社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2006－NL－172（8）
2006／3／27

情報アクセス対話に向けた質問応答技術の評価 ふたたび
– NTCIR-5 QAC3 での試み –

加藤 恒昭[1]　福本 淳一[2]　桝井 文人[3]　神門 典子[4]
東京大学[1]　立命館大学[2]　三重大学[3]　国立情報学研究所[4]

概要

NTCIR-5 QAC3 (Question Answering Challenge 3) で実施された情報アクセス対話に向けた質問応答技術の評価について，そのタスクである IAD (Information Access Dialogue) タスクとそこによせられた技術を述べる．IAD タスクは，利用者がレポート執筆を目的として対話的に質問応答システムを利用する状況を想定したもので，そこで必要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を評価の対象に含めている．IAD タスクによる評価は QAC2 の subtask 3 としての実施に次いで２度目となるが，今回は質問範囲，正解範囲の精緻化を行うとともに，その評価を自然なものとするための多段階評価の採用や正解セットという概念の導入等，様々な工夫を行っている．また，今回の実施におけるテストセット構築では WoZ 方式という新しい試みを行った．その実施結果からは，新しい試みやこのタスクに関する着実な技術進歩が伺われる．

A Second Evaluation of Question Answering Technologies for Informatin Access Dialogues
– A Report of NTCIR-5 QAC3 –

Tsuneaki Kato[1]　Jun'ichi Fukumoto[2]　Fumito Masui[3]　Noriko Kando[4]
The University of Tokyo[1]　Ritsumeikan University[2]　Mie University[3]　National Institute of Informatics[4]

Abstract

This paper describes the evaluation task of question answering technologies for handling information access dialogues and a workshop using it conducted in NTCIR-5 QAC3 (Question Answering Challenge 3). This task assumes interactive use of QA systems and evaluates, among other things, the abilities needed under such circumstances, i.e. proper interpretation of questions under a given dialogue context; in other words, context processing abilities such as anaphora resolution and ellipses handling (we refer to the task as the IAD task, where IAD stands for Information Access Dialogue, and to the whole workshop as QAC3). The IAD task in QAC3 is based on QAC2 subtask 3 with several improvements, including elaboration of the scope of questions and answers and introduction of multi-grade evaluation and the concept of a correct answer set. In addition, a new WoZ method was devised and applied in the test set construction. New trials and advances in existing methods were observed from the submission results to the workshop.

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 [17]．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を中心に研究が始められたが，近年は対話性の重視等，様々な面での展開が見られる．質問応答技術を牽引してきたといってよい TREC[16] では，TREC2001 において対話的な利用を前提とした文脈処理の能力を評価する試みがなされている [18]．その後，TREC2004 から，相互に独立した質問ではなく，あるトピックに関する一連の質問の集まりという形で課題を与えるようになっている [19]．そこでは，その一連の質問のトピックが何であるかが与えられているために代名詞等の処理は殆ど自明であり，文脈処理の能力を評価するものではないが，このような利用の場面が自然であるとの認識が共有されている点は重要である．この利用形態は質問応答システムの進むべき方向のひとつとしても議論されており，例えば，新人レポータがある事件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形で，アナリストやレポータが利用しうる質問応答システムへの発展が提案されている [2]．また，ARDA の AQUAINT program[1] ではアナリストが分析的に用いる質問応答システムの構築がその目的とされており，積極的に対話的な質問応答の研究が進められている．質問の分解を含めて，分析的説明的な質問にどう答えるか，clarification 等の利用者とのやりとりはどうするか等が研究の関心となっている [4][9][15]．

日本語における質問応答技術の評価ワークショップ QAC では，NTCIR-3[14] で行われた QAC1 から現実的な場面での事実に関する (factoid) 質問応答への能力を定

量的に評価することを目的に進められており，その当初から，対話的な利用を視野に入れサブタスクを設定していた[3]．その後，QAC2 でタスク設計の大幅な精錬を行ない，情報アクセス対話の参加者としての質問応答システムの評価を行えるタスク，IAD タスクを確立した [5][8]．IAD タスクは，利用者があるトピックについてのレポートを作成する目的で，そのレポートに含めるべき情報を質問応答システムを用いて収集するという場面を想定している．本稿では，NTCIR-5 QAC3 での IAD タスクの 2 回めの実施について報告する[*1]．

以下，まず，QAC3 における IAD タスクの定義を述べる．特に，QAC2 での実施時に較べ，質問範囲，正解範囲の精緻化を行うとともに，その評価を自然なものとするために多段階評価の採用や正解セットという概念の導入等，様々な工夫を行っているので，この点について詳しく述べる．次に QAC3 での IAD タスクの実施について報告する．テストセット構築において採用した WoZ (Wizard of Oz) 方式という新しい試みについて説明するとともに，得られたテストセットの特徴を述べる．その後，評価報告と考察を行う．

