情報学基礎 64-18
自然言語処理 145-18
（２００１．９．１１）

係り受けの制約と優先規則に基づく数量表現抽出

藤畑 勝之　志賀 正裕　森 辰則
横浜国立大学
E-mail: {fujihata,shig,mori}@forest.eis.ynu.ac.jp

本稿では，質問応答に用いる数量表現の取り扱いについて考察する．数値表現はそれ自身は単独では意味をなさず，どの事物のどの属性に関しての値であるかが判明して初めて有用な情報になり得る．そのため，それぞれの数値表現に関して，数値が対応する事物とどのような係り受け構造で結び付くかを考慮する必要がある．そこで，数量表現の出現する文脈について考察し，数値情報の構造を考慮した係り受けの制約と優先規則に基づく抽出規則を提案する．さらに新聞記事を対象とする実験により，本規則が数値情報を含む文書中から数値およびそれに付随する情報を抽出するうえで精度の向上に役立つことを示す．

Extraction of Numerical Expressions by Constraints and Default Rules of Dependency Structure

Katsuyuki Fujihata　Masahiro Shiga　Tatsunori Mori
Yokohama National University

In this paper, we study the treatment of numerical expressions in question-answering systems. Numerical expressions themselves do not convey enough information, because they are only some attribute values. In combination with the objects and the attribute names, numerical expressions convey useful information. Therefore, we propose a set of extraction rules for numerical expressions, which extract numerical expressions along with the supplementary expressions. The set of rules consists of the constraints and preference rules of dependencies among numerical expressions and the supplementary expressions. An experiment of extraction we conducted with news paper articles shows that our rules are useful for improving the accuracy of extraction.

1　はじめに

近年，膨大かつ多様な電子化文書が利用可能になっているのにともない，利用者にとって必要な情報を効率良く入手する手段を要求する声が高まっている．これを受けて情報抽出や情報検索の研究が盛んに行なわれるようになり，特に質問応答システムに関心が高まりつつある [TRE00] [福本 01]．質問応答システムは利用者の問いに対してその解となる文書部分を文書データベースから発見し提示するシステムである．その質問においては利用者が距離，時間などの数量表現に関する情報を要求する頻度は高いと思われる．

数量表現は固有表現と異なり，事物を表す物ではなく，物事の性質を記述するいわゆる属性である．すなわち，単体では意味を持たず、別の物事と結び付いて特定の表現となるため，ある数量が「何について」,「どのような観点」の値を示しているのかが判明して初めて有用な情報になりうる．これらは「長さ」,「高さ」,「幅」,「速さ」,「重さ」のように，各種尺度にあわせて存在し，多様である．

よって，数量表現だけでなく，対応する事物と組にして文書中から抽出する事が必要である．また，そのようなシステムを質問応答システムに組み込むことで，解の同定が容易になるといったことが考えられる．

本稿では，上記の背景の下，数量表現の出現する文脈について考察し，数値情報の構造を考慮した係り受けの制約と優先規則に基づく抽出規則を提案する．さらに新聞記事を対象とする実験により，本規則が数値情報を含む文書中から数値およびそれに付

随する情報を抽出するうえで精度の向上に役立つことを示す.

2 数値表現の分類

数値表現は事物と結び付くことによって意味をなす．よって、数値情報とはその数値と事物の間の関係と捉えることができる．ここでは，これを n 項組で表現する．数値情報には以下の 6 種類が考えられる．

1. 物 (object) の属性値を表すもの.

< 物, 属性, 数値 > 3 項組で特徴付けられる.

(1) a. 東京タワーの高さは 333m です.
b.< 東京タワー, 高さ,333m>

2. 物の数量を表すもの.

< 物, 数値 > の 2 項組で特徴付けられる.

(2) a. 新型 PC ○○を 100 台出荷した
b.< 新型 PC ○○,100 台 >

3. 物の集合の中の特定の物を表すもの.

