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

磯貝 直毅　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
{n_isogai,r_nishimura}@afc.ryukoku.ac.jp, {watanabe,okada}@rins.ryukoku.ac.jp

1　はじめに

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。

Q&A サイトがさかんに利用されている理由の 1 つに、そこでは「教えてほしい」「助けてほしい」と考えている人と「教えてあげたい」「だれかのためになれたら」と考えている人が出会い、心豊かなコミュニケーションが行われていることが考えられる。こうした心豊かなコミュニケーションを促進するためには、質問を作成するのにかかる手間や時間を軽減することと、質問に答えるために十分な情報が記述されている質問が投稿されることが重要である。そこでわれわれは、自然な文で表現されているユーザの質問で不足している情報や確認するのがのぞましい情報を Q&A サイトに投稿する前に示し、回答するのに十分な情報が記述されている質問を作成するのを支援するシステムを開発することをめざしている。

われわれはこれまでに、Yahoo! 知恵袋の「パソコン、周辺機器」カテゴリの質問と回答から、機械学習による方法で質問の作成に役立つ情報の抽出が行えることを示した [1] 。本研究では、「パソコン、周辺機器」カテゴリに加えて「健康、病気、ダイエット」カテゴリ、「レシピ、調理法」カテゴリの質問と回答からも機械学習による方法、具体的には、サポートベクトルマシン (SVM) と最大エントロピー法 (MEM) を用いて質問の作成に役立つ情報の抽出について検討を行う。また、Q&A サイトでは内容や文体が異なる複数のカテゴリがあり、それぞれのカテゴリに個別で学習データを作成するのは、手間と時間がかかるため、むずかしい。そこで本研究では、複数のカテゴリから質問の作成に役立つ情報を抽出するために必要な学習データについても検討を行う。

2　質問と回答に含まれる質問の作成に役立つ情報

本研究では、十分な情報が記述された質問の作成を支援するシステムで用いるための知識を抽出する方法について検討を行う。具体的には、以下の 2 つを取り出す。

- 質問から、質問の中心となる文 (重要文)
- 回答から、十分な情報が記述されている質問の作成に役立つ情報

質問を作成するのに役立つ情報について、質問と回答の例を利用して考察する。

(質問 1) と (回答 1) は、「健康、病気、ダイエット」カテゴリにおける例である。

(質問 1) 最近よく眠れません。ぐっすり眠れる方法を教えてください。

(回答 1) あなたの年齢、職業、部屋の採光くらいは書いてください。きちんと回答できません。

(回答 1) は、詳しく回答するのに重要な情報が質問で述べられていないことを指摘しているだけで、回答そのものはない。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質問を作成している時に与えられれば、(質問 1–a) のように質問することはむずかしくない。

(質問 1–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。ぐっすり眠れる方法を教えてください。

(質問 2) と (回答 2) は、「パソコン、周辺機器」カテゴリにおける例である。

(質問 2) PC が起動しません。どうしたらいいでしょうか。

(回答 2) windows XP を利用されているのでしたら、以下のようにすればよいと思います。(以下略)

(回答 2) は、質問者が述べていない OS の種類を仮定して問題解決の方法を説明しようとしている。しかし、この仮定が誤っていれば (例えば質問者の OS が windows XP ではなく Vista である場合)、問題を解決できないおそれがある。もし、OS の種類は述べておくのがのぞましいという情報が質問を作成している時に与えられれば、(質問 2–a) のように質問することはむずかしくない。

(質問 2–a) windows Vista を使っていますが PC が起動しません。どうしたらいいでしょうか。

(質問 3) と (回答 3) は、「レシピ、調理法」カテゴリにおける例である。

(質問 3) 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？

(回答 3) 氷に対応したミキサーをつかえば作れます。(以下略)

(回答 3) は、氷に対応したミキサーをつかえば作る方法があると回答している。しかし、氷に対応したミキサーがなければ、(回答 3) の方法は利用できない。氷に対応したミキサーがあるかどうか確認し、もしなければ、(質問 3–a) のように質問して (回答 3) 以外の解決方法を求めることができる。

(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もっていません。

われわれはこれまでに、(回答 1)、(回答 2) を用いて考察した「どんな情報を質問で述べたらいいのか判断する手がかりになる情報」と、(質問 3) を用いて考察した「確認して質問で述べるのがのぞましい情報」の２つを「質問の作成に役立つ情報」とし、Yahoo!知恵袋の「パソコン、周辺機器」カテゴリから機械学習 (SVM) による方法で抽出する実験を行った [1] 。本研究では「パソコン、周辺機器」カテゴリに加えて「レシピ、調理法」カテゴリと「健康、病気、ダイエット」カテゴリにおいても抽出が行えるか実験を行う。

