任意の型の記述的回答が可能な日本語 Web 質問応答システム

A Japanese Web question-answering system that can answer any class of non-factoid question

石下　円香（いしおろし　まどか・Madoka Ishioroshi）[1]、佐藤　充（さとう　みつる・Mitsuru Sato）[2]、
森辰則（もり　たつのり・Tatsunori Mori）[3]

[1]横浜国立大学大学院 環境情報学府 博士課程後期、[2]同 環境情報学府 博士課程前期[i]、
[3]同 環境情報研究院　教授
{ishioroshi,mitsuru,mori}@forest.eis.ynu.ac.jp

[Abstract]

In this paper, we propose a method of non-factoid Web question-answering that can uniformly deal with any class of Japanese non-factoid question by using a large number of example Q&A pairs. Instead of preparing classes of questions beforehand, the method retrieves already asked question examples similar to a submitted question from a set of Q&A pairs. Then, instead of preparing clue expressions for the writing style of answers according to each question class beforehand, it dynamically extracts clue expressions from the answer examples corresponding to the retrieved question examples. This clue expression information is combined with topical content information from the question to extract appropriate answer candidates.

The experimental results showed that the clue expressions obtained from the set of examples improved the accuracy of answer candidate extraction.

[キーワード]
Web 質問応答システム、Q&A コミュニティサービス、non-factoid 型質問応答

１．はじめに

近年、計算機の高性能化やネットワークの発達等によって電子化された文書が増大しており、大量の文書群から利用者が必要な情報を効率良く得るための情報アクセス技術が必須となっている。情報アクセス技術の一つである質問応答は、利用者の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。固有名や数量を問う factoid 型の質問応答と、定義や理由を問う 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 NTCIR-6 の QAC-4[1]においてタスク設定されるなど、近年広く研究されているものの、依然として難しいタスクである。

non-factoid 型質問は、その正解の内容から、「定義型」、「理由型」、「方法型」といった型に分類することができる。回答の手がかり表現は、大抵それぞれの質問型に特有のものであるので、多くの先行研究では、型分類を行なって個別に処理を行なう手法がとられている。質問の型分類を行なう手法では、型ごとに個別の処理を用意するのが非効率であるし、質問の型分類の精度が解抽出の精度に大きく影響してしまう。さらに、factoid 型質問においては、その型は概念辞書等に記載されるモノのカテゴリによって定義できるのに対して、non-factoid 型質問の型は「定義型」、「理由型」、「方法型」といった典型的な型以外は、包括的に全ての型を定義し、識別することは難しい。そこで、本研究では、Q&A コミュニティサービスの質問・回答事例集合を Q&A コーパスとして用いることにより、質問の型分類を行なわない新しいアプローチを提案する。入力された質問文に対して、記述スタイルが類似する質問事例をコーパスから収集し、対応する回答文の事例集合から回答の記述スタイルに関する特徴表現を動的に取得する。この特徴表現を利用して解抽出を行なう。ここで、Q&A コーパスを利用する目的は、そこから質問の回答を直接見つけるためではなく、同種の質問に対応する回答に特徴的な記述スタイルに関する表現を得るためであることに注意されたい。評価型ワークショップ NTCIR-6 の QAC-4[1]タスクにおける質問文の集合を用いてシステムの性能評価を行ない、提案手法の有効性を示す。

２．研究背景

本節では、non-factoid 型質問応答手法の特徴、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 型の質問は定義や理由、方法等を問う質問であり、数文にまたがる比較的長い文章表現による記述

的な回答が想定される。non-factoid 型質問には様々な種類があるが、大まかに分類すると表 1 に示す通りとなる。

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 (definition)	～とは何 ～って何	～とは…である ～は…のこと
理由型 (why)	なぜ～ ～の理由は何	～ため ～から
方法型 (how)	～にはどうしたらいい ～の方法は何	～するにはまず… ～のやり方は…
その他 (other)	X と Y の違いは何	X は～だが、Y は…
	～したらどうなりますか	～した場合、…
	どのような影響が	～影響を及ぼす。

non-factoid 型質問に対しては、キーワード等による検索によって得られた文書中から抽出してきた数文が解候補となるが、このような解候補の適切性は、

【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度合

の組合せで見積もることができると考えられる。質問の「内容」とは、質問の話題(トピック)を指す。「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。この両者は厳密には独立ではないと考えられるが、本研究においては両者を分離できると近似して話を進める。