< 物の集合，数値 > の 2 項組で特徴付けられる.

(3) a.3 台目の PC を購入した.
b.<PC,3 台目 >

4. 事 (event) の属性値を表すもの.

< 事, 属性-属性値 > の 2 項組で特徴付けられる.

(4) a.1997 年，香港が中国に返還された.
b.< 香港に中国が返還された, 年-1997>

5. 事の数量を表すもの.

< 事, 数値 > の 2 項組で特徴付けられる.

(5) a. ○○大統領は 3 回来日した.
b.< ○○大統領は来日した,3 回 >

6. 事の集合の中の特定の事を表すもの.

< 事の集合，数値 > の 2 項組で特徴付けられる.

(6) a. 富士山に登るのは 2 回目だ.
b.< 富士山に登る，2 回目 >

数値情報には，単位を表す表現 (m，台，回など) もしくは属性を表す表現 (年など) が付加されている．また，数値表現は何らかの係り受け構造を介して対応する事物と関連を持つはずである．よって，これらの情報より，ある数値情報がどの分類の構造を持ち，どの事物と関連するかが推定可能であると考えられる．そこで．本稿ではそれぞれの分類において，数値情報構造 (上記 n 項組) を抽出する手法を提案する.

型 4 ならびに型 5，型 6 については,「事」(evant) に纏わる数値情報であるから，対応関係を明らかにするには文全体を解析し命題構造を抽出する必要がある．質問応答システムの一部として数値情報抽出を行なうことを考えると，命題構造解析は質問文との対応において行なわれるので，あらかじめ文書に対して行なえるのは数量表現自身の抽出だけである.

一方，型 1 ならびに型 2 については，命題中のある特定の物，および型 3 についてはそれが属する集団に関する情報構造を表している．よって，数値の周囲の文脈を解析することによって，前もって事物，属性との対応関係を明らかにすることができると考えられる.

そこで，次節では，型 1 ならびに型 2，型 3 について，それぞれ，上記 3 項組，2 項組の各項目が実際の文の中でどのように現れるかを考察する.

3 数値情報の表現の類型

本節では，まず，各関係が係受けによりどのように構成されるかを考察し，次に，数値を含む表現における言語上の関係を示す．これらを組み合わせることにより，各数値に対して，属性，物を文書中に見つける手掛かりとなる.

3.1 数値表現における係受け構造制約

3 項関係 < 物, 属性, 数値 > は，言語表現の上では 2 項関係である係受け関係の組み合わせで構成される．各組み合わせを次に示す．ただし，各矢印部分においては，左右どちらか一つの方向であり，また，一つの項は高々一つの項にしか係らないという制約がある.

(a) 属性⇔数値⇔物

(b) 物⇔属性⇔数値

(c) 数値⇔物⇔属性

このうち，(c) の係受け構造は，数値-属性の係受け関係が直接ないために，現実の文としては現れないと考えられる．また，ゼロ代名詞により 2 つの文で一つの 3 項関係を表現している場合には，(a) もしくは (b) の一部のみが得られる．