2 IAD タスク

2.1 IAD タスクの基本

IAD タスクを含む QAC のタスク全般が対象とするのは，疑問代名詞を含む文の形式を持ち，名称を正解とする質問である．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，数値を含み，種の名称，機械や身体的部品等の一般名称を含む．統語的には複合名詞が正解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．システムはこれらの名称をそれを含んだ部分でなく，過不足なく抜き出すことを求められ，その文字列とそれを抜き出した新聞記事の ID を返却する．この新聞記事はその回答の根拠となるものでなければならない．回答に利用される文書セットは新聞記事（QAC3 では 2 紙各 2 年分，2000,2001 年の毎日新聞と読売新聞を用いている）で，それを使って分野に依存しない質問に回答する．

IAD タスクでは，システムに一連の質問（シリーズと呼ぶ）を与え，それに次々と回答させてゆく．それぞれの質問は前述の範囲にとどまるが，シリーズの先頭以外の質問には，それ以前の質問の一部もしくはその回答を参照する参照表現（省略や 0 代名詞を含む）を含んでいる．この一連の質問とそれへの回答が情報アクセス対話を構成する．実際の利用場面ではシステムは対話的に質問に応答することが期待されるが，本タスクではその対話性は模擬されるだけで，複数のシリーズをバッチ的に与えられそれに回答することをシステムに求める．対話の展開があらかじめ定められていることは対話本来のダイナミクスを失わせているが，その一方で，タスクに参加したシステムがすべて同じ質問に回答するので，相互比較可能な評価結果が得られることに加え，正解をプーリングすることでテストセットが再利用可能となるという利点を有している．

情報アクセス対話が，利用者があるトピックについてのレポートや要約を作成するための情報を収集する等の目的でそれに関する一連の質問を行なうような対話（収集型）と，利用者の興味の赴くところに従って対話の進行と共にトピックが変わっていくような対話（ブラウジング型）との 2 つの極を持つという直観から，それに応じたふたつの型のシリーズを設定している．IAD タスクが想定している情報アクセス対話は与えられたトピックについての様々な情報を収集するような場面で見られるもので，収集型の対話が支配的であるが，実際の場面ではその部分部分にブラウジング的な要素を含むことが知られており，これが本タスクにブラウジング型を含めている理由である．タスクにおいて，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．

システムは，各質問に対して，存在しないことを含めていくつ存在するかわからない正解を過不足なく収集することを求められ，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られる．正解数は問題毎に異なり事前に与えられていないので，個々の質問に関する評価は精度と再現率の両方を考慮した F 値を原則とする．システムの総合評価はその評価の質問全体の平均である．情報検索一般とは異なる質問応答の特殊性から評価には様々な配慮が必要となるが，これについては後述する．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して扱われる．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．

IAD タスクでは，システムの文脈処理能力や質問の文脈依存性に関する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によるランを正式ランと同時に行う．第一のテストセットは，正式ラン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からなるセットである．第二のテストセットは，そこ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．後者は意味的には大半の質問が誰のかを指定しないで誕生日を訊ねるような特定化が不充分なものとなるが，統語的には整ったものである．第一の参照用テストセットの結果は，もちろん近似的にであるが，

[*1] 本稿では QAC3 を評価ワークショップ全体をさす用語として用い，そのタスクは IAD タスクと区別する．また一部の文献では IAD タスクとそれによる評価を QACIAD と呼んでいる．

文脈処理の上限，第二の結果は，文脈処理なしで回答できる上限を示している．また，第一の参照用ランにはプーリングを充実させるという役割もある．

2.2 QAC3 における IAD タスク

今回の実施では，前回の経験をふまえて IAD タスクに以下の改善を行った．第一は正解範囲の再定義で，名称を範囲とするという定義では漠然としていた部分について説明を加えると共に質問範囲拡大の足がかりを作った．第二は評価方法の洗練で，評価を自然なものとするために多段階評価の採用や正解セットという概念の導入を行った．

2.2.1 正解範囲の再定義

今まで，名称や数量を正解の範囲とするということでその範囲に含まれるかが不明確であった以下の表現について，正解範囲であることを明確とした．

- 数値表現に属性の詳細化具体化を行うための表現が付属したもの：「年間３００台」「タテ５０ｃｍヨコ３０ｃｍ」「一人当たり３リットル」「重さ３トン」
- 範囲表現（定型的，慣用的なもの）：「１０～１２％」「８世紀後期から９世紀初期」「四国から九州まで」「３０人以上」「３０人以上５０人以下」．「東京大阪間」「羽田－千歳」「千葉県内」等，空間的な範囲表現（区間表現）も含む．
- 概数表現（蓋然表現）：「約１００人」「３億円程度」．「シカゴ近郊」「東京都近辺」「舞浜駅前」「大使館裏」等，空間的な蓋然表現も含む．

