

Web 情報を利用した QA データの自動生成 とベスト回答の選定

早川 晃央[†] 韓 東力[†]

インターネットの普及に伴い、個人が Web 掲示板やブログなどを利用して容易に情報発信を行うことができるようになり、ある商品や組織について評価・評判の情報を収集する際に個人が発信源となっている情報を重視する傾向にある。中でもトピックごとに開設されている Web 掲示板に有用な情報が含まれることが多く、利用者にとって大きな手助けとなっている。しかし、このような情報が不規則な形式で Web 掲示板内に乱雑に置かれているため、有益な情報を素早く取得することは難しく時間を要してしまう。そこで我々は Web 掲示板内の記事を回答記事と非回答記事に分類し、さらに回答記事から質問記事を探索することにより、質問回答対の自動作成を試みた。また、Web 掲示板から直接抽出された質問回答対の集合には、一つの質問に対して複数の回答が付与されているものがあるため、ベスト回答の選定手法を提案した。実験結果によって、本研究で提案した質問回答対の作成手法及びベスト回答の選定手法に一定の有効性があることを確認できたが、十分な精度が得られていないことが今後の課題として残っている。

Automated Generation of QA Data Based on Web Information and Best Answer Selection

AKIO HAYAKAWA DONGLI HAN

Individual information transmission on the Internet has been one of main information resources for general users. More and more users prefer to utilize information provided by individuals on web message boards or blogs rather than particular or formal information when carrying out an investigation on a product or an organization. In this paper, we focus on web message boards where almost all posted messages are expected to be related to a particular topic. A web message board is supposed to contain plenty of useful information. However it is not easy to extract useful information from it as messages are usually posted irregularly in time and format. In order to make the process of information extraction from web message boards more efficient and less time-consuming, we classify all the messages in a web message board into two groups, queries and answers, and then find all the answers for each query to generate a QA entry. During this process, we have proposed a method to select the best answer among all the answers for a particular query. Finally we show the results of some experiments that have been conducted to examine the effectiveness of our methods.

1. はじめに

近年の急激なコンピュータの普及により、誰もが簡単にインターネットを利用できるようになり、インターネットは我々にとって情報収集という点で欠かせない存在となってきた。そしてインターネットの普及は「情報を得る」という目的のみならず「個人が情報を発信する」ということも可能にし、現在では Web 掲示板やブログや SNS などの登場により、個人が情報を発信することはごく当たり前になってきた。ユーザがある商品や組織について評価・評判の情報を収集する際に公式な情報よりもこのように個人が発信源となっている情報を重視する場面が急激に増加している。中でもトピックごとに開設されている Web 掲示板に有用な情報が含まれることが多く、利用者にとって大きな手助けとなっている。しかし、このような情報が不規則な形式で Web 掲示板内に乱雑に置かれているため、有益な情報を素早く取得することは難しく非常に時間を要してしまう。

本研究では Web 掲示板内の情報整理とそれに基づく QA データの自動作成に注目する。関連研究には、Web 上の QA データを利用することで質問回答システムの構築を目指しているものがあるが、QA データとしてすでに質問と回答の対応関係が整理された FAQ サイトや Q&A サイトのものを利用している[1]。また、鈴木らによる研究では、Web 掲示板における質問記事と、それに対応する回答記事によって構成される質問回答対をあらかじめデータベースに格納しておき、質問を入力するとそれに関連する回答を出力するシステムを作成している[2]。ここでも質問記事と回答記事があらかじめ分類されている「OKWave」や「人力検索はてな」などの Web 掲示板を用いている。他にも類似している研究が複数あり[3,4]、いずれも質問回答対がすでに整理されている Web 掲示板を情報源として用いており、その利用目的は質問回答システムの回答内容に利用するというものである。

そこで我々はインターネットに基づく情報収集の一環として、質問記事と回答記事の対応関係が明確でない Web 掲示板のデータを整理し、質問回答対 (QA 対) を作成するシステムを構築し、さらに複数の回答が付与されている QA 対においてベスト回答を選定する手法を提案した。QA 対の作成においては、Web 掲示板内の記事を回答記事と非回答記事に分類し、さらに回答記事から質問記事を探索することにより、質問回答対の自動作成を試みた。ベスト回答の選定においては、投稿記事の「興味度」、「詳細度」、「類似度」と「新鮮度」の指標を定義し、それらを統合的に用いている。

