質問応答システムにおける数量表現の取り扱い

山下 竜之　藤畑 勝之　角田 久直　志賀 正裕　森 辰則
横浜国立大学 工学部 電子情報工学科
E-mail: {yard,fujihata,tsunoda,shig,mori}@forest.dnj.ynu.ac.jp

1 はじめに

情報検索と情報抽出を融合した技術として，近年，質問応答 (QA) システムが注目されている．QA システムとは，例えば「現在の世界の人口は何人か?」という自然言語による問に対して「約 60 億人」というように，望みの情報を含む文書ではなく，文書から取り出したその情報自身を提示するものである．現在提案されている最も一般的な QA システムは，4W1H 型の質問文を受け付け，情報検索技術に基づくパッセージ検索と，固有表現抽出など情報抽出を組み合わせて解を発見し，出力する [TRE99, TRE00]．

ここで，4W1H 型の質問における，解を分類すると次の 2 種類となる．一つめは，who,where,what,when に対する解で，人名，地名，日時などいわゆる固有名が主な値として期待される．二つめは，how に対する解で，距離，時間などの数量表現である．

前者については，直接対応する技術として固有表現抽出が存在し，解候補の抽出に貢献している．

一方，後者の数量表現は，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，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．それゆえ，質問の種類としても，(どれくらいの) 長さ，高さ，幅，速さ，重さのように，各種物理尺度に併せて存在し，多様である．これは，解自身についてその周りの文脈を考慮する必要があることを示しており，QA システムの構築の観点においては，それ以外の解の場合と異なる．

そこで，本稿では，QA システムにおいて数値表現を解とする場合の精度向上を目指し，数量表現の出現する文脈について考察し，そこから得られた制約を QA システムに応用することを検討する．

2 QA システムにおける数値解

本節では，基本的な QA システムの構成・手順について述べ，数値情報を解とする場合がその中でどのような位置付けになるかを考える．

2.1 基本的な QA システム

基本的な QA システムは以下の手順から構成される [TRE99, TRE00]．

1. パッセージ検索により，質問文中の語を含むパッセージを対象文書群からみつける．
2. 質問文中の疑問詞相当表現から，質問の問うている物の種類 (質問型) を推定する．
3. その質問型に照合しうる表現をパッセージ中に見つけ，初期解候補とする．この過程においては，例えば固有表現抽出技術などが利用される．固有表現の種類のうち，質問型と矛盾なく対応がとれるものを対象として抽出を行なう．
4. 各初期解候補に対して，文書中においてそれが関与する命題構造を抽出する．抽出された各命題構造を質問文の命題構造と照合し，矛盾なく対応が取れたものを二次解候補とする．
5. この方法では，複数の二次解候補が得られることと，解候補の抽出の精度の低さを補うために，最終的な解を一つに決定する．これには多数決などの方法が用いられる．

その中でも解の抽出精度を左右する段階は Step 3(初期解抽出) と Step 4(命題照合) である．このうち，Step 3に比べて，Step 4の方が次に述べる通り計算コストがかかる．

Step 4については，係受け構造 (依存構造) の変形操作が行なわれる．すなわち、質問文から得た係受け構造，パッセージ中の初期解候補を含む係受け構造のいずれか，もしくは，両方を変形しつつ，表現の間の対応づけが矛盾なく行なわれることを判定する．この変形ならびに照合の操作は係受け関係が複雑になるほど，計算コストがかかる．

実時間応答を目指す QA システムにおいては，処理時間が短いことが重要であることは言うまでもない．そのためには，Step 3においてできるだけ解候補を絞り込むか，Step 4における照合操作を何らか制約を考慮することにより削減することが効果的である．

2.2 数値表現に関する QA

数値情報について考えると，係受け解析を行なわずに抽出で，数値に必ず付随する情報は単位情報だけである．この場合，Step 3(初期解抽出) としては，同系列の単位を持つ数値を抽出するだけにとどまる．しかし，Step 1(パッセージ抽出) においては，すべての語を含むパッセージが必ず存在するとは限らず，また，一部の語のみを含むパッセージにも，当然，解が含まれることもあるので，解候補を確実に得るためには，制約を緩くせざるを得ない．よって，初期解候補として，同系列の単位を持つ数値が数多く取り出され，コストが高い Step 4(命題抽出) での絞り込に持ち越されてしまうことが考えられる．

そこで，固有名に対する固有表現抽出のように，数値情報についても多少のコストを払っても候補を十分に絞りこめる情報を抽出する機構を考える．これは，

Step 4における作業の一部を，数値情報に特化した方法で行なうものである．

3 QA 向け数値情報抽出

数値情報については，考慮すべき係受け構造が限定されるので，これを用いて，一般的な命題構造の照合より制限の強い照合を行なうことを考える．

3.1 数値表現の分類

QA の解となる数値は，ある事物に関する数値情報である．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事ができる．ここでは，これを n 項組で表現する．

