Q&A サイトを利用した質問の作成を支援するための情報の抽出

谷口 祐亮　小島 正裕　西村 涼　渡辺 靖彦　岡田 至弘
龍谷大学大学院 理工学研究科 情報メディア学専攻
{t11m081,t10m101}@mail.ryukoku.ac.jp, r_nishimura@afc.ryukoku.ac.jp,
{watanabe,okada}@rins.ryukoku.ac.jp

1　はじめに

大量の電子化されたドキュメントからユーザの質問に対する答えそのものを取り出そうとするオープンドメイン質問応答システム (以下 Q&A システム) についての研究がさかんに行われている [1] [2] 。一方、回答を得るのにかかるコストや時間などで不利であるにもかかわらず、Yahoo!知恵袋や[*1]、教えて!goo [*2]、人力検索はてな[*3]、OKWave [*4] などの質問を投稿しておくと他のユーザが答えてくれるコミュニティベース質問応答サービス (以下 Q&A サイト) がさかんに利用されている。例えば、Yahoo!知恵袋は 2004 年から 1 年あまりの間に 300 万件以上の質問が投稿されている。このように、大量の質問が投稿されている Q&A サイトでよい回答を得るためには、回答者にとって回答する負担が少ない質問、特に回答するのに重要な情報が説明されている質問を投稿することが重要である。例えば、以下の (質問 1) は、(回答 1) で指摘されているように、回答するのに重要な情報が説明されていない。

(質問 1)　PC が青い画面になって起動しません。どうしたらいいでしょうか？
(回答 1)　OS は何を使っていますか？

この例では回答が投稿されているが、回答するのに重要な情報が質問で説明されていない場合、回答が投稿されないおそれがある。回答するのに重要な情報 ((質問 1) では「OS は何か」) を説明されていないことが質問を投稿する前に指摘されていれば、この質問を投稿したユーザはその情報を質問で説明していただろう。

そこで本研究では、質問での説明が不十分であるかもしれない情報を指摘するシステムを作成するため、質問作成を支援するための情報を Q&A サイトに投稿された質問と回答から機械学習を用いて抽出し、質問作成支援ができるかの検討を行った。

2　質問作成を支援するための情報

本研究では、Q&A サイトの例として Yahoo!知恵袋を取り上げる。Yahoo!知恵袋は質問のテーマやジャンルごとにカテゴリが用意されている Q&A サイトである。この Yahoo!知恵袋に 2004 年 4 月から 2005 年 10 月までに 286 種類のカテゴリに投稿された約 311 万件の質問と約 1347 万件の回答が国立情報学研究所から公開されている[*5]。この Yahoo!知恵袋のデータのすべてのカテゴリに投稿された質問と回答から質問作成を支援するための情報の抽出を行う。具体的には、以下の文を抽出する。

- 質問から、質問の中心となる文 (重要文)
- 回答から、回答するのに重要な情報を含む文

なお、重要文は、システムを作成する際、回答するのに重要な情報が不足している質問を判定するための手がかりとして利用する。

回答するのに重要な情報が不足している質問とその回答の例を以下に示す。

(質問 2)　最近よく眠れません。グッスリ眠れる方法を教えてください。
(回答 2)　あなたの年齢、職業、部屋の採光とかを書いてくだされば更に詳しく回答できると思います。

(質問 2) と (回答 2) は、「健康、病気、ダイエット」カテゴリにおける例である。(質問 2) は、第 2 文 (「グッスリ眠れる方法を教えてください。」) が重要文である。(回答 2) は、第 1 文 (「あなたの年齢、職業、部屋の採光とかを書いてくだされば更に詳しく回答できると思います。」) が回答するのに重要な情報を含む文である。この文には、質問者の年齢、職業、部屋の採光が回答するのに重要な情報として含まれる。ただし、この (回答 2) には、回答するのに重要な情報が質問で説明されていないことを指摘しているだけで、問題解決の方法については述べられていない。したがって、質問者は指摘された情報を追加して質問を再度投稿しなければならない。もし、年齢、職業、部屋の採光は述べておくのがのぞましいという情報が指摘されていれば、(質問 2–a) のように質問することができる。

(質問 2–a)　25 歳、大学生です。部屋の採光には気をつかっているのですが、よく眠れません。グッスリ眠れる方法を教えてください。

以下に「レシピ、調理法」カテゴリでの質問と回答の例を示す。

(質問 3)　スタバにあるフラペチーノを家で作るにはどうしたら良いですか？
(回答 3)　濃い目に淹れたコーヒー＋氷をミキサーでひたすらクラッシュです。ただし氷に対応したミキサーでないと、刃やモーターが厳しいです。

