What 型 Q&A システムの構築

藤岡 秀明　　浦谷則好
東京工芸大学工学部コンピュータ応用学科

1. はじめに

Q&A システムとは，自然文で記述された質問に対して構文解析を行い，大量の文書データから的確な答えを出力するシステムである．Q&A システムの多くは質問に対して短く端的に答える factoid 型質問応答システムであり，文章で答えることのできるシステムは少ない．

しかし「～とは何ですか」に代表される質問，いわゆる What 型の質問には，名詞でなく文章で回答するシステムが必要とされる．このような質問に対しては既存の文章を用いるしかない．そこで Web 上の文章データを用いる手法が数多く提案されている［2］．近年の情報通信技術の発達に伴い，今や Web 上には大量の文章データが存在する．不特定多数の者が様々な情報を自由に公開することができるため，その信憑性には不確かな部分も多いが，それらの文書中には有用となる情報も数多く存在する．

そこで我々は，Web 上に存在する文章を利用した Q&A システム，その中でも回答候補に factoid 型・non-factoid 事実型の両方の可能性を秘めた，What 型の質問文に対する Q&A システムの研究を行った．

2. 期待される質問応答

我々はあらゆる質問に対する Q&A システムの構築を進めているが，ここでは What 型 Q&A システムに限定してシステムを説明する． What 型質問には，回答に名詞を要求する質問と文を要求する質問が混在しているため，適切な回答を返すには質問文から回答が文になるのか名詞になるのかの判断をする必要がある．以下に予想される質問文の例とユーザが期待する適切な回答を示す．

Ex.1)ヘモグロビンとは何ですか
　　ヒトを含む全ての脊椎動物や一部のその他の動物の血液中に存在する赤血球の中にある蛋白質である．

Ex.2)赤血球の役割は何ですか
　　肺から全身へ酸素を運搬することです

Ex.3)赤血球の色は何ですか
　　赤色

Ex.1 のような「(名詞 a) とは何ですか」という，質問文に含まれる名詞が１つ，文節が２つという簡単な質問は，名詞 a そのものの意味や説明を求める質問であり，自ずと回答は文になる．この例の場合では名詞“ヘモグロビン”で Web 検索をし，その説明がなされている文を回答として提示することが理想である．

Ex.2 と Ex.3 は文法的にはまったく同じ質問文であるが，回答として要求するものは文と名詞で異なる．このような質問文は構文解析だけでは回答の種類を判断できないため，我々は質問文中の名詞に着目した．複数の名詞を組み合わせた複雑な質問の場合，“何”の直前または直後にくる名詞を質問文中の重要語とし，それを利用して回答種類を判断する手法を新たに考案した．この例の場合，重要語が“役割”ならば回答は文を，“色”ならば回答は名詞を要求することがわかる．どんな単語が重要語になると回答が文になるのか否かを調べ，回答種類の判断をシステムに自動的に行わせる．回答の判別方法や選出方法は，次項以降で詳しく述べる．

3. 回答種類の判別方法

回答種類の判別は“文節の数”と“回答に文を要求する特定語の有無”で判別する. 以下の表１は我々が独自に定めた回答に文を要求する特定語一覧である.

表１：回答に文を要求する特定語

役割	役目・機能・定理・論・法則・公式
目的	意図・意味・狙い
原因	理由・きっかけ・訳
違い	相違・差異
長所	利点・特徴・特長・強み・メリット
短所	欠点・弱点・弱み・難点・デメリット

６つのカテゴリとその類語，合わせて３０語を特定語とした. 選定基準は，我々が Web 上で見つけたごく一般的な質問や，Q&A コミュニティ「教えて goo(http://oshiete.goo.ne.jp/)で見つけた質問を参考に， What 型質問で使われることが多いと思われるものを３０語選んだ.

まず，文節の数が２つならばその時点で回答は文と判断を下す．なぜなら文節が２つ，つまり質問文に含まれる名詞が１つしかない質問文は，名詞そのものの意味・説明を要求する質問文であり，ユーザの求める回答は明らかに文である.

