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

2004－ＦＩ－ 76　(16)
2004－ＮＬ－163　(16)
2004／9／17

# リスト型質問応答の特徴付けと評価指標

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

**概要**

質問応答におけるリスト型課題とは，与えられた質問に対して０個以上存在する正解のすべてを過不足なく求める課題である．この課題設定はごく自然なものであると思われるが，それにもかかわず，あるいはそれゆえ，多くの事例を眺めていくと質問応答に関する様々な問題が浮かび上がる．本稿ではリスト型質問応答の特徴について考察し，特にその評価についての問題を整理し，その一部を解決した評価手法を提案する．

# Characterization of List-Type Question Answering and its Evaluation Measures

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

**Abstract**

The list-type task in open-domain question answering evaluation workshop is the task in which systems are requested to enumerate all and only correct answers to a given question, the number of which is more than or equal to zero. This setting seems realistic and natural. It, however (or therefore), reveals many problems on the framework of current question answering to investigate several examples in this list-type question answering. In this paper, we examine the characteristics of list-type question answering, especially discuss its evaluation, and propose its new framework that solved some of the problems.

## 1　はじめに

質問応答におけるリスト型課題とは，0 個から複数個までの正解を持つような質問に対して，それら正解のすべてを過不足なく求める課題である．順位付きで決められた数の回答をし，その中のできるだけ高順位に正解 (複数の正解がある場合はそのうちの任意のひとつ) を持ってくることを目的とする初期の課題設定 (以下，順位課題)[8] に較べると，この課題は利用者の現実の要求に即した実用的かつ自然なものと考えられる．また，システムにとっては，解答候補を網羅し，それらのどれ (どこまで) を解答とするかを判断する必要があるためにより困難な課題となる．

自然で実用的なリスト型質問応答であるが，その評価では順位課題に較べて多くの問題が生じる．これらの問題は質問応答技術が今後何を求めていくかだけでなく，質問応答そのものの語用論にも関わるように思われる．本稿ではこれらリスト型質問応答の特徴と問題について議論する．まず議論の前提となる質問応答の枠組みに触れた後，リスト型課題の前提となる複数の解答を持つ質問について整理する．それらをふまえてリスト型質問応答における評価の問題を列挙する．続いて，その一部を解決する評価方法を提案し，その評価指標の振る舞いを QAC2 Subtask 3 を例に説明する．最後に残された課題について今後の検討の方向性を述べる．

## 2　質問応答の枠組み

本稿での議論の多くは NTCIR4[1] で行われた QAC2 Subtask 3 での経験に基づいている．QAC[2] は日本語による質問応答の評価チャレンジであり，3 つの Subtask からなる．それらは共通して，事実に基づき名称を解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加えて，日付け，数値を含み，更に種の名称，機械部品や臓器などの一般名称を含む．統語的には複合名詞がそ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ある．システムは，それを含んだ部分でなく名称それ自体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集合は新聞２紙各２年分の記事 (QAC1 では１紙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．

システムは解答と合わせてそれを抜き出した根拠記事を提示する．ある解答が正解であるかは，解答とその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り，逆にその根拠記事の中でその解答が正当化されていれば，その内容が他の記事と矛盾していても正解となる．

ここで議論する問題の一部はこれらの枠組み，つまり，解答の範囲が名称であること，1 記事内で訂正されない情報は正しいとすることと関係する．ただ，問題の本質は QAC の枠組みを越えた一般的なものであり，本稿での議

論は今後の様々な質問応答に有益であると考えている．

QAC2 Subtask 3 は，情報にアクセスするための対話における質問応答を模擬した質問応答技術の評価タスクであり，その主眼は参照表現の理解等，質問応答における語用論処理の評価にある [4]．ただし，本稿ではこの点は問題とせず，個々の質問について，リスト型質問応答としての側面にのみ注目する．従って，その議論はリスト型質問応答一般で成り立つものである．なお，この Subtask 3 でリスト型質問応答の問題が浮き彫りになった理由は，本 Subtask での質問セットの収集が，あるトピックに関するレポートを作成するという目的で必要となる情報を問い合わせるという現実的な状況を想定して行われており [5]，得られた質問がある意味でルーズで自然であり，質問応答技術を評価する目的で緻密に計算された人工的なものでなかったことによると考えている．

## 3　リスト型課題における質問の分類

そもそもリスト型課題という設定が必要となるのは，ある種の質問に複数の解答が存在するからである．意見や予測を求める質問であれば様々な解答が可能なのは当然であるが，ここで前提としている事実を解答とする質問で解答が複数となるのはどのような場合なのか．その理由を分類する．以下の議論で，質問文の例の殆どは QAC2 Subtask 3 の参照用 run (代名詞や省略を含む対話的な質問に人手で文脈処理を行った質問からなるテストセット) からとっている．本文中では質問 ID で参照し，質問文は付録にまとめた．

