Web 掲示板を対象とした質問・回答対応の自動抽出手法の提案

鈴木 佑輔，　酒井 浩之，　増山 繁

豊橋技術科学大学 知識情報工学系

{y_suzuki, sakai}@smlab.tutkie.tut.ac.jp, masuyama@tutkie.tut.ac.jp

1　はじめに

質問応答（QA）システムは，ユーザが自然言語文で入力した質問に対して自動的に回答を返すシステムである．しかし，従来の QA システムでは，対応できる質問や回答の種類が限定されており，Why 型質問や How 型質問に対する回答を高精度で返すシステムは実現されていない．

そこで我々は，Web 掲示板における質問記事，及び，それに対応する回答記事によって構成される対（質問回答対応）をあらかじめデータベースに格納しておくことで，質問を入力とし，対応する回答を出力するシステムを構築することを考えた．データベースに同一の質問があれば，それに対応する回答を出力することで，複雑な回答を要求する質問にも対応できる．しかし，Web 掲示板において質問と回答の対応は，あらかじめ取れているわけではない．そのため，Web 掲示板を，そのまま情報源として有効活用するには，人間が全ての記事に目を通し，対応関係を付与する必要があり，多大な人手と時間がかかる．これを自動的に行うことで，効率的に，かつ，大量に質問回答対応を獲得することができると考えられる．

そこで，我々は Web 掲示板を対象とした質問回答対応の自動抽出手法を提案する．Web 掲示板では，多種多様なジャンルの情報交換が行われているため，質問や回答の種類を限定することなく，経験に基づいた情報や，多人数の意見等，様々な，かつ，詳細な情報を獲得することができる．また，質問回答対応を自動的に，かつ，大量に獲得することで人手によるコストを削減し，多くの質問に対応可能である．

2　関連研究

QA システムの例として，佐々木ら [1] が提案する質問タイプ，回答判定を学習させた分類器を用いたシステムや，諸岡ら [2] が提案する Why 型質問に対応したシステム等が挙げられる．これらは共に新聞記事を知識源としており，名称や日付等，事実に基づく質問にしか対応していない．また，佐々木らの提案したシステムの正解率は 55.7% であった．QA システムとして実用化するため，提案手法では，より高い正解率を目標とする．

麻野間ら [3] は，How-to 型質問に対応するために，WWW から検索によって得られた「方法に関する文書」を分析し，回答が質問内容に合致した方法かどうかの評価尺度を付与している．また，付与した評価尺度によって効率良く満足度の高い回答が検索できるかどうかの検証を行っている．この手法は方法説明に特化しているため，全ての質問，回答には対応していない．

Burke ら [4] は，Web 上に存在する，よくある質問（FAQ）を収集し，収集した FAQ の検索を可能としたシステムを提案している．しかし，このシステムは FAQ の収集を人手で行っているため，大規模なシステムを構築するには膨大なコストがかかる．

表 1　情報要求表現 [6] の例（一部抜粋）

項目	表現
確認要求	～ですよね，～ますよね
判定要求	～ですか，～ますか
選択要求	～ですか（それとも）～ですか
説明要求	どんな，どうして
積極的行為要求	～してください
消極的行為要求	～してほしい，お願いします
情報取得希望	聞きたい，知りたい
情報不足表明	分かりません，困っています

提案手法では，Web 掲示板の記事を横田 [5] の手法を用いて質問記事とそれ以外の記事に分類した後，質問記事と回答候補記事の記事対を取得する．得られた記事集合の中から，対応付けを行うことで質問回答対応を獲得する．得られた質問回答対応を自然文検索が可能な QA システムに用いることで，理由，方法等，様々な種類の質問に対応し，効率的に情報を獲得できると考える．