上記の二つの尺度で解候補のスコア付けを行なうことを考えた時、【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

本節では、質問の型ごとに個別の処理を行なうアプローチについて関連研究を紹介する。

Han ら[2]は英語の定義型質問応答において、前述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。【尺度 1】に関する確率は検索文書から計算し、【尺度 2】の計算には定義文コーパスを用いている。

磯崎ら[7]は、BACT[3]を用いて原因表現パタンをコーパスから獲得し、得られたパタンや質問文との各種類似度を素性として SVM[4]で採点関数を学習する手法を提案している。

三原ら[8]のシステムでは、検索された文書内から、質問者がすべき行動であると思われる、名詞句と動詞からなる「行動表現」を抽出して回答とする。また、よく聞かれそうな質問に対する方法表現(「X が Y したら、どうする？」等)を含む Web 文書をあらかじめ獲得し、それに対する回答を事前に FAQ としてデータベース化することで処理の高速化を図っている。

以上のように質問の型が限定された場合には、型ごとに難易度の差はあるものの、型に応じた回答表現のパタンや個別のルールを作成したり、専用のコーパスを学習データとして用いたりす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[9]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、質問の型分類を行なわずに【尺度 2】を判定する手法を提案している。トレーニングデータの準備に関しては、正例は対応する質問と回答のペアであるが、負例は質問と、他の質問に対応する回答を組み合わせて人工的に作らなければならない。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[5]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。Feng[6]は英語 non-factoid 型質問応答において、FAQ サイトの質問・回答事例集合を学習データとして用い、(a)質問文中

の各キーワードの重みの算出、(b)質問と回答の間の語彙的結束性のモデル化、(c)質問と回答の間の意味的な結束性のモデル化の三つを学習手法により行なっている。(c)においては、質問型を用いる代わりに、質問文を特徴付ける Question Phrase(QP、質問文の初めの数語)を用い、QP と回答文中に現れる固有表現の統計的な結束性を調べることにより【尺度 2】を判定している。上記の 2 手法では、【尺度 1】と【尺度 2】の両方に関係することを同時にコーパスから学習しているため、質問の内容語の網羅性を保証するために、大量の質問・回答事例集合が必要となる。さらに、回答の範囲があらかじめ決まっている必要があるため、水野ら[9]の手法と同様の問題がある。

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(複数文からなる文書小部分)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[9]や Soricut ら[5]、Feng[6]の手法と異なる。これは機械学習を基にしておらず、情報検索に基づくアプローチを採用している。

水野ら[9]の手法では、新しい Q&A コーパスが利用できるようになった際、再度時間のかかる学習を行なう必要があるが、提案手法ではその必要がなく、不自然な負例を用意する必要もない。Soricut ら[5]や Feng[6]の手法では【尺度 1】と【尺度 2】の両方に関係すること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、【尺度 1】に相当する、コーパス中での質問や回答の内容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[5]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も無い。

さらに、提案手法では、回答の範囲を質問に応じて動的に決められることも利点として挙げられる。

3．提案手法

本節では、本研究における提案手法について説明する。

提案システムは、non-factoid型質問に対する回答をWeb文書から抽出する。提案手法では、質問の型分類や、事前に型毎に特徴表現を用意することは行なわない。質問が入力された時点で記述スタイルが類似する質問をQ&Aコーパスから収集し、対応する回答文集合から回答の特徴表現を動的に生成し、解抽出に利用する。

3.1 提案手法の概略

提案手法の概要を図1 に示す。提案手法では、疑問詞を含む質問文を入力とし、質問の回答として得られたパッセージをそのスコアの降順に出力する。パッセージは文の連続であり、可変長である。回答を得るための情報源は、Web文書である。Web文書の検索には既存のWeb検索エンジンを用いている。それ以外の外部知識として、Q&Aコーパス、ならびに、Web検索エンジンが出力する要約表示であるスニペットの文面を用いる。本手法においても、解候補のスコア付けを2.1 節で述べた以下の二つの尺度(再掲) に基づいて行なうことには変わりがない。