**いわゆるリスト型質問の場合 (事物の列挙)**　スポーツ大会への参加国 (078) やある作家の過去の著作 (213) など，ある集合を構成する要素となる事物の列挙を求めるような質問の場合は，当然，解答が複数となる．これがリスト型質問の典型である．ただし，このような場合でも質問者が常に解答が複数であると期待しているわけではない．例えば，ある人物が師事した先生 (095)，発見や調査の日付 (056)(196) 等は，解答が複数の場合とただ一つの場合とがある．このため，このような質問であっても，文型として他と区別されるわけではない．

**質問の内容が曖昧もしくは漠然としている場合**　ある人物の年齢 (242) のように文書集合が複数年にまたがる場合に当然解答が複数になるものがあり，どの時点の値なのかを指定しないと解答が一意とならないものは多い．他にも販売台数 (032) のように，年間なのか月間なのか国内のみか海外を含むのかなど，一般的な表現だけでは解答が一意化しないものもある．おもしろい例では，中国産の朱鷺が「どこから日本に連れてこられたか」という質問 (176) には，本当の意味での生息地 (捕獲地) と中国を飛び立った際の空港とのふたつの解答がありうる．

**新聞記事に異なる複数の情報がある場合 (情報の列挙)**　芝居の初日 (131)，製品の発売日 (008)，事故等による死傷者数 (004) など，常識的には一意であるものであっても，新聞記事に異なる複数の情報があり，それらの「情報」の列挙を行うことによって解答が複数になる場合がある．情報が複数となるのは，予定 (予測ではなく確定的な表現で述べられいているもの) と実際が異なる場合，時間の経過と共に情報が変化した場合，情報源 (被害者側と加害者側等) で情報が異なりそれを伝聞の形で述べている場合等がある．更に新聞記事に誤りが含まれることもある．例えば，室生寺の五重塔に被害を与えた台風 (100) は 98 年の台風７号であるが 99 年の新聞記事では９号と述べられている．また誤りとまではいえないが，投書記事で番組名でないものを番組名であるかのように扱っているもの (015) もある．

**名称という解答の範囲に収まらない場合 (特徴の列挙)**　建造物の大きさ (192) や建造の時期 (193) などに関する質問の中には，「直径１４メートル、高さ３．３メートル」「７世紀末～８世紀初頭」のように名称の範囲を超えるものが解答となる場合がある．場所についても名称で表現しきれない「タイ南部スラタニ空港近くの沼地」のようなものが解答となる場合 (038) がある．これらの場合，それぞれのうちで，意味があり名称である部分をすべて解答としているため，これらの例は，「１４メートル」「３．３メートル」，「７世紀末」「８世紀初頭」，「タイ南部」「スラタニ空港 (近く)」「沼地」という複数の解答を持つものとして扱われる．また，「どのような人でしたか」という質問 (109) も「四十歳代」「女性」を解答のリストとしている．これらは，名称で表現できる範囲を単位とした「特徴」の列挙を求めていると捉えられる．

## 4　リスト型質問応答の評価とその問題点

リスト型課題における評価には数々の難しい問題があり，それら各々に対する方針の決定が要求される．どのような方針をとるかは質問応答に求めるものの反映となる．難しさのひとつは，リスト型質問応答は正解のすべてを過不足なく求めるものであるからその点を評価しなければならないということにある．過不足のなさという点で文書検索と同様に精度 (適合率) と再現率のふたつを考慮して評価されるのが適当であろうが，質問応答では文書検索では生じない様々な問題が生じることになる．

**個々の解答に関する正解判断**　ある解答が正解であるかの判断は質問応答の基本でありながら，そこにも難しい問題がある [6]．姓のみ名のみの人名を正解とするか (「ジュリー・テイモア」における「ジュリー」や「テイモア」(130))，概数表現の有無を正解に影響させるか (「４０歳代」(113)，「約３０人」(003)) 等，解答が表現として充分かという問題や，解答の粒度 (詳細さ) に関する適切性の議論 (キトラ古墳の場所 (191) として「日本」を正解とするか，逆に表現の問題と関連して「奈良県明日香村阿部山」の「阿部山」のみを正解とするか)