2 項関係 < 物, 数値 > および < 集合，数値 > は当然ながら次に示す一通りしかない．

(d) 数値⇔物/集合

3.2 数値に関する係り受け表現

数値に纏わる 2 項間の係り受け関係には，以下の表現が考えられる．

1. [物/属性] の表現を含む文節が数値表現を含む文節に係る．

(1) 例 新型 PC100 台

2. 数値表現を含む文節が [物/属性] の表現を含む文節に係る．

(2) 例 333m の東京タワー

3. 数値表現を含む文節と [物/属性] の表現を含む文節のそれぞれが述語に係る．

(3) 例 東京タワーは 333m の塔だ．

4. 述語の連体形 (関係節) が [物/属性] の表現を含む文節に係る．

(4) 例 100 台出荷される新型 PC

2 項間の係り受けにおいて，2 項が数値と属性であった場合は，さらに物が，数値もしくは属性と上に述べた 2 項間の係り受け関係にある．

数値に纏わる 3 項間での係り受け関係には，以下の表現が考えられる．

1. 属性の表現を含む文節が数値の表現を含む文節に係り，数値の表現を含む文節が物の表現を含む文節に係る．

(5) 例 幅 10m の道路

2. 数値の表現を含む文節が属性の表現を含む文節に係り，属性の表現を含む文節が物の表現を含む文節に係る．

(6) 例 1t の重さの自動車

3. 属性の表現を含む文節が数値の表現を含む文節に係り，数値の表現を含む文節と物の表現を含む文節がともに述語に係る．

(7) 例 エベレストは高さ 8848m を誇る．

4. 物の表現を含む文節が属性の表現を含む文節に係り，属性の表現を含む文節が数値の表現を含む文節に係る．

(8) 例 箱の重さは 10kg だ．

5. 数値の表現を含む文節が属性の表現を含む文節に係り，属性の表現を含む文節と物の表現を含む文節がともに述語に係る．

(9) 例 エベレストは 8848m の高さを持つ．

6. 物の表現を含む文節と属性の表現を含む文節がともに数値の表現を含む文節に係る．

(10) 例 東京タワーは高さが 333m だ．

7. 属性の表現を含む文節が数値の表現を含む文節に係り，動詞の連体形 (関係節) が物の表現を含む文節に係る．

(11) 例 高さ 8848m を誇るエベレスト

8. 数値の表現を含む文節が属性の表現を含む文節に係り，動詞の連体形 (関係節) がものの表現を含む文節に係る．

(12) 例 8848m の高さを誇るエベレスト

よって，以上のような係り受け関係にある数値と物または属性の格助詞，述語の品詞および態を解析することによって数値表現と対応する物または属性を絞り込むことができる．

3.3 属性の抽出

3 項関係の場合，係り受け構造制約によって属性と物とを判別することは難しい．これは，構造制約において，物と属性が可換な構造をしている箇所が

あることや，文節単位の係り受けを解析するツールを使用した場合，詳細な係り受け構造が解析できないために，構造制約を適切に適用できないことが原因である．

現在のところ，我々はこの問題に対して，表現が属性であるか否かを別途判定することにより対処している．まず，属性と物に関する構造制約を緩め，属性と物を区別せずに係り受け構造のみで候補を決定する．次に，その候補の各々について概念辞書などを用いて属性となり得るかを調べ，なり得る場合に属性として扱い，さらに属性の周辺の係り受け構造を解析することで物の候補を決定し 3 項を抽出している．我々の実験では概念辞書として EDR 概念辞書 [日本 93] を用いている．

4 数値情報抽出システム

前節までに述べた，数値と物との間にある係り受け構造制約を考慮して数値情報を抽出するシステムを構築した．

4.1 システムの概要

1. 文書を形態素解析システム juman[黒橋 98b] を用いて形態素解析する．このとき名詞と判定された形態素のうち，本来助数辞となるべきものの品詞を助数辞に修正する．この修正は Step 4 において，数値文節内の名詞として誤って助数辞が選ばれるのを防ぐためである．さらに，修正した形態素列を入力として KNP[黒橋 98a] を用いて構文解析を行なう．

2. 数詞を含む文節 (数値文節) を見つけ，各数値文節に対して Step 3 以降の操作を行なう．

3. 数値文節が係る用言を見つけ，これらの周辺の名詞を物の候補とする．物の候補となるのは，

   - 数値文節に係る名詞
   - 数値文節が係る名詞
   - 用言の文節に係る名詞
   - 用言の連体形が係る名詞
   - 文書の提題
   - 数値文節内で数詞の直後にある名詞

