Web 文書を情報源とする 記述的な回答が可能な質問応答システム

佐藤 充† 石下 円香† 森 辰則‡
†横浜国立大学 大学院 環境情報学府 ‡横浜国立大学 大学院 環境情報研究院
E-mail: {mitsuru,ishioroshi,mori}@forest.eis.ynu.ac.jp

1 はじめに

電子文書の増大に伴い，情報アクセス技術の一つとして質問応答が注目されている．質問応答とは自然言語による質問に対して検索文書から回答そのものを抽出する技術である．質問応答のタスクは大きく二つに分けられる．固有名や数量等を問う factoid 型の質問応答と定義や理由等を問う non-factoid 型の質問応答である．日本語における non-factoid 型の質問応答は，評価型ワークショップ NTCIR6 QAC4[1] においてタスク設定される等，近年広く研究されているものの，依然難易度の高いタスクである．

質問応答の情報源としては，新聞記事等のローカル文書を対象とする場合と Web 文書を対象とする場合とがある．後者は文書の前処理ができない点で不利であるが，新鮮かつ大量の情報が得られる点で実用的である．

本稿では Web 文書を対象とする日本語 non-factoid 型質問応答を扱う．Web 文書を単に検索対象として扱うだけでなく，Q&A コミュニティサービスの質問・回答事例集合をコーパスとして用い，そこから得た回答の記述スタイルに関連する統計情報を解抽出の際に役立てる手法を提案する．

2 non-factoid 型質問応答

本章では，non-factoid 型質問応答手法の特徴と関連研究，及び提案手法の位置付けについて述べる．

2.1 質問の種類と処理方法

non-factoid 型の質問を大まかに分類すると，表 1のような種類がある．

表 1: non-factoid 型質問の分類

質問の型	質問の記述スタイル例	回答の記述スタイル例
定義型	～って何	～とは…である
理由型	なぜ～	～ため
方法型	～にはどうしたらいい	～するにはまず…
その他	X と Y の違いは何	X は～だが，Y は…

non-factoid 型質問に対する回答としては，数文にまたがるような比較的長い文章表現による記述的な回答が想定される．このような質問の解の適切性は，

【尺度 1】 質問の内容との関連性

【尺度 2】 質問の型に応じた記述スタイルを満たすか

の組合せで見積もることができると考えられる．記述スタイルとは，表 1に示したような質問や回答の特徴表現を指す．【尺度 1】は簡単には質問文と解候補文の類似度で計算できる．【尺度 2】は人手で作成した語彙統語パタンや機械学習により判定することが多い．

2.2 質問の型ごとに処理を分ける手法

Han ら [2] は英語の定義型質問応答において，上記の二つの尺度をコーパスから推定した確率モデルに基づいて計算している．このように質問の型が限定された場合は，型に応じた回答表現のパタンを用意したり，専用のコーパスを学習データとして用いることが有効である．質問の型を限定しない場合でも，入力された質問を予め用意した型に分類し，型ごとに処理を分けることで同様のアプローチを実現できる．

2.3 質問の型分類を行なわない手法

前節で質問の型ごとに処理を分けるアプローチについて述べたが，実際には質問の型が何種類あるかは不明であるし，型ごとに個別の処理方法を用意するのは非効率である．また，質問の型分類の精度が回答精度に大きく影響してしまう．可能ならば質問の型に依らない統一的な手法が望ましい．

水野ら [4] は日本語 non-factoid 型質問応答において，Q&A コミュニティサービスの質問・回答事例集合を学習データとし，質問と回答の型の一致を判断する分類器を作成することにより，質問の型の分類を行なわずに【尺度 2】を判定している．この手法では回答の範囲を先に決め (1 段落)，得られた解候補を質問と型が一致するかどうかで分類するため，解候補の範囲を質問に応じて動的に変更できない．

