

## 質問者が求める答えの種類を意識した質問タイプの同定

峯本 祐麻 磯貝 直毅 西村 涼 渡辺 靖彦 岡田 至弘

龍谷大学 理工学部 情報メディア学科

t050569@mail.ryukoku.ac.jp, {n\_isogai,r\_nishimura}@afc.ryukoku.ac.jp,

{watanabe,okada}@rins.ryukoku.ac.jp

### 1 はじめに

質問タイプの同定を行うことは、自然文で与えられた質問の意図を正しく判定するために重要である。そして、質問の意図を正しく理解することは、ユーザが要求する意図を正しく遂行するために重要である。

これまでの質問タイプ同定の研究では、質問の分野に着目して「趣味」「ビジネス」「宇宙学」などに分類する研究 [1] や、質問の種類に着目して「What 型」「Where 型」「How 型」「Symptom 型」などで分類する研究 [2] がさかに行われている。しかし、答えの種類に着目して質問を分類する研究はこれまでほとんど行われていない。

質問者が Q&A サイトに質問を投稿する時、どんなタイプの答えが欲しいのか、あらかじめ考えていることが多い。したがって、Q&A サイトに投稿された質問からユーザが求める情報を検索するときに、「趣味」「ビジネス」「宇宙学」などの質問の分野や「What 型」「Where 型」「How 型」「Symptom 型」などの質問の種類だけでなく、質問者が求める答えの種類の情報も利用して検索できれば、より求める内容に近い質問とその回答を得られることが期待できる。

このため、質問者がどのような答えを欲しがっているのかという点に着目して Yahoo!知恵袋<sup>\*1</sup>を調査すると、質問は大きく以下の 3 つに分けられることがわかった。

回答は 1 つあればよい質問

複数の回答を求めている、その中から 1 つを選びたい質問

複数の回答を求めている質問

そこで本研究では、Q&A サイトに投稿された質問がこの 3 つのタイプのどれであるのか同定する方法について提案する。質問タイプを同定するのに、サポートベクトルマシン法 (以下、SVM) と最大エントロピー法 (以下、MEM) を用いる。実験には、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ている Yahoo!知恵袋のデータ<sup>\*2</sup>を利用した。

### 2 答えの種類を意識した質問のタイプ同定を行うための実験データの作成

本研究では、Q&A サイトに投稿された質問を求める答えの種類を意識した以下の 3 つの質問タイプに機械学習の手法を用いて同定することを目指す。

#### 質問タイプ A [回答は 1 つあればよい質問タイプ]

回答が 1 つしかないと考えられる質問

回答を 1 つとにかく至急欲しいという質問

などがこの質問タイプに同定される。例えば、「What 型」や「Yes/No 型」の質問などは回答が 1 つしかないと考えられるので、この質問タイプに同定される。

#### 質問タイプ B [複数の回答を求めている、その中から 1 つを選びたい質問タイプ]

回答が何通りもあることが考えられる質問で、それらを比較検討して質問者にとって一番良いものを選ぶようにする質問がこの質問タイプに同定される。例えば、「How 型」の質問は回答が何通りもあることが多く、質問者は複数の回答の中から 1 つを選んで問題を解決しようとする。このため、「How 型」の質問はこの質問タイプに同定されることが多い。

#### 質問タイプ C [複数の回答を求めている質問タイプ]

回答が何通りもあることが考えられる質問であるが、それらから 1 つを選ぶようにしたりしないもの、むしろさまざまな回答を収集することが目的であるものがこの質問タイプに同定される。例えば、意見収集 (評判の問い合わせなど) の質問などがこの質問タイプに同定される。

そこで、Yahoo!知恵袋の

レシピ、調理法

健康、病気、ダイエット

パソコン、周辺機器

の 3 つのカテゴリに投稿された質問を利用して、求める答えを意識したタイプ同定を行うための実験データを以下の手順で作成した。

#### step 1 Yahoo!知恵袋のカテゴリ

レシピ、調理法

健康、病気、ダイエット

パソコン、周辺機器

から質問とその回答を無作為に取り出す。ただし、取り出した質問が以下の場合、その質問と回答は不

\*1 <http://chiebukuro.yahoo.co.jp/>

\*2 <http://research.nii.ac.jp/tdc/chiebukuro.html>

どこ、だれ、なんで、なぜ、正しい、良い、どんな、どっち、いい、どちら、どう、みなさん、おいしい、上手な、うまく、簡単な、面白い、おすすめ、すすめ、方法、法、お気に入り、教えて、ください、楽しい、思い、ました、ます、です、か

