Web 上の質問応答対の自動抽出に有用な表現

工藤智佳 * 辻慶太 †

Abstract

Web 上の質問応答対を集めたデータベースの自動構築を目指し、それら質問応答対をサーチエンジンでヒットさせるのに有用な言語表現を同定した。まず質問応答サイトと FAQ ページに挙げられている質問応答対を調べ、質問・応答それぞれに偏って出現する表現を特定し、それらを組み合わせて Google での検索精度が高い組合せを同定した。調査の結果、「方が」「できます」といった応答に特徴的な表現を、「どなたか」「あるのですか」といった質問に特徴的な表現と組み合わせる有効性が示された。

1 はじめに

Web 上の質問応答対を集めたデータベースの自動構築を目指し、それら質問応答対をサーチエンジンでヒットさせるのに有用な言語表現を同定した。Web 上の質問応答対は、大きく分けて掲示板、ブログ、質問応答サイト、FAQ ページという 4 種類のサイトに存在している。前 3 者における質問は、実際に一人の質問者が答えを欲したという点で、また FAQ における質問は、潜在的に多くの者が答えを求めているという点で、一定の価値がある質問であり、その応答との対はニーズのある情報と考えることができる。それらを集めて提供する質問応答対データベースは有用であろう。

質問応答対データベースでは、同一の、あるいは類似した質問とその応答達を一括して表示し、利用者に吟味・検討させることが比較的容易に実現できる。また近年 FAQ 等の質問応答対を利用して応答を行う手法・システムがいくつか提案されているが、本研究が目指すデータベースはそうした質問応答システムの情報源として利用することができる。さらにこのデータベースは、質問応答の実際に関する計量言語学的研究の基盤コーパスとしても利用することができるだろう。[1] 本研究はそうした質問応答対データベースの自動構築に向けた第一歩として、Web 上に分散して存在している質問応答対を、サーチエンジンで効果的にヒットさせるのに有用な検索キーワードの同定を目指すものである。

本研究では、まず一定量の質問応答対を手作業で収集し、そこに含まれる表現のうち、主に頻度の点において、一般の Web ページのテキストに比べて、特徴的なものを同定する。次にそれら特徴的表現の組合せをサーチエンジンにかけて、Web 上の質問応答対を高い精度でヒットさせる組合せを同定する。手作業で収集する質問応答対は、(a) 質問応答サイト、(b) FAQ ページ、の 2 つから抽出した。(a) の利用は質問応答サイト、掲示板、ブログにおける対話的な質問応答対の抽出を目的とし、(b) の利用は整理して書かれたモノローグ的質問応答対の抽出を目的とする。

*筑波大学 図書館情報専門学群 s0312144@ipe.tsukuba.ac.jp
†筑波大学大学院 図書館情報メディア研究科 keita@slis.tsukuba.ac.jp

[1] 図書館のレファレンスデータベースは実際ニーズのあった質問を収めているという点で本研究が目指すデータベースに似ている。また応答は専門の図書館員によるもので信頼できる内容である。だが質問のテーマは比較的狭い。それに対して本研究が目指すデータベースは、回答の質は保証されていないものの、幅広いテーマを扱うことになる。

2 先行研究

近年質問応答に関しては様々な研究が行われている。その中には FAQ などの質問応答対を用いて、質問応答を行う手法を提案するものもある（Hammond et al. (1995)、Burke et al. (1997)、堤 & 牛島 (1997)、Sneiders (1999)、松村ら (1999)、陳ら (2005)）。本研究が目指す質問応答対データベースは、そうした質問応答手法の情報源として有用であろう。また横田ら (2006) は掲示板から質問記事を自動抽出する手法を提案している。横田ら (2006) の抽出対象は質問であり、また抽出の手がかりは文末の表現である。それに対して本研究手法は、質問と応答を対で抽出することを目指し、文末に限らない特徴的表現を用いるものである。

3 抽出手法

本研究では、質問応答対を含むページを高精度でヒットさせる「質問応答対自動抽出に有用な表現」を、以下の手法で抽出する。なお以下の説明では質問応答サイトを取り上げているが、FAQ ページの場合も同様である：

(1) 質問応答対に特徴的な表現の抽出：
  (a) 質問応答サイトから質問のテキストと応答のテキストを抽出する。
  (b) Web からテキストを無作為に抽出する。
  (c) 上記 3 つのテキストを構文解析して句単位の言語表現に分割する。
  (d) 上記 3 つのテキストにおける各句に関して、後述の特徴度 C_Q、C_A を算出し、これらが高い句をそれぞれ質問・応答に特徴的な表現とする。
(2)「質問応答対自動抽出に有用な表現」の抽出：
  (a) 質問・応答それぞれに特徴的な表現上位 10 句計 20 句を選ぶ。
  (b) 20 句それぞれを Google で検索し精度を算出し、精度が高い上位 15 句を選ぶ。

