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 ～

桝井文人 †，福本淳一 ‡，加藤 恒昭 ††，神門典子 ‡‡
三重大学 工学部 †　立命館大学 理工学部 ‡
東京大学大学院 総合文化研究科 ††　国立情報学研究所 ‡‡
masui@ai.info.mie-u.ac.jp†

1　はじめに

質問応答技術は，1998 年に TREC における QA Track の設定によって注目されるようになったもので，大規模テキスト集合を知識源として，オープンドメインの質問応答を行う技術である [1, 2]．例えば，「ビッグ・アップルとして有名なのは，米国の何という都市ですか。」という質問に対し，「ニューヨーク」という文字列，または，それを含む箇所を回答することが目的である．

日本では，村田らによる質問応答システムの提案 [3] や，佐々木らによる実験的なタスク実施 [4] 前後から，急速に同技術が注目されるようになった．そして，2002 年には第 3 回 NTCIR ワークショップ [5] のタスクとして，QAC1(Question Answering Challenge) が実施された [6]．NTCIR は，評価ワークショップである．評価ワークショップの目的の一つに，テストコレクションの構築がある [7]．例えば，MUC(Message Understanding Conference) では，情報抽出に関する共通テストコレクションに基づく議論を行うことで，要素技術の有効性や問題点が整理された [8] ．QAC でも，質問応答技術の研究を活性化させ，同技術の評価手法を見極めるために，再利用可能な質問応答システムテストコレクションの構築を目指している．

本論文では，QAC における，質問応答システム評価用テストコレクションの構築に関して報告する．以下，2 章で QAC の概要を紹介し，3 章でデータセットについて，4 章でデータセット作成方法について説明する．5 章では，テストコレクション構築における問題点や QAC の対応について述べる．

2　QAC タスクの概要

QAC では，参加システムに対し，質問文に対する回答を，特定の知識源から抽出することを求め，得られた回答と解答データの比較によって，システム性能を評価する．QAC1 では，知識源として，毎日新聞コーパスの 98，99 年が指定され，QAC2 では，さらに読売新聞コーパスの 98，99 年が追加された．質問に対する回答は，何らかの名称もしくは時間・数量表現 (人名や組織名等の固有表現，金額や温度等の数値表現，作品名，日付け，種やカテゴリの名称等) であり，かつ，知識源中の表現そのものでなければならない．

QAC1，QAC2 ともに三つのサブタスク[1] を設定している．task1 では，与えられた質問に対して，優先順位をつけた回答候補を上位 5 位まで回答する．評価は，最も高順位にある正解文字列を対象とし，それらの MRR(Mean Reciplocal Rank) で評価する．task2 では，与えられた質問に対して，質問の回答と判断されたものすべてを列挙して回答する．評価は，回答された正解文字列全てを対象とし，それらの MF(Mean F-value) で評価する．task3 では，連続して入力されたと想定される複数の質問 (枝問) に回答する．評価は，task1(QAC1) または task2(QAC2) と同様である．

テストコレクションは，質問とそれに対応する解答 (正解) 群によって成る質問解答データセットと，各サブタスクの評価基準に基づいて動作する自動採点ツールによって構成される．QAC1 では，最終的に 1219 の質問と，それらの解答データ 7839 が作成・公開され，QAC2 でも，同規模のものが作成されつつある．

3　質問解答データセット

質問解答データセットは，表 1 に示す通り，15 のカラムから構成される．データセットの例を表 1 に示す．

第 1，第 3，第 4 カラムは，それぞれ，task1,2,3 における質問文 ID が記述される．質問文 ID は，“QACX-YY...YY-ZZ” という形式で記述され，“X” の部分には，QAC の実施回を示し，“YY...YY” の部分には，質問文のサブタスク別の通番が記述される．“ZZ” の部分は，質問文の枝問の通番を示す．現状では，task3