Soricut ら [3] は英語 non-factoid 型質問応答において，FAQ サイトの質問・回答事例集合をパラレルコーパスとみなし，回答が質問に「書き換え」られる (*translation*) 確率を計算するという，質問の型に依らない手法を提案している．この手法でも回答の範囲を 3 文に固定しているため，水野ら [4] の手法と同様の問題がある．また，質問の長さから回答の長さ (語数) を推定する必要があるのが難点である．

2.4 提案手法の位置付け

我々の提案手法では，日本語 non-factoid 型質問に対する回答として，2.1節で述べた二つの尺度の各々について高い値を持つパッセージを解候補として Web 文書から抽出する．【尺度 2】の見積もりには Q&A コミュニティサービスの質問・回答事例集合をコーパスとして用いるが，入力された質問に適合する回答の特徴表現をコーパスから動的に取得するという点が水野ら [4] や Soricut ら [3] の手法と異なる．

先行研究に対する優位点としてはまず，回答の範囲を質問に応じて動的に決められることが挙げられる．次に，Soricut ら [3] の手法では【尺度 1】と【尺度 2】の区別をせずに質問と回答中の語の対応をコーパスから学習しているのに対し，我々の提案手法では【尺度 2】に関係する特徴表現のみをコーパスから取得するた

め，コーパス中での質問や回答の内容 (【尺度 1】に相当) の網羅性を考慮しなくてよいという利点がある．また，Soricut ら [3] の手法では質問の長さから回答の長さ (語数) を推定する必要があるが，提案手法においてはその必要が無い．

3 質問・回答事例集合の分類

提案手法に先立って試みた手法について説明する．Q&A コミュニティサービスの質問・回答事例集合を質問・回答双方の記述スタイルという観点からクラスタリングし，質問応答に利用しようとしたが，あまり良い結果が得られなかった．

3.1 Q&A コーパス

我々は Q&A コミュニティサービスである Yahoo! 知恵袋[1]の質問・回答事例集合を Q&A コーパスとして利用した．このコーパスは，2004 年 4 月から 2005 年 10 月までの間に蓄積された約 311 万件の質問と約 1347 万件の回答から成る．一つの質問には複数の回答が対応するが，回答としては「ベストアンサー」(質問者が選んだ最良回答) のみを使用した．質問とベストアンサーの対を以下では「ペア」と呼ぶことにする．質問が 1 文で構成され，かつ回答文内に URL が表記されていないものを採用した結果，70 万件程のペアが得られた．回答は複数文で構成されるが，余分な情報も多く含まれる．質問の答えは前半に書かれることが多いため，回答が 2 文以上の時は前半分のみを使用した．

3.2 ペアの分類

様々な種類の質問に対応するため，コーパス中のペアを質問の型に応じて分類し，学習データとすることを考えた．質問応答システムに質問が入力された時は学習データ中のいずれかの型に分類し，型ごとに学習データ中の回答文から取得した特徴表現に基づいて【尺度 2】を見積もる．分類する型の種類は表 1に示したように人手で設定することも考えられるが，種類がいくつあるのかは不明なため，クラスタリングにより類似したペアのグループを構成し，そのグループ群に基づいて質問を分類することにした．その際，質問と回答双方の特徴を考慮してクラスタリングするため，質問の類似度と回答の類似度をそれぞれ別の空間で計算し，両者を合わせたものをペアの類似度とした．

通常の内容語によるクラスタリングでは，質問の型ではなく内容によって分類されてしまう．先に述べた【尺度 2】の見積りを実現すべく，記述スタイルによって分類するため，疑問詞や助詞，助動詞等の機能語と一部の内容語 (後述) を表層表現 (読み) のまま用い，その他の語は品詞で代表させることにより，質問と回答の双方を一般化した．表層表現のまま用いた内容語とは，質問の焦点になりやすい名詞 (「理由」「方法」「意味」「違い」等) やコーパス中で出現頻度が高い動詞と形容詞である．文中の語の共起をある程度考慮し柔軟に類似度を計算するため，語のスキップ 2-gram(数語の間隔を許した 2-gram) を素性に用いた．