(c) 15 句から 2 句の組み合わせ計 105 個を作り、それぞれの精度を調べる。
(d) 高精度の組合せ上位 20 組を選んで、共通の 1 句を含む 2 句の組合せから 3 句の組合せを作り、それぞれの精度を調べる（例えば、「教えて下さい。／宜しく」という組合せと「教えて下さい。／ご存知の」という組合せから、「教えて下さい。／宜しく／ご存知の」という組合せを作る）。
(e) 高精度の組合せ上位 30 組を選んで、共通の 2 句を含む 3 句の組合せから 4 句の組合せを作り、それぞれの精度を調査する。

上記のような句の組合せを用いるのは、組合せ爆発を防ぐ為である。本研究では「精度の低い組合せを含む組合せの精度は低い」ことを仮定し、データマイニングにおけるアプリオリアルゴリズムのアナロジーで、上記のように句を組み合わせた。

先ほどの特徴度 C_Q、C_A は以下のように定義した。まず $m(x,S)$ はテキスト S における句 x の出現頻度、$N(S)$ はテキスト S の延べ句数とする。また Q は質問のテキスト、A は応答のテキスト、Z は Web からの無作為抽出テキストを表すものとする。ここで母比率の差の検定に関する次の尺度 T を導入する：

$$T(x,S_1,S_2) = \frac{m(x,S_1)/N(S_1) - m(x,S_2)/N(S_2)}{\sqrt{p(1-p)(1/N(S_1)+1/N(S_2))}}$$

ただしここで $p = \frac{m(x,S_1)+m(x,S_2)}{N(S_1)+N(S_2)}$ とする。例えば $T(x,Q,Z)$ は「句 x が、無作為抽出されたテキストに比べて、質問テキストに偏って出現する度合い」を表すと言える。本研究では、句 x が質問テキストに特徴的である度合い $C_{Q(x)} = T(x,Q,Z) + T(x,Q,A)$ と定義する。同様に句 x が応答テキストに特徴的である度合い $C_{A(x)} = T(x,A,Z) + T(x,A,Q)$ と定義する。

上記のような和をとらず、$T(x,Q,Z)$、$T(x,A,Z)$ を単独で質問／応答テキストに特徴的な度合いとした場合、質問と応答のどちらにも出現する表現が「応答に特徴的な表現」とされる可能性がある。Web には質問だけ記されたページも多いが、本研究は Web から質問と応答の「対」を抽出しようとするものである。そこで「質問に比べて応答に特徴的な表現」を明確にし、それを検索に用いることで Web から質問のみを抽出してしまう事態を避けようと考えた。上記のように 2 つの T の和を特徴度としたのはその為である。ただしこの点についてはいくつかの検証が必要と考えている。

4 使用データ

まず Web からの無作為抽出テキストとしては、サーチエンジンで「は」を検索語としてヒットしたページのテキストを用いた。ページ数は 938 である。[2] サーチエンジンには Google を、構文解析ソフトには CaboCha を用いた。以下では本研究が用いた質問応答サイトにおける質問応答対（以下「質問応答サイト QA」）と FAQ について述べる。

[2] このような抽出が無作為である保証はないが、その点については今後の課題とする。

4.1 質問応答サイト QA

日本の質問応答サイトとしては、教えて!goo (http://oshiete.goo.ne.jp/)、Yahoo!知恵袋 (http://chiebukuro.yahoo.co.jp/)、OKWave (http://okwave.jp/) などがある。[3] 本研究では OKWave における質問応答対をデータとして用いた。OKWave は質問応答サイトの中では草分け的な存在であり、提携している質問応答サイトも多いので採用した。[4]

OKWave では 11 のカテゴリーを設け、それぞれの下にサブカテゴリーを設けている。本研究では、この 11 カテゴリーと欄外にある「この Q&A コミュニティーについて」というカテゴリーを入れた 12 個それぞれの下にあるサブカテゴリー 2 個ずつから、30 個ずつ質問応答対を抽出することを試みた。全部で 636 個を調査用質問応答対 QA とした。[5]