3　用語定義

質問記事　質問記事は，田中 [6] が定義している情報要求表現（表 1）のような何らかの質問が含まれている，図 1 ＜質問記事＞のような記事とする．

回答記事　回答記事は，質問記事に対して何らかの返答をしている図 1 ＜回答記事＞のような記事とする．ただし，「自分で調べろ」のような質問者が具体的な情報を得られない記事は，獲得しても得られる情報が有益とはならないため回答記事としない．

回答候補記事　回答候補記事とは，回答記事になり得る可能性がある記事のことである．記事集合には回答記事になり得る記事とそうではない記事が存在するが，何も行わなければ質問記事ではない全ての記事が回答候補記事となる．一方，後述するスレッドツリーを用いれば，回答候補記事は限定される．

記事対　記事対とは，質問記事と回答候補記事の組み合わせである．質問回答対応である記事対は図 1 のような組み合わせであり，これを記事対の正例とする．逆に，質問回答対応ではない記事対は質問記事と本来対応しない記事の組み合わせであり，これを記事対の負例とする．

スレッドツリー　本研究で対象とする掲示板は，複数のスレッドによって構成されている．スレッドとは，あるテーマについて投稿された，複数の記事から構成される系列である．これらの記事は，それぞれ返信先記事へのリンク情報を保持している．そこで本研究では，スレッドに含まれる記事集合から質問記事 q_i を抽出し，質問記事 q_i を根，リンク情報に基づき回答候補記

＜質問記事＞
対西武戦を西武ドームに初めて見に行くのですが、駐車場はたくさんありますか？またデーゲームの時には何時ごろまでに行けば駐車場にとめることができるでしょうか？

＜回答記事＞
正規の駐車場は早いウチからけっこういっぱいになってしまいますが、コンビニや民家でも貸してくれる場所があります。自分はココ最近は公共機関を利用するのでわからないですが、3 年くらい前に車で行った時は民家で借りました。駐車場の持ち主の人がライオンズの帽子被って旗振って誘導してくれるのですぐわかると思います。

図 1　質問回答対応の例（Yahoo!掲示板より）

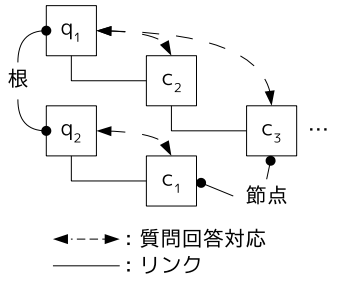

図 2　スレッドツリーの例

事 c_j を節点とした木を構成する．これをスレッドツリーと呼び，上記の手順で作成したスレッドツリーの例を図 2 に示す．

4　提案手法

提案手法では，事前準備（5.1 節），Web 掲示板の記事を質問記事とそれ以外に分類する処理（5.2 節）を行った後，各記事から特徴語を抽出（4.1 節）し，それを用いて質問記事 q_i と回答候補記事 c_j の類似度を計算（4.2 節）する．記事の類似度と記事間の距離による重み（4.3 節）を用いて，対応付けにおけるスコアを算出（4.4 節）し，閾値 γ より大きい記事対を質問回答対応として獲得する．

4.1　特徴語の抽出

記事の内容を特徴付ける上で重要な語である特徴語は，対応付けにおいても有用であると考えた．しかし，助詞や助動詞等の語は多くの記事で用いられており，記事の内容とは無関係であるため，全ての語を特徴語として用いることは好ましくない．そのため，本手法では，名詞，未知語，動詞の原形，形容詞の原形の 4 種類の品詞を特徴語とする．

次に，得られた特徴語に対して重み付けを行う．少数の記事でしか使用されていない語は記事を対応付ける上で特徴的な語であると考え，少数の記事でしか使用されていない語の重みを高くする．一方，多数の記事で使用されている語の重みを低くする．今回は，文書頻度の逆数である IDF（式 1）のみを用いる手法と，該当記事における語の出現頻度である TF を乗算した TF · IDF（式 2）を用いる手法の 2 種類を実験した．

$$\mathrm{idf}(w_i) = \log_{10} \frac{N}{n_{w_i}} \tag{1}$$

$$\mathrm{t\ df}(w_i, a) = \mathrm{tf}(w_i, a)\mathrm{idf}(w_i) \tag{2}$$

ただし，記事集合 $A = \{a_i \mid a_1, a_2, \cdots, a_N\}$，$N = |A|$，$n_{w_i}$ は記事集合 A のうち語 w_i を含む記事数，$\mathrm{tf}(w_i, a)$ は語 w_i の記事 a における出現頻度である．