【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度合

提案システムでは、まず、キーワード群を質問文から抽出した後、そのキーワード群に関連する語群をWeb検索結果のスニペットから収集する。質問文中のキーワード群、ならびに、それらの関連語群について、検索文書中での密度を調べることで【尺度1】を見積もる。なお、本論文では質問文内の内容語をキーワードとしている。キーワードの抽出と関連語の収集については、3.6.1 節で詳しく述べる。

【尺度2】については、提案手法では入力された質問の型分類を行なわず、Q&Aコーパスから記述スタイルが類似する質問を検索し、質問事例を収集する。そして、収集した質問事例に対応する回答事例から特徴表現を取得し、【尺度2】の見積りに利用する。すなわち「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」に適する回答の特徴表現を動的に生成するのである。Q&Aコーパスのデータ量の多さを活かすことにより、入力された質問文に記述スタイルが良く似たものに限定したとしても、Q&Aコーパスから十分な量の質問事例と回答事例を集められる。この方式の利点として、以下のものが期待できる。

- 型分類の失敗を考慮しなくてよい。
- 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ため、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。
- Q&Aコーパスは口語表現による質問・回答事例集合であるため、ここから取得した回答の特徴表現を用いることで、同じく口語表現で書かれた個人ホームページやブログ等から回答が抽出できると期待される。

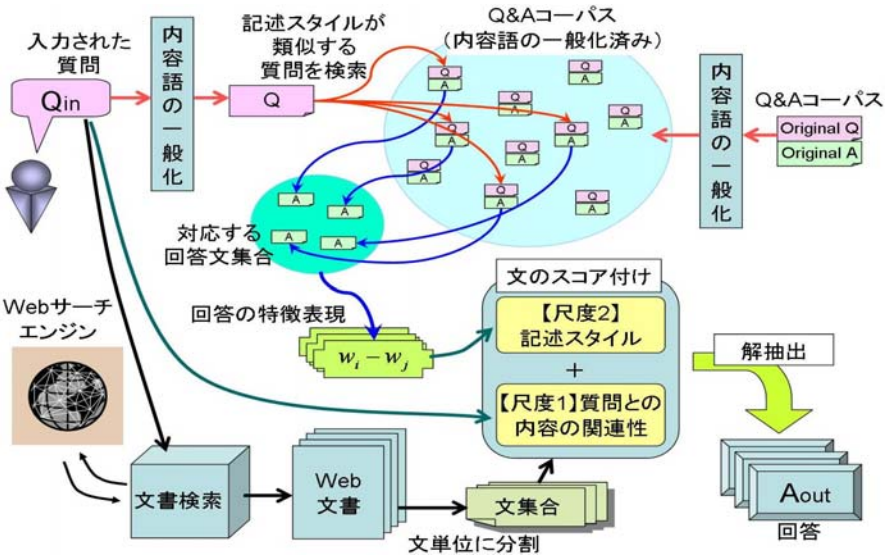

図1: 提案手法の概要

3.2 Q&Aコーパス

Q&Aコーパスとして「Yahoo!知恵袋」[ii]を利用する。Q&Aコミュニティサービスである「Yahoo!知恵袋」には利用者同士によってなされた大量の質問・回答事例集合が存在する。このうち、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1347万件の回答が研究利用のために国立情報学研究所より提供されているので、これを利用する。一つの質問には複数の回答が存在するが、このうち「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「Q&Aペア」と呼ぶことにする。

回答の文書内にURL が表記されているものは、参照先のURL が書かれているだけで、回答そのものは書かれていない場合があるために、使用するQ&Aペアからは除く。一般に一つの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。

同コーパスには疑問詞を含まない質問文も多く存在するが、質問文に頻出する「～を教えて」、「～を知りたい」、「～は?」という表現は「～は何ですか」というように疑問詞を補った表現に置き換えた。この時、不自然な日本語になる場合もあるが、特に考慮はしていない。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、最終的には約90万件のQ&Aペアが得られた。