4.2 FAQ

Google で「FAQ」「Q&A」「よくある 質問」「一問一答」という検索語でヒットしたページのうち、3 個以上 20 個以下の箇条書きの形で質問応答対を提示しているページを本研究では FAQ ページとみなし、そこに含まれる質問応答対をデータとして用いた。全部で 630 個を調査用 FAQ とした。「20 個以下」という制限を設けたのは、少数の特定テーマのサイトが、結果に大きく影響を与えることを防ぐ為である。[6]

5 結果と考察

5.1 有用な表現

3 節で述べた手法によって、質問応答それぞれに特徴的とされた表現の上位 10 個は表 1・2 の通りである。FAQ の方では「年金が」「スパイスの」など、各サイトのテーマと密接に関連した表現がいくつか抽出された。その意味で FAQ の方が質問応答サイト QA よりも、質問応答に真に特徴的な表現というものが少ないのかもしれない。

「質問応答対自動抽出に有用な表現」は表 3 〜 8 の通りである。各表右端の件数は、Google でヒットした上位 20 件中に質問応答対を含むページが何件ヒットしたかを表している。表から分かるように質問応答サイト、FAQ ページともに、2 句、3 句と組み合わせると件数が増え、その意味で抽出精度が高くなっている。だが質問応答サイトに関しては 4 句の組合せにおける精度は、3 句の場合よりも低くなっている。質問応答サイトに関しては 3 つ程

[3] 海外には Yahoo!Answers(http://answers.yahoo.com/)、Wondir(http://www.wondir.com/wondir/jsp/index.jsp)、answerbag(http://www.answerbag.com/) などがある。

[4] OKWave では、質問や回答をする場合には会員登録をすることになっている。登録は無料であるが、これがユーザに責任感を持たせ、信頼できる応答が寄せられていることを本研究では期待している。

[5] カテゴリーによっては 1 個しかサブカテゴリーがなく、また質問応答対が 30 件に満たないサブカテゴリーなどもあった為、総数は 720 になっていない。

[6] サイトによってはクレジットカートに関する 100 近い質問応答対を提示しているものもある。

質問表現	C_Q	応答表現	C_A
よろしく	53.3	方が	21.9
宜しく	30.1	私は	18.2
どなたか	29.0	私の	16.9
ご存知の	23.1	思います。	14.4
方、	22.4	言う	12.9
教えて	22.3	ほうが	12.2
何か	20.7	と	11.4
教えてください。	19.3	あります。	11.3
そこで、	18.5	私も	10.7
最近	18.5	もし	10.3

表 1: 質問応答サイト QA の質問と応答それぞれに特徴的な表現

質問表現	C_Q	応答表現	C_A
どう	46.9	あります。	23.4
Q.	45.1	ことが	23.4
何ですか	37.4	できます。	22.5
どのような	34.5	また、	14.9
あるのですか	33.3	可能です。	14.8
どんな	29.1	なります。	14.6
過ぎましたが、	27.4	著	14.6
年金が	25.8	ことも	14.5
振り込まれていません。	25.3	スパイスの	13.9
＜問＞年金の	25.3	著作物を	13.5

表 2: FAQ の質問と応答それぞれに特徴的な表現

度の句を組み合わせるのが良いのかもしれない。

表 4・5 を見ると、表 1 で応答に特徴的な表現とされたもの（「方が」「私も」など）が含まれていることが分かる。同様に表 7・8 にも、応答に特徴的な表現（「あります」「ことが」など）が含まれている。従って質問応答対をヒットさせるには、質問に特徴的な表現だけでなく、応答に特徴的な表現も組み合わせて検索した方が良いことが言えた。

さて抽出の評価においては一般に、精度だけでなく再現率も検証する必要がある。だが Google を含む多くのサーチエンジンでは、検索結果は 1,000 件程度しか表示されない。従って厳密な再現率検証は難しい。今回のような 20 件ではなく 1,000 件における精度を算出し、その値を評価尺度とするのが現実的と思われるが、作業コストを考え、この点については今後の課題とした。

教えて	どなたか	17 件
教えてください。	言う	17 件
教えてください。	どなたか	17 件
教えてください。	ほうが	17 件
教えて	思います。	16 件
教えてください。	思います。	16 件
教えて	宜しく	15 件
教えてください。	方が	15 件
教えてください。	宜しく	15 件
教えてください。	私も	15 件
教えて	ご存知の	14 件
教えてください。	ご存知の	14 件
教えてください。	と	14 件
教えてください。	何か	14 件
教えてください。	私は	14 件
どなたか	ご存知の	14 件
どなたか	方が	13 件
どなたか	言う	12 件
どなたか	と	12 件
どなたか	方、	12 件