4.2　記事間の類似度の計算

4.1 節で得られた特徴語とそれらの重みを用いて，質問記事 q_i と回答候補記事 c_j の類似度を計算する．ここで，ベクトル空間モデルに基づいて，記事中の特徴語の重みを要素とした記事ベクトルを作成する．質問記事 q_i に対して質問記事ベクトル $\mathbf{q}_i = [W(w_1, q_i), W(w_2, q_i) \cdots W(w_M, q_i)]^t$ を作成し，同様に回答候補記事 c_j に対して回答候補記事ベクトル $\mathbf{c}_j$ を作成する．ただし，M は特徴語の総数である．

作成した記事ベクトルを用いて，質問記事 q_i と回答候補記事 c_j の類似度を算出する．なお，類似度の算出には cos 尺度（式 3）を用いる．cos 尺度とは，質問記事ベクトル $\mathbf{q}_i$ と回答候補記事ベクトル $\mathbf{c}_j$ の 2 つのベクトルがなす角度であり，この角度が小さい程，類似度が大きいと言える．

$$\cos(\mathbf{c}_j, \mathbf{q}_i) = \frac{\sum_{k=1}^{M} W(w_k, c_j) W(w_k, q_i)}{\sqrt{\sum_{k=1}^{M} W(w_k, c_j)^2} \sqrt{\sum_{k=1}^{M} W(w_k, q_i)^2}} \tag{3}$$

ただし，IDF のみを語の重みとして用いる場合は $W(w_k, a) = \mathrm{idf}(w_k)$ であり，TF · IDF を語の重みとして用いる場合は $W(w_k, a) = \mathrm{t\ df}(w_k, a)$ である．また，$\mathbf{c}_j$ は回答候補記事 c_j から作成した回答候補記事ベクトル，$\mathbf{q}_i$ は質問記事 q_i から作成した質問記事ベクトルである．

4.3　距離による重みの計算

3 節で述べたスレッドツリーの根である質問記事 q_i と節点である回答候補記事 c_j の距離（ホップ数[*1]）を $\mathrm{Dist}(c_j, q_i)$ とする．今回人手で作成した正解データのうち，ランダムに 100 件（記事対の正例 50，負例 50）を抽出し，予備実験として，距離が質問回答対応の有無に与える影響を調査した．その結果，記事対の正例はスレッドツリー上で近い位置に出現することが分かった．また，質問記事 q_i と回答候補記事 c_j の距離が奇数である記事対の方が，偶数である記事対に比べて正例数が多いことが分かった．図 3 に実際の出現分布を示す．

これより，対応付けに用いるためのスコアに距離 $\mathrm{Dist}(c_j, q_i)$ を使用することで，回答候補記事 c_j が質問記事 q_i から遠ざかるにつれてスコアが低くなるような重みを表現する．また，質問記事 q_i と回答候補記事 c_j の距離が偶数である記事対は正例ではないことが多いため，上記の重みに加え，質問記事 q_i と回答候補記事 c_j の距離が偶数である記事対のスコアが低くなるような重み $\mathrm{Even}(c_j, q_i)$ を与えた．ただし，$\mathrm{Even}(c_j, q_i)$ は $\mathrm{Dist}(c_j, q_i)$ が偶数のとき α を，$\mathrm{Dist}(c_j, q_i)$ が奇数のとき 1 を返す関数である．