次に, 文節の数が３つ以上の質問文を対象として，“何”の直前または直後の名詞，つまり前項で述べた重要語に当たる名詞が“回答に文を要求する特定語”に当たるか否かを調べる．この特定語が重要語の位置に来た場合，回答は文と判断する．そうでない場合は回答を名詞と判断する.

つまり，最終的に回答種類は次の３つに分類される.

“文節が２つで回答が文”

“文節が３つ以上＋特定語で回答が文”

“文節が３つ以上で回答が名詞”

4. ユーザに提示する回答の選出方法

回答の選出方法は、前項で述べた３つの回答種類それぞれにあった形で定める.

4-1“文節が２つで回答が文”の場合

Web 検索の際の検索ワードは「(名詞 a) ＋とは」とする．これは名詞 a が広く一般的に使われている単語の場合，名詞 a だけを検索ワードとしてはユーザが求める回答候補文が抽出されにくいため，「(名詞 a) とは～である」という説明文を抽出することを目的としている．回答の選出方法は，「(名詞 a) とは～」で始まる一文を抜き出す.

4-2“文節が３つ以上＋特定語で回答が文”の場合

Web 検索の際の検索ワードは，質問文に含まれる全ての名詞・動詞・形容詞・未知語とする．回答の選出方法は，検索ワードに使った名詞・動詞・形容詞・未知語全てが含まれる一文を抜き出す.

4-3“文節が３つ以上で回答が名詞”の場合

Web 検索の際の検索ワードは前述の 4-2 と同じく，質問文に含まれる全ての名詞・動詞・形容詞・未知語とする．回答の選出方法は，検索して得られた文書群中に使われている名詞の数を合計し，質問文中の重要語を含む固有名詞・一般名詞でもっとも合計数の多いものを出力する．重要語を含む固有名詞・一般名詞が存在しない場合は，純粋に合計数の多いものを出力する．ただし、検索ワードそのものは除外する.

5. 実験と考察

本システムの回答種類の判別方法をもってした回答の選出方法の精度を調べるため，実際に質問文を用意して精度実験を行った．実験に用意した質問文は，同研究室のメンバーで作成した正解が明らかな質問文５０個である.

構文解析器は CaboCha[4]，検索エンジンは YahooAPI[5]を使用し，検索結果上位３００件の Snippet 情報を対象として回答を探す．ページタイトル・ページ内文章全体は探索対象に含めない．

その結果，正しい答えを返すことができたのは５０問中８問であった．以下の表２は，その８問の質問と出力された正しい回答である．

表２：正しい答えが得られた質問

	質問と回答
1	日本一低い山は何 →天保山
2	日本で一番狭い県は何 →香川県
3	1919 年にフランスで締結された条約は何ですか →ヴェルサイユ条約
4	イチロー選手の本名は何ですか →鈴木一朗
5	裁判員制度の目的は何ですか →裁判員制度導入の目的.国民の感覚が裁判に反映.裁判が速くなる
6	フランス語でル･クルーゼの意味は何ですか →「ル・クルーゼの名前の由来は?<クルーゼ>はフランス語で坩堝(るつぼ)を意味します
7	錬金術って何 →錬金術とは一般の物質を「完全な」物質に変化・精錬しようとする技術のことであり、さらには人間の霊魂をも「完全な」霊魂に変性しようという意味を持つこともあった(=神に近づく、神になる、神と合一する方法ともいえよう)
8	テルミンとは何ですか →テルミンとは.テルミンは 1920 年にロシアの物理学者レフ･テルミンによって発明された世界最古の電子楽器です

質問１・２・３は，回答候補群から重要語を含む語を出力することで正しい結果が得られた例，質問４は重要語を含む語が存在しなかったため出現頻度の多い語をそのまま出力し正解となった例である．特に質問３に関しては，重要語である「条約」を含む語を探すことで，出現頻度は１６位であった「ヴェルサイユ条約」を見事導き出すことに成功した．