表 3: 質問応答サイト QA に特徴的な 2 句の組合せ

どなたか	教えて	方が	20 件
教えてください。	言う	ほうが	19 件
教えてください。	思います。	ご存知の	19 件
教えてください。	思います。	私は	19 件
教えてください。	思います。	私も	19 件
教えてください。	ほうが	ご存知の	19 件
教えてください。	ほうが	方が	19 件
教えてくださ い。	宜しく	と	19 件
どなたか	教えて	と	19 件
どなたか	教えてください。	と	19 件
どなたか	教えてください。	方が	19 件
どなたか	教えてください。	言う	18 件
どなたか	教えてください。	方、	18 件
どなたか	教えて	方、	18 件
教えてくださ い。	思います。	宜しく	18 件
教えてください。	方が	私は	18 件
教えてください。	方が	私も	18 件
教えてください。	宜しく	方が	18 件
教えてください。	宜しく	私は	18 件
教えてください。	宜しく	何か	18 件

表 4: 質問応答サイト QA に特徴的な 3 句の組合せ

教えてください。	思います。	私も	と	19 件
教えてください。	ほうが	方が	何か	19 件
教えてください。	ほうが	方が	ご存知の	18 件
教えてください。	ほうが	方が	と	18 件
教えてください。	ほうが	ご存知の	と	18 件
教えてください。	ほうが	ご存知の	何か	18 件
教えてください。	思います。	ご存知の	と	18 件
教えてください。	思います。	方が	ご存知の	18 件
教えてください。	思います。	方が	と	18 件
教えてください。	思います。	私も	私は	18 件
教えてくだ	思います。	と	私は	18 件
どなたか	教えてください。	教えて	方、	18 件
どなたか	教えてください。	方、	と	18 件
教えてください。	思います。	方が	私も	17 件
教えてください。	思います。	方が	私は	17 件
教えてください。	思います。	宜しく	と	17 件
教えてください。	思います。	私も	ご存知の	17 件
教えてください。	方が	私も	ご存知の	17 件
どなたか	教えてください。	教えて	と	17 件
教えてください。	ほうが	と	何か	16 件

表 5: 質問応答サイト QA に特徴的な 4 句の組合せ

あるのですか	あります。	20 件
あるのですか	Q.	20 件
あるのですか	できます。	19 件
何ですか	できます。	19 件
あるのですか	何ですか	18 件
あるのですか	また、	18 件
Q.	あります。	18 件
Q.	どのような	18 件
あるのですか	どう	17 件
何ですか	ことが	17 件
Q.	年金が	17 件
Q.	できます。	17 件
あるのですか	著作物を	16 件
あるのですか	どんな	16 件
どのような	あります。	16 件
何ですか	どのような	16 件
何ですか	年金が	16 件
あるのですか	ことが	15 件
あるのですか	ことも	15 件
あるのですか	年金が	15 件

表 6: FAQ に特徴的な 2 句の組合せ

あるのですか	あります。	また、	20 件
あるのですか	できます。	何ですか	20 件
あるのですか	でできます。	また、	20 件
あるのですか	どう	何ですか	20 件
あるのですか	Q.	ありとます。	20 件
あるのですか	Q.	ことが	20 件
あるのですか	Q.	こととも	20 件
あるのですか	Q.	できます。	20 件
あるのですか	Q.	何ですか	20 件
あるのですか	Q.	また、	20 件
何ですか	ことが	でできます。	20 件
Q.	ありとます。	でできます。	20 件
あるのですか	ありとます。	でできます。	19 件
あるのですか	ありとます。	どう	19 件
あるのですか	ありとます。	何ですか	19 件
あるのですか	でできます。	ことが	19 件
あるのですか	でできます。	こととも	19 件
あるのですか	でできます。	どう	19 件
あるのですか	できます。	どんな	19 件
あるのですか	どう	ことが	19 件

表 7: FAQ に特徴的な 3 句の組合せ

あるのですか	あります。	また、	できます。	20 件
あるのですか	できます。	何ですか	また、	20 件
あるのですか	できます。	また、	ことが	20 件
あるのですか	できます。	また、	こととも	20 件
あるのですか	できます。	また、	どどんな	20 件
あるのですか	Q.	あります。	どんな	20 件
あるのですか	Q.	あります。	何ですか	20 件
あるのですか	Q.	できます。	何どう	20 件
あるのですか	Q.	できます。	何ですか	20 件
あるのですか	Q.	できます。	また、な	20 件
あるのですか	Q.	できます。	どこととが	20 件
あるのですか	Q.	できます。	どこととも	20 件
あるのですか	Q.	こととが	どこととも	20 件
あるのですか	Q.	どんな	ことも	20 件
Q.	あります。	できます。	年金が	20 件
あるのですか	あります。	できます。	どどう	19 件
あるのですか	あります。	また、	どう	19 件
あるのですか	あります。	また、	何ですか	19 件
あるのですか	Q.	あります。	できます。	19 件