   がある。

4. 物の候補から，4.2 節で述べる優先規則を基に物を決定する．この時点では，2 項関係と 3 項関係のいずれであるか判定を行なっておらず，物は係り受け構造のみを参照して決定されているので属性にもなりうる．

5. Step 4 で決定した物が属性であるかどうか判定を行なう．属性でなかった場合はそのまま数値と組になる物として提示する．属性であった場合は属性の周りの名詞を新たに物の候補に加える．この中から新たに優先規則を用いて物を決定し，数値と物と属性を提示する．

4.2 係り受け優先規則

節 3.2 で述べた要素を組み合わせ，実際の文書における正解率と出現頻度を基に抽出規則を作成した。この規則は図 1 のように小規模な決定木となっており，各葉には優先規則を表す決定リストが接続されている．

図 1 では名詞文節の表記を簡略化しており，格助詞は KNP の解析結果によるものを用いている．また，b,f,p の記号はそれぞれ

b - 数値文節に係る名詞

f - 数値文節が係る名詞

p - 数値文節が係る述語に係る名詞

を表している．

この規則では，まず数値文節の格助詞を判定する．数値文節が「ノ格」,「ヲ格」あるいは文末だった場合以外は，さらに用言が判定詞であるか動詞の能動態・受動態であるかによって適用する優先規則を変えている．例えば「A 社が自動車を 100 万台出荷した。」と「A社によって自動車が 100 万台出荷された。」という文章は同じ意味を持ち，動詞の態が変わると主格と目的格が入れ替わるため,「自動車」をどちらの文章においても正しく抽出するには優先順位を変える必要がある．

ノ格と文末の場合を除いたのは，数値文節が文末だった場合は用言が判定詞となることと，数値文節がノ格の場合は，述語に関係なく，数値と直接係り受けする名詞文節が物となる可能性が高いことに基づいている．