図 1 答えの種類を意識した質問のタイプ同定の手がかりになると考えられる表現

表 1 作成した実験データの内訳

	レシピ、 調理法	健康、病気 ダイエット	パソコン、 周辺機器
質問タイプ A	528	692	853
質問タイプ B	375	182	91
質問タイプ C	97	126	56
合計	1000	1000	1000

適切であるとして捨て、他の質問とその回答をあらたに取り出す。

質問が含まれていない

質問文の大部分がひらがなやカタカナで構成されている

複数の質問タイプに分けられる

**step 2** 取り出した質問を構成するそれぞれの文を以下の 2 つに分類する。

質問タイプの同定に重要な文

質問タイプの同定に重要でない文

これは、質問タイプの同定に重要な文とそうでない文を区別して扱う方が、質問タイプを同定するタスクにおいては有効であることが多いと考えたからである [3]。さらに、質問およびその回答に図 1 に示す手がかり表現が含まれていないか確認する。

**step 3** step 2 の結果を考慮して、質問を以下の 3 つの質問タイプに同定する。

**質問タイプ A** [回答は 1 つあればよい質問タイプ]

**質問タイプ B** [複数の回答を求めている、その中から 1 つを選びたい質問タイプ]

**質問タイプ C** [複数の回答を求めている質問タイプ]

以上の手順を、各カテゴリについて 1000 件ずつ実験データがそろそろまで繰り返し行う。表 1 に作成した実験データの内訳を示す。表 1 に示すように、3 つのカテゴリの質問の大部分が質問タイプ A [回答は 1 つあればよい質問タイプ] に同定されている。特に「パソコン、周辺機器」のカテゴリの質問は、質問タイプ A に同定されたものが多い。これは、このカテゴリの質問には「パソコンの OS って何ですか」など答えが 1 つしかないと考えられるものや、とにかく答えを 1 つ至急欲しいというものが多いためである。これに対して、「レシピ、調理法」のカテゴリの質問は、質問タイプ B [複数の回答を求めている、その中から 1 つを選びたい質問タイプ] に同定されたものが多い。これは、「パソコン、周辺機器」のカテゴリの質問に比べて、

S1	質問に含まれる質問タイプの同定に重要な文の自立語
S2	質問に含まれる質問タイプの同定に重要な文の付属語
S3	質問に含まれる質問タイプの同定に重要な文の特殊 (句読点・記号・括弧など)
S4	質問に含まれる質問タイプの同定に重要な文の接頭辞・接尾辞
S5	質問に含まれる質問タイプの同定に重要な文の未定義語
S6	質問に含まれる質問タイプの同定に重要な文の形態素の 2-gram
S7	質問に含まれる質問タイプの同定に重要な文の形態素の 3-gram
S8	質問に含まれる質問タイプの同定に重要な文の形態素の読み
S9	質問に含まれる質問タイプの同定に重要な文の形態素の原形
S10	質問に含まれる質問タイプの同定に重要でない文の自立語
S11	質問に含まれる質問タイプの同定に重要でない文の付属語
S12	質問に含まれる質問タイプの同定に重要でない文の特殊 (句読点・記号・括弧など)
S13	質問に含まれる質問タイプの同定に重要でない文の付属語
S14	質問に含まれる質問タイプの同定に重要でない文の接頭辞・接尾辞
S15	質問に含まれる質問タイプの同定に重要でない文の形態素の 2-gram
S16	質問に含まれる質問タイプの同定に重要でない文の形態素の 3-gram
S17	質問に含まれる質問タイプの同定に重要でない文の形態素の読み
S18	質問に含まれる質問タイプの同定に重要でない文の形態素の原形
S19	質問に含まれる形態素の数
S20	質問に含まれる形態素ごとの数
S21	質問を構成するそれぞれの文
S22	質問を構成するすべての文
S23	質問を構成するすべての文の文字数
S24	質問を構成する文の数
S25	回答 (Best Answer) の形態素の自立語
S26	回答 (Best Answer) の形態素の付属語
S27	回答 (Best Answer) の形態素の特殊
S28	回答 (Best Answer) の形態素の接頭辞・接尾辞
S29	回答 (Best Answer) の形態素の未定義語
S30	回答 (Best Answer) の形態素の 2-gram
S31	回答 (Best Answer) の形態素の 3-gram
S32	回答 (Best Answer) の形態素の読み
S33	回答 (Best Answer) の形態素の原形
S34	回答 (Normal Answer) の形態素の自立語
S35	回答 (Normal Answer) の形態素の付属語
S36	回答 (Normal Answer) の形態素の特殊 (句読点・記号・括弧など)
S37	回答 (Normal Answer) の形態素の接頭辞・接尾辞
S38	回答 (Normal Answer) の形態素の未定義語
S39	回答 (Normal Answer) の形態素の 2-gram
S40	回答 (Normal Answer) の形態素の 3-gram
S41	回答 (Normal Answer) の形態素の読み
S42	回答 (Normal Answer) の形態素の原形
S43	質問に対するすべての回答に含まれる形態素の数
S44	質問に対するすべての回答に含まれるそれぞれの形態素の数
S45	質問に対する回答の数
S46	回答を構成するすべての文
S47	質問・回答に含まれる手がかり表現の有無
S48	質問・回答の両方に含まれる同一の形態素の原形

