質問応答システム
～情報検索と情報抽出の頂点へ～

池野　篤司

Google*1) に代表されるインターネット検索エンジンの普及により情報検索はなじみ深い技術となった。しかし，情報検索は情報を含む文書を提示してくれるだけで，必要としている情報を与えてくれるわけではない。必要とする情報を入手するためには，検索結果の文書を自分で読まなければならない。情報抽出技術は，文書中の重要な情報を取り出す技術であり，検索結果からの情報の把握を容易にするのに役立つ。

質問応答システムは，ユーザが普段使っている言葉をそのまま入力するだけで必要とする情報を過不足なく提示することを目的として研究されており，情報検索と情報抽出の技術を組み合わせたものである。さらにこのシステムは，機械との対話インタフェースへの第一歩としても注目を集めている。

本稿では，まず我々が取り組んでいる質問応答の基本システム[1)] の概要と構成について解説する。次に，本システムの性能評価のために参加した国立情報学研究所主催のワークショップNTCIR（NII-NACSIS Test Collection for IR Systems）の概要と，その評価結果について述べる。さらに，Webページを対象に試作したインターネット質問応答システム，および実用化の事例についても述べる。

質問応答のアルゴリズム

質問応答システムは，「○○の発売日はいつですか」のような質問文を入力として，「××日（です）」のような回答を返すソフトウェアとして実現される。我々が開発したシステムでは，現在，固有名詞（人名・地名）や日時・数量など固有表現と呼ばれるものを中心とした30種類程度の表現による回答を実現している。

処理はおおよそ以下の順に行われる。

（1）質問解析

入力された質問文が，何に関する質問か，何を回答すればよいかを解析する。

（2）情報検索

質問解析の結果を利用して，情報源となる文書群から回答が含まれている可能性の高い文書を選び出す。

（3）情報抽出

情報検索の結果選ばれた文書中から，求められている回答の候補となる語（表現）を抜き出す。

（4）回答選択

情報抽出により抜き出された複数の回答候補から，統計量を考慮して最適なものを選択し回答とする。

基本システム構成

図1に質問応答の基本システムの構成を示す。システムは，前述のアルゴリズムに対応した，大きく4つのモジュールから構成される。各モジュールの詳細は以下の通りである。

（1）質問解析部

質問文を解析して，「いつ」「どんな」などの疑問詞とその周囲の語を手がかりに，どのような固有表現を回答にすればよいか（回答タイプ）を判定する。

たとえば，「『○○』を発売した会社はどこですか」の場合は，「会社」と「どこ」の2語から，組織名が回答タイプであると判定する。

同時に，質問文中の疑問詞を除いた名詞・形容詞を検索語として選択し，その検索語を次工程の情報検索部に出力する。このとき，「～の人数は何人ですか」といった質問文における「人数」のように，削除しても質問文の内容に影響を与えない語は，検索効率向上のため削除語彙リストを設けて除去する。

（2）情報検索部

情報検索部は，質問解析部から渡される検索語から検索式を作成する検索式構成部と，その式により実際に文書検索を行う文書検索部とから構成される。出力は「質

＊1）Google は Google Inc. の登録商標です。

問文に関係の深い文書」すなわち「回答が存在する可能性が高い文書」の集合として情報抽出部に入力される。

検索式は全ての検索語のANDを基本とする。各検索語には（検索対象文書群の）統計量から計算される重要度を付与することもある。

「じゃがいも」「馬鈴薯」「メークイン」のように，同一の物事を指す場合にさまざまな表現が用いられることも多いので，検索語そのものでは検索がうまくいかないことがある。この問題には類義語辞書（シソーラス）を利用して表現の差異を吸収することにより対処できる。

文書検索部は独立して動作するため，Namazu[2] やGETA[3] などの一般に公開されている文書検索ソフトウェアを利用することも可能である。

(3) 情報抽出部

情報抽出にはさまざまな処理モジュールが想定できるが，ここで抽出される語（表現）のみが質問応答システムの回答となれるので，固有表現（Named Entity: NE）抽出が最も基本的な要素となる。

