

照応解析を含む質問応答用テストコレクションの構築と評価

白地 隆弘[†]

[†] 横浜国立大学 大学院 環境情報学府

E-mail: {shirachi,mori}@forest.eis.ynu.ac.jp

森 辰則[‡]

[‡] 横浜国立大学 大学院 環境情報研究院

1 はじめに

質問応答システムは、自然言語で与えた質問に対し、文書群より適切な回答を導くシステムである。その中でシステムの新たな形態として対話を理解し、回答ができる対話型質問応答が期待されている。ここで対話型質問応答とは、「日本の首相は誰ですか。」という質問に続く「何歳ですか。」のような関連する質問群に答えられるシステムである。質問応答システムの評価型ワークショップ NTCIR-5 QAC3 では、対話理解を模し、このような照応表現を含む関連質問群に答えられるシステムの評価が行われた。

QAC3 で想定される質問応答は、幾つかのモジュールから構成される複合的なシステムにより実現されることが想定されている。特に、質問文に含まれる照応表現の解析が重要である。一方で、QAC3 では、質問とそれに対する回答が含まれた公式ラン用のテストセットと照応表現を代表的な先行表現に置き換えた参照ラン用の質問が提供されていたが、各モジュール毎の性能評価を行うためには、参加者が個別にモジュールの入出力を調べて独自に評価する必要がある。

そこで、システム内の代表的な処理モジュールを想定し、その入出力に関する情報を QAC3 のテストコレクションに付与することにより、モジュール毎の詳細な評価をし、失敗解析を容易にする拡張テストコレクションを構築する。さらに、QAC3 に参加したシステムの評価を実際に行いテストコレクションの有効性を確かめる。

2 先行研究

2.1 関連質問群に対する質問応答

国立情報学研究所主催の評価型ワークショップ NTCIR-5 QAC3 では、一連の関連質問群 (質問シリーズ) に対し、回答を過不足なくリストにして返す質問応答 (リスト型質問応答) の評価が行われた [1]。QAC3 には 16 のシステムが参加し、そこで各種提案がなされた。提案された手法は大別すると二手法に分けられる。

一つは、文書やパッセージ (文書の断片) を検索する際に、以前の質問に現れる語を加えるなどして、対象文書を絞り込む手法である。例えば、Murata et al. [2] は、過去の質問文の一部や質問の回答を「履歴」と考え、回答を探索する範囲となるパッセージを、その履歴との類似度が高くなるように検索する手法を提案している。

もう一つは、質問文中に現れる照応表現について、その先行詞で補完し、一問一答型の質問応答システムの入力とする手法である。Matsuda et al. [3] は質問文における参照関係を、代名詞によるもの、ゼロ代名詞によるもの、修飾表現もしくは被修飾表現の省略によるものの 3 つに分類し、選択制限や文内での登場位置等の情報に基づき先行詞を一意に決定している。そして、得られた解釈に基づき一つの補完質問文を生成、質問応答を行っている。

本研究で構築したテストコレクションは、照応解析を行う後者の手法のシステムを評価の対象としている。

2.2 照応解析に情報源を利用した質問応答

川口ら [4] は、関連質問群に対して照応解析を行った後、曖昧性の残る幾つかの補完候補から、システムが内蔵している情報源と補完候補との整合性を考慮して補完文の決定を行うという手法を提案している。このシステムは節 2.1 で述べた QAC3 に参加したシステムである。手法で大別すると、質問文に対して照応解析を行う手法に分類される。

このシステムは、照応表現の検出、照応表現の補完された質問文候補群の生成、情報源と質問文候補との整合性による補完質問文決定・回答の取得、次の質問のための先行詞候補リスト作成の 4 つの処理を行い、関連質問群の個々の質問に回答する。

情報源と質問文候補との整合性による補完質問文決定は、回答スコア分布を考慮した。石下ら [5] は、優先順位型質問応答 [6] の回答スコア分布により、入力された質問の回答が適切に見つけられているかを判断することが可能であると述べている。これより、最も回答が適切に見つけられる質問を、情報源との整合性がとれた質問であるとして補完質問文と決定している。