図 2 質問タイプの同定に用いる素性

このカテゴリの質問には「カレーの美味しい作り方を教えてください」など、さまざまな回答を求めて、その中から 1 つを選びたいものが多いからである。「健康、病気、ダイエット」のカテゴリの質問は、質問タイプ B だけでなく、質問タイプ C [複数の回答を求めている質問タイプ] に同定されたものも他のカテゴリに比べて多い。これは、このカテゴリの質問には、「頭痛がひどいのですが、どうしたらいいですか？」などのさまざまな回答を求めて、その中から 1 つを選びたいものや、「他の人でもこんな症状の方いますか？」といったさまざまな回答を求めている質問だが、特にその中から 1 つを選びだそうとしたりしないものが含まれているからである。

### 3 Q&A サイトに投稿された質問をタイプ同定するために用いる素性

本研究で用いる機械学習の素性を図 2 に示す。S1~S24 はタイプ同定する対象の質問から取り出す素性である。こ

表 2 すべての素性を用いた場合の実験結果

カテゴリ	分類器	クローズドテスト				オープンテスト			
		精度	タイプ A	F 値 タイプ B	タイプ C	精度	タイプ A	F 値 タイプ B	タイプ C
レシピ、調理法	SVM	71.4%	0.79	0.67	0.04	70.0%	0.78	0.66	0.11
	MEM	71.8%	0.80	0.68	0.04	69.8%	0.78	0.66	0.08
健康、病気、ダイエット	SVM	70.6%	0.81	0.50	0.17	78.2%	0.87	0.43	0.23
	MEM	71.0%	0.82	0.50	0.14	77.6%	0.87	0.41	0.20
パソコン、周辺機器	SVM	83.4%	0.91	0.00	0.06	87.8%	0.94	0.13	0.00
	MEM	83.4%	0.91	0.04	0.00	87.8%	0.94	0.10	0.00

表 3 素性選択を行った後の素性を用いた場合の実験結果

カテゴリ	分類器	クローズドテスト				オープンテスト			
		精度	タイプ A	F 値 タイプ B	タイプ C	精度	タイプ A	F 値 タイプ B	タイプ C
レシピ、調理法	SVM	72.0%	0.80	0.68	0.07	69.6%	0.77	0.66	0.11
	MEM	71.8%	0.80	0.68	0.04	69.8%	0.78	0.66	0.08
健康、病気、ダイエット	SVM	71.6%	0.82	0.50	0.32	78.2%	0.87	0.46	0.22
	MEM	71.4%	0.82	0.50	0.20	77.6%	0.87	0.42	0.21
パソコン、周辺機器	SVM	84.4%	0.92	0.14	0.11	87.8%	0.94	0.13	0.00
	MEM	84.0%	0.91	0.08	0.11	87.8%	0.94	0.13	0.00

