4I1-1

コミュニティQAにおける文章の表層的特徴に基く質問文と回答文の分析

An Analysis Method of Questions and Answers in QA Communities Using Text Features

朱 成敏　　武田 英明
Sungmin JOO　　Hideaki TAKEDA

国立情報学研究所
National Institute of Informatics

In this paper, we discuss the question classification that considers the superficial characteristic of questions and answers, and try to classify questions and categories in QA communities using textual characteristics.

1. はじめに

コミュニティQA は投稿された質問に対して他の参加者が回答を投稿するコミュニティサービスであり，活況を呈している知識共有コミュニティである．多くのコミュニティQA では，質問内容によって様々なカテゴリが用意されていて，参加者は質問内容に合わせて適切なカテゴリで質問を投稿する仕組みになっている．

しかし，質問の中では電子製品の機能や使い方のように知らない知識や情報を求める質問や，人生や恋愛に関して相談を求める質問も存在する．場合によっては回答の正確さより質問者の主観によって最も良い回答として選択されることもあり，良質な回答の推定において多様な判断基準を用いて対応する必要がある．

そこで，本研究ではコミュニテイ QA で良質な回答の推定を目的とし，質問タイプという観点から回答文と質問文の特徴を用いて分類を行う．また，質問カテゴリとの関連性を分析し，カテゴリ単位での分類を行う．その結果から分類結果の妥当性を検討し，分類において要求される良質な回答の特徴を考察する．

2. 質問タイプの分類

三浦ら [三浦 08] は知識共有コミュニティでの参加動機を分析する過程で質問には「正解あり」と「正解なし」の二つのタイプが存在していることを指摘した.「正解あり」タイプの場合は，質問者が決まった答えや正解を求める，すなわち信頼性が高い知識や情報を求める場合である.「正解なし」タイプの場合は，決まった答えを求めるのではなく，質問者の悩みに対して他の人に経験や意見を求める相談に近い質問タイプである．そして，質問タイプによって回答文の特徴が異なる可能性がある．例えば，知識や情報を求める「正解あり」タイプでは引用，根拠など事実に対する説明が回答になる場合が多いと予想される．また，経験や意見を求める「正解なし」タイプの場合は回答者の考えや主張が含まれていると考えられる．このような質問タイプの特徴が回答の特徴に影響を与える場合を考慮し，その特徴に適合するベストアンサーの推定方法を選んで適用することでより適切なベストアンサーを推定することが期待される．

また，コミュニティQA では質問内容によって様々なカテゴリが存在する．質問者は質問内容によって適切なカテゴリを選び，質問文を投稿する．従って，質問カテゴリは質問タイプが持つ特徴を表す可能性が高い．例えば，平均回答数も質問タイプを分類する一つの手法として考えられる.「正解あり」タイプの場合，客観的な正解が存在しており，この正解が投稿された場合は回答数に関係なくベストアンサーとして選択されることが多く見られる．一方,「正解なし」タイプはベストアンサーの選択に質問者の主観が強く影響をしている場合が多く，質問者が共感する回答が投稿されるまで進行される質問も多く存在する．また，特定な知識や情報を知っている参加者が投稿する「正解あり」タイプに対して「正解なし」タイプの場合は誰でも自分の意見を投稿することができるため回答数が多い特徴も考えられる．実際「Yahoo!知恵袋*1データ (第 2 版)」の質問カテゴリ 453 件の中から投稿数が多い上位 20 件を抽出し，平均回答数を確認してみると，上位 5 件が「Yahoo!オークション」,「恋愛相談，人間関係の悩み」,「Yahoo!知恵袋」,「正治，社会問題」,「妊娠，出産」であった．そして，下位 5 件は「Windows 全盤」,「インターネット」,「病気，病症，ヘルスケア」,「パソコン」,「健康，病気，病院」であった．このように質問カテゴリの特徴を質問タイプの特徴をから判断することも考えられる [朱 14].