3.3 Q&Aペアの一般化

記述スタイルのみの情報に着目するために、Q&Aコーパス中の質問文と回答文書の双方について、表現の一般化を行なった。この一般化では、形態素解析の後、機能語と一部の内容語をについて、その表層表現を「読み」に置き変え、一方で、それ以外の語については表層表現を品詞名に置き換えた。読みの情報に置き換える内容語として、疑問詞、頻度の高い名詞、動詞、形容詞を用いた。また、質問の焦点になりやすい語も読みの情報に置き換える内容語とした。

3.4 記述スタイルが類似する質問の収集

入力された質問文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、まず、Q&Aペアの場合と同様の手法により、入力された質問文を一般化する。そして、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文を上位N件取得する。7-gramの一致の度合いは、まずは、疑問詞が同一であることを前提とし、それに加えて7-gram中で同じ位置にある単語が一致する場合に加点をするという方式で求める。7-gramを用いるのは、評価型ワークショップNTCIR-6 QAC-4[1]のサンプル質問データ30問を分析した結果、疑問詞を中心とする語の7-gramによって質問の種類をほぼ特定できるという観察結果が得られたためである。

疑問詞としては、Q&Aコーパス中に出現するものを調査した結果、以下のものを対象とした。この分類は使用する形態素解析器(MeCab[10]) に依存することに注意されたい。

代名詞	ナニ、ドコ、ダレ、ナン、ドチラ、ドレ、ドッチ、イツ、ドナタ、イクツ、ドッカ、イズレ、ナアニ
連体詞	ドノ、ドンナ、ドウイウ、イカナル
副詞	ドウ、ナゼ、ドウシテ、イクラ、イツノマニ
その他	ッテナ、ナニモノ

以下の質問文が入力された場合を例として説明をする。

【質問文】「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。

この質問文の場合、疑問詞を中心とする語の7-gramとして、

タ_リユウ_ハ_ナニ_デス_カ_〈記号、 句点、*、*〉

が得られる[iii]。ここで、Q&A コーパス中で疑問詞を中心とする語の7-gram の類似度が高い質問文を検索すると、

【質問(Q&A コーパス内)】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。

3.5 回答の特徴表現の取得

我々は、ある言語表現*b*が回答に対する特徴表現として有効であるか否かが、前節で収集した質問に対応する回答文書の集合と、その言語表現との間の相関の度合いにより推定できると考えている。語間の関係を表現できる最小のユニットが単語2-gramであるため、本論文では言語表現の単位として単語2-gram *b*を採用する。そして、相関の度合いを測る尺度として、ここでは、式(1)に示される、次の二つの事象に関するχ二乗値 $\chi^2(b)$ を用いる。

事象α： 入力された質問を用いて検索・収集された上位*N*件の質問事例に対応する回答事例であること。この事象の回答事例集合を*A*とする。

事象$\beta(b)$：ある2-gram *b*を含む回答事例であること。この事象の回答事例集合を*B*(*b*)とする。

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

ここで、*n*はQ&Aコーパス中の全Q&Aペア数(=|A∪Ā|)である。また、式(1)に登場する項の説明を表2に示す。今回の場合では、収集した回答文書集合に特有の2-gramほど、式(1)のχ二乗値が高くなる。

表2: χ^2 値の計算

	2-gram *b* を含む回答文書集合	2-gram *b* を含まない回答文書集合	合計
収集した回答文書集合	$\left\|A \cap B(b)\right\|$	$\left\|A \cap \overline{B(b)}\right\|$	$\left\|A\right\|$
残りの回答文書集合	$\left\|\overline{A} \cap B(b)\right\|$	$\left\|\overline{A} \cap \overline{B(b)}\right\|$	$\left\|\overline{A}\right\|$
合計	$\left\|B(b)\right\|$	$\left\|\overline{B(b)}\right\|$	n