がある．これらに加えて，リスト型課題では，ある解答を回答したことによる報償と回答しなかったことによるペナルティが対称的であることからくる問題がある．つまり，ある解答を正解と判断することはそれを回答したシステムの精度を上げるが，それを回答しなかったシステムの再現率を下げる．これはある解答を正解としてもそれを回答しなかったシステムの得点を下げない順位課題との大きな違いである．それを答えることは間違いではないが，それを答えないことでペナルティを与えることに抵抗を感じる解答，例えば，上述の誤報により複数の情報が存在するものや，名前を訊ねる質問 (219) に対する正式名称「0 系」と愛称「夢の超特急」のように位置づけが異なるもの等があり，それが判断を更に難しくする．

**解答の同一性の判断**　ひとつの事物を指示する複数の表現がある．「ゴン・中山」と「中山雅史選手」は同じ人物 (084) であるし，99 年の記事を参照した「昨年五月」と「九八年五月」は同じ日を指示している (064)．これら同じ事物を指示する表現は同じ解答であると判断したい．これらを異なる解答とするとある事物を指示する表現の網羅を要求することとなり，それは質問応答の趣旨からはずれる．円やドル等の貨幣単位の違い，日本時間と現地時間の違いも表現の違いと考えられ，例えば，「３００万ドル」と「約３億６０００万円」は同じ金額を意味するとしてひとつの解答として扱うことが適当と思われる (203)．このため，複数の解答があった場合，それらが正解であるかの判断に加えて，それらが異なる事物なのか，同じ事物を指示する異なった表現なのかを判断する必要がある．前者の場合は両方回答することを要求する (一方のみの場合は再現率が下がる) ことになるし，後者の場合は両方を解答に含めた場合に重複があるとして，場合によってはペナルティを課すことになる．この判断が容易でない以下の場合がある．

**一般物の場合**　ある国の北西部がその国の海岸地域と同じ区域を指している場合，「北西部」「海岸地域」は同じものの異なった表現なのか，異なる事物なのかの判断は難しい．NATO 軍が中国大使館を誤爆したのはこれを「武器供給機関ビル」「補給・調達本部」「軍事関連施設」と見誤ったためである (006) が，この３つは同じものを指す別の表現なのか，異なるものなのかの判断は難しい．加えて，特徴の列挙ということで「タイ南部スラタニ空港近くの沼地」というひとつの場所に対して「タイ南部」「スラタニ空港 (近く)」「沼地」の３つを正解として，すべてを求めている点との整合性も問題となる．

**概数の場合**　「２２６キロ」「約２３０キロ」(187)，「数億円」「４億円」(104) は，同じ情報か異なる情報かの判断が必要になる．上述のように列挙するのは事物のみでなく情報の場合もあるため，情報として同じかという判断が必要になる．

**体系の違い**　年の表現において，西暦と元号による違いは同一年を指示しているとして問題ないだろうが，「平安時代初期」「九世紀前半」のような組み合わせ (101) も同じ扱いでよいかは自明でなく，異なる特徴を列挙していると捉えることも可能である．

**重複した解答の扱い**　上述のような理由から，一般的な文書検索 (www ページ検索では同様の問題があると予想されるが) とは異なり，システムの回答リストの中に同じものを指示している重複した解答が含まれる可能性がある．このような重複した解答の存在を評価にどう反映するかが決められなければならない．１つの極は重複の有無に関わらず個々の判断に基づいて正解とする，もう一つの極は重複した解答はそのうちの１つだけを正解とし他は誤答として扱うというものである．後者の場合，重複を含んだ回答リストはそれだけ精度が低くなる．

**解答列挙のシステムの扱い**　解答の列挙に複数の方法 (システム) がある場合がある．大人と子供，男性と女性のようにある集合を分割する複数の異なるシステムがあると，「北西部」「南東部」と「海岸地域」「平野部」のような異なる列挙で同じ地域を網羅できることになる．この時「北西部」「海岸地域」という回答リストをどう評価するかが問題となる．また，このような異なる列挙のシステムを同等に扱ってよいかが問題となる場合がある．典型例は粒度の問題と関連して生じる．例えば「９８年初めに行われたノルディックスキーW杯ジャンプの開催地」は，プラニツァ (スロベニア)，ガルミッシュパルテンキルヘン (ドイツ)，ラムソー (オーストリア)，インスブルック (オーストリア)，ビショフスホーフェン (オーストリア) の３カ国５都市であるが，「スロベニア」「ドイツ」「オーストリア」という列挙と「ラニツァ」「ガルミッシュパルテンキルヘン」「ラムソー」「インスブルック」「ビショフスホーフェン」という列挙に同じ評価を与えるべきかが問題である．「プラニツァ」「ラムソー」「オーストリア」という回答リストをどう扱うかは更に複雑である．このような例は図 4 に示すように豊富で，明らかに情報の量が異なってしまうものもあれば，詳細な解答が優れていると思えないものもあるところに難しさがある．