そこで，本研究では，回答文と質問文の表層的特徴を用いてカテゴリ毎に機械学習を行う．その結果を分析し，質問カテゴリ毎にどのような特徴が影響を与えているのかを確認する．また，コミュニテイ QA の質問カテゴリを「正解あり」と「正解なし」の 2 つのタイプに自動分類を行う．

3. 回答文の分析

本章ではカテゴリ毎に回答文が持つ特徴を発見するためにベストアンサーとして選ばれた回答文と対象として分析を行う．まず，回答文の表層的特徴を用いてベストアンサーを決める要素を相関係数を用いて確認する．そして，機械学習を用いてカテゴリ毎にベストアンサーの推定を行い，その結果からカテゴリが持つ特徴を検討する．

3.1 ベストアンサーの分析

まず，質問カテゴリ毎に回答文が持つ特徴を把握するために文章の表層的特徴を用いて確認を行う．回答文のテキストから特徴要素を抽出し，相関係数を用いてベストアンサーの選択に影響を与える要素をカテゴリ毎に確認する．

連絡先: 朱 成敏, 国立情報学研究所, 〒 101-8430 東京都千代田区一ツ橋 2-1-2, joo@nii.ac.jp

*1 Yahoo!知恵袋, http://chiebukuro.yahoo.co.jp/

回答文の分析を行うために実験データとして「Yahoo!知恵袋データ (第 2 版)」から「正解あり」の質問が多く含まれていると予想される「Windows 全盤」,「一般教養」の 2 つ,「正解なし」タイプが多く投稿されていると予想される「恋愛相談, 人間関係の悩み」,「生き方，人生相談」の 2 つ，計 4 つのカテゴリの回答文を抽出した．回答文の数はそれぞれ 2 万件を基準として収集した．実験データの詳細は表 1 に表す.

表 1: 実験データ

カテゴリ	質問数	回答数	BA 数
恋愛相談，人間関係の悩み	4,337	20,333	4,230
生き方，人生相談	2,024	20,482	2,026
Windows 全般	9,618	21,023	9,490
一般教養	4,801	20,133	4,789

そして，収集された回答文に対して文章の表層的特徴要素を抽出した．特徴要素は文章表現と文法的特徴などの文法的要素，分数や外部リンク，参照記号などの構造的要素，そして挨拶のような意味的要素など 18 種である (表 2)．これらの特徴予想を回答文から抽出し，ベストアンサーとの関係性について相関係数を用いて確認をした.

その結果を表 3 に示す．全般的に弱い相関を見せたが,「恋愛相談，人間関係の悩み」,「生き方，人生相談」では敬語形語尾の出現数，理由節の出現数，主張動詞の出現数，文数が影響を与えると思われる結果となった．これは質問者に対して丁寧に回答をしていた回答文がベストアンサーとして選ばれたと考えられる.「Windows 全般」,「一般教養」では参照記号の出現数と外部リンクの出現数が影響を与えていたことが分かった．これは質問者が求めていた情報，または回答の根拠となる情報を提示する回答文がベストアンサーとして選ばれたと考えられる．これにより回答文の表層的特徴がカテゴリ毎に異なることが確認でした.

3.2 機械学習によるベストアンサーの推測

本節では回答文のテキストから抽出した特徴要素を用いてカテゴリ単位で機械学習によるベストアンサーの分類を行う．分類結果からカテゴリが持つ特徴を発見し，カテゴリの特徴がベストアンサーの推定に与える影響について考察を行う.

ベストアンサーを推測するために学習させる特徴は表 2 の特徴要素を，機械学習手法はサポートベクターマシン (SVM) 手法を用いた．抽出した特徴要素は平均正規化を用いて特徴量正規化を行い，SVM は SVM light[*2]を，カーネルは線形カーネルを用いた．実験結果に対する検証は 4 分割交差検定を用いて行った.

$$\text{適合率} = \frac{\text{正解と判定された議論と正解の共通数}}{\text{正解と判定された議論}}$$

$$\text{再現率} = \frac{\text{正解と判定された議論と正解の共通数}}{\text{正解数}}$$

$$F\ \text{値} = \frac{2 * \text{適合率} * \text{再現率}}{\text{適合率} + \text{再現率}}$$

その結果を表 4 に示す．実験結果，カテゴリ「Windows 全盤」の推定結果では最も高い F 値とり．同じく知識や情報を求めるカテゴリとして予想された「一般教養」の F 値が 0.670,