本研究では、評価対象のシステムに上記のシステムを選び、評価実験を行った。

3 提案するテストコレクション

3.1 概要

3.1.1 システムの代表的な処理の想定

照応解析の処理は、照応表現 (照応詞・ゼロ代名詞) を検出する処理、照応表現の照応対象となる先行詞同定、次の質問のための先行詞候補を収集する処理から構成されると想定した。また、これらの処理で生成した質問文を、一問一答型質問応答システムの入力として回答を取得し、次の質問の処理を行う、という処理の流れを想定している。

3.1.2 テストコレクションが持つ情報

節 3.1.1 で想定した各部分処理の全てに対して、その入出力に関する情報をテストコレクションに付与しておけば、多くのシステムで、かつシステム内の各モジュールで詳細な評価が可能となる。

テストコレクションは以下のような情報を持っている。

- 質問文、質問文 ID
- 質問文の形態素解析結果
- 質問文の係り受け解析結果
- 補完情報 (照応表現、先行詞の情報)
- 補完した質問文の一例

なお、テストコレクションは XML 形式で記述をしている。テストコレクションの一部を図 1 に示す。タグなどそれぞれの情報は、節 3.3 で詳しく述べる。

```

<q id="QAC3-30003-02">
  <sentence id="QAC3-30003-02-Q" qtype="日付">デビューしたのはいつですか。</sentence>
  <tree id="QAC3-30003-02-T1" phrase="いつですか。">
    <bunsetsu>
      <morpheme pos="副詞">いつ</morpheme>
      <morpheme pos="判定詞">です</morpheme>
      <morpheme pos="助詞">か</morpheme>
      <morpheme pos="特殊">。</morpheme>
    </bunsetsu>
    <subtree>
      <tree id="QAC3-30003-02-T2" phrase="デビューしたのは">
        <bunsetsu>
          <morpheme pos="名詞">デビュー</morpheme>
          <morpheme pos="動詞">した</morpheme>
          <morpheme pos="名詞">の</morpheme>
          <morpheme pos="助詞">は</morpheme>
        </bunsetsu>
        <subtree>
          <tree id="QAC3-30003-02-T3" attr="動作主" particle="が">
            </tree>
          </subtree>
        </subtree>
      </tree>
    </subtree>
  </tree>
  <completion>
    <reference id="QAC3-30003-02-R1">
      <referer rtype="ゼロ" refid="QAC3-30003-02-T3">
        </referer>
        <antecedent refid="QAC3-30003-01-Q" start="1" end="5">浜崎あゆみ</antecedent>
      </reference>
      <c-sentence id="QAC3-30003-02-C">浜崎あゆみがデビューしたのはいつですか。</c-sentence>
    </completion>
  </q>

```

図 1: テストコレクションの一部

3.2 想定される利用方法

各モジュールが処理を終了する度に、モジュールの出力とそれに対応するテストコレクション中の正しい情報を比較して評価することを想定している。

また、照応解析を含む質問応答システムは複合的なシステムであるため、途中の処理で失敗してしまうとそれ以降の処理の失敗解析ができなくなる。そこで、評価対象処理の直前の処理まで正しく処理できている状態を再現する。システム内のある部分の処理を評価したとき、処理が失敗していると判断すれば、失敗しているシステムの出力を、それに対応するテストコレクション中の正しい情報と入れ替えて次の処理を続ける。これにより、処理が失敗した箇所を的確に見つけることができる。

3.3 テストコレクションの情報の説明

ここではテストコレクションに含まれるタグなどの情報について詳しく述べる。

<q> タグで囲まれている情報は、質問文 1 問当たりの情報である。この属性 id の値は、QAC3[1] で決められた質問文 ID である。

また、<sentence> タグで囲まれた文が質問文である。このタグ内の属性 id の値は、質問文自身の ID である。もう一つの属性 qtype は質問型である。図 1 では「日付」