第 2 章では QA 対の作成から表示までの流れを述べ、第 3 章ではベスト回答の選定機能について解説を行い、さらに実験結果を述べる。本研究で用いる Web 掲示板として、「みんなの就職活動日記」[6]を選定した。このサイトは就職活動の支援サイトで

[†] 日本大学大学院総合基礎科学研究科
Graduate School of Integrated Basic Sciences, Nihon University

あり、会員登録をするとサイト内の企業掲示板を閲覧することができる。

2. Web 掲示板からの質問回答対生成

本システムは大きく分けて4つの段階から成っている。以下、各段階についての概要を述べる。図1はシステムの流れを表している。

まず、『記事の収集』では、ユーザに指定された情報をもとに Web 掲示板から記事のデータを収集する。次に、『記事の分類』では、Support Vector Machine (以下 SVM) [7]を用いて、収集された記事の「A&!A」分類を行う。「A&!A」分類については2.2節で説明する。そして、『質問回答対の作成』では、記事のデータと「A&!A」分類の結果を利用して、収集した記事から質問回答対の作成を行う。最後に、『質問回答対の表示』で、作成された質問回答対をユーザに提供するため、表示の処理を行う。

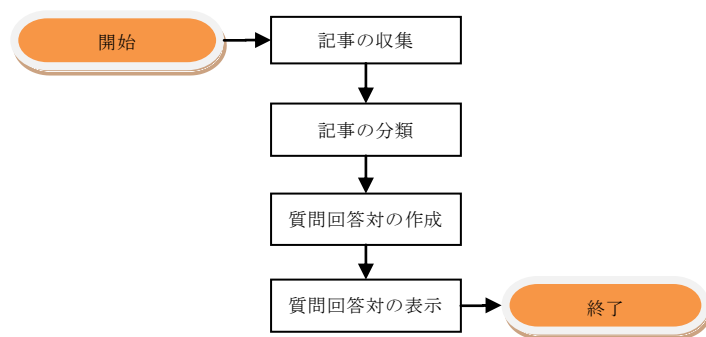


図1 システムの流れ

2.1 記事の収集

記事の収集は「みんなの就職活動日記」の企業掲示板から、ユーザが質問回答対を作成したい任意の企業を指定し、掲示板の記事から投稿日時、投稿者、本文の3つの要素を取得する。

2.2 記事の分類

本研究の研究対象「みんなの就職活動日記」は先行研究のように質問記事と回答記事があらかじめ分類されていないことに加え、本文中に質問と回答が混ざっている記事や独り言など意味を成さない記事も含まれており、明確に質問記事と回答記事に分

類することはできない。

そこで、2通りの分類方法を用いて記事分類を試みる。

- ・方法1 質問と質問以外の記事に分類する
- ・方法2 回答と回答以外の記事に分類する

本研究では、方法1を「Q&!Q分類」、方法2を「A&!A分類」と名付ける。「Q&!Q分類」の定義は、返信が存在する質問記事をQとし、それ以外の記事を!Qと定める。従って、質問記事でも返信をしてもらえずに流されてしまったものは!Qに分類される。一方「A&!A分類」の定義は、質問記事との間の明確な回答関係が認められるかつ、ユーザにとって有意義な情報が含まれている記事をAとし、それ以外の記事を!Aと定める。この場合、独り言、つぶやきの記事は!Aに分類される。また、本文中に質問表現がある記事でも、質問記事に対して明確な回答が含まれていれば、Aに分類する。

この2つの分類手法によりSVMを用いて記事分類を試みる。SVMの素性として日本語形態素を用いる。記事本文をJuman[8]で形態素解析し、同一の形態素が複数存在する場合でも素性値は常に1とし、存在しない場合は素性値を0とする。

任意に選定した7社の記事データ1200件に対して人手でそれぞれQと!QおよびAと!Aを付与したのち、全データをAグループ、Bグループ、Cグループのそれぞれ400件に分ける。そのうちの2グループを学習データに、残りの1グループをテストデータとして扱い、交差検定法で分類実験を行う。またSVMにおける素性を2種類設けた。1つ目は全ての形態素であり、2つ目は助詞を除く全ての形態素である。実験結果は表1の通りである。

