特集　自然言語による情報アクセス技術

4 質問応答技術

ー大量のデータをもとに任意の質問に答えるー

福本 淳一
立命館大学 情報理工学部
fukumoto@media.ritsumei.ac.jp

桝井 文人
三重大学
masui@ai.info.mie-u.ac.jp

質問応答技術は，自然言語で記述された任意の質問に対して大量の組織化されていない情報から答えを得る技術である．たとえば，「恐竜が絶滅したのはいつですか」や「ワールドカップで優勝したのはどこですか」のような質問に対して，数年分の新聞記事から答えとなる情報を取り出す技術である．質問応答に関する研究としては，古くはSHRDLUのように積み木の世界に対して質問を与えることで積み木を操作するものやLUNARのように月の岩石や土壌の組成に関する問合せを行うものがあった．これらを代表とする対話処理の多くは，特定の分野に関してあらかじめ記述された知識を対象とした，自然言語による問合せを実現したものであった．

質問応答技術は，本特集「自然言語による情報アクセス技術」においても紹介されている情報検索，情報抽出，自動要約などの自然言語処理の研究と関連がある．まず，質問応答システムが与えられた質問に答えるためには，質問文中の情報を利用して大量のテキストからある程度の情報まで絞り込む必要がある．大量のテキストから情報の絞り込みを行うためには情報検索技術が不可欠である．さらに，絞り込まれた範囲から答えとなる情報を取り出す技術として情報抽出技術があり，Message Understanding Conference（MUC）やInformation Retrieval and Extraction Exercise（IREX）などの評価プロジェクト☆1を中心に研究が進められてきた．従来の情報抽出において必要な情報を得るためには，あらかじめ用意された抽出用のテンプレートを用いる必要があり，分野を絞らない一般的な分野での任意の要素の抽出は困難なものであった．質問応答では，実際にどのような質問が行われるのかが分かっていないため，答えとなる要素がどのようなものであるかが不明である．そこで，質問応答において答えとなる要素をテキストから抽出するために一般分野での情報抽出技術が重要な技術となると考えられる．

質問応答においては，得られた答えが正しいものであるという証拠を示すため，答えが得られた根拠となる情報を提示することも重要である．根拠となる情報としては，答えが得られたテキスト全体を示す方法やテキスト中の一部を提示する方法が考えられる．さらに，根拠となる情報がテキスト中の広範囲に分散している場合には，それらを要約して示すことも考えられる．したがって，文書要約技術も質問応答には重要な技術であると考えられる．また，「なぜ」や「どのようにして」などの

☆1 MUC: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/, IREX: http://cs.nyu.edu/cs/projects/proteus/irex/ の詳細については本特集の「情報抽出」を参照していただきたい．

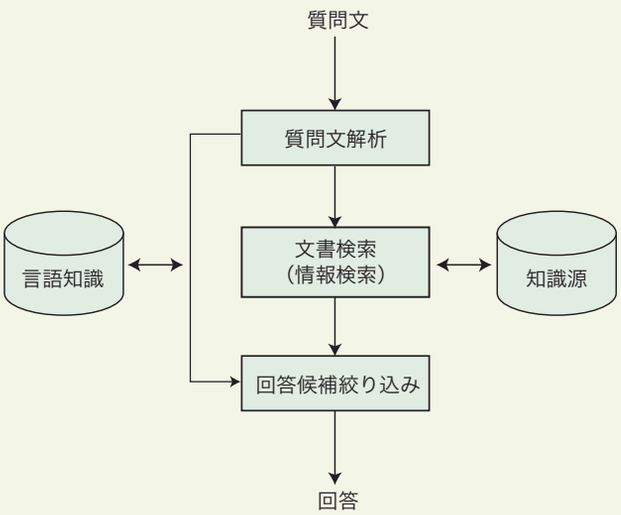

図 -1　一般的な質問応答処理の流れ

質問に対する答えを求めている場合など，質問の種類によっては答えをテキスト中からそのまま抽出するだけでなく，要約したかたちで答える場合も考えられる．この点からも文書要約技術は重要な技術である．