のうち、S1~S20 はタイプ同定する対象の質問を 1 文ずつ形態素解析した結果から取り出す素性で、S21~S24 はタイプ同定する対象の質問の全文から取り出す素性である。S25~S46 はタイプ同定する対象の質問に対する回答から取り出す素性である。このうち、S25~S44 はタイプ同定する対象の質問に対する回答 (Best Answer および Normal Answer) の全文を形態素解析した結果から取り出す素性である。S46 はタイプ同定する対象の質問に対する回答から取り出す素性である。S47 と S48 は、質問とその回答から取り出す素性である。S47 は図 1 の手がかり表現が分類対象の質問あるいはその回答に含まれているかどうかを示す素性である。S48 はタイプ同定する対象の質問とその回答の両方にあらわれる形態素を素性としている。なお、形態素解析には JUMAN[4] を用いた。

## 4 実験

各カテゴリの実験データ 1000 件を時系列順に並べ 2 分割した。得られたデータの前半分をクローズドデータ、後半分をオープンデータとする。クローズドデータは素性選択の実験に利用し、オープンデータは評価用データとして利用した。クローズドテストの評価にはクローズドデータを用いて 10 分割のクロスバリデーションを利用し、F 値のマクロ平均を求めた。F 値のマクロ平均 [6] を求めたのは、本研究では質問タイプ A、質問タイプ B、質問タイプ C の F 値をバランス良く上げることが重要だからである。オープンテストでは、学習データにクローズドデータを用いて、オープンデータで精度を求めた。素性選択には、村田らが採用している手法 [5] を用いた。素性選択の実験を行うのは、図 2 の素性には分類に有効でない素性が含まれている可能性があるからである。また、実験では機械学習として SVM と MEM を用いる。SVM には SVM multiclass<sup>\*3</sup> の線形カーネルを利用し、オプションは  $c=500000$  で実験した。MEM には maxent<sup>\*4</sup> を利用し

<sup>\*3</sup> [http://svmlight.joachims.org/svm\\_multiclass.html](http://svmlight.joachims.org/svm_multiclass.html)

<sup>\*4</sup> <http://www2.nict.go.jp/x/x161/members/mutiyama/software.html#maxent>

表 4 有効であると判断された素性

カテゴリ	分類器	有効な素性
レシピ、調理法	SVM	S40 以外の素性
	MEM	すべての素性
健康、病気、ダイエット	SVM	S15,S39,S43 以外の素性
	MEM	S18,S40 以外の素性
パソコン、周辺機器	SVM	S7 以外の素性
	MEM	S39,S40 以外の素性

ている。

表 2 に、すべての素性を用いて、それぞれのカテゴリごとにクローズドテストとオープンテストを行った結果を示す。また、それぞれのカテゴリごとに素性選択を行い、表 3 にクローズドテストとオープンテストを行った結果を示す。これらの結果のクローズドテストを比較すると、素性選択を行うことによって分類ごとの F 値を向上させることができた。特に「健康、病気、ダイエット」カテゴリにおける質問タイプ C では、SVM において F 値が 0.15 向上している。また、「パソコン、周辺機器」カテゴリの MEM における分類では、質問タイプ C の F 値は 0 だったが、素性選択を行うことによって 0.11 という F 値が得られた。表 4 に素性選択を行った結果、有効であると判断された素性を示す。それぞれのカテゴリを比較すると、有効であった素性が異なっていることがわかる。また、カテゴリごとの機械学習法を比較しても有効であった素性が異なっていることがわかる。

F 値のマクロ平均を向上させることができた素性選択が有効であったか評価するために、すべての素性の組み合わせと素性選択を行った後の素性の組み合わせをオープンテストの結果で比較する。表 2 と表 3 から精度と F 値ともに素性選択による効果はあまり得られなかった。したがって、本研究で用いた素性には特に分類精度を低下させるような素性はなかったと思われる。この結果は、できるだけ多くの素性を用いたほうがよいとする平ら [7] の実験結果に沿うものである。