固有表現抽出部では，情報検索部から入力された文書から文書中の固有表現などを抽出して回答選択部に出力する。

我々は従来から固有表現抽出の研究に注力している[4]

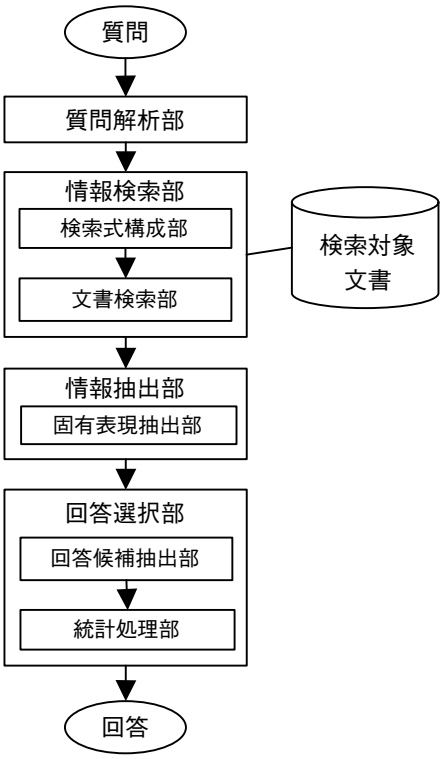

図1　システム構成

表1　固有表現の例

固有表現	例
人名	ブッシュ，山田
地名	ローマ，東京
組織名	○○株式会社
サブ組織名	工学部，営業部
日時	9月8日，平成12年
数量表現	25メートル、100グラム

が，現在のシステムでは，

- 一つの文字列に複数の固有表現を割り当てることを可能にする
- 抽出結果を原文とは別に保持する

などの変更により，システムの拡張性およびメンテナンス性を向上させたモジュールを用いている。**表1**に固有表現の例を示す。

(4) 回答選択部

回答選択部は，情報抽出部から入力された固有表現のうち，回答候補となりうるものを選ぶ回答候補抽出部と，それら回答候補の統計情報を利用して最適な候補を決定する統計処理部とから構成される。

1つの文書中にはさまざまな話題が書かれている場合もあるので，情報抽出部で抜き出された表現がすべて回答候補としてふさわしいわけではない。回答候補抽出部では，質問文に関する話題が述べられている範囲（＝検索語を多く含む範囲）に存在する表現のみを回答候補とする。

統計処理部では，文書中での検索語との距離が近い回答候補が高い値を持つように各々の回答候補にスコアを与える。さらに，同一の表現（文字列）の候補を統合する処理を行う場合には，出現数・全文書における分布度合いを加味した重み付けを行う。

それらの値を元に回答候補を順位付けした結果，上位の候補がシステム全体の出力として提示される。

NTCIR ワークショップ

NTCIR ワークショップとは，国立情報学研究所（NII）が主催するテキスト処理技術の研究発展を目的とした評価会議で，1998年に開始されて以来ほぼ1年半ごとに開催されている[5]。このワークショップは，実験用データセットの拡充やテキスト処理技術の評価の枠組みづくりを目的として，参加者には処理対象文書データが配布され，一定期間の後に設定された課題に対する結果を参加

者が提出し，その結果を評価している。

NTCIRのタスクは，まず言語横断検索から始まり，第2回では文書要約が加わり，2002年12月に報告会が行われた第3回（NTCIR-3）ではさらに特許検索，WEB検索と質問応答（Question & Answering Challenge: QAC）のタスク[6] が加わった。上記5つのタスクが2004年6月に報告会が予定されている第4回（NTCIR-4）でも継続実施されている。

我々はNTCIR-3の質問応答タスク発足時より参加している[7]。以下にタスク課題と我々の取り組みについて述べる。

（1）NTCIR-3 QAC （QAC-1）

自然言語形式の質問に対して，「何らかの名称もしくは値」および「その根拠となった文書（部分）」を回答するという課題である。検索対象文書は毎日新聞から提供されている1998年と1999年の記事（総数約24万）であり，課題は以下の3種類設定された。