∗2 SVM light, http://svmlight.joachims.org/

表 2: 表層的特徴要素

#	特徴要素	#	詳細
1	質問形語尾の出現数	10	主張動詞の出現数
2	勧誘形語尾の出現数	11	挨拶の出現数
3	敬語形語尾の出現数	12	参照記号の出現数
4	推測形語尾の出現数	13	外部リンクの出現数
5	命令希望形語尾の出現数	14	文数
6	意見要求の出現数	15	「?」の出現数
7	理由節の出現数	16	「!」の出現数
8	条件節の出現数	17	平均文字数
9	逆接語の出現数	18	Byte 数

1-16:文章別平均

表 4: SVM を用いたベストアンサーの推測結果

データ	適合率	再現率	F 値
恋愛相談，人間関係の悩み	0.540	0.851	0.661
生き方，人生相談	0.488	0.792	0.603
Windows 全盤	0.587	0.900	0.710
一般教養	0.546	0.826	0.670

そして相談や経験を求めるカテゴリだと予想された「恋愛相談，人間関係の悩み」,「生き方，人生相談」が 0.661, 0.603 であった．このように質問に対して「正解あり」の質問カテゴリの場合，その特徴が比較的に明確であり,「正解なし」カテゴリに比べて推測においても良い性能が期待できると考えられる.

3.3 考察

回答文を分析した結果，質問タイプの特徴は質問カテゴリでも反映されていることが分かった．例えば,「Windows 一般」のカテゴリには知識や情報を求める「正解あり」タイプの質問が,「恋愛相談，人間関係の悩み」では相談や経験を求める「正解なし」タイプの質問が多く含まれていたと考えられる．また，回答者たちはその質問タイプに応じて回答を投稿したので，回答文からもその特徴が発見できたと思われる.

4. 質問文の分析

前章では回答文の表層的特徴を用いてカテゴリの分析と分類を行った．本章では質問文の特徴を用いて質問タイプを分類し，カテゴリにどういった特徴が存在するのかを考察する.

4.1 予備実験

本節では前章で「正解あり」タイプとして判断された「Windows 全盤」のカテゴリを他のカテゴリと比較し，その結果から質問タイプによるカテゴリの分類を試みる．そして,「正解あり」タイプの質問文が持つ特徴を発見し，その関連性について検討する.

まず,「Windows 全般」から任意で 5,000 件の質問文を収集し，表 2 の特徴要素を抽出したデータを訓練データとする．そして,「恋愛相談，人間関係の悩み」,「生き方，人生相談」,「一般教養」,「Windows 全般」から任意で 5,000 件の質問文を収集し，同じように特徴要素を抽出したデータを評価データとす

表 3: 回答文の表層的特徴要素とベストアンサーとの相関係数

データ	特徴要素								
	1	2	3	4	5	6	7	8	9
恋愛相談，人間関係の悩み	-0.005	0.119	**0.239**	0.027	0.019	0.194	**0.247**	0.017	0.032
生き方，人生相談	-0.016	**0.203**	**0.266**	0.104	-0.002	0.142	**0.242**	0.025	0.031
一般教養	-0.021	-0.027	0.184	-0.008	0.023	-0.009	0.106	-0.014	0.000
Windows 全般	-0.028	-0.003	**0.227**	0.210	0.010	-0.004	0.153	0.008	0.031
全体	0.009	0.084	0.153	0.007	0.004	0.007	0.115	0.014	0.008
データ	特徴要素								
	10	11	12	13	14	15	16	17	18
恋愛相談，人間関係の悩み	**0.226**	0.033	0.105	0.102	**0.286**	0.014	-0.001	**0.281**	**0.224**
生き方，人生相談	**0.333**	0.022	0.087	0.036	**0.245**	-0.007	-0.034	**0.276**	**0.238**
一般教養	-0.005	0.006	**0.229**	**0.368**	0.096	-0.023	-0.035	**0.341**	**0.256**
Windows 全般	0.013	0.001	**0.321**	**0.375**	0.004	-0.042	-0.071	**0.341**	**0.305**
全体	0.015	-0.002	0.185	0.140	0.107	-0.006	0.005	**0.221**	**0.247**

る.「Windows 全般」から収集した評価データは訓練データと重複されないように収集した．そして，今回用いられる質問文はベストアンサーが選ばれ「解決済み」となった質問である．訓練データを SVM を用いて学習を行い，評価データの分類を行った．SVM の詳細は前章と同一である．その結果を表 7 に示す.

