非 Factoid 型質問に対応した質問応答システム

諸岡 心† 福本 淳一‡

†立命館大学大学院理工学研究科 ‡立命館大学情報理工学部メディア情報学科

E-mail: †k_morooka@nlp.is.ritsumei.ac.jp, ‡fukumoto@media.ritsumei.ac.jp

1. はじめに

膨大な情報から必要とする情報を見つけ出す技術として質問応答（QA）がある．QA に関する評価ワークショップとして QAC[1]や TREC[2]の QA トラックがある．これらのワークショップでは，名称（固有名称や一般名称）や日付・数値など事実に基づく回答を求める Factoid 型質問が主に行われてきており，現在 QAC4 ではそれを超えるものとして Why や How などの非 Factoid 型質問を対象に評価が行われる予定である．

我々はこれまでに Factoid 型の質問応答システムに加え，Why 型質問の回答を抽出する手法の提案を行ってきている．今回はそれに加えて Definition 型および How 型の質問応答モジュールの追加を行った．Why 型については[3], Rhetorical Structure Theory（RST）[4]の関係を元に回答抽出の手法を提案してきている．Definition 型質問については QAC4 において作成されたサンプル質問応答データの分析に基づき，定義表現の抽出パターンを記述することで回答抽出を行った．How 型についてはこれまで Nishimura らの研究[5]において FAQ データから回答の範囲を特定する手法が提案されている．我々は Definition 型と同様に QAC4 の質問応答サンプルデータの分析から，手順などの表現を抽出するためのパターンの記述を行い，新聞記事データを対象にした回答抽出を行った．

以下，２章で質問応答システムの概要について述べ．３章で質問タイプの判定方法について述べ，４章で Why 型，Definition 型，How 型の各回答抽出処理について述べた後，５章で本システムを用いて QAC4 の Formal Run で用いられた質問セットを用いた実験結果について考察する．

2. 質問応答システムの概要

本システムでは，まず，与えられた質問文の解析を行い，質問文の表現から質問タイプの決定と記事検索のためのキーワード抽出を行う．今回の非 Factoid 型の質問に適応させるため，これまでのシステムにはなかった Why 型，Definition 型，How 型の各タイプを判定するための質問文パターンの拡張を行った．次に，質問文から抽出されたキーワードを用いた検索精度の向上のため，検索キーワードを用いて Web 文書の検索を行い，検索結果の上位 10 位の snippet を用いて，再度記事検索を行う改良を行った．最後に検索結果の記事から各質問タイプに応じた回答候補の抽出処理を行い，最後にシステムの回答として上位 100 位のものを表示する．今回のシステムでは，抽出する回答範囲は最大で１つのパラグラフとしている．これは，非 Factoid 型の質問応答データの分析から正解の範囲として１つのパラグラフを超えるものが少なかったためである．

3. 質問タイプの判定

非 Factoid 型の質問として追加したタイプとして Why 型質問，Definition 型質問，How 型質問の３つがある．各質問タイプの判定は，質問文の表層表現のパターンをとらえることで判定している．各質問タイプの表層パターンを以下に示す．

- Why 型質問・・「なぜ」「何故」「どうして」
- Definition型質問・・「とは(何|なに|。)」「どのような＋名詞」「どんな＋名詞」「名詞＋って何」
- How型質問・・「どういう」「どうしたら」「どうする」「どうすれば」「どうやったら」「どうやって」「どうなりますか」「どのように」「どのような」

質問タイプの判定は従来の固有表現など名称を対象とした質問タイプでもなく以上の質問タイプにも当てはまらない場合，システムでは Definition 型質問であるとして回答抽出を行う．

4. 回答抽出処理

文書検索によって検索された文書から判定された質問タイプに応じて回答抽出を行っている．回答の抽出範囲は，記事検索で絞り込まれた記事中の各パラグラフが対象になる．以下では，非 Factoid 型の各質問タイプの Why 型，Definition 型，

How 型のそれぞれにおける回答抽出手法について述べる.

4.1. Why 型質問の回答抽出

Why 型の質問に対する回答については，質問と回答間に存在する意味的関係として「原因・理由」「背景」となるものを手がかりに回答の抽出を行う.このような抽出パターンとして「ため」「から」「ので」「ことで」「ことにより」「ことによって」などの表現を利用した．質問文表現と類似する文からこれらの抽出パターンを利用して回答部分の抽出を行う．これらの意味的関係を示す手がかり語のパターンとして「ため」についての抽出パターンを図 1に示す．また，この抽出パターンに当てはまった場合でも，図 2に示す非抽出パターンに当てはまった場合には抽出されない.