① 課題1

回答候補を，優先順位をつけて5つ返す。最初に現れた正解のReciprocal Rank（1／順位，以下RRと略す）により評価する。

我々は，この課題に対しては，前述の回答選択部において同一表現の回答候補のスコアを累積する計算方法を用いた。出現頻度が高い回答候補は正解である可能性が高い，という仮定に基づいたためである。結果を見る限り，この計算方法はRRへの直接的な効果が少なかったため，システムの評価を上げることはできなかったが，スコア累積が再現率*2) の向上に寄与することがわかった。

② 課題2

回答と考えられるものを過不足なく列挙したリストを1つ返す課題である。評価は再現率と精度*3) を組み合わせたF値*4) により行われる。

我々は，課題1と異なり，同一表現の回答候補が存在する場合には最高スコアの候補を出力する方法をとった。誤答が減点対象となるので，出現頻度よりも個々の候補のスコア（検索語との距離の近さ）を重視したためである。実験の結果，この方法は誤った回答候補を除去する効果があり，精度の向上に寄与することが確認できた。

③ 課題3

2つの連続した質問（本問と枝問）にタスク2の形式で回答する。枝問の設定は，対話を意識した難易度の高い課題である。たとえば，本問として「20XX年○月に合併する会社はどこですか」という質問が与えられたとすると，枝問には「新会社の名前は何ですか」というように，本問とその回答において既知となっているはずの事項を利用しなければ答えられない質問が用意される。

我々は枝問への回答のために2つのアルゴリズムを追加した。1つは質問文補完に関するものである。枝問においては，本問やその回答で述べられた事項は省略される。補われる語は，

- 枝問に指示詞（「それ」「その～」）があるときは，本問あるいはその回答から，参照される語を推定する
- 指示詞も省略されているときは，本問の回答の全てを枝問の質問文に追加する

ことによって決定することとした。

もう1つは検索対象文書の絞込みに関するものである。枝問の回答は，（i）本問の回答が存在した文書，（ii）本問の回答候補が存在した全ての文書，の順序で探索し，それでも回答候補が現れない場合にのみ全文書を対象とした処理に移行する方法を採用した。

上記アルゴリズムを取り入れた方式は，正解データから判断する限りではおおむね良好な結果を導いたと言える。

④ 全体評価

上流工程の質問解析部・情報検索部のチューニングをあまり行わないシンプルなシステムで参加した結果，各課題とも評価スコアには改善の余地が残ったが，細部の計算方法・アルゴリズムの有効性は確認することができた。一方で，速度に関しては，実用に耐えうるもの許容範囲内の評価結果が得られた。

全体評価の結果，我々の開発したシステムは，スコア，速度ともさらなるチューニングは必要であるが，実用レベルまで向上させることが可能であるといえる。

（2）NTCIR-4 QAC （QAC-2）

2度目のQACとなった今回は，検索対象文書に読売新聞の1998年，1999年の記事が加えられた。検索対象の規模が倍になり，かつ重複した情報が加えられたことになるので，その扱いにより結果が変動する。

課題1と課題2は前回と同一であるが，課題3において本問に対する枝問が複数（7～8問）設定された。

前回の結果を踏まえて質問解析部と情報検索部を改善したため全体にレベルは上がっているが，本稿執筆時点で正解データが公開されていないので，評価結果の詳細は稿を改めて報告することとしたい。

インターネット質問応答システム

QACを含めて従来の質問応答システムは新聞記事を対象としているが，より実用化が望まれる対象文書として我々はインターネットに着目しインターネット質問応答システムを開発している。Webページは今や多くの人の

*2）再現率（recall）とは正解の取りこぼしがどれだけ少ないかを表す指標であり，次の式で表される。再現率R＝回答のうち正解の数 / 正解の総数
*3）精度（precision）とは回答に誤りがどれだけ少ないかを表す指標であり，次の式で表される。精度P＝回答のうち正解の数 / 回答の総数

情報源であるが，情報量が膨大であること，質が玉石混交であることなどから，誰にでも容易に情報を探せるという状態にはない。質問応答システムなら，コンピュータに慣れない人がインターネット上の大量の情報の中から必要なものを選び出すのに適したインタフェースになりうると考えている。

