疑似訓練データを用いたQ&Aサイトの質問分類

大森 勇輔　　森田 和宏　　泓田 正雄　　青江 順一
徳島大学大学院 先端技術科学教育部

1　はじめに

近年，OKWave[1]やYahoo!知恵袋[2]をはじめとしたQ&Aサイトの利用者が増加している．Q&Aサイトにおける質問は主に情報検索型と社会調査型の2つのタイプが存在する[1]．情報検索型とは，サーチエンジンや図書館のレファレンスサービスを利用して回答を探すことが可能な質問である．社会調査型とは，適切な回答の基準が存在せず，特定の個人あるいは集団に対してアンケート調査をおこなうことで各回答者の主観に基づく回答を得るような質問である．例として，「iCloudとは何ですか？」という質問は情報検索型で，「衆議院解散についてどう思いますか？」という質問は社会調査型である．

多くのQ&Aサイトには，質問のテーマやジャンルごとにカテゴリが用意されており，回答者は初めにカテゴリを選択し，回答する質問を決めることが基本となる．しかし，回答するカテゴリを決めていたとしても，情報検索型のように知識が必要な質問や，社会調査型のように比較的誰でも回答することができる質問が混在しているため，回答者が自身に適した質問を選択することは容易ではないと考えられる．また，その結果として，質問者が客観的な事実や情報を問う質問に対し，主観的な回答だけが寄せられるといった問題や，多くの回答を期待しても数件の回答しか寄せられないといった問題が生じると考えられる．現に，甲谷ら[2]がおこなったQ&Aサイトの利用状況に関する調査によると，ユーザの約50%が寄せられた回答の質に不満を抱いており，約25%が寄せられた回答の量に関して不満を抱いていることが報告されている．

そこで本研究では，Q&Aサイトにおける回答者が自身に適した質問を容易に選択可能となるように，Q&Aサイトの質問を情報検索型，社会調査型に分類をおこなう．また，質問を分類するにあたり，Q&Aサイトの特性と各質問タイプにおける特徴的な表現を利用した効率的な分類手法を提案する．

[1]http://okwave.jp/
[2]http://chiebukuro.yahoo.co.jp/

2　関連研究

Q&Aサイトの質問を機械的に分類する研究はいくつか存在する．田中ら[3]は機械学習と質問の特徴を利用して情報検索型質問の自動抽出を試みている．渡邊ら[4]はQ&Aサイトの質問を「事実」「根拠」「経験」「提案」「意見」の5つのタイプに分類できることを示し，経験則によって選択した70語を素性として機械学習により自動分類をおこなっている．また，林ら[5]は渡邊らの5つの質問タイプの再定義をおこない，Q&Aサイトから複数の質問が含まれていない質問文を人手で抽出後，キーワードによる手法と語の頻度を利用したスコア付けによる手法の2つを用いて自動分類をおこなっている．これら3つの研究の共通点として，質問に分類の正解情報を人手で付与した訓練データを作成し，訓練データを利用して未知の質問の分類をおこなっている．しかし，多くの質問に正解情報を人手で付与することは膨大なコストがかかるため，本研究では訓練データの作成コストを抑えた効率的な手法を提案し，質問の分類をおこなう．

3　特徴表現を用いた質問分類

本研究では，特徴表現を「情報検索型質問もしくは社会調査型質問に偏って出現する言語表現」と定義し，この特徴表現を利用してQ&Aサイトの質問分類をおこなう．

3.1　特徴表現の種類

特徴表現には2つの種類があると考えられる．1つは，Q&Aサイトにおけるすべてのカテゴリで共通して特徴表現となるものである(以下，共通表現と表記)．もう1つは，特定のカテゴリのみで特徴表現となるものである(以下，特有表現と表記)．特有表現を自動収集する場合，カテゴリ毎に訓練データを作成する必要があるため，個々の質問に正解情報を人手で付与することは膨大なコストがかかる．そこで，コスト削減のためQ&Aサイトの質問に疑似的な正解情報を付与したものを作成し(以下，疑似訓練データと表記)，共通表現と特有表現の自動収集を試みた．次節で疑似訓練データの作成手法について述べる．

【情報検索型】
可能ですか，できますか，本当ですか，理由は，何が原因，何故ですか，どうやって，情報
【社会調査型】
どうですか，気に入り，お勧め，皆さん，貴方は，アドバイス，意見，感想，好き

