質問応答から対話理解へ
― NTCIR QAC Task3 の提案 ―

加藤 恒昭　東京大学
福本 淳一　立命館大学
桝井 文人　三重大学
神門 典子　国立情報学研究所

1　はじめに

質問応答技術は自然言語によって表現された質問に文書でなくその情報そのもので回答する事を可能とするもので，情報アクセスの新しい形として期待されている [6]．事実に関する独立した質問に一問一答形式で回答するものが現在の研究の中心であるが，アナリストやレポータが利用しうる枠組みへの展開も期待されている [2]．例えば，新人レポータがある事件の記事を執筆するために，彼の記事で答えられるべき大きな質問をより簡単な質問の集まりに翻訳してシステムに訊ねるという利用形態が考えられている．

一方，質問応答と複数文書要約との関連も指摘されている．Hovy は彼の講演の中で複数文書の要約を一連の質問応答に還元する可能性を指摘しているし [1]，SUMMAC では，要約対象文書のトピックに関する質問を複数用意しそれにどの程度回答できるかを要約の評価指標とすることが試みられている [5]．このことは，与えられた一連の質問に回答できるような質問応答システムが文書要約者を支援できることを示唆する．

これらの関心から，質問応答システムを一連の質問に回答できるものとすることは質問応答技術のひとつの重要な目標であるといえる．本稿では，質問応答システムのそのような能力を定量的に評価するためのタスク QAC Task3 を提案する．このタスクでは利用者が対話的に質問応答システムを利用して情報にアクセスする状況を想定しており，システムはこの情報アクセス対話の対話相手を務める．この場合，システムには様々な能力が必要となるが，ここでは，対話の実現のための基本となる対話文脈を考慮した適切な質問の解釈，つまり照応解消や省略処理等のいわゆる文脈処理に焦点を当てている．なお，対話的状況を想定してはいるが，実際にはこれらの状況は模擬されるに留まりタスクの実施はバッチ的に行なわれるため，そのテストセットは再利用可能である．

本稿では，まず，QAC Task3 の設計について述べ，その後，そのためのテストセットの構築と，得られたテストセットの特徴について述べる．

2　タスク設計

QAC の枠組み

QAC Task3 は日本語による質問応答の評価チャレンジ QAC を構成する 3 つのタスクのひとつである [4]．

QAC では共通して，名称を回答とする質問を対象としている．ここで，名称というのは，人名や組織名等いわゆる固有表現に留まらず，日付け，数値を含むだけでなく，種の名称，機械や身体的部品の名称等を含む．統語的には複合名詞が回答の範囲とほぼ重なるが，小説や映画のタイトル等そこから外れるものも含まれる．システムはこれらについてそれを含んだ部分でなく，名称そのものを正確に抜き出すことを求められる．回答に利用される文書セットは新聞 2 誌各 2 年分の記事（前回の QAC1 では 1 誌 2 年分）で，それを使って分野に依存しない質問に回答する．

また Task3 を含むふたつのタスクでは，現実的な問題設定ということで，存在しないことを含めていくつ存在するかわからない回答を過不足なく収集することをシステムに求めている．

情報アクセス対話

このような質問に答えるシステムが対話的に利用されることを想定し，その能力を測定するタスクが QAC Task3 である．対話的場面として，以下の 2 種類に分類される情報アクセス対話を考えた．

収集型 あるトピックについてのレポートや要約を作成ために情報を収集する等の目的で，利用者がそのトピックに関する一連の質問を行なうような対話．すべての質問に共通するトピックが存在し，その結果として各々の質問も局所的な文脈を共有する．

ブラウジング型 固定したトピックが存在せず，対話の進行と共に利用者の関心が変わっていくような対

Series 14
小沢征爾さんはいつ生まれましたか。
どこの生まれですか。
大学はどこを卒業しましたか。
師事した先生は誰でしたか。
誰に認められましたか。
９８年にはどこで指揮を行っていましたか。
２００２年からどこの指揮者になりますか。