これらの表現は質問への自然な回答として必要となる．「どのくらい利用されていますか」に「３００台」と回答しても年間なのか月間なのかが不明確であるし，「どのくらいの大きさですか」に「５０ｃｍ」と長さで答えることには違和感がある．これらを許さないことは正解の網羅的な列挙や重複の判断でも問題となる．「タテ５０ｃｍヨコ３０ｃｍ」と回答できずに「５０ｃｍ」「３０ cm」の両方を挙げる必要があるとか，「１０～１２％」において，「１０」だけでは単位が含まれないので正解とできないとか，「約１００人」は「１０２人」と同一の情報としていいかもしれないが「１００人」はどうか等の問題の源である．

加えて，理由，状況を訊ねるような「なぜですか」「どうなりましたか」についても一歩踏み出すことを目的として，名称とは呼ぶことが自明でないような，事象（出来事）に関する表現，例えば「施工ミス」「送電線切断」「墜落炎上」等も名詞連続複合語であれば原則として対象範囲とすることを明示し，更に特徴を訊ねる「何ですか」のように名称（名詞）のみで回答するよりも，それに説明を加えた記述（名詞句）による回答がより適切となる質問も含まれることを示して，そのような場合の回答として，その主辞である名詞を正解とした．また，その記述である名詞句に含まれる名詞で意味的にその名詞句と置き換えられるものも正解に含めることとした．例えば，「98 年，タイのどこで飛行機事故が起こりましたか」について「タイ航空機が墜落したタイ南部スラタニ空港付近の沼地で」という根拠記事から，正解は「沼地」とし，「タイ南部」「スラタニ空港付近」を正解に含める場合もあるとした．

2.2.2 評価基準

可能な正解すべてを列挙するしたリストをひとつ返すことを求める課題，リスト型課題の評価には以下のような様々な難しさがある [6].

- 重複の扱い：同じ事物を指す複数の表現，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）等があるため，同じ事物を指すこれらの表現を複数個回答のリストに含めた（回答に重複がある）場合の扱いを決めなければいけない．
- 回答の質に関する問題：同じ事物を指す上記の表現の中には，フルネームと略称のように情報の質が異なるものがある．日付や場所の場合は「００年」「００年１月３日」，「日本」「千葉県浦安市」のように詳細度（粒度）の異なるバリエーションがある．これら表現の質の問題を扱い，評価に反映させる必要がある．加えて，表現の問題ではなく，回答自体（指示されている事物）の質が異なると思える場合がある．例えば，記事中で事実もしくは伝聞として述べられているが，誤報もしくは発表者側の誤りにより事実と異なる数値や日付や，記事中では確定的な予定として述べられているがその後に変更となった日付等を正統な正解と同じように扱っていいのかには疑問が残り，その質の差にみあった評価が求められる．
- 列挙のシステムの問題：可能な正解すべてを列挙するといっても，その列挙の方法が複数ある場合がある．「東海三県」と「三重県」「愛知県」「岐阜県」のように（一定の知識を前提とすれば）同じ情報が違う形で伝えられる場合がある．例示を含んだ「川魚，エビ，カニ等の魚介類」において．「川魚」「エビ」「カニ」「魚介類」は明らかに並べられるものではないが，「川魚」「エビ」「カニ」という列挙と「魚介類」という回答とでどちらが優れているかは自明とはいえない．この問題は粒度と関連して生じることが多い．あるイベントの開催地をそれが行われた国名で列挙するか都市名で列挙するかの選択もある．また，あるイベントが「１２月１０日」と「１２月２０日」の２回行われたとき，その開催日を「１２月」と答えてしまうと２回行われたという情報は伝わらない．この場合「１２月」と「１２月１０日」のふたつを答えても，伝わる情報は「１２月１０日」だけを答えた場合と同じである．表現の粒度の問題は表現の質の問題であるが，この例のようにその粒度が荒くなって他の回答と区別できなくなった時，表現の問題にとどまらなくなる．範囲表現等を正解範囲に含めたための問題もある．例えば，「８世紀後期から９世紀初期」をひとつの回答とするリストと「８世紀後期」「９世紀初期」をふたつの要素とするリストとを等しく扱わなければならない．

$$P_{CAS_i} = \frac{\sum_{ES \in CAS_i} \begin{cases} \max_{e \in O \cap ES} f(e) & if\ O \cap ES \neq \phi \\ 0 & otherwise \end{cases}}{|O| - |(O - \bigcup_{ES \in CAS_i} ES) \cap \bigcup_{ES' \in \bigcup_{j \neq i} CAS_j} ES'|}$$

$$R_{CAS_i} = \frac{h(CAS_i) * \sum_{ES \in CAS_i} g(ES) * \begin{cases} \max_{e \in O \cap ES} f(e) & if\ O \cap ES \neq \phi \\ 0 & otherwise \end{cases}}{\sum_{ES \in CAS_i} g(ES)}$$

$$F_{CAS_i} = \frac{2 * P_{CAS_i} * Q_{CAS_i}}{P_{CAS_i} + Q_{CAS_i}}$$

$$MF1 = \max_i F_{CAS_i}$$

これらの難しさを考慮し，可能な限り直観に合う評価を行うため，以下のような評価の体系を用いることとした．中心となるのは，正解セットという概念の導入と回答の 2 種類の質に基づく多段階評価である．