3.4 節の質問文の例の場合、

【回答(Q&A コーパス内)】表向きは値段を分かり易くするため。

のような回答を含む回答文書の集合が得られる。この集合から式(1) によって2-gram のχ二乗値を計算すると、以下のような特徴表現とそのχ二乗値(括弧内) が得られる。これに基づいて、次節で述べる方法によって、解候補文のスコア付けを行なう。

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)、…

3.6 提案手法を用いた質問応答処理の流れ

最後に提案手法を用いた質問応答処理の流れを説明する。

3.6.1 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、単名詞の連続が複合名詞を構成することがあるので、K_n中で複合名詞を作れる場合には、それらを複合名詞として登録した場合の集合をK_cとする。

質問文には一般にそれほど多くのキーワードが存在するわけではなく、【尺度1】の見積りには不十分であることがある。そこで、【尺度1】の見積りに使用するために質問内容に関連する語をWebから収集し、関連度の計算を行う。K_cから3つの語の組を取り出し、それぞれの組からブーリアンANDの検索クエリを構成し、Web検索エンジンにおいて検索をする[iv]。それぞれのクエリに対して、検索結果の要約であるスニペットの集合を得る。このスニペット集合中の各語w_jを関連語とする。次にその関連度をスニペット頻度に基づき求める。クエリq_iに対して得られたスニペットの件数[v]をn_i、同スニペット集合中での語w_jのスニペット頻度を$freq(w_j,i)$とする時、語w_jの内容関連度$T(w_j)$を次式で定義する。

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

また、入力された質問文中のキーワード$k \in K$に関しては、関連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。

3.4 節の質問文の例の場合、次のような関連語と内容関連度(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)、…

3.6.2 情報源となる文書の検索と整形

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。

3.6.3 解候補の抽出

3.5 節で述べた方法で回答文書の特徴表現である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のスコアを式(3)で見積もる。

$$\text{Score}\left(S_i\right) = \frac{\left\{\sum_{j=1}^{l} T\left(w_{i,j}\right)\right\}^{\gamma} \cdot \left\{\sum_{k=1}^{m} \sqrt{x^2\left(b_{i,k}\right)}\right\}^{1-\gamma}}{\ln\left(1 + \left|S_i\right|\right)} \quad (3)$$

ここで、l は文S_i中の語w_{ij}の異なり数、mは文S_i中の2-gram $b_{i,k}$の異なり数、γは【尺度1】と【尺度2】の混合比を決めるパラメータである。

式(3) で計算される値は、文内に【尺度1】を満たす語(キーワード及び関連語) や【尺度2】を満たす2-gram(回答の特徴表現) が多いほど高くなる。文内でのこれらの語や2-gramの密度を見るために文の長さで正規化してい

るが、回答として不適切な場合が多い短い文を優遇し過ぎないために文の長さの対数をとっている。

文書中でスコアが高い文が連続した場合、極大値を持つ文を起点としてその前後の文のうち、極大値の1/2 以上のスコアを持つ文をまとめたパッセージを一つの解候補とし(図2)、その解候補のスコアは極大値のスコアとする。他の場合は1文で解候補とする。