以上のように方針として決めなければならない，決めるのが難しい問題に加えて，再現率を用いることに関連して以下のような問題がある．

**テストセットの再利用と比較可能性の減少**　新しい正解が見つかった時点，今までどのシステムも回答しておらずプーリングされていなかった解答が新たに現れ，それが正解と判断された時点で正解総数が変化し再現率が変わるので，それが見つかる以前の結果と正確な比較ができなくなる．これは文書検索でも同様であるが，質問応答では正解が比較的少数なので，それまでひとつしか正解がないとされていた質問にマイナーな別解が見つかるとその途端に再現率が半分になることがある．順位課題の

(008) 当初発表の「11 月 20 日」と実際の発売日の「11 月 27 日」との２つの正解があるが，「昨年十一月に日本で先行発売された」と「十一月」という解答も可能．
(052) 「主食の魚介類が必要だが」「川魚やエビ、カニなどを食べる」「好物のウナギやモクズガニなど」「魚や貝類を主食として」と様々な粒度の正解がある．
(056) 「９２年３月に…が、９４年１月と３月にその近くで、…が見つかっている」と９４年には２回の発見があるが，「９２、９４年に…、ふんが見つかった」から「９４年」という解答も可能．
(101) 「八世紀後半の創建」「天平時代末―平安時代初期にさかのぼる」「平安時代初期 (九世紀前半) の優美な姿」とあり，世紀による列挙，時代による列挙が可能な上，「天平時代末―平安時代初期」を１つの解答とする選択肢もある．
(142) 「「マロリー＆アービン捜索隊」によって…発見された」「遺体を発見した「マロリー＆アービン捜索隊」隊長のエリック・シモンソン氏が」と捜索隊名と人名の選択肢があることに加え，人名については「遺体の最初の発見者、アンカー隊員も」ような解答もある．
(175) 「中国西安から…到着した」「中国陝西省洋県から約三千六百キロの空の旅を終えて」との２つの正解があるが，「…を贈った中国や地元関係者からは」と「中国」という解答も可能．
(238) 「頭骨や手足の骨、石器使用の跡など」「頭蓋骨や歯、大たい骨、上腕骨、前腕骨など」とあり，「手足(の骨)」が「大たい骨」「上腕骨」「前腕骨」を含んでいる．

図 1: 複数の解答列挙システムと解答粒度が問題となる例

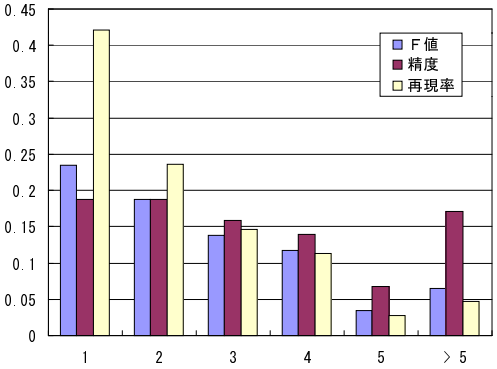


図 2: 質問の正解数と $F$ 値との関係

場合は新しい正解が見つかった場合でも既に評価された結果の点数が変わることはない．

**正解総数と問題の難易度** 経験的事実として正解数の多い質問の方が高い $F$ 値を得にくい，いわゆる難しい質問となる．図 2 に QAC2 Subtask 3 の参照用 run での状況として，参加 14 システムの平均 $F$ 値と質問の正解数との関係を示す．図からもわかるように，これは正解数の多い質問の方が再現率が下がるためである．なお，この傾向は $F$ 値における精度と再現率の寄与率を調整することでは解決しない．質問毎に難易度が異なるのは当然のことであるが，それが正解数とシステマッティックな相関を持っていることは問題であるように思う．ただ，このような傾向は決して数学的根拠のあるものではなく，技術の現状に帰因した一時的なものである可能性もある．

## 5 QAC2 Subtask 3 における評価

本節では，上記の問題に関する QAC2 Subtask 3 での基準について述べる．上で述べた問題のうち，個々の解答の正解判定と解答の同一性の判断の方針については，一定の方針を設けそれに従って行ったが，個々の事例の積み上げになるのでここでは述べない．それを答えることは間違いではないがそれを答えないことによってペナルティを与えることに抵抗を感じる解答についても，現時点では何の提案も行えていない．ただ，これらの問題に関連した判断を通じて整理されたのが，3 節で述べた複数解答を持つ質問の分類である．事物の列挙，特徴の列挙，情報の列挙という質問の分類によって，同一性の判定の基準も変わると考えている．