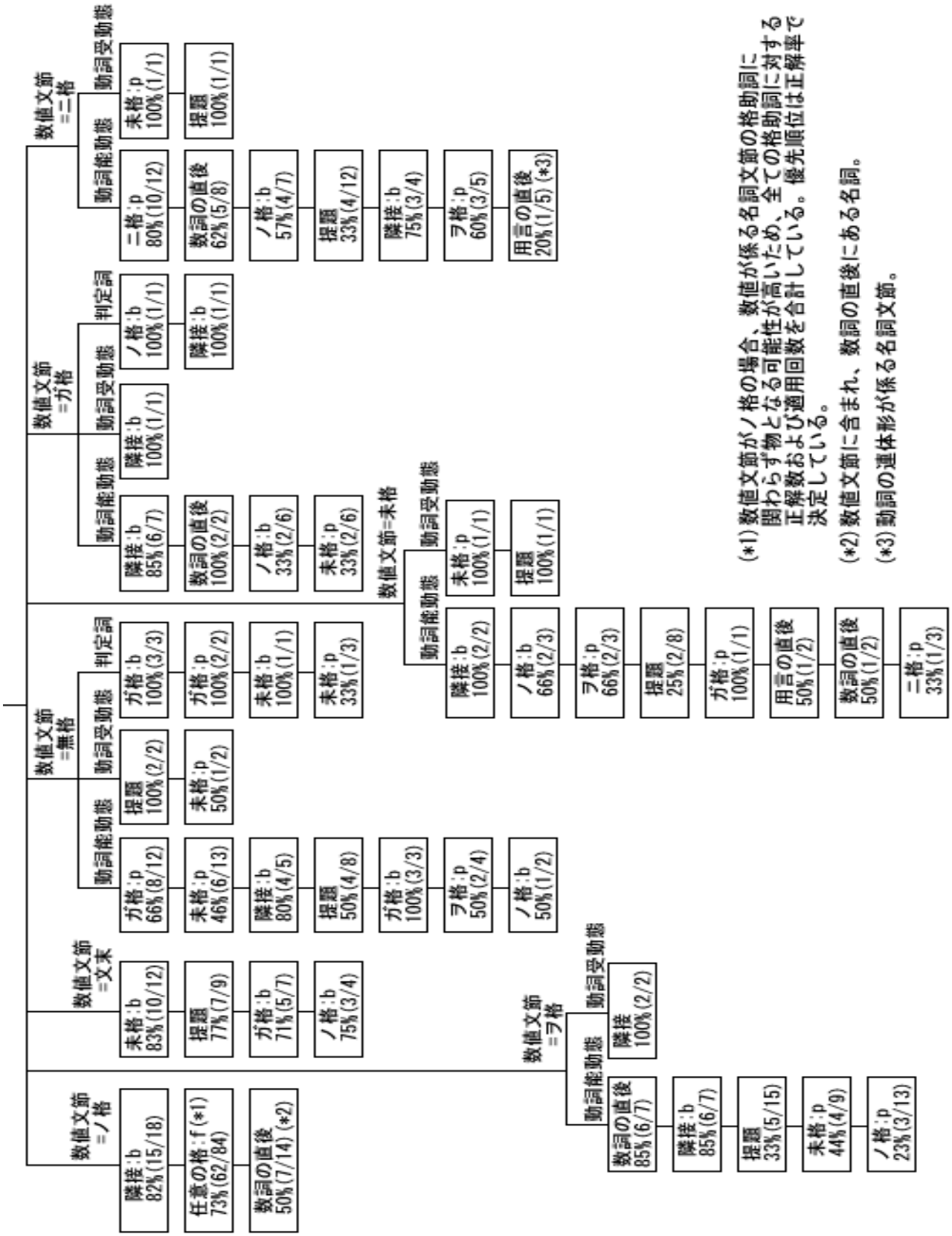
数値文節=ノ格
隣接:b 82%(15/18)
任意の格:f(*1) 73%(62/84)
数詞の直後 50%(7/14)(*2)
数値文節=文末
未格:b 83%(10/12)
提題 77%(7/9)
ガ格:b 71%(5/7)
ノ格:b 75%(3/4)
数値文節=無格
動詞能動態
ガ格:p 66%(8/12)
未格:p 46%(6/13)
隣接:b 80%(4/5)
提題 50%(4/8)
ガ格:b 100%(3/3)
ヲ格:p 50%(2/4)
ノ格:b 50%(1/2)
動詞受動態
提題 100%(2/2)
未格:p 50%(1/2)
判定詞
ガ格:b 100%(3/3)
ガ格:p 100%(2/2)
未格:b 100%(1/1)
未格:p 33%(1/3)
数値文節=ヲ格
動詞能動態
数詞の直後 85%(6/7)
隣接:b 85%(6/7)
提題 33%(5/15)
未格:p 44%(4/9)
ノ格:p 23%(3/13)
動詞受動態
隣接 100%(2/2)
数値文節=ガ格
動詞能動態
隣接:b 85%(6/7)
数詞の直後 100%(2/2)
ノ格:b 33%(2/6)
未格:p 33%(2/6)
動詞受動態
隣接:b 100%(1/1)
判定詞
ノ格:b 100%(1/1)
隣接:b 100%(1/1)
数値文節=未格
動詞能動態
隣接:b 100%(2/2)
ノ格:b 66%(2/3)
ヲ格:p 66%(2/3)
提題 25%(2/8)
ガ格:p 100%(1/1)
用言の直後 50%(1/2)
数詞の直後 50%(1/2)
二格:p 33%(1/3)
動詞受動態
未格:p 100%(1/1)
提題 100%(1/1)
数値文節=二格
動詞能動態
二格:p 80%(10/12)
数詞の直後 62%(5/8)
ノ格:b 57%(4/7)
提題 33%(4/12)
隣接:b 75%(3/4)
ヲ格:p 60%(3/5)
用言の直後 20%(1/5)(*3)
動詞受動態
未格:p 100%(1/1)
提題 100%(1/1)
(*1) 数値文節がノ格の場合、数値が係る名詞文節の格助詞に関わらず物となる可能性が高いため、全ての格助詞に対する正解数および適用回数を合計している。優先順位は正解率で決定している。
(*2) 数値文節に含まれ、数詞の直後にある名詞。
(*3) 動詞の連体形が係る名詞文節。

