Q&Aサイトにおける質問文自動分類のための特徴分析

○鳥巣 寿明*　寺本 優香**　馬場 睦也**　波多野 賢治*
* 同志社大学 文化情報学部　** 同志社大学大学院 文化情報学研究科

1．研究背景

■情報要求が曖昧な検索者への支援方法

・CQAリソースを用いたクエリ拡張[1]

質問文：紀子様が外出される時にいつも乗っている グレーの車の名前が知りたいです

回答文：日産のフーガです

紀子様　外出　グレーの車 ⇒ フーガ　具体的！

■Q&Aサイトの質問文分類

・Yahoo!知恵袋の質問文の分類[2]

質問の型	説明
情報検索型	サーチエンジンなどを利用し回答を探すことが可能な質問
社会調査型	アンケート調査を行うことで回答を得るような質問
非質問型	自分の主張に対する反響・反応を求めている記述表現

社会調査型の質問文の例
「超ロマンチストでナルシストな男の人をどう思いますか？」

予想される回答文の例
「自意識過剰だと思います．」

曖昧な情報要求を具体化したキーワードが含まれていない

既存のクエリ拡張：質問文＋回答文 → 質問回答ペア → 拡張クエリ抽出

質問の型を考慮したクエリ拡張：質問文 → 不要な質問文を省き有効な質問文のみ利用 → 有効な質問文＋回答文 → 質問回答ペア → 拡張クエリ抽出

クエリ拡張に有効な質問文のみを扱うことでノイズを削減

クエリ拡張の精度向上

クエリ拡張に用いる質問文は膨大で質問の型ごとに人手で分類するのは困難 ⇒ 自動分類が必要

【本研究の目的】

CQAリソースを用いたクエリ拡張の精度向上のための質問文自動分類法の提案

[1] Gao, L., Lu, Y., Zhang, Q., Yang, H., Hu , Y.(2016). Query expansion for exploratory search with subtopic discovery in Community Question Answering, in 2016 International Joint Conference on Neural Networks(IJCNN), pp.4715-4720, IEEE.
[2] 栗山 和子, 神門 典子(2009).『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』, 2009-FI-95(19), pp.1-8.

2．関連研究

■Yahoo!知恵袋の質問文の分類[3]

・質問の類型を抽出することを目的とし，Yahoo!知恵袋の質問文について統計的に調査

・「求解型」「共感型」「調査型」「釣り型」「その他」を定義

質問の型	定義
求解型	具体的な答えを探す質問
共感型	読者の共感や事例を求める依頼文
調査型	世論調査に類する質問
釣り型	釣りや笑いを誘うネタ型の質問文
その他	上記の型に当てはまらない文

各型の定義が，回答文に曖昧な情報要求を具体化したキーワードが含まれているかを明確に分類できていると判断

【関連研究の問題点】

自動分類法を提案できていない

[3] 劉 舒暢, 伊藤 栄典, 中島 幸子, 廣川 佐千男(2015).『Yahoo!知恵袋の質問文分類のための質問文分析』,『言語処理学会第21回年次大会発表論文集』, pp.357-360.

3．提案手法

■提案手法の流れ

Yahoo!知恵袋(第二版)[4]

質問データ 約1600万文 → 標本抽出 → 質問文 508文 → 型情報付与 → 型情報付き質問文 → 分析結果 → 自動分類

各質問文の型を判定し3名の判定が一致した質問文にその型を付与（協力者5名）

ランダムフォレストなどを行い質問の型に影響する特徴を分析

質問の型に強く影響する特徴を用いて多項ロジスティック回帰分析などで自動分類

■使用データ

・Yahoo!知恵袋(第二版)の質問文約1600万文から層化無作為抽出法により抽出した508文を使用

■質問文の抽出方法

・層化無作為抽出法

母集団を特定の性質でいくつかの層に分け，あらかじめ決めた標本の割合で各層から独立に標本を非復元単純無作為抽出する方法[5]

母集団 1000人 → 標本 100人

500人	50人
350人	35人
150人	15人

・カテゴリ情報ごとに質問の型の分布が異なる[3]

・中分類あるいは小分類を層とした場合，標本数が膨大になる

⇒ カテゴリ情報の大分類を層とみなす

■正解データの作成

・劉らの分類法をさらに厳格化した定義に基づき，実験協力者5名に与えられた質問文がどの型に当てはまるのかを判定してもらう

質問の型	定義
求解型	固有名詞や具体的な方法などを含む回答文が一つでもあれば解決するような質問
共感型	質問者の意見に対する回答者の意見を求める質問
調査型	ある事物事象に対する回答者の意見をより多く求めるような質問
釣り型	閲覧者の笑いを誘うような質問，あるいは問題を解決するのではなく，明らかに注目を集めることや回答数を増やすことを目的としている質問
その他	自然文以外の情報（画像，URL，ソースコード，etc.）を求めるような質問，あるいは上記の型に当てはまらない文

■分析手法

特徴量の重要度を可視化することが可能

・ランダムフォレスト

型情報付き質問文 → 無作為 → 標本1 … 標本m → モデル作成 → tree1 … tree$m$ → 結果1 … 結果m → 多数決で分類決定

目的変数：質問の型（求解型，共感型，調査型，釣り型，その他）

説明変数：Yahoo!知恵袋(第二版)で各質問文に付与されているデータ（カテゴリ情報，回答数，etc.）
JUMAN++[6]で形態素解析を行い，JUMAN形式の出力で得られた単語や品詞の出現頻度などのデータ（bag-of-words，内容語の出現頻度，etc.）

■自動化の手法

・多項ロジスティック回帰分析
目的変数が離散値の多値分類として一般的な手法
質問の型に強く影響する特徴を用いて自動分類を行う

[4] Yahoo!知恵袋データ, https://www.nii.ac.jp/dsc/idr/yahoo/chiebkr2/Y_chiebukuro.html (2018/7/25 閲覧)
[5] 鄭 躍軍, 金 明哲(2011).『Rで学ぶデータサイエンス17 社会調査データ解析』, 共立出版.
[6] JUMAN++, http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN++ (2018/11/27 閲覧)

■検討中の評価指標

・評価指標として適合率 再現率，F値を検討

・各型ごとにこれらを算出しすべての型のF値の平均で提案手法の性能を評価

$$\text{適合率} = \frac{\text{分類された質問文のうち正解の質問文数}}{\text{分類された質問数}}$$

$$\text{再現率} = \frac{\text{分類された質問文のうち正解の質問文数}}{\text{各型の正解となる質問文数}}$$

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

4．今後の課題

・分析および自動分類の手法の実装
・提案手法の評価
・本研究と比較する先行研究の調査
・分析に用いる説明変数の検討
・回答文がある状態での型情報付与作業の実施

連絡先：bip0270@mail4.doshisha.ac.jp