表 8: FAQ に特徴的な 4 句の組合せ

5.2 ヒットしたページとノイズ

ヒットしたページやノイズのタイプは以下のようになった。

5.2.1 質問応答サイト QA に特徴的な表現

質問応答サイト QA に特徴的な表現では、掲示板、ブログ、質問応答サイト、FAQ ページ等がヒットした。ノイズの主なものは、ニュース、アンケート、インタビュー記事であった。また少数ではあるが、個人間で何かを譲る・譲ってもらいたいと呼びかけているページもノイズとしてヒットした。ニュースには見出しが疑問形で始まるものがあり、これが質問に特徴的な表現と一致してヒットする場合が多かった。同様に、アンケートはアンケートの質問部分に、インタビュー記事はインタビュワーの言葉の中に、それぞれ質問に特徴的な表現が多く含まれてヒットしていた。これらについては、本研究と同様の手法でニュースやアンケート、インタビュー記事に特徴的な表現を特定しておき、それらを論理差で用いることでヒットさせないようにできるかもしれない。

5.2.2 FAQ に特徴的な表現

FAQ に特徴的な表現では、企業のサポートページ、地方公共団体のページ、フリーソフト等のサポートページ、子ども向けのページ、など様々なテーマの FAQ ページがヒットし、また質問応答サイトもいくつかヒットした。企業や地方公共団体のページには「です。」「ます。」といった丁寧な表現が多く、これらは先ほどの表 6 〜 8 にも表れている。また、子ども向けのページには「FAQ」や「Q&A」などの表現はあまり見られず、従ってこれらを検索語としてもヒットしないと思われるページが多かった。即ち、これらのページに関しては、FAQ に特徴的な表現で検索するという本研究手法の有効性が感じられた。ノイズの主なものは、インタビュー、アンケート、ブログ、Q&A 方式で書かれている書籍の紹介、などであった。インタビューやアンケートについては前節で述べた手法でヒットさせないようにできるかもしれない。ただし稀にインタビュー形式の FAQ があり、これと本当のインタビュー記事とを分ける必要が感じられた。

6 おわりに

質問応答対データベースの自動構築を目指し、本研究では、質問応答サイト、FAQ ページから「質問応答対自動抽出に有用な表現」を抽出した。その結果、例えば「どなたか／教えて／方が」といった表現の組合せが、高い精度で質問応答対を含むページをヒットさせることが示された。今後は、ヒットさせたページから、質問応答対を適切に抽出する手法の開発に取り組みたい。具体的にはまず、質問の始めと終わりに特徴的な表現、また同様に応答の始めと終わりに特徴的な表現がないかを調査する。その結果を援用しながら、質問応答対に特徴的な表現を高い割合で含む文を抽出するといった手法の有効性を検証する予定である。また本研究では、サーチエンジンでヒットした上位 20 件における精度のみを考えたが、範囲を上位数百件に拡張し、再現率も視野に入れた形で検証したい。

参考文献

[1] Burke, R. et al. (1997) “Natural Language Processing in the FAQ Finder System: Results and Prospects,” *Working Notes from AAAI Spring Symposium on NLP on the WWW*, p.17-26.

[2] Hammond, K. et al. (1995) “FAQ Finder: A Case-based Approach to Knowledge Navigation,” *Proceedings of the 11th Conference on Artificial Intelligence for Applications*, p.80-86.

[3] Sneiders, E. (1999) “Automated FAQ Answering: Continued Experience with Shallow Language Understanding,” *Question Answering Systems. Papers from the 1999 AAAI Fall Symposium.* Technical Report FS-99-02, p.97-107.

[4] 陳亮ほか (2005) “良くある質問（ FAQ ）のコンテンツ・統語検索を組み合わせた自然言語質疑応答システム”, 情報処理学会研究報告 NL-167(11), p.69-79.

[5] 堤豊, 牛島和夫 (1997) “電子メールを用いた日本語文による質問応答システムにおける類似質問の抽出について”, 自然言語処理 117-22, p.161-166.

[6] 松村真宏ほか (1999) “AAS: 過去の回答の自動組み合わせに基づく質問応答システム”, 電子情報通信学会技術研究報告. AI, 人工知能と知識処理, 98(498), p.17-24.

[7] 横田隼ほか (2006) “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”, 言語処理学会第 12 回年次大会発表論文集, p.332-335.