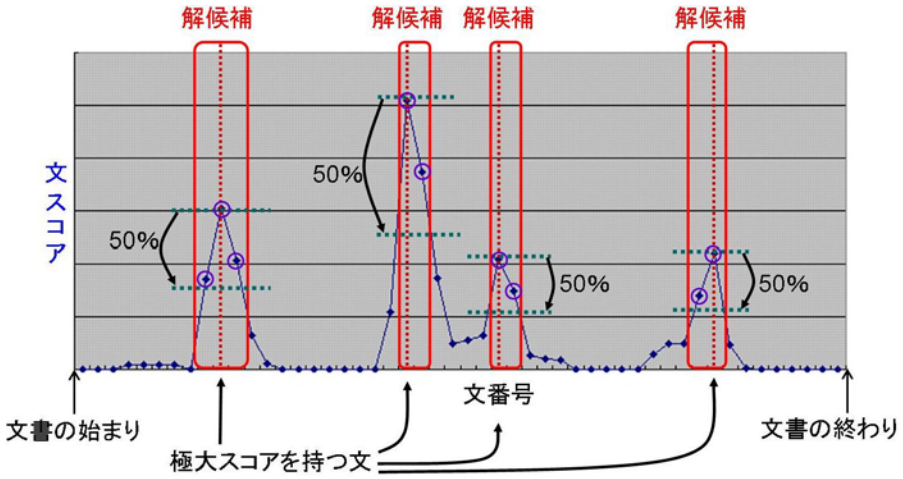

図2: 文スコアと解候補

ここまでの処理で得られた解候補の中には、似通った表現のものが存在する可能性がある。類似した解候補を複数出力するのは冗長であるので、語の頻度を成分とするベクトルのcosine類似度に基づく完全リンク法によってクラスタリングし、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代表解として取り出し、代表解をスコアの降順に並べて出力する。

4. 実験と考察

評価型ワークショップNTCIR-6のQAC-4タスク[1]におけるFormalRunの質問文100問を用いて提案システムの性能評価を行なった。同質問集合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、方法型質問は1割に満たず、残りは他の型のnon-factoid 型質問であった。なお、QAC-4においては、新聞記事を情報源として利用しているのに対し、我々は、Web文書を情報源にしている点に注意されたい。そのため、QAC-4に参加したシステムと、我々の提案手法に基づく手法を直接比較することはできない。

4.1 実験方法

まず、提案手法において、有効な【尺度1】と【尺度2】の混合比を調べるために、式(3)におけるγの値を変えた実験を行なった。次に、ベースライン手法として質問の型分類を行なう手法を用意し、提案手法との精度比較を行なった。

Web 検索エンジンはYahoo! JAPAN[vi]を利用した。一つの質問に対して3.6.2 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。Webから取得できる文書は時間とともに変化するので、一度取得した文書を保存しておき、同一の文書群を用いて実験を行なった。文書の取得は、NTCIR-6のQAC-4 タスク[1]におけるFormal Run のテストセット100 問の1 問目～70問目までは2008年6月11日、71問目～100問目までは2008年6月12日に行なった。

質問内容の関連語を収集する際のスニペットの取得件数は1クエリにつきn_i= 100件とし、関連語の収集は文書の取得と同一の日に行なった。入力された質問と類似する質問のQ&A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。今回は解候補の上位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。型分類する手法との比較では、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。

4.2 実験結果

4.2.1 【尺度1】と【尺度2】の混合比に関する実験

式(3)におけるγの値を0から1まで、0.1刻みで変えた場合のおのおのについて、MRRの平均値を求めた。その結果を図3 に示す。質問の型によって有効な混合比を見るために、全体の結果だけではなく、質問を「定義型」、「理由型」、「方法型」、「その他」に分けた結果も合わせて示す。

