

共起単語に着目した自治体 FAQ のための質問生成

保延 渉[†] 菊井 玄一郎[‡]岡山県立大学大学院 情報系工学研究科[†]岡山県立大学 情報工学部[‡]

1. はじめに

大企業、教育機関、官公庁などの大型のポータルやウェブサイトでは、利便性向上のために“よくある質問” (FAQ) が提供されていることが多い。市町村といった各種自治体のサイトでも FAQ が提供され、納税、教育、各種手続きなどのジャンルごとにまとめられている。しかし、次章でも述べる通り、FAQ の充実度は自治体によって差があり、FAQ そのものがウェブサイトには存在しない自治体もある。

そこで、本研究では、自治体における FAQ の現状について調査し、特に FAQ が存在しない、あるいは内容に乏しい自治体向けに、FAQ の作成を支援する手法について検討する。

2. 自治体 FAQ の現状

研究を進めるにあたり、岡山県内にある 27 の基礎自治体 (市町村) を対象に FAQ の整備状況を調査した。その結果を表 1 に示す。

全てのジャンルについて充実している自治体はわずか 2 か所しかなく、およそ半数の自治体で FAQ そのものが存在していないことが分かった。これには様々な要因が考えられるが、少なくとも FAQ を作成するのに手間やコストがかかることが挙げられる。FAQ を作成するには、よくある、あるいはあると想定される質問とその回答を用意しておく必要があるが、特に質問を用意することが課題である。表 1 内の“全てのジャンルにおいて充実した自治体”の 1 つである倉敷市は、質問を人手以外で生成する方法として、コールセンターと提携し、市民の方から寄せられた質問に回答すると同時に、FAQ として登録するシステムを採用している。しかし、すべての自治体で導入するには、コールセンターやシステムの保守といったコストがかかるため、予算の限られた小さな自治体では難しい。この問題を解決するためには FAQ の作成を自動化、あるいは、省力化するシステムの実現が望ましい。

Question Generation for Local Government FAQs Using Co-occurrence Words

[†]HONOBE Wataru and KIKUI Genichiro (Okayama Prefectural University)

表 1: 岡山県の自治体 FAQ の調査結果

全てのジャンルにおいて充実した自治体 [⊖]	2 か所(7.4%) [⊖]
ジャンルは豊富だが質問数が少ない(質問不足である)自治体 [⊖]	6 か所(22.2%) [⊖]
用意されているジャンルは限られているが、 [⊖] その FAQ は充実した自治体 [⊖]	5 か所(18.5%) [⊖]
用意されているジャンルが限られており、 [⊖] その FAQ もあまり充実していない自治体 [⊖]	1 か所(3.7%) [⊖]
FAQ が存在しない自治体 [⊖]	13 か所(48.1%) [⊖]

3. 関連研究

FAQ の自動生成については先行研究がある。松本ら[1]は電子掲示板の質問スレッドから 5 種類の意図表現「挨拶」「状況」「行為」「質問」「意思」を付与し、取捨選択ルールを用いて FAQ の質問文としてふさわしい部分を抽出する手法を提案している。この方法は有用であるが自治体については「電子掲示板」ようなものがないため、そのまま利用することはできない。

倉田ら[2]は多くの文書に共通して記載されている要点を自動的に抽出することで、文書群の概要の把握を可能にし、コンタクトセンターでの FAQ 作成の支援を提案している。この方法は表現の頻度に依存しているため本研究の分野への適用は難しいと思われる。

4. 提案手法

本研究では、自治体の FAQ は自治体間で共通部分が多く、また、充実した FAQ を持つ自治体がいくつか存在することに注目して、そのような自治体から収集した FAQ の質問部分を入力とし、質問テンプレートを出力する方法を提案する。ここで質問テンプレートとは次に示すように、空欄に対応する文字を埋めることで質問文として用いることができる形式のことである。

“[場所]市の法人市民税の税率について知りたい”