質問応答においては，ユーザによって与えられた質問に対して，システム側が返した答えでは不十分な場合や質問があいまいで適切な答えを返すことができないといった場合がある．このように，質問応答システムを実際に利用するためには，ユーザのあいまいな質問内容やシステム側が不十分な回答を返した場合に，質問応答システムに対していくつかの補足的な情報を与える必要がある．これをユーザとシステムとのインタラクションによって実現することも考えられる．また，システム自身がユーザに問い合わせて質問のあいまいな部分を解消することも考えられる．さらに，ユーザは関連する一連の項目について質問したい場合もあり，得られた答えに対して，その詳細情報や関連情報を追加的に質問することも考えられる．そのような一連の質問応答に対応するためには，対話処理技術が必要となる．

質問応答処理の流れ

質問応答技術は，大量の文書を対象とし任意の質問に答えるものであると述べてきた．そこで，そのような質問応答システムの質問応答処理の一般的な構成と処理の流れを概観しながら，質問応答処理に必要な要素技術について解説する．図 -1 に，一般的な質問応答処理の流れを示す．

まず，質問文解析では，質問表現のパターンから回答が何を尋ねているのかの属性を予測するとともに，特徴的なキーワードを抽出する．次に，文書検索により前記キーワード集合に適合するテキスト集合を知識源であるコーパスから抽出する．回答絞り込み処理は，絞り込んだテキストから回答属性に合致する回答候補を見つけ出す操作である．そして，得られた回答候補に対してキーワードとの単語間距離などの尺度を用いてスコアを計算し，上位のスコアのものを回答として選択する．

以下では，質問応答技術の各段階で用いられている要素技術について解説する．

質問文解析と固有表現抽出

質問文解析によって，質問文が回答として求める情報のタイプとして，たとえば，会社名，大臣の名前，商品名や各種の数値情報などを判断し，それらの情報を対象文書から探し出す必要がある．このような情報を探し出すための技術として固有表現抽出が必要である．日本語の固有表現抽出は，IREXの開催以降研究が盛んになり，人名，組織名，場所名，日時，価格などの情報抽出について非常に精度の高いレベルに達してきている．しかしながら，質問応答においては，IREXなどで設定された上記のような固有表現情報のみでは，任意の質問に対する回答を抽出するためには不十分であり，多くの質問応答システムが固有表現情報のタイプの詳細化や拡張を行ってきている．拡張されたものとしては，IREXなどの8種類程度のものから70種類や200種類に拡張されたものがある．さらに，質問文のタイプとして詳細化されたどのような固有表現情報を回答として求めているのかについても認識する必要がある．固有表現抽出の詳しい解説は，本特集の「情報抽出」を参照していただきたい．各種の質問応答システムにおいて用いられている固有表現抽出処理はそれぞれ独自のものが用いられているが，フリーのツールとして，筆者らはパターン駆動型の固有表現抽出ツールNExT[1)]を公開している．

情報検索

質問応答システムが対象としているのは，数年分の新聞記事やWebのような巨大なテキスト集合である．したがって，そのような大量のテキストすべてから回答を探し出すことはなく，情報検索によってある程度絞られた文書に対して回答を探すことになっている．検索結果の文書中に回答が含まれていなければ回答抽出の精度がよくても無駄になってしまうため，情報検索の精度は非常に重要になってくる．質問応答システムの多くは，与えられた質問文の表現からキーワードを抽出し，それらのキーワードを含む文書集合を汎用の全文検索システムや，独自の文書検索エンジンを利用して取り出してい

る．これらのシステムでは，TF*IDF☆2のように単語や文字列の頻度情報に基づくモデル，文書中の単語を利用したベクトルの余弦尺度による文書の類似度（図-2に示すように文書中の単語の出現頻度を要素とする文書ベクトル間で計算される），テキスト中の単語の共起情報を利用したモデルなどさまざまな技術が用いられている．検索システムとしては，Namazu（http://www.namazu.org/）などフリーで公開されているものもある．

