質問応答システムのための Web 掲示板からの質問記事抽出

横田 隼 †　　酒井 浩之 †　　増山 繁 ‡

豊橋技術科学大学 知識情報工学系

E-mail: † {yokota, sakai}@smlab.tutkie.tut.ac.jp

‡ masuyama@tutkie.tut.ac.jp

1　はじめに

現在利用されている検索サイトのほとんどがキーワードを用いた検索システムを用いている．そのような検索システムでは，ユーザが質問内容を適切にキーワードで表現できなければ，ユーザが想定していた質問に対して的確に回答することはできない．

しかしながら，ユーザにとって質問内容を適切にキーワードで表現するのは熟知した分野を対象とする場合以外は困難である．そこで，実用レベルの品質の結果を出す自然言語文質問応答システムを利用したサイトが存在すれば，ユーザが質問内容を表現する負荷が軽減される．しかし，現在の自然言語文質問応答システムのほとんどは実験段階である [1].

本研究では知識源として Web 掲示板を用いる自然言語文質問応答システムを構築することを目標とする．Web 掲示板を知識源として用いる理由は，多数の質問と回答が多分野にわたり存在するためである．そこで，質問と回答を大量に収集し，かつ質問と回答を対応付けることにより，ユーザが入力した質問と類似した質問を検索することが可能となり，適切な回答を従来手法よりも容易に取得することが可能であると考えている．

本研究で構築するシステムを実現するために必要となる要素技術が多くあるため対象となる問題を分割し，以下に示す順序で研究を進めていく．また，図 1 にこのシステムの概要図を示す．

1. Web 掲示板からの質問記事抽出
2. 質問記事と回答記事との対応付け
3. 質問記事とユーザ質問の類似質問検索及び回答生成

今回はその第一段階として Web 掲示板からの質問記事抽出の研究を行なった．

なお，質問記事とは疑問文や依頼文（何かを依頼している文）が一つでも存在する記事のことを指す．依頼文を含む記事を質問記事として扱う理由は，記事中に疑問文がなく，「〜で困っています．よろしくお願いします.」というような箇所がある記事が多々存在するからである．

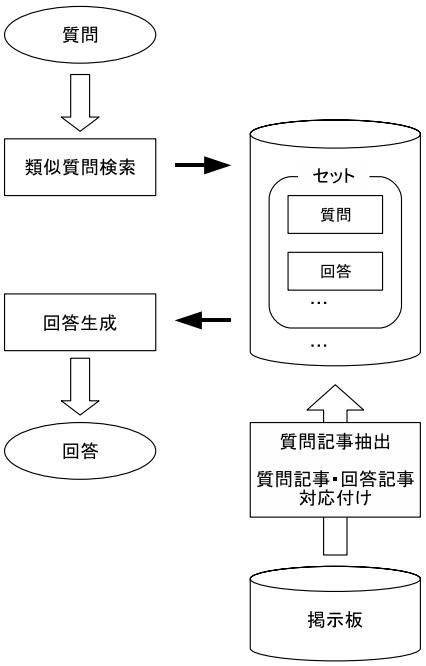

図 1　システム概要図

2　関連研究

本研究の関連研究として質問応答タスク [2] がある．一般に質問応答タスクとはユーザが入力した質問に対しての回答を新聞記事などのコーパスから取得するタスクを指すことが多く，本稿でもそれを指すものとする．このタスクで行なわれている研究のほとんどが質問のタイプを限ったものであり，回答として名称を求めるもの，事実に基づく回答を求めるもの，原因や動機を求めるもの等がある [3, 4].

本研究は質問応答タスクとは以下の点で異なって

いる．それは，質問応答タスクが質問に対応する回答をコーパスから取得するタスクであるのに対して，本研究のタスクは質問記事と回答記事の対応をとり，それら記事群をデータベースとして保存しておき，ユーザの質問に対する回答をデータベースの中から検索する．質問と回答の対応のとれたものをデータベースとして用いることが本研究と質問応答タスクの研究とで大きく異なる点である．

質問記事と回答記事の対応がとれたものをデータベースとして用いることにより，ユーザの質問と類似した質問をデータベース内から探すことで回答を得られると考えている．このように回答を探すことで、質問のタイプに応じて回答を探すための手法を変えずにすむという利点も生まれる．

また，Web に存在する FAQ サイトを人手で集め，質問を検索できる “FAQ Finder” [5] というシステムが存在する．本研究の最終目的は，“FAQ Finder” と同じように質問を検索し，回答を得ることである．ただし，質問記事と回答記事を人手ではなく掲示板から自動的に収集する点が “FAQ Finder” と根本的に異なる．

3 提案手法

本研究で提案する手法は掲示板の全記事に対して，学習器を用いて質問記事とそれ以外の記事とに分類する手法である．そして，質問記事であると分類された記事群を抽出結果とする．なお，本手法は学習器として SVM(Support Vector Machine) を用いた．なお，本手法で扱う記事は前処理として，括弧内の文の削除，HTML タグの削除，URL の削除，改行の削除を行なった記事である。