4.4　質問記事と回答記事の対応付け

4.2 節で述べた記事の類似度，及び，4.3 節で述べた距離による重みを用いて，質問記事 q_i と回答候補記事 c_j から構成される記事対の対応付けにおけるスコア（式 4）を算出する．この

[*1] 質問記事 q_i から，目的の回答候補記事 c_j に到達するまでに通過する枝の数．

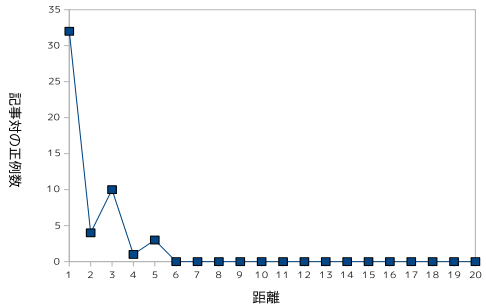

図 3　正解データにおける質問回答対応の出現分布

スコアが閾値 γ より大きい記事対を，質問回答対応として採用する．

$$\mathrm{Score}(c_j, q_i) = \frac{\cos(\mathbf{c}_j, \mathbf{q}_i)}{\mathrm{Dist}(c_j, q_i)^{\beta}\mathrm{Even}(c_j, q_i)} \qquad (4)$$

ただし，　は距離による重みの強さを示すパラメータである．

5　評価実験

Web 掲示板と QA サイトから記事を取得し，人手で正解データを作成（5.3 節）した．評価実験では，2 つのベースライン手法と 3 つの提案手法（5.5 節）に対して評価を行う．

5.1　事前準備

まず，形態素解析[*2]を行う際に使用する辞書の拡張を行う．今回使用する形態素解析器の辞書には普通名詞や一部の固有名詞が登録されているが，製品名や人名，地名等の固有名詞は登録されていない場合が多い．しかし，固有名詞は記事を対応付ける上で特徴的な語になることが多いと考えられる．形態素解析器に，これらも適切に判定させるため，Wikipedia[*3]からページタイトルを抽出し，固有名詞として辞書に登録した．

次に，Web 掲示板からタイトルと記事本文を抽出する．記事本文に加えてタイトルを抽出する理由は，Web 掲示板では重要なことを記事のタイトルに書くパターンも見られるためである．抽出された記事に対して，URL の置換，文字の正規化，文字の半角全角変換を施した後，形態素に分割する．

5.2　記事の分類

Web 掲示板では，記事が質問か，それ以外かに分類されていないため，横田 [5] の手法を用いて，事前に記事を分類した．横田は，あらかじめ記事が質問とそれ以外に分類されている QA サイトに着目し，この特徴を用いて Web 掲示板の記事の分類を試みている．形態素 1-gram と 2-gram の 2 つを組み合わせたものを素性とし，QA サイトの記事を学習データとして SVM[*4]による機械学習を行う．学習した分類器を用いて，Web 掲示板の記事を質問記事とそれ以外の記事に分類する．

なお，今回の実験において，対応付けで使用する正解データは，横田の手法で質問記事と判定され，かつ，人手で質問記事とタグ付けしたものを質問記事，それ以外を回答候補記事としている．これは，横田の手法によるゴミを除去することと，人手で作成した正解データのブレによる影響をできるだけ少なくするためである．

表 2　実験結果（単位：%，括弧内は閾値 γ）

手法	精度	再現率	F 値
ベースライン手法 1	32.9	94.7	48.9
ベースライン手法 2（0.16）	42.0	38.7	40.3
提案手法 1（0.006）	35.8	89.4	51.1
提案手法 2（0.006）	72.4	75.3	73.8
提案手法 3（0.011）	74.6	72.5	73.5

5.3　使用コーパス

今回，Web 掲示板として，Yahoo!掲示板[*5]を使用した．取得した記事を人手で分類した結果，記事の分類における正解データとして，質問記事 601，回答候補記事 1,454 が得られた．