また、カテゴリが異なる質問と回答では、書かれている文章に大きな違いがある。例えば、(質問 1)、(質問 2)、(質問 3) では、カテゴリが違うため使われている名詞が大きく異なる。また、以下の質問と回答のようにカテゴリで特有の書き方を用いているものがある。

(質問 4) 本などからイラストをスキャンしてワードにだすことは可能ですか？

(回答 4) ワードを起動して、挿入→図→ファイルからを選択してみてください。

(質問 4) と (回答 4) は、「パソコン、周辺機器」カテゴリにおける例である。このような問題解決の手順を矢印記号を用いて回答しているものは、「パソコン、周辺機器」カテゴリでは使われることが多いが、「レシピ、調理法」カテゴリ、「健康、病気、ダイエット」カテゴリではあまり使われることはない。

カテゴリの数は Q&A サイトによって異なるが、例えば、Yahoo! 知恵袋[1] では約 400 のカテゴリ[2]がある。使われている名詞や文章の違いを考慮して、それぞれのカテゴリに学習データを作るのは、手間と時間がかかるため、むずかしい。そこで本研究では、複数のカテゴリから情報を抽出するために必要な学習データについて検討する。

また、われわれは質問タイプによって質問の作成を支援する手法を変えることを検討している。Q&A サイトに投稿された質問は、質問者の求める答えの種類によって以下の３つに分類できる。

質問タイプ A: 回答は１つあればよい質問

質問タイプ B: 複数の回答を求めていて、その中から１つを選びたい質問

質問タイプ C: 複数の回答を求めている質問

質問タイプ A および B では、質問者は自らの問題や条件に適切な回答を求めている。そのため、質問者が望む回答を得るためには問題や条件についての情報が十分に記述されている質問をする必要がある。一方、質問タイプ C では質問者は適切な回答を得ることよりも、質問と回答を通してコミュニケーションを行うことや、いろいろな人の意見を聞くことが目的であることが多い。そのため、質問者はコミュニケーションを円滑に行うために適切な質問をする必要がある。したがって、質問タイプによって質問の作成を支援する手法を以下のように２つに分ける必要がある。

[十分な情報が記述された質問の作成を支援するシステム]

- 質問タイプ A および B の質問に対しユーザの質問で不足している情報を提示し、回答するのに十分な情報が記述されている質問を作成するのを支援

[円滑なコミュニケーションを行うための質問の作成を支援するシステム]

- 質問タイプ C の質問に対し円滑なコミュニケーションを行うために、回答者にとって読みやすい質問を質問者が書けるように支援

われわれは、以上のような２つの質問の作成を支援するシステムのため、質問者の求める答えの種類を意識した質問タイプの同定についても研究を行っている。本研究では、十分な情報が記述された質問の作成を支援するシステムで用いるための情報を抽出するが、そのシステムが対象とするのは質問タイプ A および B であるため、情報の抽出も質問タイプ A および B の質問とその回答から行う。

実験データの作成と SVM、MEM で用いる素性の調査には、ヤフー株式会社が 2007 年度より国立情報学研究所にて研究用に公開した Yahoo! 知恵袋のデータを用いた[3]。本研究では、Yahoo! 知恵袋データの「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つのカテゴリからそれぞれ 1000 個の質問とその回答 1000 個をランダムに抽出し、質問タイプを人手でタグ付けした。そして、質問タイプを質問タイプ A または質問タイプ B にタグ付けした質問とその回答を実験対象とした。実験データの調査結果を表 1 に示す。また、今回の実験では 3 文以内で書かれた回答を対象としている。3 文以内の回答を対象にしたのは、短い回答の方が質問を作成するのに役立つ情報が初心者にとって理解しやすい形式で表現されていることが多いと考えたからである。

3　実験で用いる素性

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 2 値分類の対象となる文のことである。これら $s1$〜$s6$ は、質問から重要文を、回答から質問を作成するのに役立つ情報を含む文を取り出すのに利用することを考えて、質問あるいは回答から取り出す素性である。一方、 $s7$〜$s12$ は、回答から質問を作成するのに役立つ情報を含む文を取り出すのに利用することを考えて取り出す素性である。なお、形態素解析には JUMAN [2] を用いた。

[1] http://chiebukuro.yahoo.co.jp/
[2] 2009 年 1 月現在
[3] http://research.nii.ac.jp/tdc/chiebukuro.html

表 1: 質問と回答の調査結果

カテゴリ	対象	テキスト数	文数	重要文の数	質問の作成に役立つ情報を含む文の数
パソコン、周辺機器	質問	944	2536	1239	-
	回答	944	1740	-	174
健康、病気、ダイエット	質問	874	2160	1250	-
	回答	874	1682	-	153
レシピ、調理法	質問	903	2204	1330	-
	回答	903	1787	-	119