### 5.1 重複した解答の扱い

QAC2 Subtask 3 では，評価は精度と再現率の両方を考慮して $F$ 値を用いた．重複した解答の扱いについては，その修正として対処している．参加者に周知した主たる評価指標 (以下，$MF1$ と呼ぶ) では，同じ解答もしくは同じものを表現する異なる表現を複数リストに含めた場合は，そのうちひとつだけを正解とし，それ以外は誤答としている．従って，重複を含んだ場合は精度が落ちることとなる．これは，システムの正解を過不足なく抽出する能力には同じものを指示する表現を判定して重複を取り除く能力が含まれるという判断からきている．もうひとつの修正として，正解のない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 としている．実際の評価に用いられるのは，$MF1$ をすべての質問に関して平均した $MMF1$ である．

採点の過程で解答の同一性の客観的判定が困難であるという感触と $F$ 値における再現率が色々な点で問題となるという観察から以下のふたつの補助指標を考案した．第一の補助指標 (以下，$MF2$ と呼ぶ) は，重複を取り除く能力を評価に含めないことを近似するもので，重複した解答を正解から除くだけでなく，解答全体からも取り除いて計算した精度を用いた修正 $F$ 値である．例えば，システムが 5 件のリストを回答し，そのうちの 3 件が正解で，更にその内の 1 件が他と重複していた場合，$MF1$ での精度は $(3-1)/5 = 0.4$ であるが，$MF2$ では $(3-1)/(5-1) = 0.5$ となる．再現率は $MF1$，$MF2$ で同じ値が用いられる．

第二の補助指標は，正解を過不足なく抽出することの困難さを考慮し，これを正解のひとつを見つければよいという課題に近似するもので，順位課題で使われる $MRR$ を順位なしのリストに拡張したものである．$RR$ は次のように解釈することができる．質問応答の場合，システムが提示した解答を根拠記事等を用いて検証するというコストがかかる．順位付リストでは順位の高いものから確認するので，1 位に正解があれば 1 回の検証で正解が得られるし，2 位にあれば 2 回の検証が必要となる．つまり，ひとつの正解を得るためのコストつまり検証回数は最初の正解がある順位に一致する．このコストの逆数が $RR$ である．これを順位なしのリストに適用して，$m$ 個の項目からなるリストを回答し，そこに $n$ 個の正解が含まれている場合に必要な検証のコスト $c(m,n)$ を考えると，まず解答のうちの任意のひとつを取り出して検証するが，その検証で正解が得られればコストは 1，そうでなければ残った $m-1$ 個について同様の検証を繰り返すことになるので，更に $c(n,m-1)$ が加わる．この最初の検証で正解が得られない確率は $(m-n)/m$ であるので，漸化式 $c(n,m) = 1 + \frac{m-n}{m}c(n,m-1)$ が得られる．この漸化式を解くと $m+1/n+1$ が得られる．この逆数を評価指標とする．$n=0$ の場合は，これらの検証作業によって得るものがないので，分子が 0 ということで，0 である．この指標を $RC$(reciprocal cost) と呼ぶ．つまり，

$$RC = \begin{cases} 0 & \text{(正解を含まないとき)} \\ \frac{\text{正解数}+1}{\text{解答数}+1} & \text{(それ以外のとき)} \end{cases}.$$

システムはすべての問題の $RC$ を平均した $MRC$ によって評価される．$MRC$ による評価の場合，その位置づけから正解数はそれが他と重複しているかどうかにに関係なく計算される．また，正解がない質問は評価の対象から外される．$RC$ は精度と関連の高い評価である (ただし精度の関数ではない) ので，正解総数と評価の相関の問題を避けることができる．図 3 に先程と同じテストセットでの正解数との関係を示す．ただし，$MRC$ は，これだけを指標とするのであれば確実な解答をただ一つだけ答えることで高い評価が期待できてしまうので，あくまで補助指標としての役割しか果たさない．

QAC2 Subtask 3 参照用 run 上位 9 システムをこれらの指標で評価した結果を図 4 に示す．横軸に示されているのはシステムの ID である．$MMF1$ を主とし，ふたつの補助指標 $MMF2$，$MRC$ を用いてシステムの様々な特徴を浮かび上がらせることを期待したが，図からわかるように，実際には，大きな差は見られていない．ひとつの理由は，参加システムは $MMF1$ で主たる評価がなされることを前提としているので，$MMF2$，$MRC$ を最大にするような設計をされていなかったことにあると思われる．