回答候補の絞り込み

情報検索において検索結果として与えられる文書に対して固有表現抽出によって得られる回答候補を絞り込む必要がある．絞り込みのために用いられる情報としては，質問文中のキーワードと回答候補との単語間距離や頻度が多く用いられている．これは，回答候補のうち本当の回答は統計的に質問文中のキーワードと近い位置にあるという傾向があるためである．また，相互情報量☆3などの意味的な距離情報や，回答候補を含む文と質問文の構文解析を行い，構文的な類似性を利用した手法を用いているものもある．また，絞り込みの際に，たとえば「自動車」を「車」や「カー」へ，「日本航空」を「JAL」へ，「神戸製鋼」を「神鋼」，「Question Answering」を「QA」とする省略表現の扱いも有効な手法である．新聞記事では，「日本」は「我が国」と言い換えて扱う場合や，「最も」のような程度を表す表現は「最良」や「一番」という表現に置き換えて処理することも必要である．ほかにも，「言い換え技術」を使った質問応答システムが研究されている[2)]．質問文の表現を文節や構造レベルで言い換え，たとえば，「ベルリンの壁が崩壊したのはいつですか」という質問文と「市民はベルリンの壁を破壊した」という候補文が相互に言い換え可能であることが分かれば，疑問詞に相当する部分を回答候補として抽出することで質問応答が可能である．

質問応答技術評価プロジェクト

質問応答技術については，いくつかの評価プロジェクトが存在し，同技術の研究を推進している．ここでは，代表的な質問応答技術評価プロジェクトとして，TREC QA TrackとNTCIR QACを紹介する．

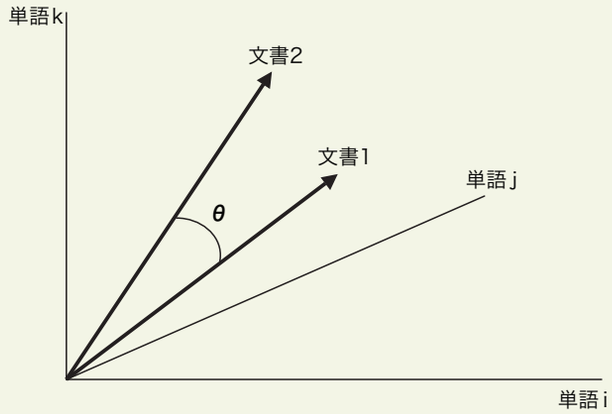

図-2　ベクトルの余弦尺度による文書の類似度

TREC QA Track

まず，質問応答の先駆的な立場として，Text Retrieval Conference（TREC）のQuestion Answering Track（http://trec.nist.gov/）が挙げられる．TRECは，米国National Institute of Standard Technology（NIST）とThe Defense Advanced Research Projects Agency（DARPA）の後援により，1990年代から現在まで継続されている情報検索技術に関する評価型ワークショップである．1999年，TREC-8において，文書検索から情報検索への技術的発展の目指す研究支援および，文書検索分野と情報検索分野の研究者にアピールすることを目的としてQA Trackが設けられた．このときのタスク設定は以下のようなものであった．まず，タスク参加者には，文書コレクションと質問テストセットが与えられる．質問文の例を以下に示す．

- In what year did Joe DiMaggio compile his 56-game hitting streak?
- Name a film in which Jude Law acted.

例に示すように質問文はある事実に関する情報を求める単純な英語の質問（fact-based, short-answer）である．各質問は，与えられた文書コレクション中に最低1つの明確な回答を持つことが保証される．

参加者は，各質問に対して解答と思われる上位5位リスト（document-id, answer-string）を返す．解答文字列は50バイトまたは250バイトに制限され，かつ，対応

☆2　文書中の単語頻度TFと文書集合中のある単語の文書頻度DFの比を計算することにより，ある特定の文書にのみ多く出現する単語に重みを与えることで文書中の重要度を計算する手法．
☆3　相互情報量：I(x, y) は，単語xと単語yとが共起する確率 P(x, y) と単語 x, y それぞれが個別に生起する確率 P(x), P(y) との比として計算される．I(x, y) = log2P(x, y)/P(x)P(y) .