であり、日付を問う質問であることが分かる。

<tree> タグで囲まれた情報は質問文の係り受け解析結果である。質問文を文節単位に分割し、文節同士の係り受け関係を明らかにしている。また、文節ごとに ID を振っている (属性 id)。タグ構造は構文木を模して木構造になっており、外側のタグの文節ほど木の根に近い文節 (受ける側の文節) である (係る側の文節は <subtree> タグ内にある)。属性 phrase にはその文節の表層表現が入っている。また、ゼロ代名詞・ゼロ連体格代名詞が存在するとき (ゼロ代名詞・ゼロ連体格代名詞については節 3.3.2 で後述) も、その要素をノードにして木構造を作っている。

<tree> タグ内の <bunsetsu> タグで囲まれた情報は、対応する文節の形態素解析結果である。形態素毎に <morpheme> タグで分割してある。属性 pos は形態素の品詞情報である。

その他のタグについては節 3.3.1 以降で述べる。

3.3.1 補完情報 (照応表現、先行詞の情報)

質問が照応解析が必要な質問文の場合、補完に関する情報は <completion> タグに入っている (図 1 参照)。

また、<reference> タグには照応表現の情報 (<referer> タグ) とそれに対応する先行詞の情報 (<antecedent> タグ) の組が入っている。属性 id によって、この情報の組に ID を付与してある。図 1 では一組の情報があることが分かる。また、補完が必要な照応表現の数だけ <reference> タグが存在する。

3.3.2 照応表現の情報

図 1 の <referer> タグには照応表現の情報が入っている。属性 rtype は照応表現のタイプである。照応表現のタイプは「ゼロ」(ゼロ代名詞・ゼロ連体格代名詞)、「指示詞・代名詞」、「名詞」(先行する表現を、一般の名詞により参照している場合) の 3 種類に分類してある。

● ゼロ代名詞・ゼロ連体格代名詞

図 1 の照応表現のタイプはこれである。<referer> タグにある属性 refid は <tree> タグのノードと対応づけるための識別子である。図 1 の例では、<tree> タグの ID 「QAC3-30003-02-T3」のゼロ代名詞を参照すると、「QAC3-30003-02-T2」に係ることが分かる。また、属性 attr は質問文における意味役割を示す。属性 attr の値は「連体修飾」、「動作主」、「道具」、「対象」、「源泉」、「目標」、「場所」、「時間」、「比較」、「相互」、「並列」、「条件」がある。属性値「条件」は、照応解析の観点では必要ないが、質問の回答を得るために必要な限定表現を補完する場合に用いる。

また、属性 particle は付加する助詞を示す。図 1 は、助詞「が」が付加し、質問文中で「動作主」の働きをするゼロ代名詞の補完が必要な場合である。

ゼロ連体格代名詞とは、名詞句に係る接続助詞「の」が省略されたゼロ代名詞のことであり、磯江ら [7] が検出手法を提案している。

● 指示詞・代名詞

```

<reference id="QAC3-30008-02-R1">
  <referer rtype="指示詞・代名詞" start="1" end="2">それ</referer>
  <antecedent refid="QAC3-30008-01-Q" start="10" end="21" tree="QAC3-30008-01-T3">藤村新一氏の発掘のなつ造</antecedent>
</reference>

```

図 2: 指示詞・代名詞の補完情報

ゼロ代名詞・ゼロ連体格代名詞との違いは、照応表現が表層に現れていることである。図2の例は、質問IDが「QAC3-30008-02」の質問文「それが最初に発覚したのは何という遺跡ですか。」の中に照応表現「それ」がある場合である。<referer> タグの属性 start と end は、照応表現の位置情報を示す。質問文中の照応表現「それ」が質問文「それが最初に発覚したのは何という遺跡ですか。」の1文字目から2文字目にあることを示している。これは同一質問文中に同じ表層表現があった場合、それらを区別できるようにするためである。