[1] 詳細については，文献 [6] または QAC ホームページ [9] を参照されたい．

```
QAC2-10000-00,0,QAC2-20113-01,QAC2-30000-00,f,“米　大リーグで完全試合を達成したヤンキースの選手は。”,“米大リーグで完全試合を達成したヤンキースの選手は誰ですか。”,0,,,,,,“×”,,
QAC2-10000-00,1,QAC2-20113-01,QAC2-30000-00,f,,,1,“デービッド・ウェルズ”,JA-980518291,“米大リーグ　３４歳・快速左腕が快挙　ヤンキースのデービッド・ウェルズが完全試合”,1,,“○”,,
QAC2-10000-00,1,QAC2-20113-01,QAC2-30000-00,f,,,1,“Ｗｅｌｌｓ”,JY-19980519J1TYMBJ1400020,“Ｄａｖｉｄ　Ｗｅｌｌｓ（ニューヨーク・ヤンキースの投手）//ヤンキース史上初の完全試合を成し遂げたヒーローは、”,1,“英語綴りは有効か。フルネームでないが。”,“△”,,
QAC2-10000-00,1,QAC2-20113-01,QAC2-30000-00,f,,,3,“ドン・ラーセン投手”,JA-980518342,“ドジャースとのワールドシリーズでドン・ラーセン投手が記録しているが、”,0,,,,
```

図 1: データセットの例

表 1: 質問解答データセット仕様

カラム	記述項目	表記
1	質問文 ID(task1)	半角英数
2	レコード属性	半角数字
3	質問文 ID(task2)	半角英数
4	質問文 ID(task3)	半角英数
5	テストラン種別	半角英字
6	質問文 (出題)	全角日本語
7	質問文 (原文)	全角日本語
8	解答 ID	半角数字
9	解答	全角日本語
10	根拠記事 ID	半角数字
11	解答根拠部分	全角日本語
12	A:解答曖昧性	半角数字 (0/1)
13	Q:解答網羅性 A:解答曖昧性属性	○／× ○／×／△
14	曖昧性コメント	全角日本語
15	(リザーブ)	–

のみに意味がある．第 5 カラムには，質問を実施した場合のテストラン種別であり，“d” は dryrun，“f” は formalrun，“a” は additionalrun を意味する．したがって，例えば，第 4 カラムに “QAC1-300011-02” という ID があり，第 5 カラムに “f” が記述されていれば，その質問は，QAC1 の formalrun において，task3 の 11 番目の一つ目の枝問として実施されたものであることがわかる．

第 2 カラムは，該当するレコードが質問文であるか，解答であるかを区別する属性を示す．質問文の場合は “0”，解答の場合は “1” となる．第 6，7 カラムは，質問文が記述される．出題時に修正された場合は，出題文を第 6 カラムへ，原文を第 7 カラムへ記述する．第 8 カラムは，解答 ID を示す．異なる解答が存在する場合は，第 9 カラムには，解答が記述され，第 10 カラムは，解答が含まれる根拠記事の ID が記述される．根拠記事 ID には，“980518342” のような数字のみの ID 表記と，“JY-19980519J1TYMBJ1400020” のような suffix 付き ID 表記が可能である．第 11 カラムは，根拠記事中から解答を含む該当個所を抜き出した，根拠部分が記述される．根拠部分が，記事中の離れた複数の文章に存在したり，改行を挟む場合は，スラッシュを 2 つ (//) 挿入して記述する．

第 12 カラムは，解答レコード (A) にのみ適用される情報で，同一の解答として記述された一連の文字列が，曖昧性を含むか否かを示す．解答としての曖昧性を含まない場合は，空欄とするか，“0” を記述し，曖昧性を含む場合は，“1” を記述する．曖昧性がある場合は，第 13 カラムでさらに詳細な属性を付与する．第 13 カラムは，質問文レコード (Q) と解答レコード (A) で意味合いが異なる．質問文レコードでは，このカラムは解答網羅性を示し，対象コーパス中の解答が全て網羅済みである場合は “○”，そうでない場合は “×” を付与する．解答レコードでは，第 12 カラムで曖昧と記述された解答に対する解答曖昧性を示している．正解として扱う場合は “○” を，正解と見なさない場合は，“×” を，判断が難しい場合は “△” を記述する．自動採点ツールでは，デフォルト採点では甘い採点 (“○” および “△” を正解として扱う) を行うが，オプションでより厳しい採点 (“○” のみを正解として扱う) が可能である．第 14 カラムは，解答レコードのみに関係する，曖昧性コメント欄である．第 12，13 カラムで，曖昧と判断された解答について，その根拠や見解を示すコメントが記述される．