表1 分類実験の結果

分類基準	形態素	助詞を除いた形態素
Q&!Q分類	79.18%	79.18%
A&!A分類	91.63%	91.96%

実験結果から、Q&!Q分類よりA&!A分類の精度が高いことが明らかであるが、2種類の素性については有意差が確認できなかった。Q&!Q分類よりA&!A分類の精度が高かった理由としては、曖昧な質問表現がされている質問記事と独り言の記事は似た文面であることが多く、記事単体だけでは質問記事と質問以外の記事かを判別するのは難しいからと考えている。

2.3 質問回答対の生成

Web 掲示板から収集した記事を A と!A に分類したのち、A と判定された記事を回答記事として、それに対応する質問記事を探索する。ここでは、掲示板の性質から異なる A 記事から同一の Q 記事に辿り着く場合もある。そこで荒牧らの研究で提案されたコメント間の関連性を示す 2 つの指標を参考に、記事間の対応関係を示す手がかりの調査を行った。

「みんなの就職活動日記」の Web 掲示板では、投稿記事に対して返信記事を書く場合、自動的に“>[名前]さんへ”という宛名表現が本文の先頭に挿入される。[名前]には、返信先の記事の投稿者名が入る。7 社各 600 件、合計 4200 件の記事の中から A と判定された回答記事 2246 件に対して宛名表現の有無を調査したところ、図 2 のように宛名表現が存在する記事が 2224 件と回答記事全体の 99% を占めていることが分かった。そこで我々は、質問記事の探索に宛名表現を利用できると判断した。ただし、この調査では“>[名前]さんへ”以外の表記も宛名表現と見なしている。

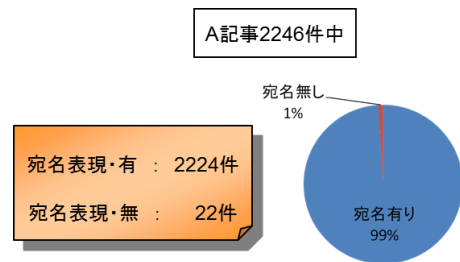


図 2 A 記事の宛名表現の有無

宛名表現を利用するにあたり、宛名表現が存在する回答記事 2224 件を確認したところ、図 3 のように“>[名前]さんへ”以外の表記法が 436 件見つかった。“>[名前]さんへ”以外の表記では、投稿者が“[名前]さん>”などと独自に宛名表現を改編している場合や、“>[名前 1]さん・[名前 2]さんへ”などと宛名が 2 つ以上存在している場合が見られた。これら全てのパターンに対応するため、本研究では全記事の投稿者名を保存した「名前のデータベース」を使用した。宛名表現を調査した結果、いかなる表記法であっても[名前]には“さん”が付いていることが確認された。そこで、名前のデータベースに格納された投稿者名に“さん”を付けたものを回答記事の本文と照合することで、本文から宛名表現を抽出し、質問記事の探索に利用する。

名前のデータベースを利用して本文から抽出された宛名表現から、回答記事から直近の、宛名と一致する投稿者名の記事を探し、該当する記事が存在した場合はその記事を質問記事として質問回答対を作成する。

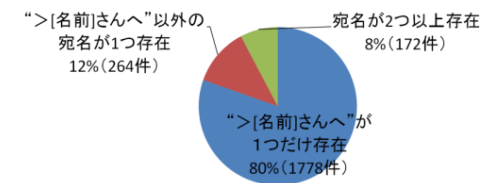


図 3 A 記事の宛名表現の内訳

2.4 質問回答対の表示

質問回答対の生成後、図 4 のように QA 対が表示される。左の赤枠で囲まれたリストには生成された質問回答対の一覧が表示されており、見出しには質問記事の本文が表示されている。また、質問回答対の左のボタンを押すことにより質問記事に連なる記事や、質問回答対の構造を調査することが可能となる。右の緑枠で囲まれたテキストボックスはリストから指定された質問回答対の詳細表示させることができる。確認できる情報は記事の投稿時間、投稿者名、記事の本文である。