数値情報には以下の 4 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 ⟩ の 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. 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．
   (3) 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⟩
4. 事の数量を表すもの．
   ⟨ 事, 数値 ⟩ の 2 項組で特徴付けられる．
   (4) a. ○○大統領は 3 回来日した．
       b. ⟨ ○○大統領は来日した,3 回 ⟩

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) が付加されているが，この情報より，ある数値情報がどの分類の構造を持ちえるかが推定可能であると考えられる．よって，それぞれの分類において，数値情報構造 (上記 n 項組) を抽出する手法を考える．

型 3ならびに型 4については，「事」すなわち命題に纏わる数値情報であるから，質問文とパッセージ中の文の照合においては，命題構造全体を対象にするしかない．

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だけに関する情報構造を表している．よって，その部分だけの処理を行ない，物に関する数値情報の構造をあらかじめ抽出することにより，質問文とパッセージ中の文の照合における命題の照合の範囲を狭めること考えられる。そこで，以下では，型 1ならびに型 2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える．

3.2 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．これは，節 3.3で述べる．

3.2.1 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せで構成される．各組み合せを次に示す．

(5) (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 ⟩ は当然ながら次に示す一通りしかない．

(6) (d) 数値⇔物

3.2.2 数値に関する係受け表現

数値に纏わる二項間の係受け関係には以下のものがある．

1. 隣接関係
   (7) 例 高さ 333m、新型 PC100 台
   制約 物 / 属性 - 数値 のみ
2. 助詞ノによる係受け
   (8) 例 333m の高さ
3. 同値を表す述語や判定詞による命題上の関係．
   二項の一方がガ格 / ハ格で，もう一方が直接，判定詞に係受け．
   (9) 例 高さは 333m だ
4. 同値関係にある名詞を介した命題上の関係．
   判定詞の直前に一般名詞．二項のうち，数値 - 属性 / 数値 - 属性がその一般名詞にノで係受け．物がガ / ハ格で係受け．
   (10) 例 東京タワーは 333m の塔だ
5. 浮遊限量子 (Floating Quantifier) による命題上の関係
   (能動形において) 物がヲ格で述語に係受け，数量が述語に直接係る．
   (11) 例 新型 PC ○○を 100 台出荷した

3.3 数値情報抽出

各数値 (数 + 単位相当表現) に対して，「物」，「属性」を見つける手続きは以下の通りである．

1. 節 3.2.2で述べた「数値に関する係受け表現」に基づき，各数値に対して係受けしている表現，ならびに，各数値が係受けしている表現を見つける．
2. 節 3.2.1で述べた「数値表現における係受け構造制約」に基づき，その候補が制約を満足しているか否かを確認する．3 項関係の場合においては，同様にして更に 1 項を見つける．

ただし，3 項関係の場合，以上の手続きによって得られた項目のうちどちらが属性か判別がつかないことがある．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、文節単位の係受けを解析するツールを使用した場合，詳細な係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．

我々はこの問題に対して，現在のところ，表現が属性である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 EDR 概念辞書で属性となり得るかを調べ，なり得る場合に属性として扱う．また，構文解析器の解析誤り等も考慮して，「数値に関する係受け表現」においては，優先順位に基づく制御をおこなっている．

4 スコア付けに基づく数値情報の照合による解候補の決定

次に考えるべき事は，抽出された数値情報と質問文から抽出された数値情報との間の照合を行ない，「適切な」最終解に絞り込むことである．

ここで考慮すべきことは，質問文から得られた数値情報と検索パッセージから得られたそれとの間で完全に照合できる場合はほとんどないということである．これは，表記の不一致や記述の詳細さの差異の他に，数値情報抽出における解析誤りが存在するためである．

そこで我々は各々の数値情報について照合スコアを算出する手法を採用した．すなわち，照合の度合に応じて加点をし，また，解の確からしさが低い場合には減点を行なうという方法により，照合スコアを算出する．そして，スコアの高い数値情報内の数値を解とする．以下に各数値情報に対する加点項目ならびに減点項目を示す．

- 加点項目
  1. 質問文中の語がパッセージ内に存在する．
  2. 質問文中の語が，注目している数値情報が取り出された文中に存在する．
  3. 共通する語の数値表現との関係 (属性 / 物の関係を問わず) が決定されている．
  4. 上記において，共通する語の数値表現との関係が同一である．
- 減点項目
  1. 加点項目 3 の状態で，対応するもう一方の文中に同じ数値表現との関係を持つ異なる語が存在する．
  2. 伝聞や予定など断定的ではない文から抽出されている．
  3. 数値に対して概数表現が付随している．

5 評価実験

前節までに記述した手法に基づき、実験システムを試作し、小規模な評価実験を行なった．

質問文は，数値表現に関する疑問詞を一つだけ含むものを 20 個用意し，文書データベースとしては WWW 上の文書 (文書検索には goo を使用)，及び毎日新聞記事 (94，95，97，98 年) をそれぞれ用いた．また，各解候補に対する照合スコアにより，1 位のみを正解とする場合と 3 位までに解があれば正解とする 2 種類の方法を考慮した．比較を行なうベースラインとして、加点項目 1 だけを用いた場合についても実験を行なった。これは、一致する語句の出現のみを判定するものであり、いわゆるキーワードのみに注目する QA システムに相当する。