表 5: 機械学習による「正解あり」タイプ質問の分類結果

評価データ	正解あり	割合
恋愛相談，人間関係の悩み	1,291	0.258
生き方，人生相談	1,331	0.264
Windows 全般	3,847	0.769
一般教養	3,182	0.636

そして，相関係数を求めて「正解あり」タイプに影響を与える特徴要素について確認を行った．その結果，弱い正の相関 (0.2-0.4) をみせた特徴要素は「6. 意見要求の出現数」,「8. 条件節の出現数」,「12. 参照記号の出現数」,「18.Byte 数」であった．そして,「14. 分数」「18. 平均文字数」が中間の強さ (0.4-0.6) をみせた．知識や情報を求める「正解あり」タイプの質問は簡潔な文章で質問の意図だけど伝える文章となっている可能性が高いと考えられる．また，文法的要素としては条件節が影響を与えていた．質問の意図を正確に伝えるため条件節を使って他の参加者に説明をしたと思われる．

一方,「恋愛相談，人間関係の悩み」,「生き方，人生相談」のカテゴリでは「正解あり」タイプの分類から外された質問文が多かった．これは「正解あり」タイプの質問が持つ特徴との関係性があまり存在しなかったからだと思われる．これらのカテゴリのように「正解あり」タイプの特徴が強く影響を与えない場合，これを「正解なし」タイプのカテゴリが持つ傾向である可能性が高いと考えられる．

4.2 「Yahoo!知恵袋」のカテゴリ分類

前節で行った予備実験の結果からカテゴリ毎に質問タイプの分類を様々なカテゴリを対象として行い，カテゴリ単位で「正解あり」タイプの質問の割合を求めた．対象となったカテゴリは「Yahoo!知恵袋データ (第 2 版)」の中から質問数が多い上位 50 件 (88,000 件以上) を，質問はベストアンサーが選択された「解決済み」の質問文 5,000 件を任意で収集した．各々のカテゴリから収集された質問文に対して前節の訓練データを用いて機械学習による分類を行った．その結果を表 7 に示す.

表 6: 「正解あり」タイプの質問と表層的特徴要素の関係性

特徴要素	相関係数	特徴要素	相関係数
1	- 0.003	10	0.104
2	- 0.010	11	-0.023
3	- 0.121	12	**0.203**
4	- 0.004	13	**0.278**
5	- 0.063	14	**0.438**
6	**0.210**	15	0.131
7	0.159	16	0.063
8	**0.225**	17	**0.417**
9	0.099	18	**0.264**

分類結果,「携帯型ゲーム全般」,「Windows 全般」,「インターネット」,「ゲーム」,「英語」など知識や情報を求める質問カテゴリに「正解あり」タイプの質問文が多く含まれていたと考えられる．また,「恋愛相談，人間関係の悩み」,「子育ての悩み」,「家族」,「恋愛相談」のカテゴリが下位となった．これらのカテゴリには「正解なし」タイプの質問が多く含まれており，相談や経験を求める質問が多いと考えられる．

4.3 考察

本章では質問文の表層的特徴を用いて分析とタイプの分類を行った．前章で行った回答文の分析に比べて質問文では文章の文法的要素や文章表現に関する特徴が影響を与えないことが分かった．今回の実験では構造的要素が質問文を分類する要素として用いられたと考えられる．これは本研究で用いられた文章の表層的特徴要素が質問文の特徴発見に適切ではない部分があったとも考えられる．

「Yahoo!知恵袋」のカテゴリ分類では,「健康，病気，病院」,「法律相談」など相談を求めるカテゴリが上位となった．これ

らのカテゴリは回答に専門性が要求されるため知識や情報を求める質問の形が多かったと考えられる．また，これらを専門性のある相談として「正解あり」と「正解なし」タイプの特徴を共有している第三の分類としても考えられる [渡邊 11].