3.3 クラスタリングの結果と考察

k-means 法により 5 万件のペアを 100 個程度のクラスタに分割した結果を観察すると，質問表現としては「って何」「どうすれば」「違いは何」「どういう意味」「できますか」というような表現を特徴とする様々な型に応じたクラスタが生成されていた．

次に，各クラスタに属する質問文と回答文から特徴表現を χ^2 値によって取り出した．質問文からはある程度分類に役立つと思われる表現を取得できたが，回答文からはあまり有効な表現を得ることができず，この方法を質問応答に適用するのは難しいと判断した．

失敗要因としては，いずれのクラスタ中心からも距離が遠いペアを排除した結果，十分なデータ量が得られなくなった点が挙げられる．記述スタイルが似通ったペアでも適切にマージされないことがあり，素性選択が難しく，類似度計算がうまくいっていないことが窺える．

4 提案手法

提案手法では，入力された質問の型分類を行なわず，コーパス中から記述スタイルが類似する質問を直接収集する．収集した質問に対応する回答文から特徴表現を取得し，【尺度 2】の見積りに利用する．

こうすることで型分類の失敗を考慮する必要が無くなる上に，コーパスのデータ量の多さを活かして記述スタイルが良く似た質問でも十分な量を集められる．「質問の型」に応じた回答の特徴表現を用意しておくのではなく，質問が入力された時点で「その質問」に応じた回答の特徴表現を動的に生成するのである．この方式では，入力された質問に，より適合する特徴表現を見つけることができると期待される．

4.1 Q&A コーパスの前処理

使用するコーパスは第 3章で述べたものと同様である．ただし，後述するように提案手法では質問文中の疑問詞を含む一部分のみに着目するため，使用する質問の文数には制限を設けなかった．疑問詞を含まない質問文も多く存在するが，頻出する「〜を教えて」「〜を知りたい」「〜は？」という表現は「〜は何ですか」という表現に置き換えた[2]．さらに，第 3章のクラスタリングの時と同様，コーパス中の質問文・回答文双方において機能語と一部の内容語を除く表層表現を品詞に置き換えて一般化した．疑問詞を含む質問文のみに限定し，かつ後述する質問文中の 7-gram の出現頻度が極端に少ないものを除いた結果，約 90 万件のペアが得られた．

4.2 記述スタイルが類似する質問の収集

入力された質問と記述スタイルが類似する質問を Q&A コーパスから収集する際には，質問文同士の「疑問詞を中心とする語の 7-gram」の一致の度合を類似度とし，入力された質問と類似度が高い質問を上位 N 件取得する．これは，評価型ワークショップ NTCIR6 QAC4[1] のサンプル質問データ 30 問を分析した結果，上記の 7-gram によって質問の種類をほぼ特定できると判断したためである[3]．システムの入力としては，疑問詞が存在する質問のみを対象とする．

以下の質問が入力された場合の例を示す．

> 「琉球王国のグスク及び関連遺産群」が世界遺産に登録された理由は何ですか。

この質問の場合，7-gram として

[1] http://chiebukuro.yahoo.co.jp/

[2] 不自然な日本語になる場合もあるが，考慮していない．

[3] より一般的な質問の場合はこの限りではない．第 3章ではバリエーションに富んだ質問・回答ペアを分類するためにより広い視点で質問・回答の特徴を捉えることを試みた．

タ_リユウ_ハ_ナニ_デス_カ_<記号, 句点,*,*>

が得られる[4]．コーパス中から 7-gram の類似度が高い質問文を検索し，

消費税込みの値段が表示されるようになった理由は何ですか。

のような質問が収集される．

4.3 回答の特徴表現の取得

4.2節で収集した質問に対応する回答文の集合から特徴表現を取り出す．特徴表現は，収集した回答文集合 (A) とコーパス中の残りの回答文集合 ($\bar{A}$) の間における χ^2 値の高い 2-gram 上位 M 件とする．2-gram b を含む回答文集合を B，2-gram b を含まない回答文集合を $\bar{B}$，全回答数を $n(=|A \cup \bar{A}|)$ とした時，2-gram b の $\chi^2(b, A)$ 値は次のように計算される．

$$\chi^2(b, A) = \frac{n \cdot \left(|A \cap B| \cdot |\bar{A} \cap \bar{B}| - |\bar{A} \cap B| \cdot |A \cap \bar{B}|\right)^2}{|A| \cdot |\bar{A}| \cdot |B| \cdot |\bar{B}|} \quad (1)$$