[*1] http://chiebukuro.yahoo.co.jp/
[*2] http://oshiete.goo.ne.jp/
[*3] http://q.hatena.ne.jp/
[*4] http://okwave.jp/
[*5] http://research.nii.ac.jp/tdc/chiebukuro.html

(質問 3) は、第 1 文 (「スタバにあるフラペチーノを家で作るにはどうしたら良いですか？」) が重要文である。(回答 3) は、第 2 文 (「ただし氷に対応したミキサーでないと、刃やモーターが厳しいです。」) が回答するのに重要な情報を含む文である。この回答では、質問者が氷に対応したミキサーを持っていることを前提に回答をしている。しかし、質問者が氷に対応したミキサーを持っていなければ、(回答 3) の方法は利用できない。もし、氷に対応したミキサーを持っているかどうかの情報が、問題解決の方法を知るのに重要であると指摘されていれば、(質問 3–a) のように質問することができる。

(質問 3–a) スタバのフラペチーノを家で作るにはどうしたらいいですか？氷に対応したミキサーは持っていません。

このように、Yahoo!知恵袋に投稿された質問には重要文、回答には回答するのに重要な情報を含む文が含まれていることがある。

Yahoo!知恵袋に投稿された質問と回答は、カテゴリが異なると使われている単語が異なることや、質問や回答の書き方が変わることがある。例えば、「レシピ、調理法」カテゴリでは、「茹でる」や「チャーハン」といったカテゴリならではの単語が多く使われている。一方、「パソコン、周辺機器」カテゴリでは、「書き込む」や「インストール」といった単語が多く使われている。また、(回答 4) のような問題解決の手順を矢印記号を用いた特有の書き方をしているものがある。

(質問 4) メールを送信する場合、画像を添付して送信すると大きな画像になりますが、どうしたら良いでしょうか。

(回答 4) WindowsXP なら、画像ファイルを右クリック→送る→メール受信者→「イメージをすべて小さくする」にチェックして OK、このやり方でもできますよ。

このような問題解決の手順を矢印記号を用いて書かれている回答は、「パソコン、周辺機器」カテゴリでは使われていることが多い。一方、「レシピ、調理法」カテゴリではあまり使われない。このため、機械学習で用いるための学習データは、カテゴリごとに作ることがのぞましいと考えられる。しかし、使われている単語や文章の書き方に違いがあるからといって、すべてのカテゴリに対して機械学習のための学習データを個別に作ることは、手間と時間がかかるため難しい。ただし、磯貝らは、カテゴリが異なる学習データであっても質問作成を支援するための情報を抽出できることを報告している [3] 。そこで本研究では、学習データを

- 「パソコン、周辺機器」カテゴリ
- 「健康、病気、ダイエット」カテゴリ
- 「レシピ、調理法」カテゴリ

の 3 つのカテゴリから作成する。そして、作成した学習データを用いて、Yahoo!知恵袋のデータのすべてのカテゴリの質問と回答から質問作成を支援するための情報の抽出を行う。

$s1$	対象文の形態素の 1–gram
$s2$	対象文の形態素の 2–gram
$s3$	質問/回答を構成する文の数と対象文の位置
$s4$	対象文を構成する形態素の数
$s5$	対象文以外の文の形態素の 1–gram と対象文との位置関係
$s6$	対象文以外の文の形態素の 2–gram と対象文との位置関係
$s7$	質問を構成する文の形態素の 1–gram
$s8$	質問を構成する文の形態素の 2–gram
$s9$	質問の重要文の形態素の 1-gram
$s10$	質問の重要文の形態素の 2-gram
$s11$	質問と回答の対象文に表れる同一の名詞
$s12$	質問と回答の対象文に表れる同一の名詞の数

図 1　質問作成を支援するための情報を抽出するのに利用する素性

表 1　質問作成を支援するための情報を抽出するのに用いる学習データの内訳

カテゴリ	種類	文数	重要文の数	回答するのに重要な情報を含む文の数
パソコン、周辺機器	質問文	2652	1239	-
	回答文	2000	-	1158
健康、病気、ダイエット	質問文	2411	1250	-
	回答文	2000	-	808
レシピ、調理法	質問文	2458	1330	-
	回答文	2000	-	834