記事の分類で使用する学習データは，人力検索はてな[*6]と OKWave[*7]の 2 つの QA サイトから取得した．その結果，学習データとして，質問記事 86,233，回答記事 846,892 を得た．これらの学習データと横田 [5] の手法を用いた結果，質問記事 344，回答候補記事 1,711 が得られた．

ここから，対応付けにおける正解データを人手で作成し，記事対の正例 320，負例 988 を得た．なお，ここでは 4.3 節で用いたものとは異なるデータを正解データとした．

5.4　回答候補記事の除外

今回の実験では，5.2 節の手法で正例と判定された質問記事を回答候補から除外する．また，質問者と同一ユーザによって投稿された記事は質問者から回答者への再質問，又はお礼である可能性が高いと考え，同様に除外する．

5.5　評価対象となる手法

- ベースライン手法 1
  3 節で述べたスレッドツリーを用いて獲得できる記事対全てを質問回答対応とする手法．
- ベースライン手法 2
  質問記事 q に語 w_1 が出現したときに回答記事 a に語 w_2 が出現する条件付き確率を語の重みとした，記事の類似度を用いる手法．ただし，質問記事から時系列順に近い記事 5 件を回答候補記事とし，スレッドツリーは使用しない．
- 提案手法 1
  IDF のみを語の重みとした記事の類似度を用いる手法．
- 提案手法 2
  IDF のみを語の重みとした記事の類似度と，距離による重みを併用する手法．
- 提案手法 3
  TF · IDF を語の重みとした記事の類似度と，距離による重みを併用する手法．

4.3 節の α は 5，4.4 節における式 4 の　は 3 とした．これらのパラメータは，4.3 節で用いたデータから，F 値が最も大きくなったときの値を選択した．また，ベースライン手法 2，及び，提案手法における閾値 γ は，0 から 0.2 まで 0.001 ずつ変化させたときの F 値が最も大きい値を採用した．表 2 では，そのときのスコアのみ掲載している．

[*2] ChaSen. http://chasen-legacy.sourceforge.jp/
[*3] http://ja.wikipedia.org/
[*4] SVM^{Light}. http://svmlight.joachims.org/
[*5] http://messages.yahoo.co.jp/
[*6] http://q.hatena.ne.jp/
[*7] http://okwave.jp/

表 3　提案手法によって得られた記事対の内訳（単位：%）

距離	P	FN	FP	N
1	93	24	77	8
2 以上	7	76	23	92

6　考察

今回使用したベースライン手法 1 では，再現率が 100% にならなかった．これは，人手で作成した正解データにおいて，回答を行いつつ質問をしている記事や，質問者が結果報告を行っている記事を含む記事対を正例としたが，これらは，5.4 節の手法で除外されたことによって取得できなかったためだと考えた．そこで，5.4 節の手法で除外しない場合のベースライン手法を実験したところ，精度 24.5%，再現率 100%，F 値 39.3 となった．このことから，5.4 節で述べた回答候補から除外する項目は，精度向上において有効であったと言える．

また，IDF のみを語の重みとして用いる提案手法 2 と TF・IDF を語の重みとして用いる提案手法 3 のどちらを使用しても大差ないことが分かった．Web 掲示板の記事は，1 つの記事を構成する文が新聞記事等と比べて少ないため，出現する語が少ない．よって，同じ語が何度も出現する可能性が低いため，TF が有効に作用しなかったのではないかと考えた．

エラー解析[*8]を行ったところ，誤って取得された例として図 4 のような記事対があった．質問記事に対して問い返しを行っている回答候補記事に，質問記事と同じ語が出現したため，質問記事と回答候補記事の類似度が大きくなったと考えられる．このような記事対のうち，質問記事と回答候補記事の距離が小さいものは，距離による重みによって対応付けにおけるスコアが低下しないため，結果として誤取得されると考えられる．

また，取得できなかった例として，質問記事と同じ語が回答候補記事にも出現しているが，質問記事と回答候補記事の距離が 3 である記事対があった．提案手法では距離による重みを用いているため，距離が大きい記事対は取得できなかったと考えられる．