動詞 ＋ ため
サ変接続名詞 ＋ ため
ため ＋ 助詞「に」
ため ＋ 記号「、| 。| a-z| A-Z」
ため ＋ 助動詞「だ」

図 1：手がかり語「ため」の抽出パターン

代名詞 ＋ 助詞「の」 ＋ ため
念 ＋ 助詞「の」 ＋ ため
動詞 ＋ ため ＋ 助詞「の」
サ変接続名詞 ＋ ため ＋ 助詞「の」

図 2：手がかり語「ため」の非抽出パターン

また，助詞だけでなく「理由」「原因」「背景」などの名詞を用いて意味的関係を示すもととして，新聞記事データの分析およびシソーラスを用いた類語を分析することでも抽出パターンの記述を行った．表 1にその手がかり語を，また，それらの手がかり語を用いた抽出パターンを図 3に示す.

表 1：手がかり語

理由	一因	根因	訴因
要素	遠因	罪因	動因
背景	外因	死因	導因
動機	禍因	主因	道因
事由	画因	従因	内因
根拠	起因	勝因	敗因
引き金	基因	心因	病因
ゆえん	近因	真因	副因
ゆえ	偶因	成因	福因
きっかけ	原因	善因	誘因
悪因	業因	素因	要因

助詞「に|が|を」 ＋ 手がかり語
格助詞「という」 ＋ 手がかり語

図 3：手がかり語を用いた抽出パターン

以上の手がかり語を利用して抽出された回答について文頭に特定の接続詞がある場合，直前の文も回答として含める必要がある．このような表現として，順接の接続詞「ゆえに」「それゆえ」「これゆえ」「そのため」「このため」「だから」が存在するものについて回答範囲の拡張を行った.

抽出した文は，回答となる原因部分(回答範囲)と質問内容の結果部分(質問範囲)が含まれ，これらの範囲を決定する必要がある．これらの範囲は手がかり語と接続詞を用いて特定する．手がかり語「ため」「から」や接続詞「だから」「ゆえに」などは語句以前が回答範囲，以降が質問範囲となり，「を＋理由」や接続詞「なぜなら」は語句以降が回答範囲となり，以前が質問範囲となる．回答範囲と質問範囲の特定例を図 4と図 5に示す.

１９９８年６月、米科学誌「ブレティン・オブ・ジ・アトミック・サイエンティスツ」は、印パ核開発など**を理由**に人類破滅までの時間を示す「終末時計」を１４分前から９分前に５分早めた

太字：手がかり語，下線：回答範囲，二重下線：質問範囲

図 4：「を＋理由」の範囲特定例

９９年３月から約２カ月半にわたるＮＡＴＯのユーゴ空爆で、連邦軍の不満は一層募る結果になった。**なぜなら**、「コソボはセルビアの発祥地」との掛け声でトマホーク・ミサイルの標的にされる兵士たちはもはや、セルビア民族主義を掲げた政権の号令を信じなくなっていたからだ。

太字：接続詞，下線：回答範囲，二重下線：質問範囲

図 5：文頭に「なぜなら」の範囲特定例

さらに，逆接となる節が文内に含まれる場合，冗長となる部分が存在するため冗長部分の削除を行う．逆接節とは，文頭から逆接助詞「が」までの範囲とする．冗長部分は逆接節と回答範囲の位置関係により決定する．回答範囲より前に逆接節がある場合，逆接節が冗長部分とし，回答範囲の後に逆接節がある場合と逆接節内に回答範囲が有る場合,逆接節より後が冗長部分として除去する．冗長部分の除去例を図 6に示す.