3.1 素性選択

予備的に Web 掲示板から人手で入手したデータで検討した結果，質問記事とそれ以外の記事の違いを最も良く表している部分は文末部分であった．

この部分を文末表現と定義し，以下にその表現について示す．そして，文末表現を素性として扱う．

3.1.1 文末表現について

文に対して形態素解析[*1] を実行し，形態素列を取得する．そして，文末に存在する形態素から文頭の方へ向かって波及的に形態素を追加していく．その中で出現確率が高く，かつ，含む形態素数が多い文字列を文末表現とする．

[*1] 評価実験では形態素解析器 “MeCab” を使用した http://chasen.org/ taku/software/mecab/

このようにすることで文中の最後の文節を素性とするよりも素性の種類が少なくなり，素性数が抑えられると考えている．また，文末に使われる記号列は句点，ピリオド，エクスクラメーションマーク，クエスチョンマーク，それらを組み合わせたものなど様々あり，投稿した人によって使用される頻度が異なるので，排除することにした。

3.1.2 文末表現の獲得

以下に示す手法で，文末表現を獲得する．なお，文末の形態素を核形態素，核形態素に形態素を付与して得られる文字列を派生表現と定義する．

Step 1 核形態素に直前の形態素を追加することで得られる派生表現を取得する．このとき，掲示板の全記事中において出現回数が 2 回以下の表現は扱わない．

Step 2 各表現のスコアを式 (1) で計算する．

Step 3 核形態素から 2 回以上形態素を付加して得られた派生表現の中でスコア最大の表現を文末表現とする．

$$Score(m,c) = -\sqrt{pf(m)} \times ef(m)\log_2 P(m,c) \quad (1)$$

$$P(m,c) = \frac{ef(m,c)}{Ne(c)} \quad (2)$$

式 (1), (2) で用いた各関数について以下に示す．

$P(m,c)$: 核形態素 c から派生した表現 m の生成確率．

$ef(m,c)$: 核形態素 c から派生した表現 m の生成回数．

$Ne(c)$: 核形態素 c から派生する表現の総数．

$pf(m)$: 表現 m に含まれる形態素数．

最大スコアの文字列は出現確率が高く，かつ，含む形態素数が多い文字列となる．図 2 にスコアの算出例を示す．この図に示した結果は文末が「したほうがいいでしょうか」となっているときのものであり，「でしょうか」が最もスコアが高い．そのため，例文における文末表現は「でしょうか」となる．なお，この例文を係り受け解析器にかけたとき，文節は “した | ほうが | いいで | しょうか” のように分かれる．

素性値は，同じ素性が該当素性が記事中に存在する数とした．また，質問記事を正例データとし，質問記事ではない記事を負例データとする．

形態素列	スコア
しょう:か	1649.07
で:しょう:か	1961.05
いい:で:しょう:か	1014.55
が:いい:で:しょう:か	1001.06
ほう:が:いい:で:しょう:か	555.31

図 2　スコア算出例 (網掛け部がスコア最大)

3.2　質問記事抽出

上記の素性選択手法を用いて，識別しようとする記事に対応するベクトルを取得する．そして，学習された SVM を用いて，その記事が正例か負例かを判別し，正例と判別されたものを質問記事として抽出する．

4　評価実験

提案手法を評価するために，学習データを用いて学習した SVM[*2]で正解記事を識別し，精度，再現率，負例のみの精度を計るために実験を行なった。測定値に負例のみの精度を加えている理由は，誤認識された記事の存在確率をみるためである．負例のみの精度は，“負例と判断された負例文書の数/正解データの負例文書数” で求めた．

4.1　使用データについて

評価実験で用いた学習データと正解データについて以下に示す．なお，学習データと正解データを別のサイトから取得している理由は学習データで得られた素性が他の Web 掲示板でも有効であるか否かを確かめるためである．

4.1.1　学習データ

Q&A コミュニティ “Okwave[*3]” からデータを取得した．“Okwave” では質問記事と質問記事でない記事があらかじめ分類されているため，その情報を利用し自動的に学習データを作成した．

その結果，質問記事とそうでない記事をあわせて 39,750 件取得できた．これら全てを学習データとした．学習データから本手法を用いて得られた素性の数は 11,829 個であった．

4.1.2　正解データ

“Yahoo! 掲示板[*4]” から取得したデータを手作業で質問記事とそれ以外の記事に分け，対象データを作成した．なお，質問記事は 396 件あり，その他の記事が 631 件あった．

4.2　結果

抽出した結果を表 1 に示す．なお, “ベースライン” としている手法は各記事中の文の最後の文節[*5]を素性としたときの手法である．この手法を用いて得られた素性の数は 66,333 個であった．

表 1　実験結果

手法	精度	再現率	F 値	精度（負例）
提案手法	0.8655	0.3737	0.5220	0.9635
ベースライン	0.8584	0.2556	0.3939	0.9760