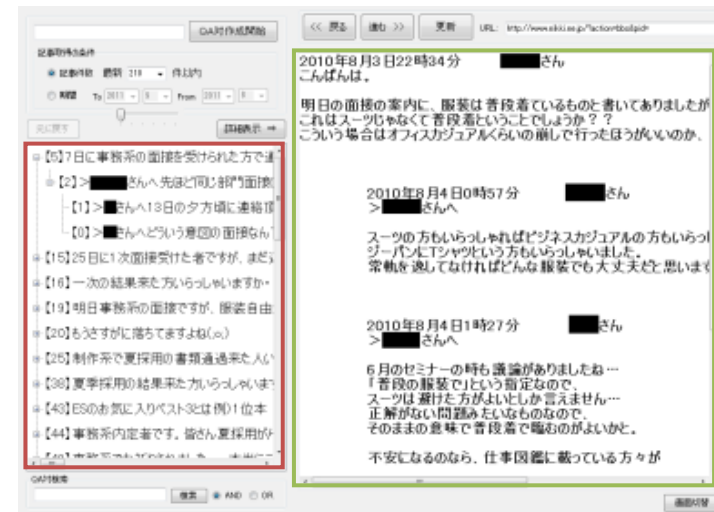


図 4 質問回答対の表示

この表示形式を採用することにより、掲示板に存在する記事の情報を簡単に取捨選択することができ、興味のある情報が存在する場合には質問記事と回答記事がまとまっているため素早く情報を入手することができる。

2.5 評価

本手法により作成された質問回答対の精度を評価する。ランダムで選んだ3社各120件の記事から質問回答対を作成したところ、合計64対の質問回答対が得られた。それらを人手で作成した質問回答対と比較しながら、以下の3段階の基準で評価した結果を表2に示す。

- 【Level 1】 Q記事もA記事の集合も完全に一致する。
- 【Level 2】 質問回答対のQ記事が一致するが、それに対応するA記事の集合が不完全である。
- 【Level 3】 人手によるA記事が全て見つかったが、正解でないA記事も含まれている。

表2 質問回答対生成の精度評価

	Level 1	Level 2	Level 3
再現率	70.77%	80.00%	86.15%
適合率	80.70%	91.23%	98.25%
F値	75.41%	85.25%	91.80%

失敗の主な原因として、質問回答対を作成する際に、回答記事から直近の記事を質問記事として採用する手法にあると考えられる。この問題を解決するためには、回答記事と質問記事と思われる記事との間に関連性があるかを判定した上で、対応する質問記事として扱う必要があると思われる。

次にシステムの有効性に関するアンケートを行った。アンケート項目は「就職活動の支援について」と「システムの使いやすさについて」の2つとなっている。「就職活動の支援について」からは就職活動における情報収集の手助けになると思うか、就職活動の時間短縮に貢献すると思うかの2つの質問となっており、「システムの使いやすさについて」からはインターフェースのデザインについて、機能の充実度についての2つの質問となっている。それぞれの質問を3段階で評価してもらう。また、被験者は本大学において就職活動を行っている3年生と4年生の11人である。

アンケートの結果を表3と表4に示す。

表3 就職活動の支援について

就職活動における情報収集の手助けになると思うか			就職活動の時間短縮に貢献すると思うか		
思う	どちらともいえない	思わない	思う	どちらともいえない	思わない
11人	0人	0人	10人	1人	0人

表4 システムの使いやすさについて

インターフェースのデザインはどうか			機能の充実度はどうか		
よい	どちらともいえない	わるい	よい	どちらともいえない	わるい
6人	4人	1人	6人	5人	0人

アンケート項目の「就職活動の支援について」では高い評価を得ることができ、本システムの有効性を示したが、「システムの使いやすさについて」では「どちらとも言えない」評価が多かったため、システムの操作性を向上させることと機能の充実化を今後の課題とする。

3. 複数回答からのベスト回答選定

ベスト回答選定は質問回答対生成後に行われる。ベスト回答とは、ある記事に対して最もふさわしい返信記事のことをいう。ここでは複数の返信が存在する記事からベスト回答のみを表示するモジュールの作成を目指す。ベスト回答の選定については、機械学習の手法を用いて試みるものもあるが[5]、本研究では学習データの規模などの問題から独自の手法を提案する。ベスト回答の選定処理は3つの手順から構成され、図5は処理の流れを表している。