4.2節の質問例の場合，

表向きは値段を分かり易くするため。

のような回答文の集合が得られる．この集合から式 (1) によって 2-gram の χ^2 値を計算し，以下のような特徴表現とその χ^2 値 (括弧内) が得られる．これに基づいて解候補文のスコアリングを行なう．

タリユウ (705)，タカラ (531)，リユウハ (219)，… ， カラ <記号, 句点,*,*> (113)，カラデス (98)，… ， タメ <記号, 句点,*,*> (42)，タノデ (34)，… ， <動詞, 接尾,*,*><記号, 読点,*,*> (19)，…

4.4 提案システムの処理の流れ

最後に提案システムの処理の流れを説明する．

4.4.1 キーワード抽出と関連語の取得

入力された質問文から内容語を取得し，キーワード集合 K とする．K を名詞キーワード集合 K_n と動詞・形容詞キーワード集合 K_p に分ける．また，K_n 中で複合語を作れる場合は複合語にした場合の集合を K_c とする．

【尺度 1】の見積りに使用する質問内容の関連語を Web から収集する．K_c から 3 つの語の組を取り出し，それぞれの組合せについて Web 検索エンジンにクエリを入力する[5]．クエリは 3 つの語の AND 検索とする．それぞれのクエリに対して，検索結果の要約である snippet の集合が得られる．クエリ q_i に対して得られた snippet の件数[6]を n_i，snippet 中の各語 w_{ij} の snippet 頻度を $freq(i, j)$ とした時，語 w_j の内容関連度 $T(w_j)$ を次式で定義する．

$$T(w_j) = \max_i \frac{freq(i, j)}{n_i} \quad (2)$$

また，キーワード $k \in K$ に関しては，他の語よりも高い重みを与えるために $T(k) = \max_j T(w_j)$ とする．

4.2節の質問例の場合，次のような関連語と関連度 (括弧内) が得られる．

２０００ (0.5)，沖縄 (0.38)，文化 (0.37)，自然 (0.34)， 日本 (0.31)，城跡 (0.25)，里 (0.25)，首 (0.25)，…

4.4.2 文書検索と整形

キーワード及びその複合語から 3 つの集合 K，K_c，$K_c \cup K_p$ を用意する．各集合ごとに，集合内の語の AND 検索をクエリとして Web 検索エンジンに入力する．得られた検索結果から各文書の URL を取得し，HTML 文書をダウンロードする．タグの除去等を行ない，HTML 文書をプレーンテキストに変換する．

4.4.3 解候補の抽出

4.3節で述べた方法で回答文の特徴表現である 2-gram b とその $\chi^2(b)$ 値のリストを取得する．そして検索文書中の文 S_i のスコアを次式で見積もる．

$$\mathrm{Score}(S_i) = \frac{\left\{\sum_{j=1}^{n} T(w_{ij})\right\}^{\alpha} \cdot \left\{\sum_{k=1}^{m} \sqrt{\chi^2(b_{ik})}\right\}^{1-\alpha}}{\log(1 + |S_i|)} \quad (3)$$

ここで，n は文 S_i 中の語 w_{ij} の異なり数，m は文 S_i 中の 2-gram b_{ik} の異なり数，α はパラメータである．

文書中でスコアが高い文が連続した場合，極大値の 1/2 以上のスコアを持つ文をまとめて一つの解候補とし，解候補のスコアは極大値のスコアで代表する．他の文は 1 文で解候補とする．こうして得られた解候補の集合を完全リンク法でクラスタリングし，冗長性を制御する．各クラスタ内でスコアが最大の解候補を取り出し，解として出力する．