図 1．疑似訓練データの作成に使用した共通表現の例

3.2 疑似訓練データ作成手法

栗山ら[1]をはじめとした多くの Q&A サイトに関する研究では，社会調査型質問は情報検索型質問よりも比較的多くの回答が寄せられるという事を指摘している．この特性を利用し，まず回答数がk以上の質問に対して社会調査型の正解情報を付与する．次に，回答数がk未満の質問に対して人手で用意した 62 個の共通表現を用いて正解情報を付与する．用意した共通表現の例を図 1 に示す．林ら[5]のキーワードを用いた質問文の分類では，各キーワードに優先度を設定している．そのため，質問文内に複数のキーワードが存在した場合は優先度が高いキーワードに該当する質問タイプに分類される．本研究では疑似訓練データの作成にあたり，この林らの手法を参考にして各共通表現に優先度を付与することで，質問に異なるタイプの共通表現が含まれる場合に対応させた．

なお，kの値は十分な実験をおこなった上で最も精度が高いものに決定することが本来望ましいが，本研究では$k = 10$として疑似訓練データを作成した．

4 提案手法

提案手法では，疑似訓練データから特徴表現を自動収集し，この特徴表現を利用して Q&A サイトの質問分類をおこなう．以下で，提案手法の具体的な説明を述べる．

4.1 χ^2検定を用いた特徴表現の収集

この処理では，疑似訓練データを用いて特徴表現を収集する．まず，疑似訓練データにおける各質問を文単位に分割し，形態素解析をおこなう．次に，分割した各文から名詞，動詞，形容詞，形容動詞を抽出して N-gram(N=1,2,3)で表現する．その後，N-gram の各要素のχ^2値を算出し，偏りに有意差があることが確認されたものを特徴表現として収集する．以下にχ^2値の計算式を示す．

$$\chi(t_i)^2 = \sum_t \sum_c \frac{(O_{tc} - E_{tc})^2}{E_{tc}}$$

O_{tc}：観測値，E_{tc}：期待値(理論値)

$t \in \{t_i, \bar{t_i}\}$，$c \in$ {情報検索型, 社会調査型}

なお，計算式におけるt_iは N-gram の各要素を表している．また，有意水準は一般的に用いられる 0.05 を使用した．

4.2 特徴表現へのスコア付与

次に，前節で収集した特徴表現にスコアを付与する．疑似訓練データにおいて情報検索型の正解情報が付与された質問での特徴表現の出現頻度と，社会調査型の正解情報が付与された質問での特徴表現の出現頻度の差分$d(s_i)$を計算し，スコアとした．以下に計算式を示す．

$$d(s_i) = \frac{P(s_i) - N(s_i)}{P(s_i) + N(s_i)}$$

$P(s_i)$：s_iが出現する情報検索型質問の数

$N(s_i)$：s_iが出現する社会調査型質問の数

なお，計算式におけるs_iは各特徴表現を表している．ここで，$d(s_i)$が正の値である場合はs_iが情報検索型質問に偏って出現していることを意味し，負の値である場合はs_iが社会調査型質問に偏って出現していることを意味している．

4.3 未知の質問へのスコア付与

最後に，正解が未知の質問Qに対し，前節で求めた$d(s_i)$を用いてスコア付けをおこなう．その後，スコアが正の値であれば質問を情報検索型に分類し，負の値であれば社会調査型に分類する．以下に式を示す．

$$Score(Q) = \sum_i d(s_i)$$

$$Score(Q) = \begin{cases} \text{正の値} \Rightarrow \text{情報検索型} \\ \text{負の値} \Rightarrow \text{社会調査型} \end{cases}$$

5 実験

提案手法の有効性を確認するため，質問分類の評価実験をおこなった．分類結果については適合率，再現率，F 値を用いて評価した．

5.1 実験に使用するデータ

実験に使用する疑似訓練データとテストデータは Q&A サイトの 1 つである Yahoo!知恵袋の質問検索 API を用いて収集した．また，対象カテゴリとして解決済みの質問が 100 万件を超えるカテゴリの中から「地域、旅行、お出かけ>国内」,「暮らしと生活ガイド>料理、グルメ、レシピ(以下，料理と表記)」,「スポーツ、アウトドア、車>自動車」の 3 つを選択した．

5.1.1 疑似訓練データについて

「国内」カテゴリでは，人口が上位 10 件の都道府県名，「料理」カテゴリでは，国内で収穫量が上位 10 件の野菜

表 1. 疑似訓練データの内訳

	情報検索型	社会調査型	分類不可能
国内	18,835	10,609	2,146
料理	16,996	9,028	1,767
自動車	20,106	8,352	1,785