● 一般名詞

先行する表現を、一般の名詞により参照している場合がある。例えば、
「日本で世界一周のクルーズをしているものには何がありますか。」
という質問文を
「日本で世界一周のクルーズをしている豪華客船には何がありますか。」
と補完する場合である。この場合、「もの」が先行表現「豪華客船」を参照している。

一般名詞や、例の「もの」といった形式名詞が補完の対象となる場合は、このタイプに分類する。

<referer> タグの表現は、指示詞・代名詞と同じ形で表されるが、システムによっては検出モジュールが分かれる可能性があるため、照応表現のタイプを分けてある。

3.3.3 先行詞の情報

図1や図2の、<antecedent> タグで囲まれた情報は、照応表現に対応する先行詞の情報である。属性 refid は先行詞が現れている質問文または回答のIDであり、属性 start,end は、先行詞が質問文中のどこにあるかという位置情報を示す。

図2の例は、質問IDが「QAC3-30008-02」の質問文「それが最初に発覚したのは何という遺跡ですか。」の照応表現「それ」に対応する先行表現が「藤村新一氏の発掘のねつ造」であり、この先行詞は「QAC3-30008-01-Q」(属性 refid)の質問文10文字目から21文字目(属性 start,end)にあることを示している。

また、先行詞は単文節とは限らない。複数の文節から成る場合はその文節間の係り受けの情報として属性 tree がある。属性 tree が「QAC3-30008-01-T3」の場合、先行詞「藤村新一氏の発掘のねつ造」の係り受け木構造はQAC3-30008-01-T3を根とした木構造と同じ構造であることを示している。

```
<reference id="QAC3-30011-05-R1">
  <referer rtype="ゼロ" refid="QAC3-30011-05-T4">
  </referer>
  <antecedent refid="QAC3-30011-04-A" set="1" no="1"
rank="A"> フェニックスリゾート </antecedent>
</reference>
```

図 3: 先行詞情報

図3は、以前の質問の回答が先行詞となった場合である。属性 refid の「QAC3-30001-07-A」は、質問IDがQAC3-30001-07の回答であることを示す。回答に関する情報は、QAC3の既存テストセットから得ている。同テストセットには、加藤[1]の提案により、回答それぞれにセット番号、回答番号、ランクの情報が与えられている。セット番号は回答群を識別する番号、回答番号は回答群の要素を識別する番号であり、同意表現があった場合これらの番号は同じになる。ランクは「S」、「A」、「B」の3種類あり、S,A,Bの順にランクが高く詳細な回答であることを示している。これらの情報が、テストコレクション内の属性 set,no,rank の値にそれぞれ含まれている。

3.3.4 補完した質問文の一例

図1の<c-sentence> タグで囲まれた情報は正しく補完した場合生成する質問文である。<sentence> タグの質問文を<reference> タグの補完情報で補完した結果である。図1の例では、「デビューしたのはいつですか。」という質問文に「浜崎あゆみが」を補完した結果「浜崎あゆみがデビューしたのはいつですか。」という質問文になる。また、属性 id で、この正しく補完した質問文にIDを振ってある。

3.4 テストコレクションの構築

QAC3[1]の全シリーズの質問文(50シリーズ、計360問)に対し、節3.3で前述した情報がある拡張テストコレクションを構築した(XML形式で記述)。

4 評価実験ならびに考察

4.1 評価実験

QAC3[1]のテストセット(50シリーズ、計360問)と本研究で構築した拡張テストコレクションを用いて評価実験を行った。評価対象のシステムとして、川口[4]の質問応答システムを利用した。情報源として2000年、2001年の毎日新聞、読売新聞の新聞記事計4年分を用いている。