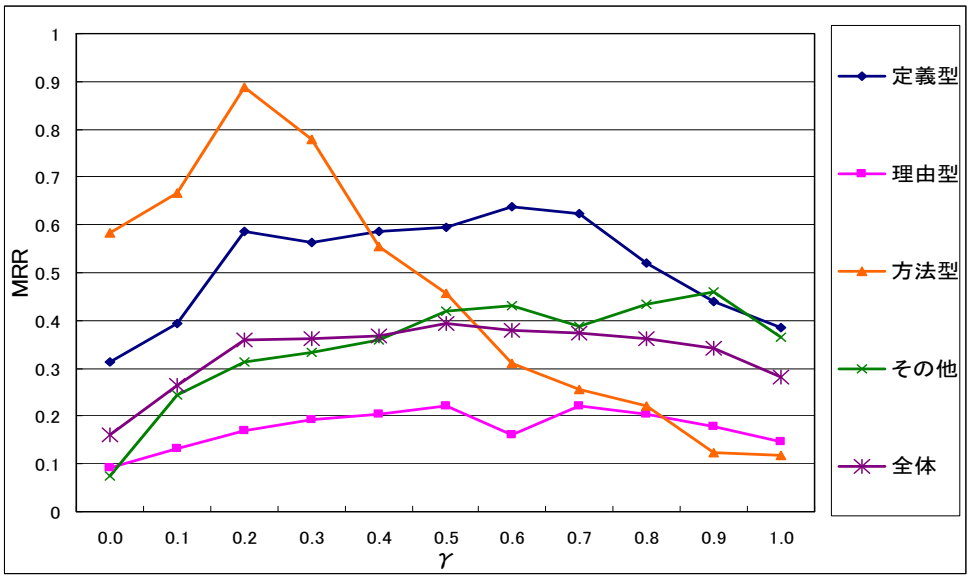

図3: 式(3) におけるγの値を変えたときのMRR

4.2.2 型分類を行なう手法との精度比較

提案手法の有効性を示すために、以下に述べる三つの手法の精度を比較した。

- 提案手法
  式(3) において$\gamma=0.5$ とした場合(【尺度1】と【尺度2】とを均等に混合した場合)。
- 【尺度1】のみ
  式(3) において$\gamma=1.0$とした場合。
- ベースライン手法
  質問の型分類を行ない、定義型か理由型だった場合には回答の抽出パタンによるスコアを、【尺度1】のみのスコアに加えた場合。回答の抽出パタンは新聞記事用に人手で作成したものである。

各実験結果を表3 に示す。

提案手法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。ベースライン手法ではシステムが質問の型分類を行なっており、人手で分類した結果とは必ずしも一致しないことに注意されたい。

4.3 考察

4.2.1 節の結果より、【尺度1】と【尺度2】の混合比γが0.5のときが、最も精度が高いことが分かる。質問の型ごとの結果を見ても、「定義型」、「理由型」、「その他」においてはおおむね【尺度1】と【尺度2】の混合比が同程度のときの精度が良い。しかし、「方法型」では、【尺度2】の混合比を大きくした方が、精度が良くなり、最終的には$\gamma=0.2$の時が最も精度が良かった。「方法型」の質問に対する正解を見ると、正解にはあまりキーワードが含まれておらず、正解に隣接する周辺のテキストにキーワードが含まれているという傾向にあった。そのために、【尺度1】の混合比が小さい方が求解精度が上昇したと考えられる。なお、文書検索の段階で、質問文中のキーワードを利用して検索をおこなっているので、【尺度1】に対応する、キーワードによるトピックの絞込みは文書検索の段階である程度、行われていることに注意されたい。

表 3：型分類を行なう手法との精度比較

	提案手法 ($\gamma=0.5$)		【尺度1】のみ ($\gamma=1.0$)		ベースライン手法 (抽出パタン使用)	
質問の型	MRR	正解質問数	MRR	正解質問数	MRR	正解質問数
定義型	0.595	17 / 21	0.384	13 / 21	0.457	16 / 21
理由型	0.220	15 / 33	0.147	8 / 33	0.199	11 / 33
方法型	0.458	4 / 6	0.117	2 / 6	0.122	3 / 6
その他	0.418	28 / 40	0.364	24 / 40	0.389	24 / 40
全体	0.393	64 / 100	0.282	47 / 100	0.325	54 / 100

また、$\gamma=0.2$～0.9 の時は、【尺度1】のみを用いた$\gamma=1.0$の場合よりも精度が高く、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。Q&A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。

4.2.2 節の結果より、ベースライン手法では、定義型と理由型に対して回答の抽出パタンを用いて【尺度2】を見積もることで、【尺度2】を見積もらない手法よりも上記の二つの質問型に対しては精度がわずかに向上している。「方法型」、「その他」型の場合に精度が落ちているのは、システムによる質問型の誤判定による。ここで用いた回答の抽出パタンは新聞記事向けに作られたものであり、上記の結果より、Web 文書を対象とした時にはあまり有効に機能しなかったことが分かる。提案手法とベースライン手法の「定義型」、「理由型」の精度を比較すると、どちらの場合においても提案手法が勝っており、提案手法で得た特徴表現を用いることによって、人手で作成した抽出パタンでは得られない正解が見つけられていることが分かる。提案手法のMRRと【尺度1】のみやベースライン手法でのMRRとの間に統計的有意差があるかどうか、ウィルコクソンの符号付順位和検定(両側検定)によって求めた。その結果、提案手法とベースライン手法との間には統計的有意差が認められなかった($p=0.091$)が、提案手法と【尺度1】のみの手法との間には有意水準1%で統計的有意差が見られた($p=0.0049<0.01$)。提案手法と【尺度1】のみの手法との間に統計的有意差が見られたことにより、Q&Aコーパスから取得した特徴表現の有効性が示されたといえる。