4　データセットの作成

データセットの作成作業は，質問作成フェーズと，詳細な解答作成フェーズの二つに大別することができる．まず，質問作成フェーズでは，質問文とその模範解答を作成する．模範解答は，あくまでも出題する質問を選択するための目安として用いたり，質問解答データセットの傾向などを把握するためのものなので，最低限の解答しか登録されていない．この作業が終了した段階で，質問文を選別し，formalrun(あるいは dryrun) を実施する．

formalrun 実施によって，参加システムの回答データが得られると，解答作成フェーズを開始する．正解の存否や網羅性を作成者の作業だけで保証するのは不可能といえる．そこで，収集された参加システムの回答データをプーリングし，明らかな不正解を取り除き，前フェーズで作成した模範解答と統合する．次に，対象となる新聞記事を参照し，プーリングデータ中の回答文字列が，正解であるかどうか，新聞記事中に存在するかどうかの確認を行う．

さらに，作業中に見つかった新たな解答や，括弧なし・仮名表記・姓のみなどの部分文字列解答などを追

加し，データの精度を向上させる．作業時には，同じ箇所を複数の作業者で確認し，作業終了後に，全員で再度確認することで，判断の揺れや，ケアレスミスを出来るだけ補正している．この手順は，次章で述べる，解答に曖昧性がある場合に有効である．

データセット作成と並行して，開発された自動採点ツールの動作試験にも用いられ，動作確認後，テストコレクションとして配布される．その後，参加者から寄せられたフィードバック情報を蓄積し，それらを反映したものを，新バージョンとして公開する．自動採点ツールは，採点対象の回答データと，質問解答データを読み込んで採点処理を行う．出力は，サブタスク毎の採点結果の他，正解や不正解の一覧や，解答の比較など，ユーザの分析作業を支援する様々な情報を出力することが可能である．

5　考察

質問解答テストコレクションを作成することによって，様々な問題点が明らかになってきた．本章では，それらのうち，特に難しい問題，今後の議論を必要とするものを挙げ，考察する．

5.1　質問の曖昧性と難しさ

同じ内容の質問であっても，質問文の表記の頑健性や，質問の意図の違いによって，回答候補を導き出すための条件に曖昧性が生じ，質問の難しさが異なると考えられる．例えば，(1)「日本の首相は誰ですか？」という質問文では,「誰」という疑問語によって，人名を回答すればよいと判断できる．しかし，同じ意図の質問文として，(2)「日本の首相といえば?」を考えた場合，回答として人名が求められているということは，質問文のみからは判断できない．(3)「米国の首都といえばワシントンＤＣですが，日本の首相は誰でしょう。」という質問を考えた場合は，前半に質問内容には直接関係しない表現が含まれている．これらは，質問文の文脈を解釈するようなシステムでない限り，質問文中の重要語特定を難しくする要因となる．

QAC1 では，質問の難しさの違いを排除するために，出題する質問文は全て (1) の形式のみに統一し，QAC2 では，少し制限を緩め，(3) の形式も認めた．

5.2　解答の網羅性 (同一性)

