自治体 FAQ の比較マイニング

伊藤拓海[1]　山口健史[1]　田然[1]　松田耕史[1]　岡崎直観[2]　乾健太郎[1,3]
東北大学[1]　東京工業大学[2]　理化学研究所 AIP センター[3]
{t-ito, k.yamaguchi, tianran, matsuda, inui}@ecei.tohoku.ac.jp
okazaki@c.titech.ac.jp

1　はじめに

税金やごみ処理など日常生活においては，様々な手続きやルールが存在する．それらの中には複雑なものもある．そのため，自治体への住民からの問い合わせも多い．札幌市をはじめ，いくつかの自治体ではコールセンターを設立し，住民からの問い合わせに対して対応している．札幌市のコールセンターへはひと月あたり 1 万を超える問い合わせがある[*1] ．これらの問い合わせに対し，より的確に素早く情報を住民に伝えるために，FAQ は重要な情報源である．FAQ は企業などでも非常に有益な情報源であり，品質向上や作成支援のための研究が行われている [3, 4]．また，FAQ の整備を進めることにより，知識に基づいた対話ボットなどへの応用も期待できる．つまり， FAQ を充実させることで行政サービスの向上につながると期待できる．

しかしながら，実際に自治体のホームページを見ると，FAQ がほとんど更新されていなかったり，そもそも QA がほとんどなかったりと整備不十分な自治体も少なくない．そこで本稿では図 1 のように複数の自治体の FAQ を横断的に解析し，**多くの自治体のウェブサイトに含まれる重要な質問や特定の自治体において不足している質問を発見**し，自治体の FAQ の質と量の改善に資するシステムを開発することを目指す．

具体的には，自治体のウェブサイトから QA ペアを収集し，質問・回答間の関係性を自治体横断で分析した．また，関係性を考慮した上で，どのような質問ペアを同一クラスタとみなすべきかを検討し，約 900 の QA ペアに対して人手でクラスタリングを行い，正解データを作成した．表現の多様性に対処するために質問をベクトルで表現し，階層型クラスタリングを適用することで，人手でのクラスタリングをどの程度再現できるか検証した．その検証結果を元に，特定の自治体に不足している質問を発見するシステムを構築した．

[*1] http://www.city.sapporo.jp/callcenter/data/operation/3month29_4-6.html

2　自治体 FAQ の収集および分析

データの収集にあたっては，75 の自治体のウェブサイトに対してクローリング・スクレイピングを行い，合計 29862 の QA ペアを獲得した．自治体のウェブサイトの構造は多様であるため，全てのウェブサイトに対応するのは困難であり，網羅的にペアを獲得するには至らなかったが，75 市のうち 6 市で 1000 個以上の QA ペアを獲得した．また，最大では 1 市から 3370 個の QA ペアを獲得した．収集した QA ペアを自治体横断で分析し，QA 間にどのような関係が存在するかを調べた．

まず，自治体間で質問も回答の内容もほとんど変わらない QA ペアが存在する．例えば，法律で定められた税金の手続きなどは自治体による違いはほとんどない．一方，自治体間で質問も回答の内容も異なるものも多い．例えば，ごみの分別は自治体によりその制度も大きく異なる．また，同じ機能をもつ施設等でも自治体によって名称が異なる．例えば，「環境センター」は焼却施設やリサイクル施設，下水処理場などの複合的施設の名称であるが，地域によっては「クリーンセンター」とも呼ばれる．

本研究では重要な質問と不足している質問の発見を目的としており，そのために似た質問をまとめあげる．まずは，質問間の関係に注目しどのような関係をまとめあげるかを決める．以下に典型的な質問間の関係を述べる．

1. 質問の仕方の抽象度が異なるが，質問の対象は同一で回答に共通部分が多い場合
   (1) 可燃ごみ，燃やせるごみ，燃えるごみどのように出せばよいか（岩手県盛岡市）
   (2) 可燃ごみ，燃やせるごみ，燃えるごみはいつ出せばよいか（岩手県盛岡市）
2. 質問の対象が複数列挙されている質問とその部分集合となる質問
   (3) カセット式ガスボンベやスプレー缶はどのよう