また、本章では2つの観点で評価を行っている。

一つ目は、照応解析を含む質問応答に対する性能評価である。それぞれのモジュールがどのような出力をしどのような精度だったかを分析する。

川口[4]のシステムが有するサブモジュールのうち、照応詞の検出、補完、ゼロ代名詞・ゼロ連体格代名詞の検出、補完、生成補完質問文群の回答取得、補完質問文決定、先行詞候補群生成の7つを性能評価した。

この時、条件を変えて2つの実験(実験1と実験2)を行った。実験1は、モジュールの処理が失敗しているとき、失敗しているモジュールの出力を、それに対応するテストコレクション中の正しい情報と入れ替えて次の処理を続ける。このように、評価対象処理の直前の処理まで正しく処理できている状態を再現し、各モジュール毎の精度を明らかにする。実験2は、質問1問を一通り処理した場合を評価する。具体的には、次の質問へ渡す入力のみテストコレクション中の正しい情報と入れ替える。回答と決定質問文がこれに相当する。質問1問毎の、照応解析から回答取得まで全体の精度を見ることができる。これらの精度を表1,2に示す。

二つ目の観点は、構築したテストコレクションに対する評価である。システムの処理を正確に想定し、それぞれを詳細に評価できるだけの情報があつたか、また、テストコレクションを用いてシステムを実際に評価可能であったかを、上記の実験1と実験2に基づき分析する。

なお、シリーズの最初の質問は照応表現が無いので、回答取得の処理のみ評価する。

4.2 実験結果

各実験1,2における結果を表1,2に示す。質問が多いため一部を抜粋してある(表1)。括弧がある箇所は実験1と2で結果が違った部分であり、括弧の外が実験1、中が実験2である。

表2は全シリーズ質問360問の結果を集計した値である。なお、ゼロ代名詞・ゼロ連体格代名詞を「ゼロ」と示してある。

先行詞収集の数値は、分母が正しい補完に必要な先行詞の数であり、分子がその中でモジュールが収集できた先行詞の数である。照応詞やゼロ代名詞の検出再現率、補完も同様であり、分母が正しい補完に必要な照応詞(またはゼロ代名詞)の数であり、分子がその中でモジュール

表 1: 実験結果 (一部抜粋)

質問 ID	先行詞 収集	照応詞検出		照応詞 補完	ゼロ検出		ゼロ 補完	(名詞句 検出)	MF 値	質問 決定
		再現率	適合率		再現率	適合率				
01-01									0.4	
01-02	1/1	1/1	1/1	1/1	N.A.	0/2			0.4(0)	0
01-03	1/1				1/1	1/2	1(0)/1		0	1
01-04	0/1							0/1	0.8(0)	1(0)
01-05	0/1				N.A.	0/1		0/1	0	0
01-06	0/1				1/1	1/2	0/1		0	0
01-07	0/1				1/1	1/1	0/1		0(0.333)	1(0)
01-08	2/3	1/1	1/1	0/1	1/1	1/1	0/1		0	0

表 2: 実験結果 (360 問集計)

実験	先行詞 収集	照応詞検出		照応詞 補完	ゼロ検出		ゼロ 補完	(名詞句 検出)	MF 値	質問 決定
		再現率	適合率		再現率	適合率				
実験 1	114/387	103/103	103/104	22/103	131/227	131/295	41/227	0/15	0.210	164
実験 2	114/387	103/103	103/104	22/103	131/227	131/295	28/227	0/15	0.167	36

ルが検出 (または補完) できた照応詞 (またはゼロ代名詞) の数である。

照応詞やゼロ代名詞の検出適合率は、分母がシステムの検出した数、分子がその中で正しく検出できた数である。

「名詞句検出」は先行表現により言い換えをすべき名詞句 (節 3.3.2 の「一般名詞」の項参照) の検出のことである。この処理は、評価対象システムに無い処理である。分母が補完が必要な名詞句の数で、分子がその中で検出できた名詞句の数 (0 となる) である。