各質問について，複数の正解セット CAS を用意する．正解セットとは，ひとつの列挙の方法にひとつ対応するもので，上の例では，{「東海三県」} がひとつ，{「三重県」，「愛知県」，「岐阜県」} がひとつのセットをなす．また，{「１２月」} がひとつ，{「１２月１０日」，「１２月２０日」} がひとつである．正解セット毎にそのセットの正解を網羅した際の係数 h（$0.0 < h \leq 1.0$）が与えられる．多くの場合，その係数は 1.0 であるが，上例の {「１２月」} のセットの場合，このセットを網羅しても他方のセットの正解を網羅した場合の半分の情報しか与えられないとして，係数 0.5 が与えられる．

正解セットは，同じ事物を指す様々な正解表現 e の集まり（表現集合 ES と呼ぶ）の集まりである．実際には正解判定は表現と根拠記事との対に対して行われるので，異なる根拠記事を持つ同じ表現も同じ表現集合に属するとして扱う．それぞれの表現集合についてそれが指すものの質に関する係数 g（$0.0 < g \leq 1.0$）が付与される．表現集合中の正解表現それぞれには表現の質に関する係数 f（$0.0 < f \leq 1.0$）が付与される．ひとつの正解セット中では，ある正解表現はただひとつの表現集合中にのみ現れる．

出力集合（システムが返却した回答であるリスト）O が与えられた時，ある正解セット CAS_i に関する精度 P と再現率 Q は上の式で与えられる．これに基づいて F 値が求められ，最も大きい F 値を与える正解セットを用いた評価がその出力集合に対する評価となる．なお，正解が存在しない質問については，回答数が 0 の場合に F 値 1.0，それ以外は 0.0 とする．この定義による F 値を MF 値，その質問全体の平均を MMF 値と呼ぶ．

この評価が意図しているのは，

- 表現の質は係数 f で表現し，質の低い表現を選んだ場合は精度再現率の分子となる正解数の当該部分に係数を乗じることでよりよい表現を回答した場合と差を付ける．
- 正解そのものの質は係数 g で表現し，再現率の分母分子の正解数両方にそれを乗じることで，再現率に正解の質を反映させる
- 同一物を指示する異表現はその同定をシステムの能力の一部と考え，同じ表現集合に属する正解を複数回答した場合は，その中で表現の質の一番よいものひとつを正解とし，それ以外は誤答として扱うことで精度を下げる．
- 正解の列挙については，ひとつのセットに基づいて回答することを期待し，それぞれの正解セットに従って採点を行い，最も高い評価となるセットの値を採用する，ただし，各セットでの採点において，そのセットでは誤答であるが，他のセットの正解であるような回答は回答数に含めないことで，誤答と区別する．これにより様々な正解セットに含まれる正解を混在させた時の精度の減少を防ぎ，ペナルティをなくす．

例を示す．「東京ディズニーランドはどこにありますか」という質問に「千葉県浦安市」「舞浜駅前」のふたつの正解があるとする．このふたつが同じ場所を指す異表現と考えるなら，同じ正解セットの同じ表現集合にこのふたつを含めることになる．その場合，一方を回答に含めればよく，両方を含めた場合，精度が下がる．これらふたつは違う情報であり，両方を列挙すべきであると判断した場合は，同じ正解セットの異なる表現集合に含める．この場合，両方を回答に含めないと再現率が下がる．このふたつは異なる回答の仕方でありどちらもひとつで充分な情報を持っているとの判断であれば，これらふたつを異なる正解セットとする．この場合，一方を回答に含めればよく，両方を含めても精度は下がらない．両方回答すべき（同じものの異表現ではない）であるが，「千葉県浦安市」の方がより適切とする場合は，「舞浜駅前」の正解そのものの質 g を落と

す．この場合，例えば「千葉県浦安市」だけで再現率 0.67，「舞浜駅前」のみで 0.33 というような重み付けが可能となる．更に「千葉県」も正解とするが，これは「千葉県浦安市」と同じものを指し，表現として劣ると判断するのであれば，「千葉県浦安市」と同じ表現集合に含め，その表現に関する係数 f を落とせばよい．

3 IAD タスクによる評価の実施

3.1 スケジュールと実施手順

QAC3 は 2004 年 6 月の NTCIR-4 Workshop Meeting で実施を宣言し，同年 9 月のラウンドテーブルミーティングで IAD タスクのみの実施等の基本方針を確認した．その後，2004 年 11 月末に CFP を提示し，正解範囲等を公開し，同年末まで参加を募った．参加者へのツールとしては過去のテストセットと Format Checker を準備した．ドライランは実施せず，正式ランを 2005 年 4 月 25 日より 1 週間の期間で行った．問題配布は www システムを用いて行い，結果の提出は電子メイルの添付ファイルを利用した．回答提出の締切は問題をダウンロードした後，48 時間以内とし，複数システムの場合はその後 1 システムにつき，24 時間が与えられた．正式ランの実施後，5 月に参照用のランを実施した．こちらには時間的制約を課していない．正解のサンプルを 6 月末，8 月初めに参加者に送付し，評価結果の配布は 8 月末に行った．報告会は 2005 年 12 月に NTCIR-5 Workshop Meeting において開催された [14].