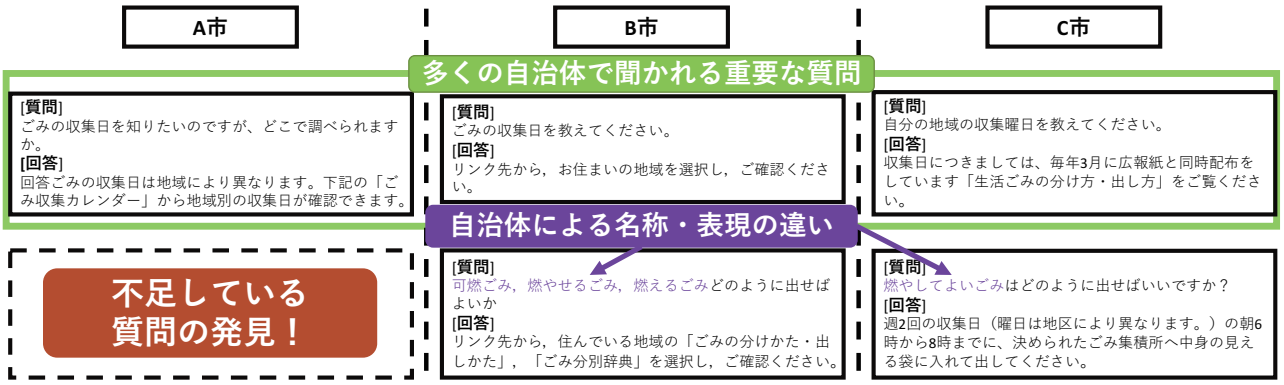

図 1　複数自治体の FAQ の比較

に出せばよいか（岩手県盛岡市）

(4) カセットコンロ用のボンベの処分に困ってます、どうすればいいですか？（岡山県岡山市）

3. 質問の対象が複数列挙されており，一部が共通している

(5) 使用済み乾電池・ボタン電池の出し方を知りたい。（愛知県一宮市）

(6) 充電式電池やボタン電池はどこに出せばよいか（岩手県盛岡市）

4. ある質問が修飾されて質問内容がより具体的になっている場合

(7) 不法投棄と思われるごみがある（岩手県盛岡市）

(8) 道路に不法投棄物がある（岩手県盛岡市）

(9) 自分の土地にごみを捨てられたらどうすればいいの。（愛知県一宮市）

5. 質問の対象に階層的な関係がある場合

(10) いろいろなごみの出し方を教えてください。（愛知県一宮市）

(11) 燃やせるごみはどのように出せばいいですか？（長崎県大村市）

1 番目の関係は，抽象度が高い質問の回答に抽象度の低い回答が含まれたり，同一の回答となっていることも多い．どういった質問の仕方をするかは自治体によって変わる．2 と 3 番目のような関係は，対象を列挙していても処理が同一である場合とそうでない場合がある．その判断をするためには回答や外部知識が必要である．また，4 番目の関係の質問は修飾の内容によって回答が変わることもある．不法投棄物の処理は投棄された土地の所有者・管理者が処分しなければならない．そのため，道路に不法投棄物がある場合と，私有地に不法投棄物がある場合では処理が異なる．しかし，修飾によって手続きが変わるかどうかも回答をみるか，知識がなければ判断できない．これらの理由により，今回は 1 番目から 4 番目の関係は同じ質問としてまとめあげることとした．

表 1　人手でまとめ上げを行った正解データの統計情報

トピック	QA 数	クラスタ数
ごみ	391	150
教育	323	113
高齢者・障害者・介護	212	48

一方，5 番目の関係のように対象物間に階層的な関係が存在する場合は，別の質問としてまとめあげることとした．「いろいろなごみ」に「燃やせるごみ」が含まれるかどうかは対象物の抽象度に関する知識が必要であり，対象物の粒度に関する規定は自治体によって異なる場合が多いためである．

このような暫定的な基準の元で，「ごみ」と「教育」，「高齢者・障害者・介護」の 3 つのトピックの質問に対して，人手でまとめあげを行い，クラスタリングの正解データを作成した．このデータの詳細については表 1 に示す．

3　実験

「ごみ」と「教育」，「高齢者・障害者・介護」に関する QA ペアに対して階層的クラスタリングを行い，前述の基準によって人手で構築した正解クラスタをどの程度再現できるか確認した．クラスタリングは階層的クラスタリングを用いる．詳細なクラスタリングのアルゴリズムについては 3.1 節で説明する．