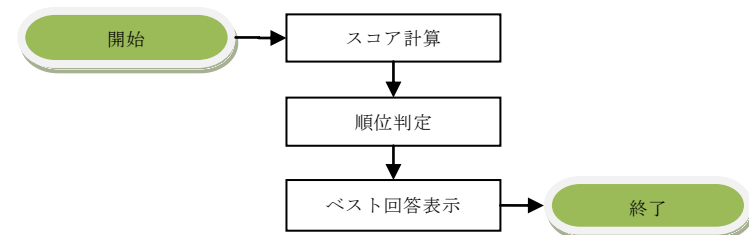


図5 ベスト回答選定までの流れ

まず『スコア計算』では、返信記事毎のスコアを出す。スコアは「興味度」、「詳細度」、「類似度」と「新鮮度」の4つの基準から計算される。各基準の詳細は以下の節で説明する。『順位判定』では、『スコア計算』で出した返信記事の4つの基準の値を記事毎に比較し、基準毎に順位付けをする。その後、基準の総合順位が最も高かった返信記事をベスト回答として採用する。また、総合順位が同率1位の返信記事が複数存在する場合、総合順位1位の返信記事を全てベスト回答に採用する。『ベスト回答表示』では質問回答対中の複数の返信が存在する記事からベスト回答のみを表示する。

3.1 スコア計算

このモジュールでは、前述の通り返信記事毎に「興味度」、「詳細度」、「類似度」と「新鮮度」の値をそれぞれ算出する。ここでは、各基準について説明する。また、以下に頻出する「返信スレッドツリー」とは、返信記事以下に連なるスレッドツリーのことを指す。

3.1.1 興味度

興味度とは、その返信記事に対して閲覧者と投稿者がどれだけ興味を持つかを表したものである。興味度は「返信スレッドツリーの深さ」と「返信スレッドツリー内の記事の投稿者数」の二つの要素からなり、次に示す手段で算出される。ここでは、任意のスレッドツリー中の並列関係を持つ返信記事を「N記事」とし、「N記事」の返信スレッドツリーの深さをグローバル変数 max_dept に保存する。N記事に対する返信記事を「 A_1 記事」、「 A_2 記事」、…、「 A_m 記事」とする(mはN記事に対する返信記事の個数を表す)。また、投稿者の名前を保存するグローバル変数の配列をPとする。以下では興味度の計算手順を示す。

- Step1 始めに、N記事を任意のQA対スレッドツリーから呼び出し、変数Dに1を代入し、変数 max_dept に1を代入する。
(変数D:現在の返信スレッドツリー中の深さを保存する変数と定義する。)
- Step2 N記事に対する返信記事が存在するか確認し、存在しない場合はStep3へ、存在する場合はStep4へ
- Step3 変数Dと変数 max_dept を比較し変数Dの方が大きい場合、 max_dept にDを代入する。
- Step4 A_1 記事、 A_2 記事、…、 A_m 記事の投稿者名を確認し配列Pに投稿者名を保存する。
- Step5 N記事をそれぞれ A_1 記事、 A_2 記事、…、 A_m 記事とし、変数DをインクリメントしてStep2に戻る。このようにStep2～Step5の処理を再帰的に行う。

以上の処理後、以下の式から興味度 (I) を算出する。

$$I = |P| + max_dept \quad (1)$$

|P|は配列Pの要素数を表す。返信スレッドツリーが深ければ深いほど、その質問について掘り下げた情報が多く存在し、返信スレッドツリー内の記事の投稿者が多いほど、多くの閲覧者と投稿者が注目している質問であると考えられる。

3.1.2 詳細度

詳細度とは、返信記事が返信先の質問記事に対し、どれだけ無駄が少なくかつ多くの情報が書かれているかを表したものである。詳細度は「返信記事の文字数」と「返信先の質問記事の文字数」の二つの要素から算出される。具体的な値の算出手順については以下に示す。ここでは、任意のスレッドツリー中の並列関係を持つ返信記事を「A記事」とし、A記事の返信先の質問記事を「 Q_1 記事」、「 Q_2 記事」、…、「 Q_m 記事」とする(mはA記事の返信先の質問記事の個数を表す)。 Q_i 記事の文字数は変数 Q_count に保存し、A記事の文字数は変数 A_count に保存する(iは1以上m以下の任意の整数を表す)。また、詳細度の値は配列valueに保存する。