図 1: 係り受け優先規則

優先規則は，成立条件と抽出判定の組を決定リストにしたもので，係り受けの種類毎に存在する．各決定リストの中の各項目の順位は，訓練文書を用いて決定した．すなわち，各項目についてその成立条件が適用されなおかつ判定が正しかった場合の数 (正解数) を数え，その降順に項目が並べてある．

このとき，正解数が等しい場合は正解率の高いものを上位としている．また，正解率が 20%を下回る項目は決定リストから省いている．

正解率は次のように求めている．例えば「A社は自動車 100 台の出荷を予定している。」という文に対する構文解析結果は図 2 のようになり，
数値文節はノ格を持ち，
図 1 における名詞文節の表記を用いて物の候補を表すと

- A社 (未格:p)
- 自動車 (隣接:b)
- 出荷 (ヲ格:f)

があることがわかる．このうち物として適切なのは「自動車 (隣接:b)」である．このような判断を人手によって行ない，訓練文書に対して係り受けの種類ごとに正解率を求める．本稿では毎日新聞記事 200 程度の文書から集計を行ない，優先順位を決定した．

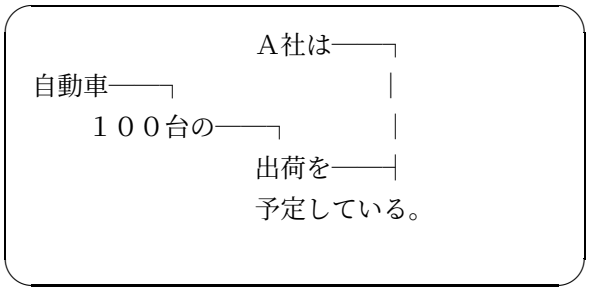

図 2: 構文解析結果の例

5 評価実験

本稿で提案する手法の有効性を示すために他の手法によるものと比較実験を行なった．ただし属性を含む 3 項組を同定するシステムは従来提案されていないので，ここでの実験は「物-数値表現」の 2 項組の抽出に限定している．

(A) 係り受け優先規則に基づくシステム
前節までに述べた手法に基づく抽出システム．(提案手法)

(B) パタン駆動型数値情報システム
構文解析を行なわず，あらかじめ作成しておいた情報抽出パターンと表層表現との照合によって抽出をおこなうシステム．[斉藤 98][SITN98] (パタン)

(C) 数値からの距離文節数が最も近い名詞を物とするシステム (ベースライン)

実験は，毎日新聞記事 (94,95,97,98 年)200 記事から数値情報の抽出を行ない，文書中に出現する数値から物を正しく推定できたものを正解とした．実験には，前節における抽出規則に用いたものとは異なる記事を用いている．実験の結果を表 1 に示す．

表 1: 毎日新聞記事からの数値情報抽出 (2 項組)

手法	適合率	再現率	F 値
提案手法	82.7% (225/272)	87.2% (225/258)	0.85
パタン	58.9% (159/270)	61.6% (159/258)	0.60
ベースライン	54.5% (181/332)	70.2% (181/258)	0.61

同様に，提案手法を用いて毎日新聞記事から 3 項関係の数値情報抽出を行なった．実験の結果を表 2 に示す．

表 2: 毎日新聞記事からの数値情報抽出 (3 項組)