する文書から抽出されたものか，文書情報に基づいて自動生成されたものでなければならない．

提出された回答データセットは，アセッサと呼ばれる専門の評価者によって正否が判断される．文字列に対する判断が与えられると，提出回答のスコアが計算される．個々の質問には，回答中の最上位の正解順位の逆数（Reciprocal of the Rank），もし正解がなければ0がスコアとして与えられ，提出回答の総合スコアとして，個々の質問における順位の逆数の平均Mean Reciprocal Rank（MRR）が与えられる．

その後，TREC2001では回答が複数ある質問に答えるlist task，一連の関連質問に答えるcontext taskが設定された，さらに，TREC2003では事物の定義を答えるdefinition taskが追加され，質問応答技術に対する興味範囲が広がってきている．

NTCIR QAC Task

2001年，第3回NII-Test Collection for IR（NTCIR）ワークショップ（http://research.nii.ac.jp/~ntcadm/index-ja.html）のタスクとして，QAC1（Question Answering Challenge 1）（http://www.nlp.is.ritsumei.ac.jp/qac/）[3]がスタートした．第1回目の参加者としては，情報抽出の研究者が多く，2003年の第2回目のQAC2では，関連分野の研究者の参加も増えてきた．これは，TRECでは情報検索技術研究の延長線上に質問応答を位置づけていたが，QACではオープンドメインの情報抽出の応用として進められた経緯があったためである．

QACでは，タスク参加者に新聞記事コーパスと質問テストセットが与えられる．質問は，「日本の総理大臣は誰ですか」や「ブラジルはサッカーW杯で何度優勝していますか」のように，何らかの名称もしくは時間・数量表現として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，種やカテゴリーの名称等を求めるものである．タスクとしては，評価や質問の与え方により，3つのサブタスクが設定された．

サブタスク1は，TREC 2001のmain task（exact answer）とほぼ同等のタスクであり，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位までとその根拠記事番号を回答する．評価は，最も高順位にある正解文字列を対象とし，各質問に対してRR（Reciprocal of the Rank）による得点付けを行い，その平均であるMRR（Mean Reciprocal Rank）をもって総合評価とする．このとき，誤った回答に対してペナルティは与えられない．質問例と正解例を以下に示す．ただし，根拠情報となる記事番号は省略している．

> 質問例：「サッカーワールドカップ第16回大会の開催期間は何日間ですか」
> 正解例：「33日間」

以上の質問文に対して，たとえば，

1位「4日間」　2位「33日間」　3位「2日間」
4位「30日間」　5位「33日間」

という回答をシステムが返してきたとする．このとき，2位と5位の答えが正解のため，より上位の2位の答えがこの質問の正解となり，2の逆数の1/2がこの質問の得点となる．

サブタスク2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と根拠記事番号の組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すべてを対象とし，適合率P（システムの出力の正解の割合），再現率R（正解のうちシステムがどれだけ答えたのかの割合），および，F-value（＝2PR/(P+R)）によって行い，その平均MF（Mean F-value）で評価する．したがって，回答順位は考慮されない．また，新聞記事などの知識源において正解の存在しない質問については，回答をしなかった場合にのみ正解とする．質問例と正解例を以下に示す．

> 質問例：「木星の衛星にはどんな名前がついていますか」
> 正解例：「イオ，カリスト，ガニメデ，エウロパ，アマルテア」

以上の質問文に対して，たとえば，

「ガリレオ，ビーナス，エウロパ，アマルテア」

という回答をシステムが返してきたとする（サブタスク2の場合には順位はつかない）．このとき，「エウロパ」「アマルテア」の2つのみが正解で後は誤りであるため適合率は2/4となる．また，正解5つのうち2つのみを回答しているため再現率は2/5となる．したがって，この質問の評価値であるF-valueは0.44となる．

サブタスク3は，連続して入力されたと想定される複数の質問（series of question）に回答する．後に続く質問（枝問）には，それ以前の質問の一部もしくは回答を参照する表現や省略を含む．QAC1では，元となる質問（幹問）1問につき枝問が1問出題されたが，QAC2では複数の枝問（5から10問程度）が出題され，1つのトピックに関連した質問群（収集型）と，質問の途中でトピックが変化する質問群（ブラウジング型）が出題された．ただし，与えられた質問文がどちらの型であるのかについては示されず，システムは自動的にそれらの型を判別