表 2. 特有表現と考えられる特徴表現の例

	カテゴリ	$d(s_i)$	χ^2値
スポット-あり	国内	-0.455	233.3
由来	国内	0.882	13.42
作り方-教え	料理	0.766	164.4
味噌汁-具	料理	-0.522	104.9
配線	自動車	0.816	88.78
試乗-し	自動車	-0.647	69.22

名,「自動車」カテゴリでは，10 件の国内自動車のメーカー名をクエリとして質問を収集し, 疑似訓練データの作成に使用した. 作成した疑似訓練データの内訳を表 1 に示す. 表 1 における分類不可能とは，回答数がk未満かつ人手で用意した共通表現を含まない質問である. 提案手法に対する実験には，これらを除いたものを使用する.

また，疑似訓練データから抽出した特徴表現のうち，特有表現と考えられるものの例を表 2 に示す.「スポット-あり」や「由来」といった表現は,「国内」カテゴリにおいて各都道府県の人気スポットやオススメスポットを募る質問や, 観光地名の由来などを問う質問で多く出現していた．また,「作り方-教え」や「味噌汁-具」といった表現は,「料理」カテゴリにおいて特定の料理の作り方を問う質問や, みそ汁の具に何を入れるかのアンケートをおこなう質問で多く出現していた.「配線」や「試乗-し」といった表現は,「自動車」カテゴリにおいて車の配線部品の取り付けた方を問う質問や, 特定の車を試乗した方に感想や意見を求める質問で多く出現していた.

5.1.2 テストデータについて

「国内」カテゴリでは徳島,「料理」カテゴリでは茄子,「自動車」カテゴリではフォルクスワーゲンをクエリとして質問をそれぞれ 300 件収集し, 正解情報を人手で付与したものをテストデータとして使用した．なお，質問だけでは正解の判断が難しい質問については, ベストアンサー及びベストアンサー以外の回答と質問者のお礼のコメントを考慮した上で正解を付与した. 十分な考慮をしても分類ができない質問，情報検索と社会調査の両方を問う質問，質問者の主張に対する反応を求めている質問, 記述として何が書かれているか理解できない書き込みは「その他」に分類した.

表 3. テストデータに人手で正解情報を付与した内訳

	情報検索型	社会調査型	その他
国内	109	154	37
料理	120	139	41
自動車	150	112	38

表 4. 情報検索型の分類結果(疑似訓練データ作成手法)

	適合率	再現率	F 値
国内	0.566	0.945	0.708
料理	0.711	0.900	0.794
自動車	0.712	0.940	0.810

表 5. 社会調査型の分類結果(疑似訓練データ作成手法)

	適合率	再現率	F 値
国内	0.986	0.442	0.610
料理	0.933	0.604	0.734
自動車	0.943	0.446	0.606

テストデータの分類の内訳を表 3 に示す. 全体の約 13%が「その他」に該当し，これらを除いたものを実験に使用した．また，疑似訓練データにおいても「その他」に該当する質問が 13%前後含んでいると考えられが，本研究では除外はおこなわなかった.

5.2 疑似訓練データ作成手法の精度確認

本手法は，疑似訓練データの質によって，結果が大きく変わると考えられるため, 疑似訓練データの質がどの程度であるかを確認するための実験をおこなった. テストデータに対して疑似訓練データの作成手法を適用し, 結果を評価した．結果を表 4，5 に示す．3 カテゴリ全てにおいて情報検索型質問は再現率が高く, 適合率が低い結果となり, 社会調査型質問は適合率が高く, 再現率が低い結果となった．また，F 値は情報検索型質問の方が社会調査型質問と比較して高い結果となった.

5.3 提案手法の精度確認

テストデータに対して，N-gram を単体もしくはいくつか組み合わせた手法(計 5 通り)を適用した．また，ベースラインに，$k = \infty$として疑似訓練データ作成手法を適用した結果を採用した．結果を表 6，7 に示す．1-gram と 2-gram を組み合わせた手法が，3 カテゴリ全てにおいて最も多くの質問を正しく分類することができた．一方で, 3-gram を用いた手法は，あまり良い結果が得られなかったため, 提案手法においては 1-gram，2-gram を用いた手法が質問を分類するにあたり最も有効であると結論づけることができると考えられる．また,「料理」カテゴリの社会調査型においては，1-gram と 2-gram を組み合わせ