Series 22
ニューヨーク・ヤンキースの本拠地となっている
　球場はどこですか。
何年に造られたものですか。
そこには何人の記念碑が飾られていますか。
１９９９年に飾られたのは誰ですか。
彼が新婚旅行で来日したのは何年ですか。
その時の結婚相手は誰ですか。
彼女をポップ・アートに描いているのは誰ですか。
彼が描く缶詰はどこの会社のものですか。

図 1: 質問シリーズの例 その 1

話．隣接する質問は局所的な文脈を共有するが，全体をカバーするトピックは存在せず，焦点は自由に推移していく．

場面設定

QAC Task3 はシステムに一連の質問に回答することを求める．この一連の質問をシリーズと呼ぶ．この質問シリーズとそれへの回答が，情報アクセス対話を構成する．後述するテストセットに含まれるシリーズの例を図 1 に示す．Series14 は典型的な収集型であり，Series22 は典型的なブラウジング型である．本タスクでは，複数のシリーズをシステムにバッチ的に与えてそれに回答させる．あるシリーズがどちらの型であるかは与えられず，システムはそれを自分で判定しなければならない．また，システムはある質問がシリーズの先頭であるという情報は利用してよいが，ある質問に回答する際にそれに続く質問を参照することは許されない．これは本タスクが対話的な状況でのシステムの利用を模擬していることからの制約である．この制約と異なった 2 種類の対話を扱うことが文脈処理を難しくしている．

評価方法

QAC Task 3 では，回答として，可能な正解すべてを列挙したリストをひとつ返すことを求めており，正解数も問題毎に異なるので，評価は精度と再現率を考慮した修正 F 値を原則としている．必要とした修正は 2 点である．第一は，同じ回答もしくは同じ回答を指示する異なる表現を複数件リストに含めた場合の扱いで，この時は，そのうちひとつのみを正解とする．従って，そのような場合は精度が落ちる．同じものを指示する異なる回答には，人名における役職の有無，外人名の異表記，貨幣単位の違い，時間帯の違い（現地時間と日本時間）等が含まれる．第二は，正解のない質問の扱いで，そのような質問には空リストを返した時にのみ 1.0 が与えられ，それ以外の場合は，すべて 0.0 となる．

ある回答が正解であるかは，回答とそれと合わせて提示される根拠記事の適切性によって判断される．質問と無関係な記事を根拠としていれば文字列として正解であっても不正解となる．また，質問の正解は判定者である人間がその文脈の下でおこなった解釈によって決定され，システムの解釈やシステムのそれ以前の回答とは無関係である．例えば，図 1 の Series22 の 2 番目の質問の正解は，ニューヨーク・ヤンキースの本拠地であるヤンキースタジアムが建てられた 1923 年であり，システムが最初の質問にシェイスタジアムと誤って答え，2 番目の質問でそれが建てられた年である 1964 年を “正しく” 回答しても不正解である．一方，その場合でも適切な根拠記事と共に 1923 年を回答していれば，2 番目の質問については正解と判断される．

3　テストセットの構築

質問収集

テストセットに用いる質問の収集は以下の手順で行った．人物，組織，出来事等の 60 のトピックとそれに関する記事を 2 年分の新聞記事から選ぶ．それらトピックを被験者に提示し，それに関する事実をまとめるようなレポートを作成する際にそこに含めたいと考える情報を質問文の形式で表現させた．質問文型は wh 型に限定し，代名詞等の自然な表現を含んだ一連の質問を作成させた．トピックの提示には以下の 3 種のバリエーションをつけた．トピックの短い記述のみを与える．そのトピックに関する代表的な記事，一定の長さより長い場合はそのリード部分を添える．そのトピックに関する記事を５件添える．30 人の被験者に協力してもらい，ひとりの被験者については，それぞれの提

示パタンについて異なる 10 トピックについて質問を作成させた．1 トピックあたりの質問数は約 10 問を目安とした．回答が記事中に存在するかは意識させていないし，提示パタンによっては記事の内容にまったく接することがないので，質問の表現もそれに影響されない自然なものになっていると期待している．