3　質問作成を支援するための情報の抽出方法

質問作成を支援するための情報の抽出には機械学習による方法を用いる。なお、機械学習にはサポートベクトルマシン (SVM) を用いた [4] 。また、SVM には TinySVM の線形カーネルを利用し、ソフトマージンパラメータを 1 とした。

機械学習で利用する素性を図 1 に示す。図中の対象文とは、機械学習による抽出を行う対象となる文のことである。なお、質問から重要文を抽出するのに $s1$〜$s6$ の素性、回答から回答するのに重要な情報を含む文を抽出するのに $s1$〜$s12$ の素性を利用する。また、形態素解析には JUMAN [5] を用いた。

学習データは、Yahoo!知恵袋の「パソコン、周辺機器」、「健康、病気、ダイエット」、「レシピ、調理法」の 3 つのカテゴリに投稿された質問と回答をそれぞれ無作為に取り出して作成した。この学習データの内訳を表 1 に示す。

4　抽出結果と検討

Yahoo!知恵袋のデータのすべてのカテゴリのうち、質問数が多い上位 10 カテゴリの抽出結果を表 2 に示す。表 2 に示すように、質問文数に対して重要文数を抽出できた割合が高いカテゴリは、「Yahoo!知恵袋」カテゴリや「テレ

表 2　質問からの重要文と回答からの回答するのに重要な情報を含む文の抽出結果 (カッコ内の数字は、そのカテゴリから抽出できた割合を示している)

カテゴリ名	重要文数	回答するのに重要な情報を含む文数
恋愛相談、人間関係の悩み	336802 (0.37)	145628 (0.03)
Yahoo!知恵袋	291378 (0.64)	78422 (0.01)
Yahoo! オークション	305590 (0.41)	260333 (0.13)
パソコン、周辺機器	232726 (0.43)	408480 (0.30)
病気、症状、ヘルスケア	112992 (0.46)	77527 (0.09)
政治、社会問題	115618 (0.58)	21245 (0.02)
テレビ、ラジオ	105512 (0.61)	14002 (0.02)
インターネット	98119 (0.48)	76542 (0.15)
言葉、語学	95209 (0.58)	15037 (0.02)
動物、植物、ペット	99882 (0.47)	46133 (0.05)

ビ、ラジオ」カテゴリ、「政治、社会問題」カテゴリなどである。一方、質問文数に対して重要文数を抽出できた割合が低いカテゴリは、「Yahoo!オークション」カテゴリや「パソコン、周辺機器」カテゴリなどである。

質問文数に対して重要文数を抽出できた割合が高いカテゴリである「Yahoo! 知恵袋」カテゴリや「テレビ、ラジオ」カテゴリ、「政治、社会問題」カテゴリなどは、問題を解決するためよりも他のユーザと議論やコミュニケーションの場として用いられることが多い。そのため、さまざまな回答を得ることが目的の質問が投稿されることが多い。以下に、「政治、社会問題」カテゴリの質問の例を示す。

(質問 5)　郵政民営化が現実味を帯びてきましたが、郵便局関係の公務員の方やそのご家族の方に質問です。民営化されると不安ですか？関係者の間では支持する意見が多数派ですか？

(質問 5) からは第 2 文 (「民営化されると不安ですか？」) と第 3 文 (「関係者の間では支持する意見が多数派ですか？」) が重要文として取り出された。この (質問 5) のように、さまざまな回答を得ることが目的の質問が数多く投稿されるカテゴリに投稿される質問は、以下の特徴をもつことが多い。

- 議論やコミュニケーションを活発にするため、1 つの質問内で複数の問いかけをしている
- あまり詳しく説明するとさまざまな回答が得られなくなるおそれがあるので、回答するのに重要な情報をあえて詳しく書かない

このため、これらのカテゴリでは、質問文数に対して重要文数の割合が高くなり、その結果、重要文数を抽出できた割合も高くなると考えられる。一方、これらのカテゴリでは、回答文数に対して、回答するのに重要な情報を含む文数を抽出できた割合が低い。以下に、「政治、社会問題」カテゴリの質問と回答の例を示す。

(質問 6)　郵政民営化について質問です。個人的な意見としては郵便局を増やして欲しいのですが、その為には民営化したほうが良いのでしょうか、しないほうが良いのでしょうか？

