6T-4

Q&Aサイトにおける情報検索型質問の自動抽出

田中 友二†　望月崇由*　八木貴史*　徳永 幸生†　杉山 精‡
芝浦工業大学†　ＮＴＴレゾナント株式会社*　東京工芸大学‡

1. はじめに

近年，大量の情報が配信されている World Wide Web(以下，WWW)上から情報を効率的に入手する手段として，種々の検索エンジンが利用されている．しかし，検索者がいつでも検索目的に適した検索語を思いつくとは限らない．

そこで，ユーザは検索エンジンで情報入手ができなかった場合，Q&A サイトでの質問の投稿や，質問回答ログの閲覧などを利用する．Q&A サイトとは，質問者が自然文で書いた質問に対して回答者が自然文で回答する，人同士の知識の共有をベースにしたサービスである．

Q&A サイトの質問は情報検索型，社会調査型，非質問型の 3 つに分類できる[1]（表 1）．その中で，情報検索型質問は客観的な情報を求めるため，情報検索型質問の質問回答ログを分析し質問者に共通の情報要求を抽出すれば，検索者の情報入手の支援に活かすことができ，また，明解な回答を作り易い情報検索型質問のみを提示することで，回答者に質問の選択支援を行い，回答を促すことができると考えられる．

そこで，客観的な情報や事実を求める質問である情報検索型質問の自動抽出を試みた．本稿では，機械学習による抽出と質問の特徴を利用した抽出を統合した自動抽出を検討した．

表 1.　Q&A サイトにおける質問の分類

質問の型	詳細
情報検索型	客観的な事実や情報を求める質問 「人名」，「エラーの解決方法」など
社会調査型	個人的な助言・意見・経験などを求める質問 「推薦」，「助言」など
非質問型	・記述として何が書かれているのか 分析者に理解できなかった質問 ・質問者の主張に対する反応を求めている質問

Automatic Extraction of Factual Information in Q&A Site
†Yuji TANAKA(ma11105@shibaura-it.ac.jp)
*Takayoshi MOCHIZUKI（mochizuki@ntt.co.jp）
*Takashi YAGI（t-yagi@ntt.co.jp）
†Yukio TOKUNAGA（tokunaga@shibaura-it.ac.jp)
‡Kiyoshi SUGIYAMA
†Shibaura Institute of Technology
*NTT Resonant Inc.
‡Tokyo Polytechnic University

2. 自動抽出アルゴリズム

情報検索型質問を自動抽出するために，機械学習と質問の特徴を利用する．

2.1 機械学習による抽出

予め人手で分類した質問回答ログ（教師データ）から情報検索型質問に特徴的な出現単語を抽出する．その結果をもとに未知の質問回答ログ（検証データ）から情報検索型質問を抽出する．本研究では，機械学習の 1 つとして，ナイーブベイズ分類器を用いた．

ナイーブベイズ分類器では，出現単語からそれぞれの質問回答ログが情報検索型質問である確率を学習する．その学習結果を用いて，未知の質問回答ログが情報検索型質問である確率を推定する．使用する単語を $w = (w_1, w_2, \cdots, w_j, \cdots, w_n)\ (0 \leq j \leq n)$，情報検索型質問を c_0，情報検索型質問でない質問を c_1 とすると，分類結果 $\hat{c}$ は式(1)となる．

$$\hat{c} = \arg\max P(c_i) \prod_j P(w_j \mid c_i) \quad \cdots (1)$$

また，質問回答ログの単語を得るために形態素解析器 MeCab[2]を用いた．そして，品詞が名詞，形容詞，動詞である単語を学習に利用した．ただし，各品詞の中でも非自立語・接尾語・代名詞・数詞・地名はどの型の質問でも出現し，特徴的な単語とならないと考えられるため学習と推定から除外した．

2.2 頻出語による除外

情報検索型質問には使われやすい単語，使われにくい単語がある．そこで，3 種類の質問ごとに頻出語を調査した．その結果，質問文に「おすすめ」「アドバイス」を含む質問は情報検索型質問でない場合が多いことがわかった．そこで，それらを含む質問を情報検索型質問から除外することとした．

3. 評価データの作成

文献[1]では表 1 のように質問を 3 種類に分類したが，1 件の質問で情報検索型と社会調査型の両方を求めている質問（複合型）であったり，質問文だけでは情報検索型か社会調査型のどちらの事柄を求めているか判定が難しく分類が困難