例えば,「三種の神器とは何ですか。」という質問に対する回答を考えてみる．一般的な解答としては,「鏡」「勾玉」「剣」なのであるが，実は，解答の網羅性に問題があった．まず，記事中に「“平成の” 三種の神器」や「“女子高生の” 三種の神器」のような変形「三種の神器」が複数個存在した．QAC1 では，これらも解答の範囲内と見なしたが，約 21 個もの解答が存在することになってしまった．仮に，task2 において，上の質問に対して 10 個回答し，うち 5 個正解した場合と，解答が 2 個の質問に 2 個回答し，うち 1 個正解した場合があったとすると，両者の質問は，明らかに回答の難しさが違うと考えられる．しかし，現状では，両者とも同じ扱いとなり，直感にそぐわないという問題がある．

この問題は，質問文が持つ曖昧性の問題として考えることもでき，質問文の表現を頑健にしておく，という対策もあり得る．しかしながら，実際の人間の場合は，上記の質問に対して，対話を繰り返すことによって，回答候補の範囲を絞り込んでいく過程を経て回答すると思われる．とはいえ，システム側に回答以外の応答を許すようなインタラクティブなインタフェイスは，現状の技術レベルでは少々敷居が高いといえるので，task3 で提案されている収集型質問 [10] のように，元の質問に関連した一連の質問を出題することで，回答に対する曖昧性を減らす工夫が現実的であろうと思われる．

QAC では，上記のような解答の曖昧性があると判断した場合，3 章で説明したデータセットの第 14,15 カラムを利用して対象を三つに分類している．この分類は，自動採点にも利用でき，ユーザの用途に応じて，正解の範囲を変更して採点することが可能である．

5.3　解答の粒度 (詳細度)

例えば,「東京ディズニーランドはどこにありますか」という質問に対して,「千葉県」,「舞浜」,「千葉県浦安市舞浜１の１」,「ＪＲ京葉線舞浜駅前」のいずれを解答と認めるかについては，質問者の意図に依存して異なってくる．例えば，質問者が地図上の位置を知りたいだけであれば「千葉県」または「舞浜」の回答で十分であるが，質問者が東京ディズニーランドへ出かけることを考えている場合には,「千葉県浦安市舞浜１の１」のような詳細な回答や,「ＪＲ京葉線舞浜駅前」のような気の利いた回答が必要である．QAC では，質問者の意図や，質問時のシチュエーションの設定が存在しないため，前述したような問題が起きてしまう．

また,「国民栄誉賞を受賞した映画監督は誰ですか。」という質問に対する解答としては,「黒沢明」も「黒沢」も受け入れられるが,「大久保利通の息子は誰ですか。」という質問に対する解答としては,「牧野伸顕」のみが受け入れられ,「牧野」は受け入れられないというケースがある．両者の違いを認識するためには，知名度の認識というような常識的判断が必要であり，現状の質

問応答システムでは対応が難しい．

関連して，さらに次のような問題も存在する．例えば，ある質問の解答が「愛知県，岐阜県，三重県」であったとする．この場合，三つの解答が存在するといえるが,「東海三県」と回答することも可能であり，この場合解答は一つになってしまう．つまり,「東海三県」は不正解ではないが，より詳細度の高い三つの県名と同じ扱いにしてもよいかという議論が生じるのである．QAC では,「東海三県」と答えた場合は，詳細度が低いとして，三つの県名のうち一つだけ正解した (1/3 の得点) と見なすことにしているが，詳細度の高い解答が 10 個存在する場合や，詳細度の低い解答が複数存在した場合への対応には議論の余地がある．

上記のような問題については，最終的には人間の判断が必要である．TREC では，assessor の判断によって対応しているが [1]，QAC ではこのような絶対的判断基準は存在せず，作業者とオーガナイザ間で相談して判断している．判断が難しいものについては，データセットの第 14, 15 カラムを用いて曖昧性を含むことを明示し，“○・△・×” に三分類している．

5.4 質問作成方法

QAC1 では，(1) 作業者が自由に質問を作成した後，対象の新聞記事から解答を探す場合と，(2) 新聞記事を読んで，目に止まったトピックに基づいて質問を作成する場合の二通りの質問作成方法を試した．(1) の場合，作業者は，質問作成においてほとんど制約を受けることはないが，新聞記事中に解答が含まれる保証がない．(2) の場合，解答の存在は保証されるが，作業者は，新聞記事を読んだ後で質問を作成するために，質問文の表現が，新聞記事中の記述を修正したような記述になる傾向があった．