~~国連安保理決議があればイタリアは公海での船舶検査を行える**が**、~~当時の対ユーゴ経済封鎖は「決議抜き」のため、イタリアの対応が注目を集めていた。

下線：回答範囲，~~打ち消し線~~：冗長部分，**太字**：逆接助詞「が」

図 6：回答範囲の前に逆接節がある場合

4.2. Definition 型質問の回答抽出

Definition 型の質問においては，「世界遺産条約とは何ですか。」のように質問文中に現れた「世界遺産条約」のように，質問によって定義を求める用語の説明表現を抽出する必要がある．また,「世界遺産条約とはどんな条約ですか。」のように，定

義を求める用語とそれに関連する言葉を用いている場合もある．そこで，質問文中の情報として，どのような格や品詞のキーワードが回答抽出の主な手掛かりとなるかを分析した．質問応答データの分析から得られたキーワードの質問文中での役割を，ハ格・ガ格，ハ格・ガ格を修飾する句，動詞・サ変接続名詞，属性語に分類した．属性語とは，回答の属性を特定する語とし，例えば上の例では「条約」が属性語となる．また，質問文中に出現した固有表現や括弧付の語（“「」”で囲まれた語を対象にする）にも注目して分析を行った．その結果，質問文からのメインキーワード（質問対象となる中心語）および属性語について次のような抽出規則を作成した．

(1) メインキーワードの抽出規則(a,b,c の順に適用される)
  a. 括弧で囲まれた語もしくは固有表現
  b. 質問文中のハ格もしくはガ格を修飾する語
  c. ハ格もしくはガ格の語
(2) 属性語の抽出
「どういった」「どのような」「どんな」などの疑問詞に付属する語

以上の手法により抽出されたメインキーワードおよび属性語を利用して検索結果のパラグラフから回答候補の選択を行う．回答候補の選択は以下のパターンにマッチするものを抽出する．

- メインキーワード＋「は」|「が」|「も」
- ～「が」＋メインキーワード＋「を」
- ～「する」＋メインキーワード
- ～「の」＋メインキーワード
- メインキーワード＋「とは」
- ～「の」＋属性語
- ～「する」＋属性語

以上のパターンにマッチした場合，メインキーワードを含んでいる場合はその１文を抽出し，属性語を含んでいる場合は属性語に対する修飾部分を抽出し，それを回答候補としている．修飾部分の判断のためには対象となる１文の構文解析結果から適切な部分の抽出を行っている．

4.3. How 型質問の回答抽出

How 型質問では何かの行動の手順や方法，条件などについての質問している．そのため，それらの動作や手順を示すものとして，質問文中に出現する動詞やサ変接続名詞が重要であると考えられる．その回答としては，質問対象となる行動についての記述であり，これは質問文中の動詞やサ変接続名詞を用いて探索することにより回答抽出が行えると考えられる．例えば，質問が「世界遺産はどのようにして決定しますか。」の場合，「世界遺産の決定には～が必要である。」「世界遺産は～し、決定している。」「世界遺産は～で決定される。」など「決定」というサ変接続名詞が回答に出現する．

How 型の質問についても Definition 型と同様の手法でメインキーワードの抽出を行い，属性語については How 型では属性語の代わりに動詞キーワードの抽出を行う．回答抽出においては，以上の情報を用いたパターンとともに回答候補として手順などを示す表現にも注目して回答抽出を行う．

- メインキーワード＋「は」|「が」＋動詞キーワード
- 「手順｜手法｜方法｜条件」＋「は」
- 「が」＋「手順｜手法｜方法｜条件」

これらのパターンとマッチする文を回答候補として抽出する．

例えば，質問が「横綱になるにはどうすればいいですか。」の場合，メインキーワードが「横綱」で動詞キーワードが「なる」になり，「横綱になるには2連続優勝しなければならない。」の１文が回答として抽出される．また，「条件は2連続優勝である。」や「手順|条件|方法」を含む節が機能語「が」の節にかかっている場合では「2連続優勝が条件である。」が回答候補として抽出される．

5. 回答抽出例と考察

以下にNTCIR6のQAC4のFormal Run質問データを用いたWhy型質問の回答例を図 7に示す．

Q：茨城県東海村で起きた臨界事故の原因は何でしたか。
A1:原子力事故は発電所内の問題とばかり考えてきたから
A2:事故は社員がこの違法マニュアルの工程をさらに簡略化したため
（QAC4-00041-00）

下線：手がかり語

図 7：Why型質問回答抽出例

図に示されている質問に対する回答候補から選択された質問範囲は「事故に対する電力業界の反応は「あんなところで臨界事故が起きるとは」というものだった。」や「茨城県庁で会見したＪＣＯ東海事業所の小川弘行・製造部計画グループ長は国に認められた正規の手順を無視した「違法マニュアル」の存在を認めた。」となっており，質問文中の「臨海」や「事故」「起きる」などの語と類似度が高くものとなっている．これらに対応する回答範囲として回答 A1,A2 が抽出され，正しく回答抽出が行われている．