4.2節の質問例の場合，次のような解が得られる．

沖縄県に点在する「琉球王国のグスク及び関連遺跡群」、東南アジア、中国、朝鮮半島、日本と経済的政治的交流をもっていたことを示す建造物群であること、失われた琉球王国の遺跡と失われつつある文化的伝統を今に伝える遺産であること、自然崇拝、祖先崇拝という沖縄伝統の信仰形態を今日まで伝えていることなどが評価され、文化遺産に登録されました。

5 評価実験

評価型ワークショップ NTCIR6 QAC4[1] Formal Run のテストセット 100 問を用いて提案システムの性能評価を行なった．テストセット中の質問の型の分布は，定義型質問 (約 2 割) と理由型質問 (約 3 割) で半数を占め，残りは他の型の non-factoid 型質問である．

5.1 実験方法

式 (3) において $\alpha = 0.5$ とした場合 (提案手法:【尺度 1】と【尺度 2】の組合せ) と $\alpha = 1.0$ とした場合 (ベースライン:【尺度 1】のみ) の結果を比較した．

Web 検索エンジンは Yahoo! JAPAN[7] を利用した．一つの質問に対して 4.4.2節で述べたように 3 種類のクエリを生成し，それぞれ 50 件ずつ文書を検索したが，実際に取得できた文書の異なり数は平均 40 件であった．

質問内容の関連語を収集する際の snippet の取得件数は 1 クエリにつき $n_i = 100$ 件，入力された質問と類似する質問のコーパスからの取得件数は $N = 500$ 件，回答の特徴表現の取得件数は $M = 200$ 件とした．

[4] MeCab(http://mecab.sourceforge.net/) で形態素解析した結果の読みと品詞を "_" で接続して表記している．

[5] 検索エンジンに入力する語数が多いとヒット件数が少なくなるため，3 語ずつ組を作る．$|K_c| < 3$ の時は全語を入力する．

[6] 取得件数には上限を設ける．

[7] http://www.yahoo.co.jp/

システムはスコアの降順に解を出力する．今回は上位5件までを評価対象とした．正解判定は人手で行ない，回答の一部に正解と思われる表現が含まれていれば正解とした．評価尺度としてはMRR(最上位の正解順位の逆数の全質問平均) を用いた．正解を1件以上出力できた質問数 (正解質問数) も調査した．

5.2 実験結果

実験結果を表2に示す．システムは質問の型分類を行なわないが，参考までに人手で分類した型ごとに結果を分けて表記する．「その他」の型に分類される質問の書かれ方の例としては，「どうなりますか」「違いは何」「どのような影響」「どういう場合に」等がある．

表 2: 実験結果

	$\alpha = 0.5$ (提案手法)		$\alpha = 1.0$ (ベースライン)	
質問の型	MRR	正解質問数	MRR	正解質問数
定義型	0.643	16/21	0.561	17/21
理由型	0.307	16/33	0.178	11/33
方法型	0.297	4/6	0.067	2/6
その他	0.513	29/40	0.343	24/40
全体	0.459	65/100	0.318	54/100

5.3 考察

実験結果より，提案手法によって【尺度2】を見積もることで回答精度が向上することが分かる．コーパス中の回答文から取り出した特徴表現が解抽出の際に有効に働いていると考えられる．

質問の型ごとの結果を見ると，定義型質問においては両手法間で精度にあまり差が無く，それ以外の型の質問の場合に提案手法を用いることで精度が大きく向上していることが分かる．定義型質問において精度にあまり差が見られないのは，質問で問われいる語の説明に必要な語が【尺度1】を見積もる際の関連語として収集され易く，特徴表現にそれ程頼らなくても解を抽出できるためと考えられる．回答の特徴表現の利用は，定義型以外の型の質問の解抽出において特に有効であるといえる．