$s1$	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: 質問と回答から情報を抽出するのに用いる素性

4 実験結果と評価

2 章で述べた実験データを対象に SVM と MEM を用いて、

- 質問を構成する各文に対して、質問の中心になる文 (重要文) として取り出すかどうか
- 回答を構成する各文に対して、質問を作成するのに役立つ情報を含む文として取り出すかどうか

という 2 値分類を行った。また、実験データはカテゴリごとに時系列順に 2 分割し、上半分をクローズドデータ、下半分をオープンデータとして用いた。本実験では、SVM には TinySVM[4] 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。また、MEM には maxent[5] を用いた。素性選択は、村田らが利用している手法 [3] を用いた。素性選択に用いる評価値は、質問の分類では精度を用いた。回答ではデータの量に対して、質問の作成に役立つ情報の量が少ないため F 値を用いた。

カテゴリごとで情報を抽出できるか検討するために、以下のデータを用いて実験を行った。

データ 1 実験データの「パソコン、周辺機器」カテゴリのクローズドデータとオープンデータ

データ 2 実験データの「健康、病気、ダイエット」カテゴリのクローズドデータとオープンデータ

データ 3 実験データの「レシピ、調理法」カテゴリのクローズドデータとオープンデータ

実験は以下の 3 つを行った。

(実験 1) 全素性を用いて、クローズドデータで学習を行い、オープンデータの分類を行う

(実験 2) クローズドデータを用いて素性選択を行い、その素性の組み合わせにおける 10 分割クロスバリデーションを行う

(実験 3) 実験 2 の素性選択の結果を用いて、クローズドデータで学習を行い、オープンデータの分類を行う

複数のカテゴリから情報を抽出するのに必要な学習データについて検討するために、以下のデータを用いて実験を行った。

データ 4 実験データのすべてのカテゴリのクローズドデータとオープンデータ

データ 5 学習データに「パソコン、周辺機器」のクローズドデータ、評価データに残り 2 つのカテゴリのオープンデータ

データ 6 学習データに「健康、病気、ダイエット」のクローズドデータ、評価データに残り 2 つのカテゴリのオープンデータ

データ 7 学習データに「レシピ、調理法」のクローズドデータ、評価データに残り 2 つのカテゴリのオープンデータ

データ 4 では (実験 1)、(実験 2)、(実験 3) を行い、データ 5、6、7 では (実験 1) のみを行った。実験結果を表 2 に示す。素性選択の結果を表 3 に示す。

われわれはこれまでに、「パソコン、周辺機器」のみを対象に Q&A サイトにおける質問とその回答からの情報抽出の実験を行った [1] 。表 2 のデータ 2、3 の実験結果で、「パソコン、周辺機器」以外のカテゴリにも質問の作成に役立つ情報があり、その情報を機械学習による方法によって抽出が行えることを示した。例えば、データ 3 の回答からの情報抽出 (実験 1) では、精度 95.1%/94.8%(SVM/MEM)、F 値 0.450/0.281(SVM/MEM) で抽出が行えた。

また、表 2 の SVM を用いた回答からの情報抽出では素性選択してオープンテストを行う (実験 3) よりも、全素性を用いてオープンテストを行う (実験 1) 方が、F 値が良い結果となるものが多かった。一方、MEM では、回答