しかしながら，(1) の場合，ある程度作成される質問文が蓄積されると，新たな質問文を作成することが困難となり，分野依存性の強い質問が増えたり，重複する内容の質問が増える傾向にあった．このことから，(2) の作成方法は，有限のコーパス中から作成できる質問文は限られており，(1) に比べて作業者への負担も大きく，効率が悪いといえる．

そこで，QAC2 では，(2) の方法のみで質問作成を行った．ただし，作成者の知識の他に，複数の新聞記事や文献の参照は認めた．また，質問の重複を防ぐために，各作業者に異なる担当分野を提示した．

5.5 自動採点

自動採点ツールは，回答データセットと，質問解答データセットを読み込み，両者を詳細に比較することで採点処理を行う．それゆえ，精度の高い採点処理実現のためには，解答データ中の解答網羅性を，出来る限り高く保つ必要がある．しかしながら，対象テキスト集合が新聞記事 10 年分に増加したり，Web テキストを扱うとなった場合，同様の方法で事前に高い解答網羅性を維持することは困難となる．このような，より大規模な知識源を用いた評価タスクも，情報検索分野では既に行われており，QAC でも現実的な問題として捉えておくべきであろう．

6 おわりに

本論文では，QAC テストコレクション構築について報告した．質問応答データセットの仕様と，その作成手順について説明し，テストコレクション構築を通して明らかになった，主な問題点について考察した．特に，質問の曖昧性や，解答の網羅性，粒度などの問題を，主に取り上げたが，これらは，データセットの作成方法や，採点方法などにも波及するため，慎重に分析を続けていく必要がある．

今後も，上記のような知見を意識しながら，データセットの精緻化，自動採点ツールの高性能化などをすすめ，テストコレクションの充実を図る予定である．

参考文献

[1] Ellen M. Voorhees and Dawn M. Tice. The TREC-8 Question Answering Track. In *Proc. of LREC 2000*, 2000.

[2] Ellen M. Voorhees. The Evaluation of Question Answering Systems: Lessons Learned from the TREC QA Track. In *Proc. of the LREC 2002 Workshop on Question Answering — Strategy and Resources*, pp. 1–4, 2002.

[3] 村田真樹, 内山将夫, 井佐原均. 類似度に基づく推論を用いた質問応答システム. NL135-24, 情処研報, 2001.

[4] 佐々木裕, 磯崎秀樹, 平博順, 広田啓一, 賀沢秀人, 平尾努, 中島浩之, 加藤恒昭. 質問応答システムの比較と評価. NLC2000-24, 信学技報, 2000.

[5] 神門典子, 安達淳. 評価ワークショップによるテキスト処理–第 3 回 NTCIR ワークショップを例として–. 人工知能学会誌, Vol. 17, No. 3, pp. 312–319, 2002.

[6] Jun'ichi Fukumoto, Tsuneaki Kato, and Fumito Masui. Question Answering Challenge (QAC1) An Evaluation of QA Tasks at the NTCIR Workshop 3. In *Papers from the 2003 AAAI Spring Symposium "New Directions in Question Answering"*, pp. 1–4, 2003.

[7] 福島孝博, 奥村学, 加藤恒昭. テキスト処理研究の動向–情報検索・自動要約・質問応答における評価ワークショップの重要性–. 人工知能学会誌, Vol. 17, No. 3, pp. 301–305, 2002.

[8] 桝井文人, 関根聡. TIPSTER Text Program Phase III 24th Month Workshop 参加報告. NLC98-57, 信学技報, 1999.

[9] 福本淳一, 加藤恒昭, 桝井文人. QAC タスク ホームページ. http://www.nlp.cs.ritsumei.ac.jp/qac/, 2004.

[10] 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 回年次大会発表論文集, 2004.