次に、学習用データと分類したいデータのカテゴリが異なる場合に質問タイプの同定ができるか実験した。学習用データとして、あるカテゴリのクローズドデータとオー

表 5 異なるカテゴリのデータを用いた場合の実験結果 (すべての素性)

学習用データ	分類器	評価用データ											
		レシピ、調理法				健康、病気、ダイエット				パソコン、周辺機器			
		精度	F 値			精度	F 値			精度	F 値		
	タイプ A	タイプ B	タイプ C		タイプ A	タイプ B	タイプ C		タイプ A	タイプ B	タイプ C		
レシピ、調理法	SVM	-				71.9%	0.83	0.33	0.00	85.6%	0.92	0.16	0.00
	MEM	-				72.1%	0.83	0.31	0.00	85.6%	0.92	0.14	0.00
健康、病気、ダイエット	SVM	64.7%	0.74	0.55	0.23	-				85.1%	0.92	0.27	0.22
	MEM	64.0%	0.73	0.53	0.22	-				85.2%	0.92	0.27	0.23
パソコン、周辺機器	SVM	58.7%	0.72	0.28	0.18	70.4%	0.82	0.19	0.11	-			
	MEM	58.0%	0.72	0.25	0.15	70.5%	0.82	0.20	0.11	-			

ブデータを組み合わせたもの (1000 件) を用い、評価用データとして学習用データに用いたカテゴリとは別のカテゴリのクローズドデータとオープンデータを組み合わせたもの (1000 件) を用いた。表 5 にすべての素性を用いて分類を行った実験結果を示す。同じカテゴリのデータでオープンテストを行った結果と表 5 の結果を比較すると、「健康、病気、ダイエット」カテゴリを学習用データに用いることで別のカテゴリを分類する場合の質問タイプ C の F 値が向上していることがわかる。特に、「パソコン、周辺機器」を分類する場合においては、学習用データに同一カテゴリのデータを用いず、「健康、病気、ダイエット」カテゴリを学習用データに用いた場合に、F 値で SVM の場合は 0.22、MEM の場合は 0.23 という結果を得ることができた。この原因としては、「健康、病気、ダイエット」カテゴリには質問タイプ C が最も多く含まれていることが考えられる。これは、分類したいカテゴリにおいて、「パソコン、周辺機器」カテゴリのように質問タイプ C の数が少ない場合には、質問タイプ C の数が多い「健康、病気、ダイエット」カテゴリのようなデータを学習用データに用いて作成した分類器で分類した方が良いことを示唆していると考えられる。

- 定, 情報処理学会論文誌, Vol.47, No.6, pp.1954-1962, (2006)
- [4] 黒橋 禎夫, 河原 大輔: 日本語形態素解析システム JUMAN version 5.1 使用説明書, 京都大学大学院情報学研究科, (2005)
- [5] 村田 真樹, 金丸 敏幸, 白土 保, 井佐原 均: 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, Vol.21, No.6, pp.165-175, (2008)
- [6] 猪野 陽子, 松井 藤五郎, 大和田 勇人: WordNet からの共通概念抽出によるテキスト分類精度の改善, 2004 年度人工知能学会全国大会 (第 18 回), 2F1-02, (2004)
- [7] 平 博順, 春野 雅彦: Support vector machine によるテキスト分類における属性選択, 情報処理学会論文誌, Vol.41, No.4, pp.1113-1123, (2000)
- [8] 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 知識共有コミュニティワークショップ, pp.19-28, (2008)

## 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!知恵袋データを利用させていただきましたことを、改めて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] 安形 輝, 石田 栄美, 久野 高志, 野末 道子, 上田 修一: WWW ページの自動分類: NDC の分類体系と Yahoo のカテゴリを使った分類, 情報処理学会研究報告, 99-DBS-118, pp.113-120, (1999)
- [2] 清田 陽司, 黒橋 禎夫, 木戸 冬子: 大規模テキスト知識ベースに基づく自動質問応答システム-ダイアログナビ- , 自然言語処理, Vol.10, No.4, pp.145-175, (2003)
- [3] 田村 晃裕, 高村 大也, 奥村 学: 複数文質問のタイプ同