して回答する必要がある．収集型の質問例を以下に示す．

幹問A　「小沢征爾さんはいつ生まれましたか」
枝問A1「どこの生まれですか」
枝問A2「大学はどこを卒業しましたか」

この質問群においては，幹問Aにおいてトピックとなった「小沢征爾さん」に関連する質問を連続する枝問A1，A2において行っている．また，ブラウジング型の質問群を以下に示す．

幹問B　「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」
枝問B1「何年に造られたものですか」
枝問B2「そこには何人の記念碑が飾られていますか」
枝問B3「1999年に飾られたのは誰ですか」
枝問B4「彼が新婚旅行で来日したのは何年ですか」

この質問群においては，幹問Bにおいてトピックとなっている「球場」について，枝問B1，B2において関連する質問を行っている．そして，枝問B3において「記念碑に飾られた人」にトピックが移り，枝問B4において新たなトピックに関する質問が行われている．

参加システムの性能であるが，サブタスク2とサブタスク3についてはもう一歩というところであるが，サブタスク1は結構使えるレベルだと思われる．QAC2サブタスク1における最高スコアは約0.6であり，出題200問中7割程度の質問に正解できた．また，各システムの得意とする質問にはばらつきがあり，参加システムが力を合わせて回答する上手い仕組みを考えれば，9割程度の質問に正解できるという興味深い結果が得られている．

評価プロジェクトの主要な目的の1つに，テストコレクションの構築がある．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．テストコレクションは，質問とそれに対応する正解群によってな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．自動採点結果の例を**図-3**に示す．

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，同規模のものが公開される予定である．

サブタスク1結果出力

平均スコア：　0.311			
質問文	回答	出力	正解数
197	385	902	108
再現率	適合率	F-value	MRR
0.281	0.120	（0.168）	0.311

＊実際の出力は英語となっている．

図-3　自動採点結果の例

質問応答技術の今後

以上述べてきたように質問応答に関する評価ワークショップを通じて質問応答についてはfactoid型の質問を中心に研究がなされてきており，答えのうちいずれか1つが5つの順位付けされた回答候補の中でより上位にあればよいというスタイルをとっている．しかしながら，実際に質問応答システムを使うことを考えた場合，答えが複数存在したり，答えが存在しない場合も考えられる．このような問題を質問応答の評価として扱ったものとしてはリストタスクがある．リストタスクは，質問の答えとして正しいと考えられるものすべてを与えるものであり，QACにおいてはサブタスク2として設定されている．しかしながら，参加者の多くは，factoid型の質問応答が中心となっており，本当に正しい答えをどのようにして判断，抽出すればよいかについてはこれからの研究を待つ段階である．また，リストタスクの評価方法としては，TREC，QACどちらもF値を評価尺度として用いているが，多くの答えを持つ質問と少ない答えの質問が等価に評価されてしまい，システムの性能を図る公平な評価尺度としてはまだ問題があると考えられる．

質問に対する回答の範囲としては，人名や会社名などの名詞語句で表現されるものを対象に研究が行われてきた．それを超える回答に対する質問範囲としては，TREC2003のdefinition taskがある．definition taskとは，たとえば「コリン・パウエルとは誰ですか（Who is Colin Powell?）」のようにある事柄の定義，説明を質問として与えるものである．TRECでは質問文の作成を検索エンジンの検索ログから選択したものを用いていたが，検索ログにはこのような質問が多くあり，その点からこのタスクが設定されている．また，実際に小さい子供がする質問にもこのようなものが多くあり，説明に困った方も多いのではないかと思う．質問の対象として名詞語句を超えるものとして考えられるものとしては，「なぜ～」のWhy型の質問がある．Whyを扱った研究としては，森本らによるものがある[4]．彼らはWhy型の質問に答えるため，Rhetorical Structure Theory（RST）の関係を用い，Why型質問と回答文の間にどのような関係が成り

【採点例（サブタスク 1 の場合）】

【質問 1】" サッカーワールドカップ第 16 回大会の開催期間は何日間ですか "
【質問 2】" 木星の衛星にはどんな名前がついていますか "
【質問 3】" 日本の国鳥はなんですか "