各質問文は，質問に含まれる単語ベクトルを足し合わせたベクトルで表現する．単語ベクトルとして，Wikipedia と複数の自治体ホームページからテキストを収集した自治体コーパスから，それぞれ word2vec を用いて学習したものを用いる．また，一般的な単語や頻出する単語による影響を小さくするために，質問文を文書とみなして計算した IDF をそれぞれの単語ベクトルに対して乗算する重み付けを行った．

また，2 節で定めた基準では（名詞で表されることの多い）質問対象が決定すると，その対象に対する質問はかなり限定されることがわかっている．そのため，質問文中の名詞を近似的に質問対象と考え，名詞の単語ベクトルのみで質問文ベクトルを作成し，クラスタリングを行うことも試みた．

クラスタリングの評価指標については 3.2 節で示す．

3.1 階層的クラスタリングアルゴリズム

群平均法の一種である，非加重結合法 (unweighted pair-group method using arithmetic averages, UPGMA)[1, 2] を用いた．クラスタ C_i と C_j の非類似度 $d(C_i, C_j)$ は式 1 のように計算する．なお，$C_i = C_{i1} \cup C_{i2}$ である．事例間の距離には，質問文ベクトル間のコサイン距離を用いた．

$$d(C_i, C_j) = \frac{|C_{i1}|}{|C_i|} d(C_{i1}, C_j) + \frac{|C_{i2}|}{|C_i|} d(C_{i2}, C_j) \quad (1)$$

3.2 評価指標

評価指標としては，Zhao ら [2] によって提案された，階層的クラスタリングの評価のために修正された F-Score を用いる．これはフラットな正解クラスタと，階層型クラスタリングによって生成されたデンドログラムの全ノードを比較しスコアを求めるものである．今回は似ている質問をまとめ，特定の自治体に含まれていない質問を探すことが目的である．クラスタ数を固定する必要がないため，この評価方法を採用した．

L_r を正解のクラスタとし，S_i を階層型クラスタリングによって生成されたクラスタのノードとする．L_r の要素数を n_r とし，S_i の要素数を n_i とし，S_i の要素のうち，L_r に含まれる要素数を n_{ri} とする．この時，$Precision(L_r, S_i) = n_{ri}/n_i$ と $Recall(L_r, S_i) = n_{ri}/n_r$ とすると，L_r と S_i の F 値は式 2 の様に定義される．

$$F(L_r, S_i) = \frac{2 * Recall(L_r, S_i) * Precision(L_r, S_i)}{Recall(L_r, S_i) + Precision(L_r, S_i)} \quad (2)$$

図 2 のように階層的クラスタリングのデンドログラム T の全ノードに対して $F(L_r, S_i)$ を計算をし，最大値を L_r に対する F-Score とする．つまり，式 3 の様になる．

$$F(L_r) = \max_{S_i \in T} F(L_r, S_i) \quad (3)$$

最終的なクラスタリング全体の F-Score は式 4 の様に，各クラスタの要素数に基づく重み付き平均で計算される．なお，c は正解のクラスタ数である．

$$FScore = \sum_{r=1}^{c} \frac{n_r}{n} F(L_r) \quad (4)$$

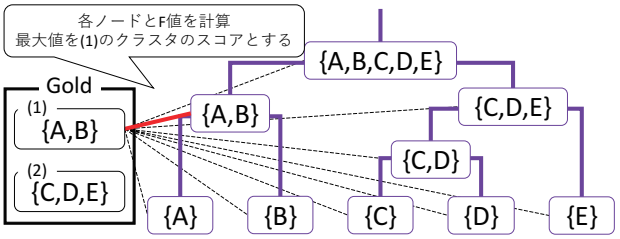

図 2　デンドログラムに対する正解クラスタとの F-Score の計算方法

3.3 実験結果

実験結果を表 2 に示す．Wikipedia よりも自治体コーパスで学習した単語ベクトルを用いてクラスタリングを行った方が性能がわずかに高い．自治体コーパスが学習に十分な量であることと，ドメインにあったベクトルが獲得できていることが理由であると推測できる．質問文ベクトルの構成法に着目すると，全単語のベクトルを用いた場合は IDF で重み付けを行うことでスコアが上がっていることがわかる．質問文の末尾は「〜ついて教えてください」など決まったパターンが多いため，IDF で重み付けを行うことで，末尾の表現によってクラスタがまとまるのを防ぐ効果があったと考えられる．また，どのトピックでも，質問文中の全単語を用いてクラスタリングをするよりも，質問文に含まれる名詞のみでクラスタリングをすることによって性能が上がっていることが確認できる．これは，想定していたように質問対象が決まることによって質問内容が限定されるためだと考えられる．今回は質問対象を近似的に名詞のみと考えたが，その仮定が妥当であることが確認できた．