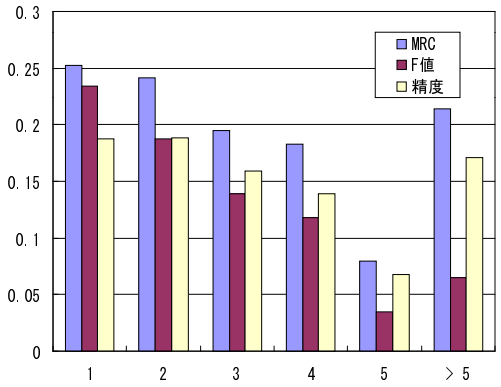


図 3: 質問の正解数と $MRC$ との関係

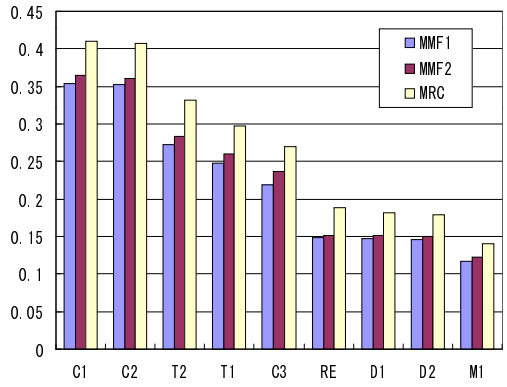


図 4: 提案した評価指標の振る舞い

### 5.2 粒度等に関連した解答列挙のシステムの扱い

解答列挙のシステムが複数あることや解答の粒度と関係してそれらシステム間に優劣があることに対しては，解答列挙のシステムに対応させて複数の正解セットを用意し，それらを用いた採点結果のうち，最も高い評価となったものを採用することで対処している．以下，具体例で説明する．

**例 1**　キトラ古墳の作られた時代 (193) の正解として，「7 世紀末」「8 世紀初頭」というふたつを答える回答と「古墳時代終末期」のひとつを答える回答の両方を許し，前者はふたつが含まれていて再現率 1.0，後者はひとつだけで再現率 1.0 としたい．

**例 2**　「12 月 20 日」「12 月 24 日」のふたつの日付けが正解となる質問 (131) で，「12 月」とだけ回答した場合，ふ

たつのうちのひとつだけ回答したものと同じ扱いとし，「12 月」と「12 月 20 日」,「12 月」と「12 月 24 日」でも一方しか回答していないものとして扱いたい．

例 1 の場合　セット 1 の内容を「7 世紀末」「8 世紀初頭」で正解総数 2，セット 2 の内容を「古墳時代終末期」のみで，正解総数 1 と与える．例 2 の場合　セット 1 の内容は「12 月 20 日」「12 月 24 日」で正解総数 2，セット 2 の内容は「12 月」で，正解総数 2 とする．ここでセット 2 では実際には正解がひとつしか含まれていないにもかかわらず，正解総数が 2 となっているので，「12 月」のみの解答は再現率 0.5 となる．「12 月」と「12 月 20 日」のふたつを回答した場合も，どちらのセットを使っても一方しか正解とならないので，意図した評価を与えることができる．この例では「12 月」のみという解答列挙を「12 月 20 日」「12 月 24 日」より劣るものとした評価が実現できる．QAC2 Subtask 3 の全質問 251 問のうち，複数の正解セットを必要とした質問は 18 問で，正解セット数はたかだか 2 つで充分であった．

解答が複数の正解セットをまたがっていた場合，例えば例 1 で「7 世紀末」「8 世紀初頭」「古墳時代終末期」からなるリストを回答した場合，セット 1 の採点では「古墳時代終末期」が，セット 2 の採点では「7 世紀末」「8 世紀初頭」が誤りであると判断される．この例では精度が高いセット 1 による採点が採用される．この問題の更に極端な場合は以下のようなものである．セット 1 は A, B のふたつからなっており，B を 4 つに細分化した別の表現があるので，セット 2 として A, B1, B2, B3, B4 を用意したとする．A, B という解答も充分ということでセット 1 の正解総数は 2，セット 2 は 5 とする．この場合，A だけを回答したシステムはセット 1 を用いて精度 1，再現率 0.5 となるが，A と B1 を回答するとセット 1 の精度 0.5，再現率 0.5 かセット 2 の精度 1，再現率 0.4 の高い方となり，いずれも A だけを回答した場合より評価が低いという直観に反する結果となる．

これは，この方針がある解答列挙のシステムに従って統一的に回答することを期待しており，複数のシステムを混在させることに価値を与えていないためである．前述のW 杯ジャンプの開催地の質問でいえば，国の列挙で回答するか都市の列挙で回答するかのどちらかで統一すべきであるという方針である．正解セットの作り方でこの問題は回避できないこともないが，その場合は正解セット数が組み合わせ的に多くなることが危惧される．また，他の正解セットに含まれる解答を誤答ではなく重複した解答と考え，先に議論した $F$ 値の修正のバリエーションを工夫することで，この問題を緩和することもできる．