- Step1 iを1する。
- Step2 A記事と Q_i 記事の文字数をそれぞれ A_count と Q_count に保存する。
- Step3 変数 A_count と変数 Q_count を比較した結果をもとに(2)により詳細度(D)を算出する。
- Step4 $i < m$ であるか確認し、真のときiをインクリメントしStep2に戻る。

以上の処理を終了したとき、配列valueにQ記事に対するA記事の詳細度の値が保存されている。

$$D = \begin{cases} \frac{A_count}{Q_count} & A_count < Q_count \\ 2 - \frac{A_count}{Q_count} & A_count \geq Q_count \end{cases} \quad (2)$$

3.1.3 類似度

類似度とは、返信記事と返信先の質問記事がどれだけ共通の話題や単語を扱っているかを表したものである。類似度(S)は「返信記事の形態素」と「返信先の質問記事の形態素」をもとにコサイン類似度により算出される。具体的な計算は以下の式に従う。

$$S = \frac{C_{AB}}{\sqrt{C_A} \cdot \sqrt{C_B}} \quad (3)$$

ただし、変数 C_{AB} を返信記事と返信先の質問記事間で一致した名詞と動詞の数とし、変数 C_A は返信記事に存在する名詞と動詞の数、変数 C_B は返信先の質問記事に存在する名詞と動詞の数とする。

3.1.4 新鮮度

新鮮度とは、その返信記事が他の返信記事に比べどのくらい新しい情報なのかを表したものである。新鮮度は「返信記事の投稿時間」を基準とする。

3.2 順位判定

ここでは 3.1 節で算出した並列関係にある返信記事内の 4 つの基準値の順位付けを行う。順位の付け方は「興味度」、「詳細度」と「類似度」に関しては値が最も高いものを 1 位とし、「新鮮度」に関しては最新の記事を 1 位とする。以上 4 つの基準の総合順位が一番高かった返信記事をベスト回答として採用する。また、同率 1 位のものが複数存在する場合は、どちらもベスト回答として採用する。

3.3 ベスト回答表示

ベスト回答の選定後は質問回答対からベスト回答以外の返信記事、またベスト回答以外の返信記事に連なる記事は全て除去される。図 6 では左がベスト回答選定前の質問回答対の表示、右がベスト回答選定後の質問回答対の表示となっている。図 5 のようにベスト回答選定後はベスト回答以外の記事は除去されている。

このように表示することで、複雑な構造を持つ質問回答対であっても容易に有益な情報を得ることができ、情報収集の時間短縮にもつながると思われる。

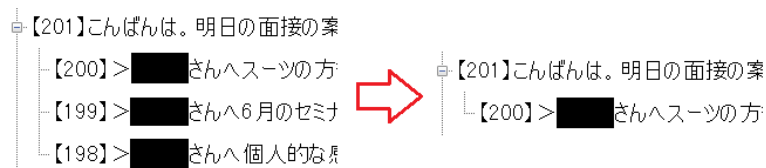


図 6 ベスト回答選定前(左)とベスト回答選定後(右)

3.4 ベスト回答の判定実験

ここでは、人がベスト回答と判定した返信記事を正解とし、現在のシステムが判定したベスト回答と一致するか評価実験を行う。今回の実験では、質問回答対 350 件中に複数回答が存在する返信スレッドツリー 100 件のデータを使用する。また、現在のシステムの性能を計るため、ベースラインとなる手法を二つ設ける。一つ目は返信記事中の最も投稿時間が新しい記事をベスト回答として採用する手法であり、二つ目は最も文字数が多い返信記事をベスト回答として採用する手法である。複数の返信記事がベスト回答に選ばれ、ベスト回答に正解の返信記事が含まれれば正解としてカウントする。表 5 に実験結果を示す。

表 5 複数回答を持つ質問回答対のベスト回答選定実験結果

	データ件数	正解数	正解率
投稿時間が新しい	100 件	46 件	46.00%
文字数が多い	100 件	44 件	44.00%
システム	100 件	55 件	55.00%

実験結果からシステムの手法がベースラインの手法より精度が少し高いことがわかる。なお質問回答対の第一階層目に複数回答が存在する返信スレッドツリーのみを実験データとし、再度実験を行っている。第一階層目とは、図 7 の質問回答対中の太線の枠のように質問回答対の根となる記事とその記事に対する返信記事群のことを指す。また点線の枠は第二階層目を表しており、第一階層目に存在する返信記事とその記事に対する返信記事群のことを指している。