3.2 テストセット構築

テストセットの構築は質問を収集作成し，それをシリーズとして構成していくという 2 つの段階からなる．特に質問の収集作成では，得られたものにリアリティがあることが重要である．

3.2.1 質問収集

準備:2000,2001 年の毎日新聞，読売新聞の見出しを参考に，事件，人物，組織，等を取り混ぜ，101 のトピックを選び，新聞記事全文検索システムを利用して，その 800 文字から 1600 文字程度の長さで得られた情報をまとめたレポート（サマリと呼ぶ）と 100 文字前後の簡単な説明（概要と呼ぶ）とを作成した．

アンケート方式による収集:ある程度の情報が集められたことを基準に 101 トピックの中から 80 のトピックを選んだ．被験者を 12 人用い，一人につき 20 トピックを割り当て（1 トピックにつき 3 人が割り当てられる），それらのトピックについて以下の手順で 2 回に分けて質問を収集した．実施は郵送で行った．第一回はトピックと概要を提示し，それについてのレポートを作成するという状況を設定し，そこに含めたい情報を一連の質問の形式で記述させた．質問は疑問代名詞を含めた文の形とし，一連の質問では参照表現等を使って構わないことを指示している．ひとつのトピックに関しての質問の数は 10 問を目安とした．第二回はトピックとサマリを提示し，前回作成した質問によって得られる情報がレポートに含めるべきものであるかの判断を行わせた．また，サマリを読んで新たに含めるべきと思う情報を訊ねる質問を追加で作成させた．

WoZ 方式による収集:上記の 80 トピックから更に 20 トピックを選んだ，サマリを作成した 4 名が質問応答システムの WoZ 役を務め，自分がサマリを作成したトピックを担当し，作際したサマリ，新聞記事全文検索システム，自分の記憶を用いて，利用者からの質問に対話的に回答した．対話はキーボードとディスプレイを用いて行った．6 人の被験者に各 10 トピックを割り当て（1 トピックにつき 3 人が割り当てられる），被験者にトピックと概要を提示し，それについてのレポートを作成するという状況設定で，質問を事前に考えたのち，情報アクセス対話を行った．被験者には事実に関する簡単な質問に回答できる質問応答システムを利用していると説明し，WoZ 役にも理由や意見を訊ねる質問については回答できないと応答するように等，その役割を教示した．

アンケート方式の第一回では延べ 2416 問が収集され，そのうち 1847 問が適切と判断された．WoZ 方式では延べ 620 問の質問が収集され，そのうち 502 問が WoZ 役によって回答された．それ以外は問い返されたり，回答が見つからない，質問が不適切等の応答がなされたりした．アンケート方式と WoZ 方式による収集とで得られた質問の内容や表現に大きな差はなかった．

アンケート方式において第一回の収集で作成され第二回で適切と判断されたもの，WoZ 方式によって集められ WoZ 役の回答があったものを中心に，502 問を選んで，正解の存在について確認した．これらはそのトピックに関する質問として適切であり，かつ質問作成者がそのトピックの詳しい情報やその表現に触れることなく作成したものであるので，内容的にも表現的にも自然であることが期待されるためである．また，これらの作業とは別に対象となる新聞記事の情報について質問を 200 問作成し，合わせて正解例の確認を行った．

3.2.2 シリーズの構成

あるトピックについて収集され正解の存在が確認された質問から選択し，文脈的に問題のないように並び替えや表現の修正を行うことで，特定のトピックに関する質問の列であるような収集型シリーズを作成した．QAC2 で収集型と名付けたシリーズでは，そこに含まれる質問はすべて最初の質問で提示されるトピックを参照する指示表現を含んでいたが，今回はその制約を緩めて，あるトピックに関して上記で収集された質問から構成されるものを収集型と呼ぶことにした．これに対して明らかに複数のトピックを渡り歩いている質問の列であるものをブラウジング型と呼ぶが，こちらについては，上記で収集された質問と作成した質問をシリーズの種として，そこからもしくはそこに質問をつなげる形で作成した．