今回はこれらのうち，40 トピックについての各 1 パタン延べ 3 名による質問計 1033 問をテストセット作成に用いた．また，この 1033 問を分析した結果，その 58%から 72%が名称を回答とするものであり，QAC で扱う範囲に含まれた．

テストセット作成

これら収集した質問を用いてテストセットを以下の手順で作成した．40 トピックから 26 トピックを選び，そのトピックに関する質問を適当に選び並べることで収集型のシリーズを作成した．代名詞表現を含めて質問文の表現については，できるだけ収集したものをそのまま用いることに努めたが，意味的，語用論的な曖昧性除去のために編集を行っている．これらのシリーズのトピックの分布は，人物 5 件，組織 2 件，事件出来事 11 件，人工物 5 件，動物等 3 件で，それらのうちの 4 件は複数の組織，複数の出来事，毎年繰り返される行事というような集合的なトピックとなっている．

ブラウジング型のシリーズは，残りのトピック中の質問と，複数文書要約に関するチャレンジである TSC[7] のために収集された評価用の質問文を種とし，そこからもしくはそこまでの適当な流れを付加することで作成した．なお，質問収集のトピックも TSC のものと重複させている．例えば，Series22 は，ヤンキースタジアムをトピックに収集した質問前半 4 問に後半 4 問を付け加えたものである．このようにして，10 シリーズを作成した．

結局，今回のテストセットは 36 シリーズ，251 質問で，収集型 26 シリーズ，ブラウジング型 10 シリーズからなる．1 シリーズを構成する質問の平均数は 6.92 である．

参照用テストセット

このようなテストセットを用いた場合，文脈処理能力だけが取り出され測定されるのではなく，情報アクセス対話という状況でのシステム全体の能力が測定されることになる．一問一答形式の質問応答にもまだ研究の余地があり，正解を過不足なく要求することが挑戦的である現状では，例えば，ある質問の正答率が低い時にその難しさがその語用論的側面にあるのかが明らかにならないという問題が残る．この点についての情報を得るための道具立てとして，2 種類の参照用のテストセットを作成した．

表 1: テストセットに現れる語用論的現象

分類	
代名詞	76 (21)
ゼロ代名詞	134 (33)
定名詞句	11 (4)
省略	7

第一のテストセットは，今回のテストセットに含まれる照応表現をすべて人手で解消し，それを補った独立の質問 251 問からなるセットである．第二のテストセットは，今回のテストセットに含まれる照応表現をすべて機械的に除去した独立の質問からなるセットである．ここに含まれる大半の質問は，意味的には誰のかを指定しないで誕生日を訊ねるような過度に一般的なものとなるが，文法的なものである．第一のテストセットの結果は語用論処理の効果の上限，第二の結果は語用論処理を行わない場合である下限を示している．もちろん，参照解消の結果得られる表現は一種類ではないし，語用論処理が悪影響を持つこともあるので，その結果は参考に留まるが，このような参照用テストセットは技術の特徴を検討するのに有益であろう．

4 テストセットの特徴

日本語には前方照応のための手段が大きく分けて 4 つある．連体詞を含む代名詞，ゼロ代名詞，定名詞句，省略である．ただし，定名詞句は表層的にそれ以外の名詞句と区別できない．テストセットの質問 251 問のうち，シリーズの先頭でない 215 問について，上記のいずれの照応表現を含んでいるかをまとめたものが表 1 である．12 問の質問は複数の照応表現を含んでいるので合計は 251 を越える．図 1 の Series22 の 6 番目の質問「そのときの結婚相手は誰ですか」はそのような質問の例である．カッコ内は照応先が出来事であるものの数である．表から様々な語用論的現象が現れていることがわかる．

含まれている照応表現から改めてシリーズを分析すると，先頭質問文中に先行詞を持つ質問だけが並んでいる狭義の収集型は 5 シリーズであった．これらのシリーズでは，先頭の質問文で述べられるトピックに関する質問だけでシリーズが構成されている．図 1 に示した Series14 はこの狭義の収集型で，先頭質問で述べられている「小沢征爾」を補うことで，すべての質問