【回答 1】"33 日間 ", JA-980715137, "4 日間 ", JA-990310146,T,12, "2 日間 ", JA-991104274, "11 日間 ", JA-990717167, "30 日間 ", JA-981202153,
【回答 2】" ガリレオ ", JA-980703323, " ビーナス ", JA-991010157, " エウロパ ", JA-980303354, " アマルテア ", JY-19980916J1TYEUG0400050,
【回答 3】" 白頭ワシ ", JA-990703218, " ハクトウワシ ", JA-990729153, " シンボル ", JA-990703218, 生息数 , JA-990729153, " キジ ", JA-990207220,

【採点】
回答 1 の得点（1 位で正解）：RR = 1
回答 2 の得点（3 位で正解）: RR = 1/3 = 0.333
回答 3 の得点（5 位で正解）：RR = 1/5 = 0.2
全体の得点 :　MRR = (1 + 1/3 + 1/5) ÷ 3 = 0.511

立つのかを分析し，Evidence，Motivation，non-Volitional Cause，Purpose，Means の関係を持つ場合に質問文と回答文の関係が成り立つことを示している．そして，新聞記事を対象とした 23 問の Why 型質問を用いた実験から，これらの関係を抽出するための接続表現などの表層的手掛かりを用いて Why 型質問に対して回答が得られることを示している．このように回答として名詞語句を超えるものを対象とした研究はまだ始まったところであり，回答抽出方法とともに，一文やそれを超えるものを回答対象とした場合の評価方法も含めてまだ解決しなければならない点も多く，今後の研究が待たれるところである．

最初の節においても述べたが，質問応答においては，ユーザによって与えられた質問に対して，システム側が返した答えでは不十分な場合や質問が曖昧で適切な答えを返すことができない場合がある．このため，曖昧性を解消するためにユーザとシステムのインタラクションによって必要な情報を補う必要がある．このようなインタラクションを仮想的に実現しようとしたものとして，QAC2 のサブタスク 3 として context task が設定されている（タスクの詳細については前章を参照）．ここでは，質問応答をタスクとして設定するという制約から質問応答をバッチ処理で行うため，与える質問をシステム側の回答に応じて切り替えるということができない．しかしながら，質問応答システムを実際に運用する際には，このような制約はなく，評価プロジェクトを通じて培ってきた技術を実際の応用システムに用いることでシステムとのインタラクションを実現したシステムが今後実現されていくことになると思われる．また，現在質問応答システムの多くは新聞記事を対象としているが，知識源として百科事典を対象としたものや[5]（これは実際に商品化まで行われている），Web の情報を対象とした研究[6]もある．このように新聞記事以外の他の知識源を対象としたシステムも今後実現されていくことになると考えられる．

参考文献

1）桝井文人，鈴木伸哉，福本淳一：テキスト処理のための固有表現抽出ツール NExT の開発，言語処理学会第 8 回年次大会，pp.176-179，言語処理学会 (2002)．(NExT 入手先 URL http://www.ai.info.mie-u.ac.jp/~next/next.html)
2）高橋哲朗，縄田浩三，乾健太郎，松本裕治：質問応答における構文的照合と言い換えの効果，言語処理学会第 9 回年次大会，pp.617-620，言語処理学会 (2003)．
3）Fukumoto, J., Kato, T. and Masui, F. : Question Answering Challenge (QAC-1) An Evaluation of Question Answering Tasks at the NTCIR Workshop 3, Proceedings of AAAI Spring Symposium on New Directions in Question Answering, AAAI Press, pp.122-133 (2003)．
4）森本格行，福本淳一：Why 型質問に対する回答抽出，言語処理学会第 10 回年次大会，pp.293-296，言語処理学会 (2004)．
5）関根　聡：百科事典を対象とした質問応答システムの開発，言語処理学会第 9 回年次大会，pp.637-640，言語処理学会 (2003)．
6）山田一郎，柴田正啓，金　淵培：Web を情報源とした Q&A システムの検討，言語処理学会第 9 回年次大会，pp.633-636，言語処理学会 (2003)．

（平成 16 年 5 月 10 日受付）