(回答 6)　郵便局を増やして欲しい理由は何ですか？

(質問 6) からは第 2 文 (「個人的な意見としては郵便局を増やして欲しいのですが、その為には民営化したほうが良いのでしょうか、しないほうが良いのでしょうか？」) が重要文として取り出された。(回答 6) からは第 1 文 (「郵便局を増やして欲しい理由は何ですか？」) が回答するのに重要な情報を含む文として取り出された。このため、(質問 6) は回答するのに重要な情報が不足していると考えられる。しかし、この (質問 6) のように、さまざまな回答を得ることが目的の質問では、回答するための情報をあまり詳しく説明するとさまざまな回答が得られなくなるおそれがある。実際、(質問 6) の場合、(回答 6) 以外にも 4 件の回答があり、さまざまな回答を得ることができている。そして、あまり詳しく情報を説明しない方がよい質問に対しては、情報不足を指摘する (回答 6) のような回答は少ないと考えられる。このため、これらのカテゴリでは回答文数に対して、回答するのに重要な情報を含む文数の割合が低く、その結果、抽出される文数の割合も低くなると考えられる。以上のことから、さまざまな回答を求める質問が投稿されるカテゴリでは、質問作成を支援するための情報はあまり重要ではないと考えられる。

質問文数に対して重要文数の抽出した割合が低いカテゴリである「Yahoo!オークション」カテゴリや「パソコン、周辺機器」カテゴリなどは、他のユーザと議論やコミュニケーションの場としてよりも問題を解決するために用いられることが多い。以下に「パソコン、周辺機器」カテゴリの質問と回答を示す。

(質問 8)　パソコンに新しいセキュリティを入れたら、テレビが見れなくなってしまいました。見れるようにするにはどうすればいんでしょうか。

(回答 8)　ソフトは何ですかね？

(質問 8) からは第 1 文 (「パソコンに新しいセキュリティを入れたら、テレビが見れなくなってしまいました。」) が重要文として取り出された。(回答 8) からは第 1 文 (「ソフトは何ですかね？」) が回答するのに重要な情報を含む文として取り出された。「Yahoo!オークション」カテゴリや「パソコン、周辺機器」カテゴリなど、ユーザ間でのコミュニケーションよりも、問題の解決を目的とする質問が数多く投稿されるカテゴリに投稿される質問は、以下の特徴をもつことが多い。

- 1 つの質問内で複数の問いかけをすることが少ない

- 回答するのに重要な情報を詳しく書く

このため、質問文数に対して重要文数の割合が低くなり、その結果、抽出できた重要文数の割合が低くなると考えられる。一方、「パソコン、周辺機器」カテゴリでは、回答文数に対して、回答するのに重要な情報を含む文数の抽出できた割合が高い。以下に「パソコン、周辺機器」カテゴリの質問と回答を示す。

(質問 8) インターネットエクスプローラーの一番上右隅の小さい３つのボタンが突然表示が変わってしまいました。判る方お願いします。
(回答 8) こんにちは。OS は何をご利用ですか？

(回答 8) からは第 2 文 (「OS は何をご利用ですか？」) が回答するのに重要な情報を含む文として取り出された。このように、「パソコン、周辺機器」カテゴリでは回答するのに重要な情報を含む文として OS の情報を確認するものが数多く取り出されていた。その他にも、「ソフト名は何ですか？」や「再起動しましたか？」、「再インストールを試しましたか？」など、質問者の環境や状況を確認する情報が回答するのに重要な情報を含む文として多く抽出されていた。このことから、「パソコン、周辺機器」カテゴリでは、環境や状況が分らないと答えられない質問が多く、また、その情報が質問で説明されていないことが多いことがわかる。このため、回答文数に対して、回答するのに重要な情報を含む文数の割合が高くなり、その結果、抽出できた文数の割合が高くなったと考えられる。以上のことから、問題の解決を目的とする質問が数多く投稿されるカテゴリでは、質問作成を支援するための情報は重要であると考えられる。

参考文献

[1] NTCIR project: http://research.nii.ac.jp/ntcir/index-en.html

[2] TREC (Text REtrieval Conference) : http://trec.nist.gov/

[3] 磯貝, 西村, 渡辺, 岡田: Q&A サイトへの質問の作成を支援するための情報の複数のカテゴリからの抽出, 言語処理学会第 15 回年次大会, P1-8, pp152-155, (2009).

[4] Kudoh: TinySVM: Support Vector Machines, (http://chasen.org/˜taku/software/TinySVM/index.html, 2002).

[5] 黒橋, 河原: 日本語形態素解析システム JUMAN version 5.1 使用説明書, 京都大学, (2005).