4 マイニングシステムについて

3.3 節の結果より，自治体コーパスから学習した単語ベクトルを用いて，質問文の名詞の単語ベクトルに IDF で重み付けをして構成した質問文ベクトルを用いて FAQ マイニングシステムを作成した．図 3 に 4 市の「ごみ」に関わる質問をシステムに入力したときのシステムのインターフェースを示す．左の質問文がクラスタに属するベクトルの平均にもっとも近い文章が代表質問として表示されている．各マスの数字は各クラスタに属する各自治体の QA の数であり，一番左の数字はその和である．実際に厚木市と函館市で不足したい質問が図 3 のように発見できる．また，クラスタをドリルダウンすることができ，ユーザーが下位クラスタに分解し適切な粒度に調節することで，より具体的な質問を発見することが可能である．その結果，表 3 のような質問が発見された．

表 2　質問文のクラスタリングの実験結果

	Wikipedia で学習した単語ベクトル				自治体コーパスで学習した単語ベクトル			
	全単語		名詞のみ		全単語		名詞のみ	
	IDF なし	IDF あり	IDF なし	IDF あり	IDF なし	IDF あり	IDF なし	IDF あり
「ごみ」関する質問	0.701	0.750	0.789	0.805	0.726	0.761	0.813	**0.818**
「教育」関する質問	0.687	0.750	0.826	0.816	0.735	0.778	0.826	**0.830**
「高齢者・障害者・介護」関する質問	0.662	0.718	0.844	0.796	0.680	0.754	**0.862**	0.834

図 3　システムのインターフェース

表 3　システムを用いて発見された神奈川県厚木市の FAQ に不足してる質問の例

カラスが集積所を荒らして困っています。何か対策はありますか？
ごみの不法投棄について。
引越しなどで、たくさんごみが出ましたが、ごみステーションに出していいですか?

5　おわりに

本研究では複数の自治体ウェブサイトの FAQ を収集・調査し，似た質問をまとめあげることにより，重要な質問や特定の自治体ウェブサイトに含まれていない質問を発見することに取り組んだ．収集した FAQ を調べた結果，質問対象が決定すると質問の内容が限定されることがわかった．そのため，自治体のウェブページから単語ベクトルを学習し，質問文中の名詞の単語ベクトルを足し合わせたものを特徴量とし，階層的クラスタリングを行う FAQ マイニングシステムを作成した．

質問間の関係は，2 節で述べたように多様である．この調査結果を踏まえて，今後は質問間の関係性の自動解析にも取り組みたい．例えば，質問対象が修飾されている場合は，どのように修飾されているかによって，適切な解答が異なることも多い．質問文間のアラインメントをとることで，質問間の差異をより明らかにすることが可能かもしれない．

また，今回は QA ペアの質問のみを解析の対象としたが，回答を解析する方向性も興味深い．同じ質問に対する回答であっても，自治体によって適切な回答は異なることも多い．回答に対しても比較に基づいた解析を行うことで，回答中の不足している情報を発見することが可能かもしれない．これは，より充実した FAQ を構築する一助となるであろう．

謝辞

本研究は JST CREST(課題番号: JPMJCR1301) の支援を受けて行った．また，本研究は JSPS 科研費 15H01702，15H05318 の助成を受けたものである．

参考文献

[1] R. R. Sokal and C. D. Michener. A Statistical Method for Evaluating Systematic Relationships. *University of Kansas Science Bulletin*, Vol. 38, pp. 1409–1438, 1958.
[2] Ying Zhao and George Karypis. Evaluation of Hierarchical Clustering Algorithms for Document Datasets. *Proceedings of the CIKM*, pp. 515–524, 2002.
[3] 早織倉田, 哲男小川, 敏行加納. 代表文生成技術と FAQ 作成の効率向上. 東芝レビュー, Vol. 66, No. 9, pp. 57–61, 2011.
[4] 木村英志, 高島俊哉, 重岡知昭, 森田豊久. 文書カテゴリを利用した文書クラスタリングのコールセンター FAQ 改善への適用. 第 76 回全国大会講演論文集, 第 2014 巻, pp. 485–486, 2014.