Definition型の回答例を図 8に示す．

Q：臨界とはどのような状態のことですか。
A1：臨界とは核分裂が連続して起こる状態をいう。
A2：核分裂性ウランの状態
A3：臨界の状態
（QAC4-00042-00）

下線：メインキーワード，箱：属性語

図 8：Definition型質問回答抽出例

ここでは，質問文解析において抽出されたメインキーワード「臨海」，属性語「状態」を用いて回答抽出が行われた．図中の回答:A1は「メインキーワード＋とは」の抽出パターンによる抽出が成功している．回答:A2は属性語を修飾する句を連結して範囲拡張を行った回答で，属性語を用いた抽出が成功している．回答A3は回答:A2と同じく属性語を用いて抽出しているが，メインキーワードと属性をそのまま回答として適切ではない．属性語を用いた抽出手法では，属性語を修飾する句にメインキーワードが含まれる場合は抽出しないなどの制限を設ける必要があると考えられる．

最後にHow型の回答例を図 9に示す．

Q：脳死はどのようにして確認されますか。
A1：主治医と他の１人の医師が脳死状態と確認後、家族から改めて脳死判定と臓器提供の承諾書を取り、２７日午前１１時４０分から再び法に基づく脳死判定を行った
A2：臓器移植法の施行規則では、脳死の判定は（１）深い昏睡（こんすい）状態（２）瞳孔（どうこう）が開いたままで縮まない（３）目に光を当てても瞳孔に変化がないなど脳幹の機能を示す各種の反応がない（４）脳波が平たん（５）自力で呼吸できない（６）前記５項目を満たす状態が６時間以上続いている――を確認することを定めている
A3：臓器移植法に基づく脳死判定では脳波が平たんか、波打っているかを確認しなければならない
(QAC4-00039-00)

下線：メインキーワード，箱：動詞キーワード

図 9：How型質問回答抽出例

ここでは，メインキーワード「脳死」，動詞キーワード「確認」で回答抽出が行われた．図中の回答:A1では，脳死確認に主治医と医師1人が必要であるということがわかるが，それ以外の部分は冗長である．動詞キーワードである「確認」の直後の接辞「後」を手掛かりに回答範囲をより限定することが可能であるかもしれない．回答:A2は，条件の全てが記述されており，得られた回答の中で一番相応しい回答であると考えられる．回答:A3は回答:A2で記述された条件の1つが挙げられている．本研究では，複数の回答候補をマージする処理は実現していないが，同一の回答についてはまとめて提示する手法も簡潔な回答を与えるために必要であると考えられる．

6. おわりに

本稿では，これまでに Factoid 型の質問応答システムに加え，非 Factoid 型質問として，Why 型，Definition 型および How 型の質問にも対応した質問応答システムの質問パターンの分析と各質問タイプについての回答抽出手法について述べた．評価のため，QAC4 の Formal Run において用いられた質問応答の結果についていくつかの例について考察を行った．現在，本手法を用いた質問応答の詳細化評価を行っており，QAC4 では人間による評価結果と自動評価結果についての報告が行われる予定であり，本手法についての有効性について NTCIR ワークショップミーティングにおいて議論されることを期待している．

これまでの QAC4 の Formal Run の詳細な評価については継続中であるが，これまでの評価でも非 Factoid 型の質問応答については質問だけでなく解答のバリエーションも多く，今度さらに多くのパターンの実装やどのような回答が適切であるのかについても分析，検討が必要であると考えられる．

参考文献

[1] J. Fukumoto, T. Kato and F. Masui, "Question Answering Challenge for Five Ranked Answers and List Answers - Overview of NTCIR4 QAC2 Subtask 1 and 2," Working Notes of the Fourth NTCIR Workshop Meeting (NTCIR4), pp.283-290, 2004.

[2] E. M. Voothees, "Overview of the TREC 2004 Question Answering Track," TREC 2004, NIST, pp.12-20, 2004.

[3] 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回年次大会発表論文集，pp.293-296，2004.

[4] W. C. Mann and S. A. Thompson, "Rhetorical Structure Theory: A Theory of Text Organization," USC ISI Technical Report ISI/RS-87-190, 1987.

[5] R. Nishimura, Y. Watanabe and Y. Okada : A Question Answer System Based on Confirmed Knowledge Developed by Using Mails Posted to a Mailing List, IJCNLP-05, 2005.