5. 考察と今後の課題

本研究では，質問を知識や情報を求める「正解あり」タイプと相談や経験を求める「質問なし」タイプの 2 つの分類を基準とし,「Yahoo!知恵袋」を対象に文章の表層的特徴を用いて分析と分類を行った．

質問タイプを分類することによって質問タイプによるベストアンサーまたは良質な回答を推定する手法を用いることが可能となる．回答文からベストアンサーを推定した実験ではカテゴリによって結果の差があった．このようにタイプ毎に推定手法を適用することによってより正確な良質な回答を推定することが可能になると考えられる．

回答文の分析では文章表現，文法的特徴要素のように文章が持つ特徴が，質問文の分析では構造的特徴が分類を判定する要素となった．特に質問文で文法的要素の特徴があまり発見されなかったことは質問文が回答文に比べて文章が短いことが原因として挙げられる．質問文に適した文法的要素，文章表現を補完することによってより厳密な分析と分類が可能になると思われる．

質問文の表層的特徴を用いて「Yahoo!知恵袋」のカテゴリを分類した実験では,「健康，病気，病院」,「法律相談」のような相談を目的とするカテゴリに「正解あり」タイプだと判定された質問が多かった．これは 2 つのタイプの特徴を共有する分類である可能性が高い．こういった特徴のカテゴリを分析して今後，質問分類の細分化の基準として検討していきたい．

6. おわりに

本稿では，コミュニティQA において質問文と回答文が持つ特徴を分析し，質問タイプと質問カテゴリにおける関連性を確認した．そして,「Yahoo!知恵袋」のデータに適用し，質問の分類を行い，その質問数を用いてカテゴリの分類も行った．今後表層的特徴の補完と分類の細分化によってより厳密な分析と推定が期待される．

謝辞

本研究の実施にあたって，ヤフー株式会社が国立情報学研究所に提供した「Yahoo!知恵袋データ (第 2 版)」を利用させて頂きました．深く感謝いたします．

参考文献

[三浦 08] 三浦麻子, 川浦康至：人はなぜ知識共有コミュニティに参加するのか： 質問行動と回答行動の分析, 社会心理学研究, 23(3), pp.233-245, (2008)

[渡邊 11] 渡邊 直人，島田 諭，関 洋平，神門 典子，佐藤 哲司:QA コミュニティにおける質問者の期待に基づく質問分類に関する一検討, DEIM Forum 2011，B5-1, (2011).

[朱 14] 朱成敏，武田英明：コミュニティQA における文章の表層的特徴に基づく回答文の分析，ARG 第 5 回 Web インテリジェンスとインタラクション研究会, WI2-2014-20, (2014).

表 7: 機械学習によるカテゴリの分類結果

#	カテゴリ名	正解あり	割合
1	携帯型ゲーム全般	3,931	0.786
2	**Windows 全般**	3,847	0.769
3	インターネット	3,754	0.751
4	ゲーム	3,712	0.742
5	英語	3,451	0.751
6	テレビゲーム全般	3,405	0.681
7	レディース全般	3,353	0.671
8	コスメ，美容	3,349	0.650
9	音楽	3,274	0.655
10	パソコン	3,231	0.646
11	病気，病症，ヘルスケア	3,256	0.651
12	ファッション	3,219	0.644
13	日本語	3,196	0.639
14	**一般教養**	3,182	0.636
15	楽器全般	3,043	0.609
16	健康，病気，病院	3,005	0.601
17	邦楽	2,968	0.594
18	国内	2,951	0.590
19	テレビ，ラジオ	2,891	0.578
20	住宅	2,803	0.561
21	鉄道，列車，駅	2,793	0.559
22	法律相談	2,764	0.553
23	家事	2,751	0.550
24	言語，語学	2,730	0.546
25	不動産	2,698	0.540
26	アダルト	2,603	0.521
27	サッカー	2,590	0.518
28	生物，動物，植物	2,585	0.517
29	ダイエット	2,524	0.505
30	芸能人	2,501	0.500
31	海外	2,488	2.498
32	コミック	2,472	0.494
33	バイク	2,409	0.480
34	プロ野球	2,399	0.480
35	アニメ	2,381	0.476
36	自動車	2,308	0.462
37	レシピ	2,235	0.447
38	ニュース，事件	2,191	0.438
39	料理，グルメ，レシピ	2,036	0.407
40	メンタルヘルス	1,987	0.397
41	料理，食材	1,901	0.380
42	妊娠，出産	1,850	0.370
43	恋愛相談	1,783	0.357
44	正治，社会問題	1,768	0.354
45	Yahoo!知恵袋	1,666	0.333
46	話題の人物	1,540	0.302
47	家族	1,498	0.300
48	子育ての悩み	1,331	0.266
49	**恋愛相談，人間関係の悩み**	1,291	0.258
50	Yahoo!オークション	1,179	0.236