[4] http://chasen.org/~taku/software/TinySVM/

[5] http://www2.nict.go.jp/x/x161/members/mutiyama/software.html#maxent

表 2: 実験結果

データ番号	データ内容	対象	全素性オープンテスト (実験 1)(SVM/MEM)		素性選択クローズドテスト (実験 2)(SVM/MEM)		素性選択オープンテスト (実験 3)(SVM/MEM)	
			精度 (%)	F 値	精度 (%)	F 値	精度 (%)	F 値
データ 1	パソコン、周辺機器	質問	85.7/85.7	0.847/0.847	83.9/82.3	0.831/0.823	85.1/85.3	0.844/0.842
		回答	91.5/91.2	0.370/0.278	93.9/94.3	0.634/0.672	91.0/91.9	0.331/0.455
データ 2	健康、病気、ダイエット	質問	81.7/82.3	0.842/0.847	83.2/83.9	0.863/0.868	82.3/80.7	0.844/0.838
		回答	91.9/92.5	0.355/0.289	90.1/91.4	0.339/0.372	90.4/91.8	0.349/0.340
データ 3	レシピ、調理法	質問	86.2/85.8	0.887/0.882	85.0/85.1	0.881/0.880	84.6/84.8	0.871/0.873
		回答	95.1/94.8	0.450/0.281	93.8/94.8	0.517/0.525	95.1/94.9	0.436/0.384
データ 4	実験データの すべてのカテゴリ	質問	83.5/83.7	0.849/0.851	83.0/82.2	0.849/0.838	82.9/83.7	0.839/0.851
		回答	92.6/92.9	0.454/0.417	93.4/93.1	0.557/0.489	91.9/92.9	0.441/0.452
データ 5	学習: パソコン 評価: 健康＆レシピ	質問	78.0/78.4	0.799/0.801	-	-	-	-
		回答	93.5/93.2	0.323/0.272	-	-	-	-
データ 6	学習:健康 評価: パソコン＆レシピ	質問	82.2/82.4	0.839/0.841	-	-	-	-
		回答	92.3/92.1	0.297/0.157	-	-	-	-
データ 7	学習:レシピ 評価: パソコン＆健康	質問	79.7/80.1	0.819/0.823	-	-	-	-
		回答	91.9/91.2	0.349/0.145	-	-	-	-

表 3: 素性選択で有効な素性の組み合わせを求めた結果

データ番号	データ内容	対象	SVM の素性選択結果	MEM の素性選択結果
データ 1	パソコン、周辺機器	質問	s1, s3, s4, s5	s1, s2, s3, s5, s6
		回答	s1, s3, s4, s5, s6, s7, s8, s9, s10, s11	s1, s2, s3, s9, s11, s12
データ 2	健康、病気、ダイエット	質問	s1, s2, s3, s5	s1, s3, s4, s6
		回答	s1, s3, s4, s5, s6, s7, s8, s9, s10, s12	s1, s2, s5, s6, s9, s11, s12
データ 3	レシピ、調理法	質問	s1, s2, s3, s4, s5	s1, s2, s3, s4, s5
		回答	s1, s2, s3, s5, s6, s8, s9, s10, s11, s12	s1, s3, s4, s5, s6, s7, s9, s11
データ 4	実験データの すべてのカテゴリ	質問	s1, s2, s3, s4, s5	s1, s2, s3, s4, s5, s6
		回答	s1, s3, s4, s9, s11	s1, s2, s3, s5, s7, s9

からの情報抽出が全素性を用いてオープンテストを行う (実験 1) よりも、素性選択してオープンテストを行う (実験 3) 方が、F 値が良い結果となるものが多かった。

データ５、６、７の質問からの情報抽出では、精度 80% 程度、F 値 0.800 程度の結果となった。回答からの情報抽出では、精度 92%程度、F 値 0.300 程度の結果となった。この評価は、データ１、２、３の結果よりも多少下がっている。そのため、学習データと評価データで異なるカテゴリを用いた場合、分類はできるが、学習データと評価データが一緒のカテゴリを用いた場合より精度が多少下がるという結果となった。

また、表 3 の素性選択で有効な素性の組み合わせを求めた結果のうち、回答では s1, s9 が必ず含まれるという結果となった。これは、質問を作成するのに役立つ情報を含む文を抽出するためには、回答文に加えて質問文も参照する必要があることを示していると考えている。

5 おわりに

本研究では、十分な情報が記述された質問の作成を支援するシステムで用いるための情報、すなわち

- 質問から、質問の中心となる文 (重要文)
- 回答から、十分な情報が記述されている質問の作成に役立つ情報

を機械学習 (SVM,MEM) を用いて抽出する方法と、複数のカテゴリから情報を抽出するのに必要な学習データについて検討を行った。

今後は、実験データの量を増やして質問から質問の中心となる文 (重要文) の抽出、回答から質問の作成に役立つ情報の抽出の実験を行う。また、十分な情報が記述された質問の作成を支援するシステムを実現するために必要な情報の抽出以外のモジュールについても検討を進める。

謝辞 本研究を実施するにあたり、ヤフー株式会社が国立情報学研究所にて研究用に公開した Yahoo! 知恵袋のデータを利用させていただきました。ここであらためて感謝とお礼を申し上げます。本研究の一部は、日本学術振興会科学研究費補助金基盤 (C) 「心豊かなコミュニケーションを促進する質問作成支援システムの作成」(課題番号 20500106) の助成を受けて行われたものです。

参考文献

[1] 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抽出, 情報社会学会, 知識共有コミュニティワークショップ, pp.19-28, (2008).

[2] 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).

[3] 村田, 金丸, 白土, 井佐原: 入力文の格助詞ごとに学習データを分割した機械学習による受身文の能動文への変換における格助詞の変換, システム制御情報学会論文誌, vol21, No.6, pp.165-175, (2008).