以下に提案手法での成功例を一つ示す。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

4.3.1 失敗要因の分析

提案手法でγの値を変えても全く正解を出力できなかった25問について、どの処理段階で失敗したかを分析すると、表4 に示す内訳となった。

表 4：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	1 / 25
文書検索の失敗	6 / 25
解抽出の失敗	18 / 25

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。また、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。

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】平成１０年に制定された特定非営利活動促進法が、ＮＰＯ法と略称で呼ばれているため、ＮＰＯはＮＰＯ法人格を取得した団体（特定非営利活動法人、通称ＮＰＯ法人）のことだと思われる方も

【質問】臨界とはどのような状態のことですか。
【回答】超臨界流体とは、気体と液体が共存できる限界の温度・圧力（臨界点）を超えた状態にあり、通常の気体、液体とは異なる性質を示すユニークな流体です。

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付けの式(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。

５．結論

本研究では、Web文書を情報源とする日本語non-factoid 型質問応答について検討をした。

提案手法は、Q&Aコミュニティサイトの質問・回答事例集合をQ&Aコーパスとして用いることで質問の型分類を行なわないという特徴がある。入力された質問と記述スタイルが類似する質問を直接Q&Aコーパスから探し出し、対応する回答文集合から回答の特徴表現を動的に生成し、解抽出に利用する。評価型ワークショップNTCIR-6のQAC-4タスクにおける質問集合とWeb文書を用いた評価実験により、提案手法によってQ&A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、解候補のスコア付け方法の改良、文書検索の改善、入力された質問文とQ&Aコーパス中の質問文の類似度計算方法の見直し、Q&Aコーパスの利用方法の工夫などがあげられる。

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] Fukumoto,J. ,Kato,T. ,Masui,F.,and Mori,T. (2007). “An Overview of the 4th Question Answering Challenge (QAC-4) at NTCIR Workshop 6.” In *Proceedings of the Sixth NTCIR Workshop Meeting*, pp. 433–440.

[2] Han, K.-S.,Song, Y.-I.,and Rim, H.-C. (2006). “Probabilistic model for definitional question answering.” In *SIGIR*, pp. 212–219.

[3] Kudo, T. and Matsumoto, Y. (2004). “A Boosting Algorithm for Classification of Semi-Structured Text.” In *EMNLP*, pp. 301–308.

[4] Joachims, T. (2002). “Optimizing Search Engines Using Clickthrough Data.” In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining*.

[5] Soricut, R. and Brill, E. (2006). “Automatic Question Answering Using the Web: Beyond the Factoid.” *Journal of Information Retrieval - Special Issue on Web Information Retrieval*,9, 191–206.

[6] Feng, J. (2008). “Question Answering with Question Answering Pairs in the Web.” In *WWW2008*.

[7] 磯崎秀樹、東中竜一郎(2008) “パターンマイニングを用いて「なぜ」に答えるシステム” 言語処理学会14 回年次大会発表論文集pp.1025–1028、言語処理学会

[8] 三原英理、藤井淳、石川徹也(2005) “行動表現に着目したヘルプデスク指向の質問応答” 言語処理学会第11回年次大会、 言語処理学会

[9] 水野淳太、秋葉友良(2007) “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

[10] 工藤拓(2007) *MeCab: Yet Another Part-of-Speech and Morphological Analyzer*, http://mecab/sourceforge/net/

---

[i] 現在、ヤフー株式会社所属

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

[iii] Mecab[10]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。

[iv] 検索エンジンに入力する語数が多いとヒット数が少なくなるため、3 語ずつの組を作る。$|K_c|<3$ の時は全語を入力する。

[v] 取得件数には上限を設ける。

[vi] http://search.yahoo.co.jp/