以下に成功例を一つ示す．単純なパタンマッチでは抽出できないような質問内容に対する記述を抽出できている．

【質問】シドニーはどんな街ですか。
【回答】オーストラリアの自然と一体化したシドニーは高層ビルが立ち並ぶ中心地からほんの少し足をのばせば、緑の森林、真っ青な海、真っ白な砂浜、透き通るような青い空を一望できてしまう都市です。英語の勉強、レジャーと、何においても多くの選択肢を与えてくれるシドニーで、将来の展望を図りつつ、のどかで優雅なオージーライフをエンジョイすることができる・・・シドニーはそんな都市です。

提案手法で正解を出力できなかった35問について，どの処理段階で失敗したかを分析すると，表3に示す内訳となった．

表 3: 失敗原因の内訳

失敗原因	質問数
質問文からのキーワード抽出の失敗	3/35
文書検索の失敗	9/35
解抽出の失敗	23/35

キーワード抽出の失敗は形態素解析のミスによるものである．提案手法では質問文の型分類を行なわないので，処理の初期段階におけるミスがほとんど無いのが利点といえる．

文書検索の失敗は少なくなかった．正解を含む文書を取得できなければ正解は得られない．検索クエリの生成方法等を再考する必要があるといえる[8]．

解抽出の失敗例を以下に示す．

【質問】ＮＰＯ法が出来ても法人化されていないＮＰＯが多く存在しているのはどうしてですか。
【回答】(省略)... ようになりました。 しかし、これらの団体の多くは、法人格を持たない任意団体として活動しているため、団体の名で銀行口座を開設したり、事務所を借りたり、不動産の登記をすることができず、様々な不都合が生じています。 特定非営利活動促進法（ＮＰＯ法）で定められた...(省略)

【質問】臨界とはどのような状態のことですか。
【回答】超臨界とは液体状態でも気体状態でもない状態で、圧力、温度が臨界点を越えた状態を一般的に超臨界流体と呼びます。

このようにキーワードとその関連語，及び質問に適合する回答の特徴表現のいずれをも含むパッセージでも適切ではない場合がある．これは解候補のスコアリングの式(3) が語や 2-gram の密度という粗い観点でしか見ていないためである．回答精度の向上のためには，文レベルでの類似度の計算，あるいは確率モデル等の導入が必要と考えられる．

6 おわりに

本稿では，Q&A コーパスを用いることで質問の型分類を行なわない non-factoid 型質問応答手法について提案した．評価実験により，提案手法でコーパスから得た回答の記述スタイルに関連する統計情報が解抽出に役立つことを示した．

今後の課題としては，解候補のスコアリング方法の改良や文書検索の改善，入力された質問文とコーパス中の質問文の類似度計算方法の見直し等が挙げられる．

謝辞

本研究の実施にあたり，ヤフー株式会社が国立情報学研究所に提供した Yahoo! 知恵袋データを利用させて頂きました．なお，本研究の一部は文科省科研費特定領域「情報爆発 IT 基盤」(課題番号 19024033) によるものである．

参考文献

[1] Junichi Fukumoto, Tsuneaki Kato, Fumito Masui, and Tatsunori Mori. An Overview of the 4th Question Answering Challenge (QAC-4) at NTCIR Workshop 6. In *Proceedings of the 6th NTCIR Workshop Meeting*, pp. 433–440, 5 2007.

[2] Kyoung-Soo Han, Young-In Song, and Hae-Chang Rim. Probabilistic model for definitional question answering. In *SIGIR*, pp. 212–219, 2006.

[3] Radu Soricut and Eric Brill. Automatic question answering using the web: Beyond the factoid. *Journal of Information Retrieval - Special Issue on Web Information Retrieval*, Vol. 9, pp. 191–206, November 2006.

[4] 水野淳太, 秋葉友良. 任意の回答を対象とする質問応答のための実世界質問の分析と回答タイプ判定法の検討. 言語処理学会 13 回年次大会発表論文集 pp.1002–1005, 言語処理学会, 3 2007.

[8] 実験で使用したテストセットが 1998 年～ 2001 年の毎日新聞を対象にしたものであり，現在の Web 文書向きではないことも原因の一つと考えられる．