	適合率	再現率	F 値
属性	86.7%(13/15)	68.4%(13/19)	0.76
物	68.0%(17/25)	94.4%(17/18)	0.79

6 考察

表 1 によれば，提案手法が再現率および適合率において他の手法に勝っている．しかし，表 2 によると，3 項関係における属性の抽出精度が 2 項関係における抽出精度に比べて低いことがわかる．これは，3 項関係の数量表現が利用した訓練文書集合にあま

り存在しなかったために，精度の良い抽出規則が作成できなかったことが原因として挙げられる．

2 項関係を正しく抽出できなかった例を挙げると，「3 人の性格が一致しない。」という文から「性格、3 人」を 2 項組として抽出しており，「人」を助数辞にもつ数値に対応する物として「性格」が不適当であることを判定する必要がある．この問題に対して，今後 EDR 概念辞書を用いた改良を行なう予定である．

次に < 物，属性，数値 > の 3 項を正しく抽出できた例を挙げる。「パワーショベルで雪を半円筒状に掘り下げた幅１９メートル、長さ９０メートルのコースはカチカチの氷状態。」という文から「コース、幅、１９メートル」と「コース、長さ、９０メートル」の 2 つを出力しており，係り受け優先規則を用いることによって 1 つの物に複数の属性と数値を含んだ表現からも正しく 3 項組を抽出することに成功した．

一方，3 項関係の抽出に失敗した例としては，「銅鐸の中でも高さ 12，13 センチを下回るものは小銅鐸と呼ばれる。」という文から「高さ、12，13 センチ」の 2 項しか出力しなかったというものがある．本来物である「小銅鐸」は数値の表現を含む文節，属性の表現を含む文節，動詞の連体形 (関係節) のいずれとも直接係り受けの関係にないため，「銅鐸の中でも高さ 12，13 センチを下回るもの」と「小銅鐸」の表すものが同一であることを解析する必要がある．

7　まとめ

本稿では，数量表現に関して，その情報構造に着目し数値情報の前後の文脈を考慮することで数値情報の抽出を行なう手法を提案した．数値情報の構造を考慮した係り受けの制約と優先規則に基づく抽出規則を適用することで，ある程度の精度で数値情報抽出を行なうことができることが，小規模ながら実際の文書に基づく実験で確認できた．

今後，より精度の高い抽出を行なうためには，さらに多くの訓練文書に基づき，抽出規則を改善していく必要がある．現在のところ，訓練文書数が少ないこともあって，抽出規則中の決定リスト (優先規則) を人手で決定しているが，より多くの訓練事例があれば，この部分を決定リストの学習アルゴリズム [Yar94] などで学習することも可能であろう．

参考文献

[SITN98] Koich Saito, Yoshihiro Iwai, Naoyoshi Tamura, and Hiroshi Nakagawa. Numerical information extraction from newspaper articles. In *Proceedings of the 3rd International Workshop on Information Retrieval with Asian Language*, Oct 1998.

[TRE00] TREC Project. *Proceedings of The Eighth Text Retreival Conference TREC 9*. http://trec.nist.gov/pubs/trec9/t9_proceedings.html, 2000.

[Yar94] D. Yarowsky. Decision lists for lexical ambiguity resolution: Application to accent restoration in spanish and french. *In 32th Annual Meeting of the Association for Computational Linguistics*, pp. 88–95, 1994.

[斉藤 98] 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 98-NL-125, 自然言語処理研究会, 情報処理学会, May 1998.

[福本 01] 福本淳一, 加藤恒昭. Question and answering タスクの提案. 言語処理学会研究報告 2001-FI-63-4, 言語処理学会, 7 2001.

[黒橋 98a] 黒橋禎夫. 日本語構文解析システム KNP version 2.0b6 使用説明書. 京都大学大学院 情報学研究科, 1998.

[黒橋 98b] 黒橋禎夫, 長尾真. 日本語形態素解析システム JUMAN version 3.6 使用説明書. 京都大学大学院 情報学研究科, 1998.

[日本 93] 日本電子化辞書研究所. EDR 電子化辞書使用説明書, 1993.