3.2.3 テストセットの特徴

今回のテストセットにおけるシリーズの例を図 1 に示す．最初の 2 つの例は収集型の例であり，3002 は「ハリー・ポッター」シリーズを 3004 は発泡酒をトピックとしている．ここで，3002 の第 5 質問はハリー・ポッター」シリーズそのものではなくその第一巻を指示する代名詞を含んでいるし，3004 のシリーズでは発泡酒の特定の銘柄から発泡酒全体に関心が移っているのが伺われる．このように QAC3 の収集型は QAC2 のそれよりも幅が広く，指示表現の現れ方で定義することは困難である．3 番目の例はブラウジング型で，ここではテーマパーク，俳優，映画とトピックが変わっていることがわかる．最後の質問はテーマパークとは何の関係もなく，ひとつのトピックの少なくとも回りを回り，そこから大きく離れない収集型と対比される．図 2 に参照用ランの質問シリーズの例を示す．これは図 1 の 30002 に対応するものである．

今回のテストセットは 50 シリーズ 360 問からなる．1 シリーズの質問数は 5 問から 10 問で，平均質問数は 7.2 問，シリーズ中，35 シリーズが収集型で，残りの 15 シリーズがブラウジング型である．収集型におけるトピックの内訳は出来事が 11，組織が 2，人工物が 9，人物が 8，動植物等が 5 である．平均正解数は 1.98 で，正解数が 1 のものが 204 問である．360 問のうち，37 問について複数の正解セットが利用された．また，360 問中 18 問が名詞句による回答が可能なものもしくは事象を訊ねるものであった．例えば，「事故当時、潜水艦は何をしていたのですか。」は「魚雷発射」「演習」，「アラビア石油はどんな会社ですか。」は「石油開発会社」を正解とする．

4 参加システムと評価

今回の評価への参加チームは 7 チームで，提出されたシステム（提出期限や知識源の制約を満たしていない参考用ランを含む）は 16 である．参照用ランもすべてのチームから提出されたが，一方のランだけのもの，利用文書に問題のあるものもあった．

各システムの評価を示す．図 3 はすべての質問についての MMF 値，および各シリーズの先頭の質問とその後の質問のそれである．図 4 にシリーズの型による評価の違いを示す．全体的傾向として収集型の評価が高いのは予測通りであるが，ブラウジング型の評価の方が高いシステムがあること，同じチームから提出されたシステムの間で収集型とブラウング型のバランスが異なること等から，それぞれのシステムで様々な文脈処理が試みられていることが想像される．図 5 に参照用ランとの比較を示す．図中の Forst1 は Forst1, Forst2, Forst3 に共通のランである．その他については，図 3 と図 5 のシステム ID が対応づけられている．QAC2 でもそうであったが，シリーズ先頭の問題はその他の問題よりも難度が低いので，参照用ランでもその間の評価には差が出る．文脈処理により評価は

Series 30002
「ハリー・ポッター」はどんなジャンルの読み物ですか。
作者は誰ですか。
主な登場人物は誰ですか。
シリーズの第１巻の発売はいつでしたか。
そのタイトルは何ですか。
２００１年までで何巻出ていますか。
何ヶ国語に訳されていますか。
日本ではどのくらい売れましたか。

Series 30004
アサヒビールが発泡酒の発売を開始したのはいつですか。
商品名は何といいましたか。
値段はいくらでしたか。
その頃、発泡酒にはどんな銘柄がありましたか。
シェア１位はどこでしたか。
ビールと比べてどれくらい売れていたのですか。
最初に作ったのはどこのメーカーですか。

Series 30024
ＵＳＪはどこにできましたか。
最寄りの駅はどこですか。
オープン初日のフィルムカット式に出席した俳優は誰でしたか。
彼が主演した０１年の正月映画は何ですか。
同じ時期に封切られたケビン・コスナー主演の映画は何ですか。
何を題材にした映画ですか。
コスナーはどんな役で出演していますか。

図 1　質問シリーズの例

「ハリー・ポッター」はどんなジャンルの読み物ですか。
「ハリー・ポッター」の作者は誰ですか。
「ハリー・ポッター」の主な登場人物は誰ですか。
「ハリー・ポッター」シリーズの第１巻の発売はいつでしたか。
「ハリー・ポッター」シリーズの第１巻のタイトルは何ですか。
「ハリー・ポッター」は２００１年までで何巻出ていますか。
「ハリー・ポッター」は何ヶ国語に訳されていますか。
「ハリー・ポッター」は日本ではどのくらい売れましたか。

図 2　参照用セットの質問シリーズの例

50% から 80% となる．今回明確にした範囲表現，概数表現についてはそれを積極的に出力に含めているシステムが少数ながらあり，今後の展開が期待される．正解範囲として名称を越えた質問については難度が高く，正解を出力できるシステムはきわめて少ない．

各チームから提出されたレポートから情報アクセス対話に向けた質問応答を可能とする技術について以下が読み取れる．質問応答における文脈処理の基本的な手法は，現在の質問に過去の質問に現れたキィーワードを結びつけて得られるものを処理の対象にするものである．最もよい結果