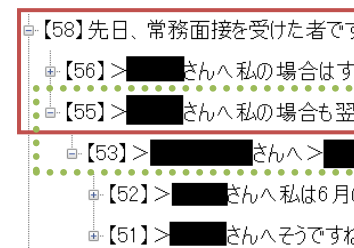


図 7 質問回答対における第一階層目と第二階層目

ここで第一階層目に対してのみ実験を行う理由としては、第一階層目の質問記事には本文中に明確な質問表現が含まれている場合が多く見られ、システムの傾向を調査できるのではないかと考えたためである。また、実験データは前回の実験で扱った100件のデータから第一階層目に複数回答を持つ質問回答対のみを抽出したデータを扱う。表6は実験データの詳細と結果を示す。

表6 第一階層目に複数回答を持つ質問回答対の
 ベスト回答選定実験結果

	データ件数	正解数	正解率
投稿時間が新しい	52件	29件	55.77%
文字数が多い	52件	30件	57.69%
システム	52件	36件	69.23%

表5と表6の結果を比較すると、第一階層目の選定精度が若干高いことが分かる。これは今回、設定した基準が質問回答表現を含む記事に対して上手く作用したからであると考えられる。以上の実験結果から研究手法が有効であることが分かるが、ベースラインの手法と比べて大きな差が見られなかった。その一因として、基準の計算方法は記事の内、必ずしも最適ではない可能性がある。例えば詳細度の計算方法は本文の内容によって変える必要がある。例えば、本文が「連絡が来た方いらっしゃいますか？」という記事に対して返信記事は「はい」や「いいえ」という短文で要求している情報が十分に含まれているが、本文が「面接の内容はどのようなものでした？」という記事に対しては長文でなければ要求している情報が十分に含まれない可能性が高くなるからである。この問題を解決するためには、質問のカテゴリををし、それぞれに対応した計算方法を考案する必要があると思われる。

4. おわりに

本研究では、質問記事と回答記事の対応関係が明確でないWeb掲示板を対象に、掲示板内の記事を回答記事と非回答記事に分類し、さらに回答記事から質問記事を探索することにより、質問回答対の自動作成を試みた。また、複数の回答が付与されている質問回答対については投稿記事の「興味度」、「詳細度」、「類似度」と「新鮮度」を新たに定義し、それらをもとにベスト回答を選定する手法を提案した。実験結果によりそれぞれの研究手法の有効性が確認できたが、ユーザインタフェースの改善が今後の課題として残っている。さらに、関連研究で提案されているように質問の型ごとに

処理を分ける方法[9]を参考に、ベスト回答選定の精度向上を今後の目標と定めている。

参考文献

- 1) 池上敬明, 竹内孔一: “Web上のQAデータの構造の抽出と利用”, 言語処理学会第11回年次大会, C2-7, 2005.
- 2) 鈴木佑輔, 横田隼, 酒井浩之, 増山繁: “Web掲示板からの質問・回答対応の自動抽出”, 人工知能学論文誌 25巻1号 SP-Q, pp 168 - 173, 2010.
- 3) 磯貝直毅, 西村涼, 渡辺靖彦, 岡田至弘: “Q&Aサイトへの質問の作成を支援するための情報の抽出”, 情報社会学会論文誌, Vol.3, No.2, pp.5-14, 2009.
- 4) 横田隼, 酒井浩之, 増山繁: “質問回答システムのためのWeb掲示板からの質問記事抽出”, 言語処理学会第12回年次大会, B2-5, 2006
- 5) 石川大介, 酒井哲也, 関洋平, 栗山和子, 神門典子: “コミュニティQAにおける良質回答の自動予測”, 情報知識学会誌, Vol21, (早期公開), 2011.
- 6) 楽天 みんなの就職活動日記 <http://www.nikki.ne.jp/>
- 7) T. Joachims: “Making large-Scale SVM Learning Practical”. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press. 1999.
- 8) 黒橋禎夫, 長尾真: “日本語形態素解析システムJUMAN version5.1”, 京都大学大学院情報学研究科 2005.
- 9) Tatsunori Mori, Mitsuru Sato, and Madoka Ishioroshi. Answering any class of Japanese non-factoid question by using the Web and example Q&A pairs from a social Q&A website. In Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence (WI-2008), pp. 59--65, 2008.