表 6. 情報検索型の分類結果(提案手法)

	国内			料理			自動車		
	適合率	再現率	F 値	適合率	再現率	F 値	適合率	再現率	F 値
ベースライン	0.566	0.945	0.708	0.694	0.908	0.787	0.706	0.947	0.809
1-gram	0.608	0.853	0.710	0.593	0.983	0.740	0.726	0.900	0.804
2-gram	0.755	0.367	0.494	0.847	0.600	0.702	0.889	0.480	0.623
3-gram	0.398	0.303	0.344	0.630	0.242	0.349	0.656	0.280	0.393
1,2-gram	0.696	0.798	0.744	0.674	0.967	0.795	0.832	0.860	0.846
1,2,3-gram	0.679	0.817	0.742	0.671	0.967	0.792	0.819	0.873	0.845

表 7. 社会調査型の分類結果(提案手法)

	国内			料理			自動車		
	適合率	再現率	F 値	適合率	再現率	F 値	適合率	再現率	F 値
ベースライン	0.986	0.442	0.610	0.940	0.568	0.709	0.959	0.420	0.584
1-gram	0.887	0.610	0.723	0.982	0.403	0.571	0.803	0.545	0.649
2-gram	0.788	0.773	0.780	0.866	0.741	0.798	0.623	0.857	0.722
3-gram	1.000	0.026	0.051	0.800	0.029	0.056	0.500	0.009	0.018
1,2-gram	0.866	0.753	0.806	0.964	0.583	0.726	0.804	0.768	0.785
1,2,3-gram	0.875	0.727	0.794	0.964	0.576	0.721	0.814	0.741	0.776

表 8. 有意水準 0.05 を満たした特徴表現の数(1-gram)

	情報検索型(割合)	社会調査型(割合)
国内	556(22.1%)	1,961(77.9%)
料理	741(42.1%)	1,019(57.9%)
自動車	1,205(21.6%)	4,361(78.4%)

た手法よりも 2-gram 単体で使用した手法の方が高い精度で分類できていた.

5.4 考察

「料理」カテゴリにおいて 2-gram 単体で使用した手法の方が高い結果を得られたことから，1-gram の特徴表現が取得できなかったことに原因があると考えられる.

表 8 にχ^2検定を用いて抽出した社会調査型質問における特徴表現の数(1-gram)を示す. 他の 2 カテゴリと比較して社会調査型の特徴表現が少なく, なおかつ情報検索型と社会調査型の割合がほぼ等しいことがわかる. また，他の 2 カテゴリでは社会調査型質問の分類においては, 1-gram 単体で使用した手法の方が共通表現のみを用いたベースラインよりも精度が高いことから,「料理」カテゴリでは 1-gram の特有表現が収集できなかったと考えられる. 解決策としては, χ^2検定の有意水準を各 N-gram で変更すること，疑似訓練データを拡充することが挙げられる.

6 まとめと今後の課題

本研究では, 訓練データに正解情報を人手で付与するコストを削減するために, 正解情報を疑似的に付与する訓練データを作成する手法を提案した. また, χ^2検定を用いて質問分類のための特徴表現を抽出し, この特徴表現にスコア付けをおこなうことで質問を分類する手法を提案した. 評価実験の結果から 1-gram と 2-gram を組み合わせた提案手法において良好な結果を得ることができた.

今後は，質問の分類精度をさらに高めるため，疑似訓練データの質を向上させることを目指す. 具体的な課題としては，人手で用意した共通表現を拡充すること，kの最適値を調査することが挙げられる.

参考文献

[1] 栗山和子，神門典子：Q&A サイトにおける質問と回答の分析，情報処理学会データベースシステム研究会，Vol.2009-DBS-148，No.19，2009.

[2] 甲谷優，岩田具治，塩原寿子，藤村考：QA コミュニティにおける複数情報源を用いた効果的な質問推薦，情報処理学会論文誌 データベース，Vol.3，No.4，pp.34-47，2010.

[3] 田中友二，望月崇由，八木貴史，徳永幸生，杉山精：Q&A サイトにおける情報検索型質問の自動抽出，情報処理学会第 74 回全国大会, 6T-4, pp.529-530, 2012.

[4] 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，第 3 回データ工学と情報マネジメントに関するフォーラム(DEIM 2011)，B5-1，2011.

[5] 林秀治，山本和英：質問意図による QA サイト質問文の自動分類，電子情報通信学会言語理解とコミュニケーション研究会, NLC2013-10, pp.51-56, 2013.