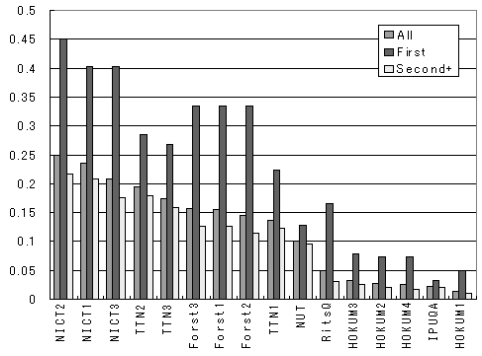

図 3　MMF 値による参加システムの評価

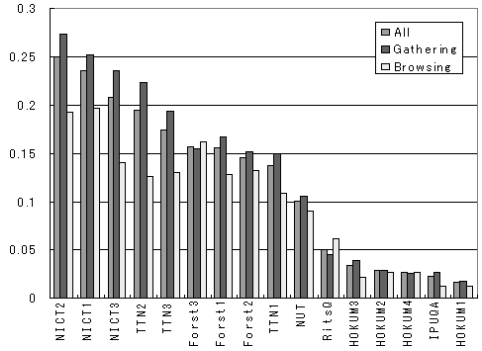

図 4　シリーズの型による差位（MMF 値による）

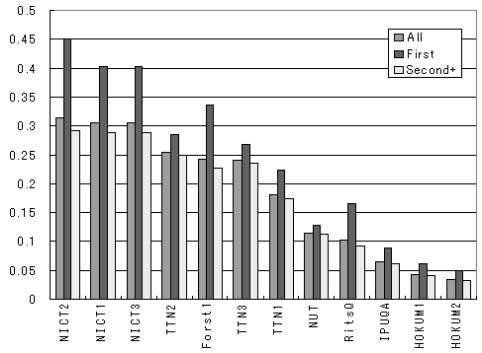

図 5　参照用ランとの差位（MMF 値による）

を出してる NICT システムは，この手法を採用しており，シリーズの先頭から現在処理している質問の直前の質問までを現在の質問に繋げて処理を行っている [12]．核システムの高性能と相まって，この手法は成功しており，高い評価を得ている．このような手法で問題として予想されるのは，その質問にとって不要かつ有害なキィーワードが混入してしまうことであるが，NICT システムはそのようなノイズに強いと推測され，それが核システムの高い評価を導いていると同時に，このような単純な文脈処理でも適切な結果が得られる理由ではないかと思われる．

対照的に RitsQ は従来の言語理解の手法を用いて文脈処理に緻密に取り組んでいる [10]．ここでは質問における参照表現を代名詞によるもの，動詞の必須格が 0 代名詞化したもの，名詞の修飾語，被修飾語が 0 代名詞化したものに分類し，後 2 者について，EDR の格フレームとコーパスを概念辞書と合わせて言語知識として用いて，それぞれ参照物を同定している．この手法は質問の形態素統語解析の失敗や言語知識の不足等の理由から現在のところあまりよい結果を出していないが，扱っている言語現象は広く，今後の充実が期待される．

今回提案された手法の中で興味深いのは，文脈処理が行った選択の適切性を質問応答システムが見つけてくる回答のもっともらしさに求めるというものである．TTN は現在処理している質問に加えるべきキーワードとして，シリーズ先頭，直前に現れたもの，あるいはその両方のどれが適切かを決定するが，その際，それぞれのキーワードを加えた質問を実際に質問応答システムで実行し，その質問に答えるもっとも適切なパーッセージが得られるものを選択している [13]．キーワードの候補については直前の回答等，それ以外の候補に関しても検討されている．文脈処理の適切性の根拠をより適切な回答が得られる所に求める点が新規的である．NICT のシステムは暗黙にこれと同じ処理を行っていると考えることもできるが，この点を明確に主張した意義は大きい．

Forst は，これと同じ考え方（彼らは “cohesion of knowledge” と呼んでいる）に加えて，格フレーム知識と中心化理論を用いた候補の絞り込みを行っている [11]．RitsQ が行っている言語理解的手法に文脈処理の適切性の根拠をより適切な回答が得られる所に求めるという考えを組み合わせているともいえる．ここでは，動詞の必須格が 0 代名詞化した場合のみを対象とし，日本語語彙大系の格フレームと中心化理論に基づいて予測される参照物の候補（シリーズ最初の質問と省略補完後の直前の質問とその回答から選ばれる）の尤もらしさを，質問応答システムがその補完の結果に対して出してきた回答の得点に求めている．興味深いのは，中心化理論を採用すると，収集型での評価は若干下がるものの一般には低いブラウジング型の評価が収集型よりもやや高くなるという点である．また，文脈処理の失敗がそのまま質問応答システム全体の誤りに結びつくわけではないという知見も得られており，この点も質問応答システムを前提とした文脈処理について考えるきっかけとなる．