な質問（分類困難型）が存在した．そのため，分類型は，情報検索型，社会調査型，非質問型，複合型，分類困難型，の5つとした．

本研究では，Q&Aサイトの2011年3月10日から2011年10月25日までに「国内旅行（全国）」カテゴリに投稿された1000件の質問回答ログを用いた．この1000件の質問を分類した結果を表2に示す．

表2.　質問の分類結果

分類の型	件数
情報検索型	235
社会調査型	626
非質問型	10
複合型	40
分類困難型	89

4. 評価実験

4.1 評価データを用いた実験

3節で作成した評価データの中で分類困難型の質問を除いた911件の質問を用いて，情報検索型質問を自動抽出し評価実験を行った．その際，評価値として，適合率と再現率を用いた．適合率と再現率は式(2)のように定義される．

$$(\text{適合率}) = \frac{R}{N} \quad , \quad (\text{再現率}) = \frac{R}{C} \quad \cdots (2)$$

R：適切に分類された情報検索型質問の件数
N：情報検索型質問と分類された質問の件数
C：評価に用いた情報検索型質問の件数

また，ナイーブベイズ分類器では教師データが必要であり，911件のデータを4分割（228件3セット，227件1セット）し，このデータに対して4分割交差検定を行った．その際，教師データの素性は名詞，形容詞，動詞を用いた7パターンとした．実験結果を表3に示す．

表3.　適合率と再現率の実験結果

パターン	適合率	再現率
名詞	0.667	0.581
形容詞	0.390	0.045
動詞	0.687	0.261
名詞＋形容詞	0.705	0.589
名詞＋動詞	0.758	0.572
形容詞＋動詞	0.748	0.292
名詞＋形容詞＋動詞（全て）	0.784	0.563

4.2 考察

質問回答ログの分析や質問の推薦においては抽出の正確性である適合率が重要である．実験結果から，全て（名詞，形容詞，動詞）の品詞を素性とした場合が最も高い適合率となった．また，形容詞のみを素性とした場合が最も低い適合率となった．

全ての品詞だけでなく「名詞＋動詞」の適合率も高いことから，名詞と動詞の出現単語が抽出における特徴となっていると考えられる．また，頻出語による除外を行うことで適合率の向上が確認できた．

次に，抽出ミスとなった原因を分析した．ナイーブベイズ分類器は教師データの出現単語を基に学習している．そのため，たとえば社会調査型質問であっても情報検索型質問で多く使われている単語が多ければ情報検索型質問と分類されていた．抽出ミスとなった社会調査型質問を図1に示す．

急行「きたぐに」指定席の混み具合は？
普段は、そんなに混雑する事は無いと聞いたのですが8月上旬に行われる長岡花火の時はどうでしょうか？
花火を見てから、大阪までの移動に利用したいと考えてるんですが、1ヶ月前の発売開始日と同時に指定席(座席・寝台共に)が売り切れるなんて事がありえますか？花火大会当日に利用した事がある方や、混雑状況に詳しい方、教えて下さい。

図1.　抽出ミスとなった質問

質問に出現する「指定」「席」という単語が情報検索型質問で多く使われていた．そのため，ナイーブベイズ分類器は情報検索型質問に分類したと考えられる．ナイーブベイズ分類器における「指定」という単語に与えられたパラメータ値を表4に示す．このような抽出ミスの対策として，抽出に有効だと考えられる出現単語に重みを与えることが考えられる．例えば，質問における疑問文の箇所の出現単語に重みを与えることが有効と考えられる．

表4.　「指定」に与えられたパラメータ値

	情報検索型の教師データ	情報検索型でない教師データ
「指定」の出現回数	34	25
全単語の出現回数の総和	7003	23873
単語の種類数	5031	5031
ナイーブベイズ分類器のﾊﾟﾗﾒｰﾀ	0.0029	0.0009

5.おわりに

本稿では，機械学習による抽出法と質問の特徴を利用した抽出法を統合し，情報検索型質問の自動抽出を試みた．その結果，名詞，形容詞，動詞を機械学習の素性とすることで，適合率0.784，再現率0.563という値を得られた．

今後は，他のカテゴリでの評価実験と情報検索型質問を利用したシステムの提案を検討する．

――― 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, 情報処理学会研究報告, Vol.2009-FI-95 No.19
[2] MeCab，http://mecab.sourceforge.net/
[3] 田中友二，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とクラスタリング，FIT2011 第10回情報科学技術フォーラム，D-013(Sep. 2011)