## 6　考察と今後の展開

TREC の QA Track でも，2003 年より回答数を指定しないリスト課題が開始されている [9]．評価は $F$ 値である．この課題の質問は 37 問とあまり多くなく，“List the names of chewing gums.”，“Who are female boxers?” 等，すべてが事物の列挙を求めるいわゆるリスト問題である．更に殆どの問題は，“What Chinese provinces have a McDonald’s restaurant?” のように解答のクラスが巧みに指定されており，粒度の問題が生じるような表現，例えば “Where in China does McDnald have a restaurant?” は避けられている (Ellen Voorhees に確認したところ，これは意図的に行われているとのこと)．質問文のみからの判断であるが，問題が出る可能性のあるのはわずかに “What foods can cause allergic reaction in people?” の 1 問だけである．このような状況であるので，本稿で議論したようなリスト型課題の問題に着目した提案は著者の知る限り全く行われていない．

個々の正解判定の困難さに着目して多段階評価を提案し，それを基に順位型課題とリスト型課題の融合を図ろうという提案がある [7]．そこでの課題は，正解を過不足なく求めるという本稿でのリスト課題とは若干異なるが，その提案は本稿での議論と矛盾するものでなく，本稿での提案と組み合わせることで評価の可能性が広がると考えている．

本稿では，質問応答におけるリスト型課題の特徴について考察し，その評価に様々な難しい問題があることを述べ，その一部を解決した評価指標を提案した．3 節で述べた複数の解答が存在する質問の分類に立ち返ってみると，ここで示した評価指標は，やはり，事物の列挙を求める「いわゆるリスト問題」を前提にしているという印象を持つ．曖昧なもしくは漠然とした質問や情報の列挙を (結果的に) 求める質問に対しては，そもそも解答の列挙が正しい対応なのかが疑問である．曖昧性解消のためのシステム側からの問い返し [3] もひとつの方法であろうし，条件や背景を付け加えた「９８年で４６歳です」「６日時点の新華社通信の発表で３名です」のような協調的応答の枠組みも必要であろう．更に進んで，複数記事の情報を総合し矛盾を解消した解答を求める枠組みも必要となるかもしれない．解答が名称の範囲に収まりきらない特徴の列挙となる質問については，やはり，名称という制限を緩めて，例えば名詞句を解答範囲とするのが本筋であろう．もちろん，そのことで問題が解決するわけではなく，文書集合からの抜き出しを前提とすれば，「スラタニ空港から南西約３キロのゴム園」と「タイ南部スラタニ空港近くの沼地」とはどの程度の情報を共有しているか等の判断が必要となる．その際は，TREC における定義質問の評価で用いられている情報のナゲットを単位とした評価 [9] が参考になるかとも思う．いずれにせよ，質問応答技術に何を求めていくかの議論が必要になろう．

$F$ 値を用いることによる評価の問題は，ともかくプーリングを充分行って，正解候補を網羅し，それら全体を俯瞰できる状況で個々の正解判断をしていくことによってしか解決しないと思われる．つまり，評価作業の手順や方法論で，評価指標が持つ問題を減じていくことが現実的であろ

う．一方で，様々な評価指標を試みることも重要である．それらも質問応答技術に何を求めるかに関連するからである．ただし，それらの指標は事前に提案され，システム設計に影響を与えるものでなくてはならない．run の実施後に複数の評価を実施しても大きな差は得にくいというのが今回の経験からの結論である．

本稿で行ったのは質問応答システムがどのような解答を出すべきか，どのような解答を出した場合を高く評価すべきかという，利用者の側からの考察である．システムを設計する側としては違う観点での評価が質問応答技術の発展のために必要になると思われる．それらを合わせてスコープに入れた議論が必要である．

## 謝辞

QAC2 に参加していただき熱心な議論を行ってくださった皆様に感謝します．特に，乾健太郎氏，秋葉友良氏には，評価に関して貴重なコメントいただきました．また，QAC2 Subtask 3 の採点をお手伝いいただいた奈良先端科学技術大学院大学の学生の皆様にも多くの示唆をいただきました．ここに感謝します．本研究の一部は，国立情報学研究所の共同研究として支援されています．

## 参考文献

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

[2] 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.

[3] Chiori Hori, Takaaki Hori, Hideki Isozaki, Eisaku Maeda, Shigeru Katagiri, Sadaoki Furui. Deriving Disambiguous Queries in a Spoken Interactive ODQA System. *Proceedings of ICASSP-2003*, Vol.I pp.624 – 627, 2003.

[4] 加藤恒昭，福本淳一，桝井文人，神門典子. 質問応答から対話理解へ – NTCIR QAC Task3 の提案 –. 言語処理学会第 10 回年次大会, D2-7, pp. 317 – 320, 2004.