5　おわりに

利用者がレポート執筆を目的として対話的に質問応答システムを利用する状況を想定して，そこで必要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を主たる評価の対象とする，情報アクセス対話に向けた質問応答技術の評価タスクを説明し，それを用いた 2 回目の評価の実施について報告した．初回と較べ，質問範囲，正解範囲の精緻化を行うとともに，その評価を自然なものとするために多段階評価の採用や正解セットの導入等，様々な工夫を行っている．加えて，テストセット構築では WoZ 方式という新しい試みを行った．

その実施結果からは，情報アクセス対話に向けた質問応答技術への新しい試みや着実な技術進歩が伺われる．

今後の課題として，タスク定義やテストセット構築の手法について一部の問題が残されている．既に自覚されていた問題であるが，正解判定においてシステムのそれまでの出力が考慮されていないこと，ブラウジング型のシリーズ構築にリアリティがないことが解決されず，そのまま積み残されている．評価の実施という点では，再利用可能なテストセット構築についてプーリングが不充分であることが危惧されると共に，質問応答技術の評価全体としては参加者の減少があったことが悔やまれる．

参考文献

[1] AQUAINT Home Page: Advanced Question & Answering for Intelligence. http://www.ic-arda.org/InfoExploit/aquaint/.

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

[3] Jun'ichi Fukumoto, Tsuneaki Kato and Fumito Masui. 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, 2003.

[4] Andrew Hickl, Johm Lehmann, John Williams, and Sanda Harabagiu. Experiments with Interactive Question Answering in Complex Scenarios. *Proceedings of HLT-NAACL2004 Workshop on Pragmatics of Question Answering*, pp. 60-69, 2004.

[5] Tsuneaki Kato, Jun'ici Fukumoto and Fumito Masui. Question Answering Challenge for Information Access Dialogue – Overview of NTCIR4 QAC2 Subtask 3 –. *Proceedings of NTCIR-4 Workshop Meeting*, 2004.

[6] 加藤恒昭，桝井文人，福本淳一，神門典子．リスト型質問応答の特徴付けと評価指標 **情報処理学会自然言語処理研究会** *2004-NL-163*, pp. 115–112, 2004.

[7] Tsuneaki Kato, Jun'ici Fukumoto and Fumito Masui. An Overview of NTCIR-5 QAC3. *Proceedings of Fifth NTCIR Workshop Meeting*, pp. 361–372, 2004.

[8] Tsuneaki Kato, Jun'ici Fukumoto, Fumito Masui and Noriko Kando. Are Open-domain Question Answering Technologies Useful for Information Access Dialogues? – An empirical study and a proposal of a novel challenge – *ACL TALIP (Trans. of Asian Language Information Processing)*, In Printing, 2006.

[9] Elizabeth D. Liddy. Preparing to Explore a New Paradigm in Information Access: A Scenario Approach to Question-Answering. http://nrrc.mitre.org/NRRC/workshop03/Scenario_BaseQAWriteup.htm. 2003.

[10] Megumi Matsuda and Jun'ichi Fukumoto. Answering Questions of IAD Task using Reference Resolution of Follow-up Questions. *Proceedings of Fifth NTCIR Workshop Meeting*, pp. 414–421, 2005.

[11] Tatsunori Mori and Shinpei Kawaguchi. Answering Contextual Questions Based on the Cohesion with the Knowledge – Yokohama National University at NTCIR-5 QAC3 –. *Proceedings of Fifth NTCIR Workshop Meeting*, pp. 386–393, 2005.

[12] Masaaki Murata, Masao Utiyama, and Hitoshi Isahara. Japanese Question-Answering Systems Using Decreased Adding with Multiple Answers at NTCIR 5. *Proceedings of Fifth NTCIR Workshop Meeting*, pp. 380–385, 2005.

[13] Yuichi Murata, Tomoyosi Akiba, Atsushi Fujii, and Katunobu Itou. Towards Question Answering Challenge 3: Automatic Lexico-Syntactic Pattern Acquisition for Answer Evaluation and Context Processing exploiting Dynamic Passage Retrieval. *Proceedings of Fifth NTCIR Workshop Meeting*, pp. 394–401, 2005.

[14] NTCIR (NII-NACSIS Test Collection for IR Systems) Project Home Page. http://research.nii.ac.jp/ntcir/index-ja.html.

[15] Sharon Small, Nobuyuki Shimizu, Tomek Strzalkowski, and Liu Ting. HITIQA: A Data Driven Approach to Interactive Question Answering: A Preliminary Report. *AAAI 2003 Spring Symposium New Directions in Question Answering*, pp. 94-104, 2003.

[16] TREC Home Page. http://trec.nist.gov/.

[17] 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.

[18] Ellen M. Voorhees. Overview of the TREC 2001 Question Answering Track. *Proceedings of TREC 2001*, 2001.

[19] Ellen M. Voorhees. Overview of the TREC 2004 Question Answering Track. *Proceedings of TREC 2004*, 2004.