

## 具体性と記述長を考慮した質問文自動生成手法の提案

木村 泰知<sup>1)</sup>, 渋木英潔<sup>2)</sup>, 高丸圭一<sup>3,4)</sup>, 森辰則<sup>2)</sup>

小樽商科大学<sup>1)</sup> 横浜国立大学<sup>2)</sup> 合同会社ディクティオ<sup>3)</sup> 宇都宮共和大学<sup>4)</sup>

### 1. はじめに

近年, ウェブ上には膨大な文書が存在しており, 情報検索技術を利用して, 利用者の望む文書を見つけることが可能になっている. しかし, 専門性の高い文書集合から, 利用者が望む文書を見つけることは難しい. 例えば, 刑事事件や交通事故の判例から利用者の望む状況に近い内容を検索するタスク [1] やマニフェストや会議録から利用者の考えに近い政党や政治家を見つけるタスクがある [2]~[5]. これらのタスクにおいて, キーワードだけで適切な文章を判断することは困難である. その理由として, 利用者が考えるキーワードが, どの文書にも含まれるようなキーワードになりやすい点が挙げられる. 交通事故の判例検索の場合, 「右折」と「交差点」の複数のキーワードによる検索で絞り込んだとしても, 数多くの文書が存在し, 「交差点を右折していた際」なのか, 「右折のために交差点内で停車」なのか曖昧性がある. そして, 曖昧性解消のために, 文書内に出現する適切なフレーズを考えることも難しい.

このような情報マッチングのタスクにおける解決策として, 興味の有無を尋ねる YesNo 型質問をシステムから能動的に行う対話型質問のアプローチが考えられる. システムから質問を行うアプローチの場合, 文章中に出現する表現を利用して, 質問文を生成することが最も単純な方法と考えられる. 文書中の表現を利用する場合, 文字数を長くすることで具体的な質問文を生成することが可能と考えられるが, 冗長な表現になる可能性が高く, 短い表現では曖昧になりやすい.

このような問題を考慮しつつ, 具体性があり, 冗長性の少ない, 質問文の自動生成手法を提案する. 我々は, 今までに質問文を用いた地方議員とのマッチングに関する研究を行っており, 決定木を利用した方法を提案している [2]. これらの決定木の各ノードにおいて, 適切な質問文を生成することで, 利用者の考えに近い政治家を見つけることを考えており, 本稿では質問表現の生成に焦点を当て, 生成手法および実験結果について述べる.

2. は, 興味を尋ねる質問文生成および関連研究,

3. は最適な質問表現の調査, 4. は質問表現生成手法 5. は質問表現生成の評価実験について説明し, 最後に結論を述べる.

### 2. 興味を尋ねる質問文生成

政治分野における興味を尋ねる質問は, 利用者に近い考えをもつ政党や国会議員の分類に焦点を当てた研究がある [5]. 上神らは, 国会議員の活動を対象としているため, 人手で作成した質問文を利用している. しかしながら, 地方議会の議員とのマッチングを考えた場合, 地域ごとに問題が異なり, 各地域に対応して, 興味を尋ねる必要があるため, 地域ごとに人手で質問文を構築するには労力がかかり, 実現することが難しい.

そこで, 我々は, 地方政治を対象に政治的カテゴリ体系を構築し, この政治的カテゴリを用いて各議員の活動を特徴付ける方法について研究を行ってきた [2]~[4]. 政治的カテゴリに対応する会議録中の記述を抽出するために, 文献 [2] では, 利用者にとって質問文として相応しい表現とはどのようなものであるかを調査した結果, 「A の B について興味ありますか.」という表現が良いことが確認された. ここで, 「A」, 「B」は名詞あるいは複合名詞を想定している. 本稿では, これらの調査結果に基づいて, 議員の発言内容から特徴付ける「A の B」のフレーズを抽出し, 興味の有無を尋ねる質問について考える.

#### 2.1. 「A の B」による質問表現

平成 19 年小樽市議会会議録に含まれる「A の B」は 14,336 パターン存在する. 会議録中に出現する「A の B」の例を表 1 に示す. 表 1 の「A の B」には, 「市民の皆さん」のように質問表現として相応しくない「A の B」が存在するため, 質問表現の候補としてチェックする必要がある.