[5] 加藤恒昭，福本淳一，桝井文人，神門典子. 質問応答技術は情報アクセス対話を実現できるか. 情報処理学会研究報告, 2004-NL-162, pp. 145 – 150, 2004.

[6] 桝井文人, 福本淳一, 加藤恒昭, 神門典子. 質問応答システム評価用テストコレクションの構築　～ NTCIR QAC の取り組み～. 言語処理学会第 10 回年次大会, D2-6, pp. 313 – 316, 2004.

[7] Tetsuya Sakai. New Performance Metrics based on Multigrade Relevance: Their Application to Question Answering. *Working Notes for the Fourth NTCIR Workshop Meeting*, Supplement Volume 2, pp. 9 – 16, 2004.

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

[9] Ellen M. Voorhees. Overview of the TREC 2003 Question Answering Track. *Proceedings of TREC 2003*, pp. 14 – 27, 2003.

## 付録

本文中で参照した質問文を示す．ここに挙げたものは全て QAC2 Subtask 3 の参照用 run からの質問である．以下の質問文には `xxx(yyy-zz)` という番号がつけられているが, 本文中では `xxx` の部分によって参照している．QAC コーパスでの正式 ID は `QAC2-32xxx-01` となる．括弧内の `yyy-zz` は対応する Formal run の質問 ID であり，正式には `QAC2-31yyy-zz` である．

003(001-03)　ＮＡＴＯによるユーゴ空爆で中国大使館が誤爆を受けた時、大使館には何人の人がいましたか。

004(001-04)　ＮＡＴＯによるユーゴ空爆での中国大使館誤爆で何名の死者が出ましたか。

006(001-06)　ＮＡＴＯによるユーゴ空爆での中国大使館誤爆は、ＮＡＴＯ軍によれば、大使館を何と間違えたために攻撃したのですか。

008(002-02)　ドリームキャストはいつ発売されましたか。

015(003-02)　野菜に含まれる高ダイオキシン濃度についての報道はどの番組で行われたのですか。

032(005-04)　トヨタのハイブリッド車「プリウス」のこれまでの販売台数はどれくらいですか。

038(006-02)　９８年、タイのどこで飛行機事故がおきたのですか。

052(008-02)　ニホンカワウソはどんなものを食べていますか。

056(008-06)　最後に生きた姿が見られた１９７９年以後で、ニホンカワウソの生息の痕跡が発見されたのはいつですか。

064(009-06)　安室奈美恵が子供を産んだのはいつですか。

078(012-02)　サッカー「フランスW杯」に参加した国や地域はどのようなところでしたか。

084(012-08)　サッカー「フランスW杯」で日本の得点を決めたのは誰ですか。

095(014-04)　小沢征爾さんが師事した先生は誰でしたか。

100(015-02)　室生寺五重塔が被害を受けた台風は台風何号でしたか。

101(015-03)　室生寺五重塔はいつごろ建てられたので

しょうか。
104(015-06)　台風の被害を受けた室生寺五重塔の修復費用はいくらかかりますか。
109(016-03)　９９年の臓器移植法施行後初めての脳死臓器移植でのドナーはどのような人でしたか。
113(016-07)　９９年の臓器移植法施行後初めての脳死臓器移植を受けた患者の年齢は何歳ですか。
118(017-05)　ハワイ・マウナケア山頂に設置されている望遠鏡は何という名前ですか。
130(019-05)　だれがライオンキングの演出をしましたか。
131(019-06)　ライオンキングの日本での初演はいつですか。
142(020-08)　ジョージ・マロリーの遺体を発見したのは誰ですか。
176(025-07)　佐渡トキ保護センターで９９年５月に生まれたトキの優優の両親、「友友」と「洋洋」はどこから日本へ連れてこられましたか。
179(026-03)　タンザニア、ケニアでの米国大使館同時爆破事件で死亡した人はどのくらいいましたか。
187(027-04)　９９年１０月、ＪＲ山陽新幹線北九州トンネルで落下したコンクリート塊はどれ位の重さでしたか。
191(028-01)　キトラ古墳はどこにありますか。
192(028-02)　キトラ古墳の大きさはどのくらいですか。
193(028-03)　キトラ古墳はいつ造られたものですか。
196(028-06)　キトラ古墳の内部調査は、いつ行われましたか。
203(029-06)　９８年のコロンビアでの志村昭郎さん誘拐で身代金はいくら要求されたのですか。
213(031-02)　瀬戸内寂聴はこれまでにどのような作品を書いていますか。
219(032-02)　東海道新幹線の初代車両の名前は何ですか。
238(034-07)　エチオピアで新種の猿人のどの部分が発見されましたか。
242(035-03)　ロシアのプーチン首相は何歳ですか。