MF 値は、システムが出力した回答リストに対する修正 F 値である [1]。

また、質問決定の数値は、正しい補完文を選択できれば 1、選択に失敗すれば 0 となる。

4.3 評価対象のシステムについての考察

表 2 を見ると、必要な照応詞を完全に検出 (再現率参照) できていることが分かる。一方、照応詞やゼロ代名詞・ゼロ連体格代名詞の補完の精度が低くなっている。これは先行詞収集の精度の低さが影響したと考えられる。

回答の評価は、実験 1 では、正しく補完した質問文に対する回答を評価している。一方実験 2 では、補完失敗した質問に対する回答を評価する場合がある。このことから、実験 2 より実験 1 の方が MF 値が常に高いと予測できる。

しかし、ID が 01-07 の質問文は、実験 1 では MF 値が 0、実験 2 では MF 値が 0.333 となっている (表 1 参照)。正しく補完できた質問文で求めた回答が間違っており、補完に失敗した質問で回答できていることが分かる。これは正しい照応解析による補完が、必ず正解を導き出すとは限らないことを示している。

4.4 テストコレクションについての考察

既存の QAC3 [1] のテストセットでは MF 値しか分からないが、本研究のテストコレクションにより、表 1 のような詳細な分析が可能となった。

照応解析を含む質問応答システム内の代表的な処理モジュールを想定し、それぞれの処理結果を全てテストコレクション内の情報で性能評価することができた。

また、節 4.3 のように処理毎の分析や、実験 1 と 2 など条件を変えた性能評価、質問毎の精度の比較、2 つのモジュール処理結果の関連性分析などを可能にしたことは、このテストコレクションの利点といえる。求解精度と照応解析精度の関連など、興味深い現象の観察ができた。

また、実験結果は処理の成功した数だけでなく、失敗した数も評価でき、いろいろな視点からの考察が可能である。例えば、照応詞検出の処理では必要な照応詞が完全に検出されているが、必要のない過剰な検出があった質問がある。このように、テストコレクションの評価結果は、照応解析処理の失敗解析の際に、大きく役立つ。

5 おわりに

本稿では、照応解析を含む質問応答システム内の代表的な処理モジュールを想定し、その入出力に関する情報を QAC3 のテストコレクションに付与することにより、モジュール毎の詳細な評価をし、失敗解析を容易にする拡張テストコレクションを構築した。

評価実験により、評価対象のシステムの各モジュールを詳細に分析できることが確認された。また、対象システムの処理結果より新たな発見があった。

本研究では、QAC3 に参加したシステムの一つしか評価していない。多くのシステムを評価し、その有効性を確かめることにより、テストコレクション内の情報が洗練されていくだろう。

参考文献

- [1] T.Kato et al. An Overview of NTCIR-5 QAC3. In *Proceedings of the Fifth NTCIR Workshop Meeting*, pages 361-372, 2005.
- [2] Y.Murata et al. Question Answering Experiments at NTCIR-5: Acquisition of Answer Evaluation Patterns and Context Processing using Passage Retrieval. In *Proceedings of the Fifth NTCIR Workshop Meeting*, pages 361-372, 2005.
- [3] M.Matsuda et al. Answering Questions of IAD Task using Reference Resolution of Follow-up Questions. In *Proceedings of the Fifth NTCIR Workshop Meeting*, pages 361-372, 2005.
- [4] 川口晋平, 森辰則. 照応解析に知識源を利用した関連質問群に対する質問応答. 言語処理学会第 12 回年次大会発表論文集, pages 927-930, 2006.
- [5] 石下円香, 森辰則. 優先順位型質問応答の解スコア分布に基づくリスト型質問応答. 情報処理学会自然言語処理研究会報告, 2005-NL-169, 2005.
- [6] Tatsunori Mori. Japanese Question-answering System Using A* Search and Its Improvement. *ACM Transactions on Asian Language Information Processing*, Vol.4, No.3, pages 280-304, 2005.
- [7] 磯江健史, 竹井光子, 相沢輝明. ゼロ代名詞連体格の自動検出システム. 言語処理学会第 10 回年次大会, pages 765-768, 2004.