What 型Q&A システムの構築

発表者：サ　ミンソン

著者：藤岡秀明（東京工芸大学
工学部コンピュータ応用学科）

1．はじめに

- What型の質問

「～とは何ですか」に代表される質問

→Web上の文章データを用いることが多い、

名詞や文章で回答（factoid型とnon-factoid型）

・Q&Aシステムの多くは質問に対し端的に答えるfactoid 型

研究の目的

- 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型の両方のWhat型の質問文に対する研究

・What 型Q&A システムの構築を目的

2. 期待される質問応答

- What型質問は、

ー名詞を要求する質問

ー文章を要求する質問

↓

質問文から回答が名詞になるのか
文章になるのかの判断が必要

質問文の例とユーザが期待する回答（1）

- 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.

「名詞aとは何ですか」 → 名詞1つ、文節2つ

名詞aの意味や説明を求める質問、
回答は文

質問文の例とユーザが期待する回答（2）

- 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

同じ文法の質問だが、回答として要求するのは文と名詞で異なる→構文解析では回答の種類が判断できない

「何」の直前または、直後の名詞を
重要語と指定
↓
回答種類を判断する

- Ex2.)の場合、重要語は「役割」→回答は文
- Ex3.)の場合、重要語は「色」→回答は名詞

3. 回答種類の判別方法(1)

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

表1:回答に文を要求する特定語

Q&Aコミュニティ「教えてgoo」で見つけた質問を参考に選定

３．回答種類の判別方法（２）

- 回答種類は次の３つに分類

ー文節が2つで回答が文

ー文節が3つ以上＋特定語で回答が文

ー文節が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（１）

- 「文節が2つで回答が文」の場合

検索ワードが名詞aだけだと、求める回答文が抽出されにくい

検索ワードは「名詞a＋とは」とする。

→「（名詞a）とは～である」という説明文を抽出するため

４．ユーザに提示する
回答の選出方法（２）

- 「文節が3つ以上＋特定語で回答が文」の場合

検索ワード：質問文にある名詞・動詞・形容詞

回答の選出方法：検索ワードに使った名詞・動詞・形容詞全てが含む一文

４．ユーザに提示する
回答の選出方法（３）

・「文節が3つ以上で回答が名詞」の場合

検索ワード：４．（２）と同じ

回答の選出方法：検索で得られた文書群中の名詞の数を計算し、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

５．実験と考察

- 実験に用意した質問文は、自ら作成した正解が明らかな質問文５０個
- 構文解析器はCaboCha
- 検索エンジンはYahooAPIを使用し、検索結果上位３００件のSnippet情報を対象

- 正しい答えを返したのは５０問中８問

正しい答えが得られた質問(1)

- 質問1．日本一低い山は何

答え:天保山

→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

正しい答えが得られた質問（2）

- 質問2．1919 年にフランスで締結された条約は何ですか

答え：ヴェルサイユ条約

→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．

正しい答えが得られた質問（3）

- 質問3．裁判員制度の目的は何ですか
- 答え：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

- →「目的」という特定語が重要語になり、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

正しい正解が得られなかった質問（１）

- 質問A. グラハム数とは何か
  →（検索結果なし）

- 質問文を構文解析器にかけると、「グラハム」と「数」に分割→検索ワードは「グラハムとは」、「数とは」の２つになり結果が得られない。

- 形態素解析後に形態素調整が必要

正しい正解が得られなかった質問（2）

- 質問B. JR の正式名称は何ですか
  →情報

- 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，検索ワードにならかった。
- 分割されたローマ字表記の名称を結合する処理が必要

正しい正解が得られなかった質問（3）

- 質問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

- 「凶器」が重要語になり上位300件の単語出現頻度だけで回答を選出し失敗
- 質問2のように重要語が回答に含むのは少ない
- 単語の係り受け関係も利用しての回答選出手法が必要

正しい正解が得られなかった質問（4）

- 質問4．錬金術における4 元素って何？→火

- 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が、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

6. おわりに

- 文節数と重要語を使っての回答種類の判別は良い結果
- 回答選出方法に関しては課題を残す結果

- 今後の課題:「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」