図 3 に質問記事として抽出された記事の例を，図 4 に抽出に失敗した記事の例を示す．なお，記事中の改行は削除した．

- 信託銀行がよく分かりません。銀行と信託銀行の違いってなんでしょうか？
- 埴輪と土偶というのは、同じものでしょうか。それとも、使い分けるべきものでしょうか。似たりよったりのものを指しているため同じものだとみえてしまいます。どなたかお願いします。

図 3　抽出記事例（成功例）

- 文法、読解問題が苦手です。アドバイスいただけませんか？
- アパートを借りたのですが、重要事項説明書に、プロパンガスだったのに都市ガスと記載されてあったり、バス・トイレ付きのはずが風呂は共有だったのに何も事前に説明がありませんでした。苦情を言っても「そんな事問題にならない」の一点張り。これってどんなんでしょうか？ちなみに入居後一ヶ月で退去しました。

図 4　抽出記事例（失敗例）

5　考察

両手法を比較すると、提案手法の方が，精度では 0.0071 高くなっており，再現率では 0.1181 高くなっ

[*2] 評価実験では SVM^{light} を使用した
http://svmlight.joachims.org/
[*3] http://okwave.jp/
[*4] http://messages.yahoo.co.jp/index.html
[*5] 係り受け解析器 “Cabocha” を使用した
http://chasen.org/ taku/software/cabocha/

ている．F 値で比較すると提案手法の方がベースラインよりも 0.1281 高くなっている．負例のみの精度では提案手法の方がベースラインよりも 0.0125 低くなっている．

提案手法とベースラインを比較した結果，精度においてはあまり差が出ておらず同程度の精度であるといえる．ただ，再現率で比較すると、提案手法とベースラインの差が明確に出ており，提案手法の方が良い結果を得ている．そのため，提案手法の方が効率良く質問記事を抽出できているといえる．

評価実験で得られた結果を検討してみると，以下に示すような質問記事のみに出現する文末表現が存在した場合に質問記事として抽出されていた．ただし，図 4 の 1 番目の記事にある「いただけませんか？」のように質問文の文末表現の中でもあまり用いられないものしか存在していなかった場合に抽出されていないケースが多かった。

- でしょうか
- よろしくお願いします
- ですか
- 教えてください

特に上記のような文末表現が複数存在している記事は多くの場合質問記事として抽出されていた記事が多かった．しかし，図 4 の 2 番目の記事のように記事内の文章が長く，上記のような文末表現が一つしか存在していない記事の場合は質問記事として抽出されなかった．そのため，質問の前提条件を細かく指定した質問記事や質問と関係ない文が多い質問記事は抽出されていなかった．

このような結果となったのは，一般的な文（「である」，「です」で終わる文）が多く存在する記事が質問記事より質問記事ではない記事に多くあるためであると考えている．そのため，一般的な文が多く存在すると質問記事ではないと判断されていると考えられる．

負例のみの精度はベースラインよりも提案手法の方が落ちているが，0.0125 の差であるため，誤差であると考えている．

6 まとめ

提案手法はベースラインと同程度の精度であったが，再現率が良くなっていた．しかしながら，質問の前提条件を細かく指定した質問記事を抽出できないことが多かった．本研究で構築するシステムにおいてユーザが前提条件を細かく指定した質問を行なうことがあると推測できるため，このような質問記事は抽出できた方が良い．そのため，文章長に左右されない手法を考案する必要がある．現在，一般的な文で用いられる文末表現は素性として用いない手法を考えている．考案中の手法により抽出結果への記事長の依存度が弱まると考えている．また，このような文末表現が排除されることにより使用頻度が低い質問文の文末表現しか存在しなかった場合でもうまく抽出できると考えている．

今回，評価実験の正解データは一人で作成した．質問記事とそうでない記事との分類は人により分類基準にゆらぎがあると考えられるため，複数人でより多くの正解データを作成し，実験を行なう予定である．

また，提案手法に対して，今後 “Yahoo! 掲示板” のデータのみではなく他の掲示板サイトからもデータを収集し手法の頑健性を確かめる予定である．

そして，満足できる結果が得られた後に，次の研究テーマである質問記事とその解答記事との対応付けの研究を行なう．

謝辞

本研究の一部は文部科学省 21 世紀 COE プログラム「インテリジェント ヒューマンセンシング」および文部科学省研究費特定領域 (B) (2)16092213 の援助により行なわれた．

参考文献

[1] 大量文書に基づく質問応答技術「SAIQA」，http://www.ntt-tec.jp/technology/A107.html

[2] Tsuneaki Kato, Jun’ichi Fukumoto and Fumito Masui. An Overview of NTCIR-5 QAC3. In Proceedings of the Fifth NTCIR Workshop, 2005.

[3] 諸岡心, 福本淳一, “Why 型質問応答のための回答選択手法”, NLC2005-106〜113, pp.7-12, 2006

[4] 森本格行, 福本淳一, “Why 型質問応答のための回答選択手法”, 言語処理学会第 10 回年次大会発表論文集, pp.293-296, 2004

[5] R. Burke, K. Hammond, V. Kulyukin, S. Lytinen, N. Tomuro, and S. Schoenberg. Question Answering from Frequently Asked Question Files. AI Magazine, 18(2):57– 66, 1997.