表 1 「A の B」の例

「A の B」	出現回数
理事者の答弁	97
市民の皆さん	63
提案理由の説明	25
質問の概要	24
医師の確保	13
人件費の抑制	13

表 2 Google 7 gram を利用したフィルタリングの例

会議録に含まれる「A の B」	出現回数	Google 7 gram に含まれる「A を B する」	出現回数
提案理由の説明	25	提案理由を説明する	43
医師の確保	13	医師を確保する	31
人件費の抑制	13	人件費を抑制する	70
計画の策定	12	計画を策定する	51
病院の建設	12	病院を建設する	20
総合計画の策定	11	総合計画を策定する	95
財政の健全化	11	財政を健全化する	71

## 2.2. 「A の B」に関する従来研究

名詞句「A の B」に関する研究は、以前から数多く行われている。黒橋らは国語辞典を用いた名詞句「A の B」の意味解析を行っている[6]。「の」で結ばれる2つの名詞の関係は、所有（私の車）、範疇（野球の選手）、道具（トランプの手品）など様々な関係があり、我々が取得したい関係は「対象-述語」である。片岡らは、言い換え可能表現の絞り込みとして、動詞型連体修飾で「A（が/を/に）Bする」が成立する「A の B」の言い換えを行っている[7]。本稿では、片岡らの考え方を参考に「A の B」をフィルタリングする。

## 2.3. Google7gram を利用したフィルタリング

2.1 で述べたように、「A の B」に質問表現に利用することが難しいパターンが存在する。我々が利用する「A の B」は「A を B する」と言い換えた場合にも、意味が成り立つパターンを想定している。そこで、言い換えても成立する表現であるか判断するために、Google 7 gram を利用する[8]。Google 7 gram はウェブ上に存在する7単語の共起の中で20回以上存在する表現がまとめられている。20回以上出現する表現は、質問表現として利用できると判断する。その結果、14,336の中で805パターン得られた。表2にフィルタリングの結果を示す。

## 3. 最適な質問表現の調査

手法を考えるにあたって、まず最適な質問表現とは何かを明らかにする必要があり、そのための調査を行う。Google 7 gram を用いて得られた質問表現の候補には、「計画の策定」のように曖昧な表現が含まれているため、「A の B」の「A」を複合名詞に限定せず、拡張した質問表現を作成し、具体的であり、かつ、冗長性のない質問表現の調査を行う。

## 3.1. 質問表現の評価データ作成

我々が地方議会会議録を分類するために利用している政治的カテゴリは96個存在する。それらの政治的カテゴリを平成19年の小樽市の市議会会議録の各発言に対して政治的カテゴリの注釈付けを行ったデータを利用する。複数の政治的カテゴリが付与されている発言内容も存在するため、本稿では単独のカテゴリを付与された1,667の発言内容を対象にする。その対象となる発言内容からGoogle 7 gramによりフィルタリングを行った「A の B」を利用する。「A の B」の拡張表現候補を作成するために、構文解析ツールCabocha[9]を用いる。Cabochaにより区切られた文節を「A」の前方に向かって段階的に長くすることで拡張表現の候補を作成する。文節を段階的に長くした質問表現の例を表3に示す。表3の「費用の試算」は、「費用を試算する」と言い換えられるため、「A の B」の候補となっている。その「費用の試算」を前方の句読点が存在する「対象区域を広げた場合のボールに係る費用の試算」という最長の表現まで段階的に質問表現候補とした。このような質問表現を20問作成した。

## 3.2. 評価基準

評価基準は3つの観点があり、「曖昧⇔明瞭」、「冗長⇔簡潔」、「読みづらい⇔読みやすい」に分けている。各項目について被験者が5段階評価を行う。まず、その3つの観点について、それぞれ相加平均を計算する。そして、それらの3つの平均結果から調和平均を求め、最も評価の高い質問表現を適切な質問表現とする。

$$\text{調和平均} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

表 3 質問表現の評価例

発言内容	次に、再開発の対象区域を広げた場合のプールに関する費用の試算でありますけれども、地価の高い当該地で区域を広げ、プールを建設することは、プールの特徴である大スパンの確保や水対策が必要になることから、土地の高度利用に制限を受けるため、採算性の問題などディベロッパー誘致が困難となり、事業が成立しなくなると考えております。このことから、対象区域を広げた場合のプールに係る費用の試算は行っておりません。						
質問表現							
質問 1	「費用の試算」について詳しく知りたいですか。						
質問 2	「係る費用の試算」について詳しく知りたいですか。						
質問 3	「プールに係る費用の試算」について詳しく知りたいですか。						
質問 4	「場合のプールに係る費用の試算」について詳しく知りたいですか。						
質問 5	「広げた場合のプールに係る費用の試算」について詳しく知りたいですか。						
質問 6	「対象区域を広げた場合のプールに係る費用の試算」について詳しく知りたいですか。						
評価項目							
評価項目 1	質問文として	成り立つ			成り立たない		
	質問文として、成り立つ場合、下記の質問について評価する。						
		←-----→					
評価項目 2	曖昧	1	2	3	4	5	明瞭
評価項目 3	冗長(無駄に長い)	1	2	3	4	5	簡潔
評価項目 4	読みづらい	1	2	3	4	5	読みやすい