ところが，新聞記事は内容・フォーマット・語彙のいずれにおいてもある程度統制されている，いわば定型文書に近いものであるため処理が容易であったのに対し，Webページにはさまざまな種類の文書があり，執筆者もさまざまで，決まった形式がない（非定型である）ので，基本システムのままでは性能が維持できない。

一方で，Webページの標準であるHTML文書の示す構造情報には質問応答の回答を定めるに有用な情報が含まれている。我々は，HTMLのタグ情報を利用して回答候補の抽出精度を向上させる方法[8]についても研究開発を行い，インターネット質問応答システムにより実験を継続中である。

インターネット質問応答システムは，基本システムに対して，以下の拡張を行ったことが特徴である。

- 情報検索の文書検索部にインターネット検索エンジンを用いる
- 情報抽出の回答候補抽出部において，回答候補の存在する範囲推定にHTMLタグのタイトル情報，表情報を加味する

質問応答システムの実用化

研究開発を通じて実用に供することのできる質問応答システムを提案している。

ニュース情報を発信するメールマガジンに適用したものがMAILPIA*5)である[9]。MAILPIAは電子メールを利用した個人向けの情報アクセスのためのサービスであり，WEBページの更新検知，メールマガジンのクリッピングなどいくつかの機能を持っているが，クリッピングされたメールマガジンのデータベースから自然言語での質問により情報を調べることができるように本稿で述べた質問応答技術を適用している。

また，質問応答の形態をとってはいないが，情報検索と情報抽出のモジュールを，産学連携支援ツールBluesilk*6)に搭載している[10]。これは，大学のホームページや論文，特許，その他さまざまな文書から，連携相手のキーパーソンや企業を探したり，関連する技術を探したりすることができるというものである。対象をホームページとして情報を抽出する機能の実現にインターネット質問応答システムの成果を利用している。

あ と が き

情報検索と情報抽出を組み合わせた，機械との対話インタフェースへの第一歩となる質問応答のシステムについて，我々の取り組みを中心に解説した。

インターネットの膨張はとどまるところを知らず，情報量も飛躍的に増大するなかで，誰にでも簡単に必要な情報を得ることができる質問応答システムへの期待は大きい。この分野の研究はまだ端緒についたばかりで，比較的統制のとれた文書に関してようやく成果が出始めた段階であるが，すでに部分的ながら一般ユーザの目に触れるところにもシステムが現れてきており，今後もますますの発展が期待される。 ◆◆

■参考文献

1）大沼宏行，池野篤司：ホームページやメールを対象とした質問応答システム，情報アクセスのためのテキスト処理シンポジウム発表論文集，pp.89–95，2003年
2）全文検索システム namazu，http://www.namazu.org/
3）汎用連想計算エンジン GETA，http://geta.ex.nii.ac.jp/
4）福本淳一，他: 固有名詞抽出における日本語と英語の比較，情処研報，98-NL-126，pp.107–114，1998年
5）NTCIR情報検索システム評価用テストコレクション構築プロジェクト， http://research.nii.ac.jp/ntcir/index-ja.html
6）NTCIR-4質問応答タスク（QAC-2）ホームページ，http://www.nlp.cs.ritsumei.ac.jp/qac/index-j.html
7）IKENO, A., OHNUMA, H.: Oki QA system for QAC-1, Proceedings of NTCIR Workshop 3 Meeting QAC1, pp.17–20, 2002
8）大沼宏行，池野篤司：HTML文書を対象とした質問応答システムにおける回答抽出方法，第63回情報処理学会全国大会予稿集，3-41，2001年
9）MAILPIA 電子メールによる個人用情報アクセスサービス，http://www.oki.com/jp/RDG/JIS/mailpia/
10）産学連携支援ツール Bluesilk，http://www.bluesilk.jp/

●筆者紹介

池野篤司：Atsushi Ikeno.研究開発本部 ユビキタスシステムラボラトリ

*4）F値（F measure）は，一般に次の式で表される。$F=2*P*R/(P+R)$　*5）MAILPIAは沖電気工業（株）の登録商標です。
*6）Bluesilk は（株）三菱総合研究所の登録商標です。