2・13、35回」と人手によっても困難な分類である。しかし、95年4月の「七三・五、八六・一 、八〇・〇、七五・二」や95年12月の「排気量は二三〇〇、二九〇〇CC」は、「後数量の文字数を2」と限定する分析項目を追加することで解決できる。

未認識は、既に『前後組合せ』の分類で述べた「一九四三、四四年の二年間」であり、文脈を考慮して分類する必要があるので単純な分類木による分類では不可能である。

『置換』全体の適合率と再現率を見ると、それぞれ88.2%、96.8%と良い結果を示している。

5－5 その他の分類

今回、最も精度が悪かった『その他』の分類は、94年1月のデータでさえ精度が30%より低いものである。この原因は、分析項目が値表現だけに限定したことにある。出現数も少ないため、94年1月で分類できた例と同様の分類は、他の月には出現しなかった。逆に、他の月では出現したものは、94年1月では見られなかった、新たな分類が出現したため全体的に『その他』の精度は低くなっている。これを改善するためには、今後さらにデータを増やし、分類木に分析項目を増やすことで改善できる。

5－6 実験結果の考察

表2を見ると今回提案したアルゴリズムによる分類はよく適合していると結論付けられる。今回用いた分析項目はすべて値表現だけで判定を行った。読点の分類は、値表現の情報だけで、実用化に近いレベルで可能であることがわかった。このことは、これまでの研究で明らかにされておらず、特筆すべき結果といえる。ただし、『その他』の認識精度が悪いため、『その他』の分類精度を向上させるためには、さらなるパターンの追加が必要である。さらに認識精度を向上させるためには、今回提案した値表現の情報だけによる自動分類に加え、従来から研究されているような数量表現前後の単語を含めて文脈を解析する必要が生じる。

6. まとめ

本論文では、数量表現を四則演算可能な数値に変換することで、数字の表記に依存しない範囲や数値の検索できることを示した。特に数量表現を数値情報に変換する際の問題において、読点の種類による分類がある。ここでは数値への変換方法の違いから読点を5種類に分類した。この読点の種類を自動的に判別するために、分類木で生成したアルゴリズムの提案を行った。計算機による実験結果では、『その他』の分類以外は、すべて83%以上の精度があり、非常に良い分類アルゴリズムであることが証明された。このことは、読点の分類が、値表現の形式情報だけで、ほぼ90%以上区別できることを示しており、特筆すべき結果といえる。

今後の課題は、分類精度の向上が考えられる。今回の計算機実験結果から、『その他』の分類精度を上げるために、分析項目を増やさなければならない。また、さらなる精度の向上を行うためには、分類木の分析項目に文脈解析を加えた修正を行う必要がある。また、「読点」とともに数量表現の認識に問題となる「範囲」を表すダッシュ「—」についても、同様に分類アルゴリズムを提案し、精度評価を行う必要がある。また、検索プログラムのプロトタイプを作成する作業が残されている。

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/

Classifying the Japanese Punctuation Mark 'Touten' for Recognition of Numerical Expressions in Newspaper Articles

Nobuyuki KOBAYASHI*, Hiroshi KIMURA** and Hiromitsu SHIINA**

Department of Applied Mathematics, Faculty of Science,
**Graduate School of Science,*
***Department of Information Science, Faculty of Informatics,*
Okayama University of Science
1-1 Ridai-cho, Okayama 700-0005, Japan
(Received November 1, 2002)

We can now access a huge corpus such as Japanese newspaper articles through WWW or other media. However, it is difficult to retrieve items including numerical expressions, for example, "personal computers 100,000 yen or less", because identical expressions do not always have unique meaning. In this paper, we propose a method to recognize the meanings of one type of Japanese punctuation mark, 'Touten', which is used in various kinds of numerical expressions. Our concern is restricted to Japanese newspaper articles as the Touten notation is the most popular notation used. We classify the meanings of Touten into five categories and propose an algorithm to convert numerical expressions including Touten into numerical values with a set of attributes to specify the meaning of values. We obtained good recognition rates in computer experiments using the proposed algorithm.