エラー解析で立てた仮定を基に，提案手法によって得られた結果から記事対の内訳を調べたところ，表 3 のようになった．表 3 において，P は取得された記事対の正例，FN は取得された記事対の負例，FP は取得されなかった記事対の正例，N は取得されなかった記事対の負例である．提案手法によって，誤って取得された記事対の負例のうち，距離 1 の割合が 77%，取得できなかった記事対の正例のうち，距離 2 以上の割合が 76% となっている．表 3 より，距離による重みが記事対の正例と負例を判別する手法として有効であることが分かった．その反面，距離による重みが誤取得や取得できない記事対の原因になっていることも分かった．

7　まとめ

本研究では，記事の類似度と距離による重みを用いた質問と回答の対応付けを行い，評価の結果，スレッドツリー，及び，距離による重みを用いることで良い結果が得られた。今後の課題として，6 節でも述べたように，記事の類似度や距離による重みが望ましくない方向へ働いているケースがあるため，記事の類似度や距離による重みの強さ　の決定に工夫が必要である．

＜質問記事＞
ホームページ に BGM を付けたいのですが、プレビュー画面で表示させても音楽がなりません。取り込んだ BGM は他の サイト から ホームページ に使えるものを 圧縮・解凍 したものです。すみませんが アドバイス のほう宜しくおねがいします。

＜回答候補記事＞
チョット聞きたいのですが、他の サイト から ホームページ で使えるものをダウンロードしたと思うのですが、圧縮 したのですか？それとも 解凍 したのですか？それとも 圧縮 した後に 解凍 したのですか？

図 4　誤って取得された記事対の例

上記以外の課題として，QA システムに応用するためには提案手法で得られた精度では十分ではないため，再現率をなるべく低下させずに精度を向上させる手法を考案する必要がある．また，正解データ作成時の人によるブレの影響を減少させるため，複数人で，より多くの正解データを作成する必要がある．さらに，今回の実験では，5.2 節で述べたように横田 [5] の手法で質問記事と判定され，かつ，人手で質問記事とタグ付けしたものを質問記事，それ以外を回答候補記事としているため，Web 掲示板の記事分類から対応付けまでを通した評価実験や質問記事の分類手法について改良を行う必要がある．

[*8] エラー解析は，提案手法 2 の結果について行った．

参考文献

[1] 佐々木裕, 磯崎秀樹, 鈴木潤, 国領弘治, 平尾努, 賀沢秀人, 前田英作. SVM を用いた学習型質問応答システム SAIQA-II. 情報処理学会論文誌, Vol. 45, No. 2, pp. 635–646, 2004.

[2] 諸岡心, 福本淳一. Why 型質問応答のための回答選択手法. 電子情報通信学会技術研究報告, Vol. 105, No. 594, pp. 7–12, 2006.

[3] 麻野間直樹, 古瀬蔵, 片岡良治. How-to 型質問応答の実現に向けた質問回答文書の特徴分析. 電子情報通信学会技術研究報告, Vol. 105, No. 203, pp. 55–60, 2005.

[4] R.Burke, K.Hammond, V.Kulyukin, S.Lytinen, N.Tomuro, and S.Schoenberg. Question Answering from Frequently-Asked Question Files: Experiences with the FAQ Finder System. *AI Magagine*, Vol. 18, No. 2, pp. 57–66, 1997.

[5] 横田隼. Web 掲示板からの質問情報抽出に関する研究. Master's thesis, 豊橋技術科学大学, 2008.

[6] 田中弥生. 電子コミュニケーションにおける「質問表現」の特徴－ Yahoo!知恵袋を対象に－. 社会言語科学会第 22 回大会発表論文集, pp. 114–117, 2008.

[7] 北研二, 津田和彦, 獅々堀正幹. 情報検索アルゴリズム. 共立出版, 2002.