質問５と６は，「目的」や「意味」といった特定語が重要語になったため回答を文と判断，１文中に検索ワードが全て含まれる文の上位１位がそのまま正解となった．質問７・８も，「～とは」で検索し得られた Snippet 情報の中で，「～とは」で始まる箇所を抽出し，見事正解している．

他にも文章を要求する質問に関しては，上位５件までを表示させた場合はさらに４問正しい答えが得られ，正解は５０問中１２問となった．

次の表３は，正しい回答が得られなかった質問のうち，５例である．

表３：正しい回答が得られなかった質問

	質問と回答
A	グラハム数とは何か →（検索結果なし）
B	JR の正式名称は何ですか →情報
C	リンカーン大統領はどんな凶器で暗殺されましたか？ →日記
D	錬金術における 4 元素って何？ →火
E	世界一高い山は何？ →富士山

質問 A は，文節が２つのため「～とは」をつけての検索を行いたかったが，この質問文を構文解析器にかけると「グラハム」と「数」が分割されてしまい，検索ワードが「グラハムとは」「数とは」の２つになってしまった．これでは十分な検索ができない

ため，形態素解析後に形態素調整が必要である．

質問 B も，構文解析器にかけるとローマ字表記の名称が全て分割されてしまい，検索ワードにならなかった．これに関しても分割されたローマ字表記の名称を結合する処理が必要である．

質問 C では「凶器」が重要語になり，上位３００件の Snippet 内の単語出現頻度だけで回答を選出しようとして失敗している．「条約」や「山」などのように回答にもその単語が含まれる場合は多くない．しかし単語出現頻度だけで回答を出そうとする手法では不十分であり，正しい回答を選出することができたのは質問４の場合だけであった．今後は単語の係り受け関係も利用しての回答選出手法を考案したい．

質問 D では，回答が複数存在する場合に対応できないことが分かる．３原則・４元素など重要語に数詞がつくものはその数だけ回答を選出する，などの手法が考えられる．しかし「エジソンの発明品は何？」という質問のように，数詞がつかなくても解答が複数ある場合も存在するため，さらなる研究が必要である．

質問 E は，提案手法によって正しい解答が得られなくなった例である．単語出現頻度を利用しての回答選出では正しい回答「エベレスト」が得られるが，重要語が含まれる回答候補検索を先に行っているため，本来出力されることのない結果「富士山」が出力されてしまった．つまり，重要語を含む回答候補語検索の後に出現頻度を用いての回答選出では不十分である．単語共起情報や単語出現頻度の差なども用いての回答選出手法が必要と考える．

6. おわりに

本論文では What 型 Q&A システムの構築を目的として，回答種類の判断方法とユーザに提示する回答選出方法の２つを提案した．「質問文の文節数」「何の直前直後にくる重要語」の２点で回答種類を３つに分類し，それぞれのパターンに別々の回答選出方法を当てはめた．文節数と重要語を使っての回答種類の判別に関しては大きな問題は見られなかったが，回答選出方法に関しては課題を残す結果となった．「形態素解析後の形態素調整」「係り受け関係や共起情報の利用」「複数存在する回答への対応」が今後の大きな課題である．

参考文献

[1] 丹波達洋　福本淳一　“質問応答システムにおける回答候補の選定手法”　言語処理学会第９回年次大会発表論文集 pp625-628

[2] 山本正範　延澤志保　大原育夫　“Web 文書の抜粋を回答とする質問応答システム”　言語処理学会第１１回年次大会発表論文集 pp1076-1079

[3] 伊藤雄　秋葉友良　“事実型オープンドメイン質問応答システムにおける周辺情報を考慮した詳解の抽出”　言語処理学会第１４回年次大会発表論文集 pp185-188

[4] CaboCha
http://chasen.org/~taku/software/cabocha/

[5] YahooAPI
http://developer.yahoo.co.jp/webapi/search/websearch/v1/websearch.html