Series 20
ジョージ・マロリーはどこの国で生まれましたか。
彼の有名な言葉は何ですか。
それを言ったのはいつのことですか。
彼が初めて山に登ったのは何歳の時ですか。
彼がエベレストの頂上付近で行方を絶ったのは何次遠征のときですか。
それは何年のことですか。
彼が最後に目撃されたのはエベレストの何メートル付近ですか。
彼の遺体を発見したのは誰ですか。

図 2: 質問シリーズの例 その 2

の照応解消が行えることになる．収集型として作成されたシリーズであっても，狭義の収集型でないものは，例えば，図 2 に示す Series20 の第 3 問，第 6 問のように先頭質問文で述べられているトピック以外の先行詞を持つ照応表現を含んでいる．

一方，ブラウジング型には，そこに含まれる照応表現の先行詞が直前の質問の回答であるような質問だけからなるシリーズは存在しない．すべてのシリーズに Series22 の第 3,4,6 問のように直前以外の回答や質問文中の要素を先行詞とする照応表現を含む質問や，複数の照応表現を含む質問が含まれている．

このようにいずれの型も焦点の推移は単純ではなくその追跡には洗練された技術が必要となる．そして，そのような焦点の追跡なしでこれらの質問に適切に回答することは不可能である．ブラウジング型についてはその必要性は自明である．ニューヨーク・ヤンキースからキャンベルスープまでを含んだ新聞記事があるとは思えない．収集型についても，それに関する記事が比較的多いトピックを選んでいるので，そのトピックに関する記事を検索してもそこから正しく回答を抽出することは，何らかの文脈処理なしでは困難である．例えば,「小沢征爾」をキーワードとする記事は 155 件あり，そのうちの 22 件が彼のウィーンフィルへの移籍を扱っているが，その中で彼の誕生日に言及しているものは 2 件のみである．

5 おわりに

本テストセットを用いた QAC Task3 の実施は 2003 年 12 月に行われ，8 チーム，14 システムの参加を得た．参照用のテストセットによる実施も合わせて行っている．結果については現在分析中であるが，現状の技術でこのようなタスクを扱うことは，決して容易ではないが絶望的な程困難でもなく，適度に挑戦的なものとなっているという感触を得ている．分析結果の報告と参加チームの報告は，2004 年 6 月の NTCIR4 Workshop Meeting[3] で行われる．

今後，タスクの設計に関しては，評価指標として提案した修正 F 値が適切であるのか，対話の流れの中でのどのような回答を正答とすべきかについて，検討を続けるつもりである．また，情報アクセス対話においてどのようなタイプの質問がよく現れ，それと利用者の特性との関係についても検討を進め，情報アクセス対話での利用で要求される質問応答システムの特徴を明らかにしていきたい．

参考文献

[1] Eduard Hovy. 2001. `http://www-nlpir.nist.gov/projects/duc/pubs/2001papers/isi_hovy_duc.pdf`.

[2] John Burger, Claire Cardie, and et al. 2001. Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A) `http://www-nlpir.nist.gov/projrcts/duc/roadmpping.html`.

[3] NTCIR4 Workshop Home Page. 2003. `http://research.nii.ac.jp/ntcir/workshop/work-en.html`.

[4] Jun'ichi Fukumoto, Tsuneaki Kato and Fumito Masui. 2003. Question Answering Challenge(QAC-1) An Evaluation of question answering tasks at the NTCIR workshop 3 *AAAI 2003 Spring Symposium New Directions in Question Answering*, pp. 122-133.

[5] Inderjeet Mani, David House, and et al. 1998. The TIPSTER SUMMAC text summarization evaluation final report. Technical Report MTR98W0000138, The MITRE Corporation.

[6] Ellen M. Voorhees and Dawn M. Tice. 2000. Building a Question Answering Test Collection *the Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 200 - 207.

[7] Text Summarization Challenge Home Page. 2003. `http://lr-www.pi.titech.ac.jp/tsc/index-en.html`.