提案手法全体は次の 3 つのステップからなる。

ステップ1：共起語を用いた質問文の選択

ステップ2：質問形式による質問の分類

ステップ3：テンプレート生成

以下、各ステップについて説明する。

ステップ1では質問文内の単語の共起関係を調べるために、質問文を形態素解析して単語の集合とし、アソシエーション分析を適用する。アソシエーション分析とは、データ間の相関関係を発見する場合に用いられるデータマイニング手法の一つであり、今回は注目単語(例：税、幼稚園など)をあらかじめ1つ設定しておき、注目単語と共起することが多い単語の組み合わせを質問文で共起しやすいとみなす。注目単語に対してよく共起2つ以上の単語の組み合わせ(例：注目単語”税”と”軽自動車”，”知る”)をリストアップした。

ステップ2では、まず、ステップ1でリストアップした単語の組み合わせをすべて含んだ質問文を抽出し、次に、その質問文を yes/no 型(はいかいいえで答えられるような質問)、疑問詞型(“どこですか?”や”なぜですか?”のような主語のある文章で答える必要のある質問)、非質問文型(“教えてください”など、文末が疑問の形でないもの)の3種類に分け、まとめたものを新たな質問群とする。

ステップ3では質問の中に固有名詞が含まれる場合、その部分を挿入可能な空白に書き換え、質問文のテンプレートとして収録する。

以上の手順により、任意の自治体で利用可能な質問を、形式ごとでまとめることができる。

5. 実験

提案手法を用いて、正しく質問文が分類され、生成できているかの検証を行った。

i. 実験方法

表1において”全てのジャンルにおいて充実した自治体”であった岡山市と倉敷市のFAQに加え、同じく充実している名古屋市、福岡市、大阪市のFAQを用いて、”税”のジャンルの質問文423件を対象に質問形式の分類と生成が正しく行われているか実験を行った。注目単語を”税”とし、”税”と共起することが多かった単語を含む質問文を分類の対象として、分類の正確性と質問のカバー率について人手で判断した。

ii. 実験結果

まず、質問を yes/no 型、疑問詞型、非質問文型の3種類に分けた後の質問件数を表2にまとめた。全体の約43%の質問がステップ1の段階でリストアップされたことがわかる。

表2：質問分類結果

yes/no 型 ^⓪	40 件(9.4%) ^⓪
疑問詞型 ^⓪	78 件(18.4%) ^⓪
非質問文型 ^⓪	63 件(14.9%) ^⓪
合計 ^⓪	181 件(42.8%) ^⓪

次に、分類された質問の比較を行う。共起した単語が、”税”，”市民”，”通知書”で振り分けられた疑問詞型質問の例を表3に示す。元の質問文に地名があった場合は、各自治体で書き換えができるよう[場所]という形の空白に置き換わっている。この生成された質問文を4章のステップ3のテンプレートとして、各自治体に利用されることを想定している。

表3：質問の例(”税”，”市民”，”通知書”を含む「疑問詞型」の質問)

転出して他市町村に住んでいるのに、[場所]から市民税・県民税の納税通知書が届いたのはなぜですか ^⓪
昨年退職して、今年は働いていないのに市民税・県民税の納税通知書が届いたのはなぜですか ^⓪

最後に、ステップ1でリストアップされなかった質問について調べた。その例を表4に示す。

表4：リストアップされなかった例

納税貯蓄組合とは何ですか? ^⓪
市県民税の所得割とは何ですか? ^⓪
地籍図を見ることはできますか? ^⓪

リストアップされなかった質問は、文が短すぎて共起できる単語が少ない、かつ”所得割”のように珍しい単語が含まれる場合が多かった。また、当たり前ではあるが、注目単語の”税”が含まれない質問もリストアップされなかった。

6. まとめ

本稿では、文章内の単語の共起性に着目し、FAQが充実している自治体の質問群を質問の型ごとに分類し、そこからどの自治体でも利用できるように質問テンプレート集の作成を目的とし、実際のFAQを用いて質問の分類と生成を行った。

今後は別カテゴリにおいても同様の実験を行うことや、より正確な質問収集ができるよう改善に努めていきたい。

参考文献

[1] 松本 和之ほか，”FAQ 生成機能を持つ新しい電子掲示板の提案”，情報処理学会第66回全国大会，pp.2-185 2-186，2004.

[2] 倉田 早織ほか，”代表文生成技術とFAQ作成の効率向上”，東芝レビュー，pp57 61，2011