得られた解の精度を表 1ならびに表 2に示す。また、各実験において得られる平均解個数を表 3ならびに表 4に示す．

表 1: WWW 文書に対する正解率

	WWW 文書		
	1 位 (1 件のみ)	1 位 (複数)	3 位まで
加点項目 1 のみ	25%(5/20)	60%(12/20)	90%(18/20)
全項目	50%(10/20)	60%(12/20)	85%(17/20)

表 2: 新聞記事に対する正解率

	新聞記事		
	1 位 (1 件のみ)	1 位 (複数)	3 位まで
加点項目 1 のみ	24%(4/17)	70%(12/17)	100%(17/17)
全項目	59%(10/17)	59%(10/17)	82%(14/17)

6 考察

表 1および表 2によれば、同スコアで 1 位となっている解候補に正解が含まれている場合を正解とする評

表 3: WWW における解の平均個数

	WWW 文書	
	1 位	3 位まで
加点項目 1 のみ	2.7	14.8
全項目	1.8	10.0

表 4: 新聞記事における解の平均個数

	新聞記事	
	1 位	3 位まで
加点項目 1 のみ	5.0	17.5
全項目	1.5	5

価においては、WWW 文書においてはベースライン手法と同じであるもの、新聞記事における実験では、ベースライン手法が勝っている。しかし、これは、表 3、表 4を見れば、解の平均個数の差に依存することがわかる。すなわちベースライン手法では、上位のスコアの差があまりなく、解の絞り込みが不十分である。実用的な QA システムとして考えると、解候補が複数存在することは望ましくない。そこで、1 位となっている解候補が 1 つに絞り込まれており、なおかつ、それが正解である場合を正解として、評価を行なう。同じく表 1ならびに表 2を見ると、WWW 文書ならびに新聞記事のいずれにおいても、提案手法がベースラインに対して、倍以上の精度で解の同定に成功している。これは、本手法の有効性を示すものである。

スコアが 3 位までの解候補まで考慮した場合には、当然、正解を含む比率は高くなるものの、どちらの手法においても、解候補が多過ぎて絞り込みが十分にできないことがわかる。

文書の違いによる解精度の違いについては、文の記述の仕方に関する均一性より、新聞記事における解の抽出精度の向上のほうが、WWW 文書の場合よりも大きいと期待していたが、今回の実験ではその差異がみられなかった。

次に照合により抽出を誤った例を示す。

- 数値情報抽出における誤抽出

  たとえば、

  (12)質問 東京タワーの展望台の高さは何メートルですか。

  文書 高さ３３３ｍ，１０００人を収容する展望台（１２０ｍ地点）まで１分間で . . .

  においては、⟨ 展望台, 高さ,333m⟩ と誤抽出した。

- 表形式であった為に解析できなかったもの

  ベースライン手法では構文解析などを使わないため、表形式でも対応できるが、本手法においては、数値情報の一部が一致するものの正しくない解が加点項目のために相対的に上位となることがあるので、精度が悪くなった。

7 関連研究

Srihari et al.[SL99] に代表される基本的な QA システムにおいては、固有表現抽出が重要な役割を果たしている。我々の研究は、4W1H 型の質問において固有表現抽出で対応しづらい数値表現を扱っている点で従来の固有表現抽出に基づく手法と相補的であると考えられる。

一方、命題レベルの照合に重点をおいた研究としては、村田ら [村田 00] の研究がある。その手法は、質問文ならびにパッセージ中の文の両者に変形規則を適用し，照合可能な構造を見つけ、解を得るというものである。これは、命題の照合を行なうための一つの手法であるが、やはり、コストが高いようである。詳細な命題レベルの照合を行なう前に、我々の手法による数値情報抽出を組み合わせれば、数値表現を解とする場合のコストの削減が期待できる。

8 まとめ

本稿では、QA システムの初期解を絞り込むという目的で、数値情報抽出を行なう手法について考察した。この技術は、固有表現抽出技術を補うものとして、位置付けられるが、数値情報に関する検索質問において、完全な命題照合を行なわず、数値情報同士の構造的な照合を行なうだけでも、それ単独で、ある程度の精度の QA 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。

今後の課題としては、以下のものがある。まず、数値情報の構造解析の精度がまだ十分ではない点である。今後、実例に基づき、その向上を目指す。さらに、今回実験で利用した加点の点数はひと手で割り振ったものであるが、実例に基づき最適な点数を見つける必要がある。

参考文献

[SL99] Rohini Srihari and Wei Li. Information extraction supported question answering. In *Proceedings of The Eighth Text Retreival Conference TREC 8*, 1999.

[TRE99] TREC Project. *Proceedings of The Eighth Text Retreival Conference TREC 8*. http://trec.nist.gov/pubs/trec8/t8_proceedings.html, 1999.

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

[村田 00] 村田真樹, 内山将夫, 井佐原均. 質問応答システムを用いた情報抽出. 言語処理学会第 6 回年次大会発表論文集. 言語処理学会, 3 月 2000.