表 4 各発言内容に対して最も調和平均の高かった表現

	調和平均の最も高い質問表現	調和平均
1	プールに係る費用の試算	3.60
2	市場をめぐる環境の変化	3.94
3	サービスの向上*	3.41
4	市道の適正な機能の保持	3.80
5	文化芸術に触れる機会の拡充と人材の育成	3.57
6	子育て親子の交流などを促進する事業の拡充	4.00
7	75歳以上の高齢者への過酷な保険料の負担	4.17
8	市民からの要望のあるバス路線の新設	4.30
9	企業の育成*	3.62
10	学校の安全にかかわる環境の変化	4.17
11	土産品の購入*	3.38
12	人材の登用*	4.41
13	病原体が体内に侵入して感染して増殖し発症する疾患の総称	2.24
14	撤去の指導*	3.08
15	介護予防プランの作成*	4.01
16	部の再編*	3.57
17	保険証が使えないような資格証明書の発行	3.60
18	赤字の解消と財政の健全化	4.30
19	企業債元利償還金の減少や維持・管理の効率化	3.74
20	地域の教育力を生かした豊かな教育活動の推進	3.72
	合計(評価実験の指標として利用する) **が付いている表現は拡張していない「AのB」である。	74.63

### 3.3. 拡張表現の評価結果

男性 2 名、女性 4 名の大学生に対して 3.1 で作成した質問表現に対して、「A」の拡張に関する調査を行った。その結果、最も調和平均の高かった

表現を表 4 に示す。ここで、「質問文として成り立たない」と選択された表現は、評価結果を「0」として、計算している。

表 5 正解率と調和平均による実験結果

手法	正解数/問題数	正解率	調和平均合計 /最大調和平均	調和平均による 評価尺度
Baseline (Google 7 gram)	7/20	0.35	61.84 / 74.63	0.8286
手法 1 (助詞の制約)	15/20	0.75	66.62 / 74.63	0.8927
手法 2 (助詞と副詞の制約)	16/20	0.80	70.24 / 74.63	0.9412

### 3.4. 拡張表現の考察

表 4 の「\*」が付いた拡張なし「A の B」は 7 つ存在し、全てを拡張する必要がないことがわかった。今回の実験で使用した質問表現は、拡張なし「A の B」の 20 フレーズを含め、90 フレーズ存在する。その 90 フレーズの中で被験者 6 名のうち、2 名以上が質問文として成り立たないと判断したフレーズは 71 フレーズ存在した。その 71 フレーズの特徴は次の通りである。

1. 「が」、「は」などの助詞を含む文節は質問表現に適さない。
2. 副詞を含む表現は質問表現に適さない。

### 4. 質問表現生成手法

3.4 の考察結果を踏まえて、次の 2 つの手法を提案する。

手法 1 助詞の制約による手法

次の助詞 ("で","が","も","は","から",""における","により") が文節の末尾になる場合は拡張せず、最長の表現を選択する。

手法 2 助詞と副詞の制約による手法

手法 1 の制約に、副詞を含む文節が出現した場合は拡張せず、最長の表現を選択する。

### 5. 評価実験

#### 5.1. 評価方法

本節では 3. の調査で利用した 20 の発言内容から最適な質問表現を選択する評価実験を行う。評価は 2 つの観点から行う。1 点目は最も調和平均が高かった表現を正解の質問表現とする評価であり、表 4 の調査結果を質問表現の正解として評価を行う。2 点目の評価は、提案手法により選択された表現の調和平均の合計を最も高い調和平均の合計（最大となる調和平均の合計）で割った評価である。この 2 つの観点を用いて、Google 7 gram で得られた拡張なし「A の B」をベースラインとして、提案手法との比較を行う。

#### 5.2. 実験結果

表 5 は Google 7 gram による手法と提案手法の実験結果である。表 5 からわかるように、手法 2

の結果が、正解率および調和平均の評価尺度において、最も高い結果となり、具体的であり、冗長性の少ない質問表現を選択できることを確認した。

### 6. まとめ

「A の B」を拡張した最適な質問表現を選択するために、助詞と副詞の制約による手法を提案し、有効性を確認した。今後は提案手法を用いたシステムを実装する予定である。

### 謝辞

本研究の一部は総務省 SCOPE 補助金 (No. 082301004) の支援により行われた。

### 文献

- [1] Rzepka, R., Shibuki, H., Kimura, Y., Takamaru, K., Matsuhara, M., and Murakami, K., "Judicial Precedents Processing Project for Supporting Japanese Lay Judge System", Workshop on Semantic Processing of Legal Texts, LREC2008, pp. 33-41, May 2008.
- [2] 洪木英潔, 木村泰知, 高丸圭一, 森辰則, "地方議員マッチングシステムのための質問表現の検討", 信学技報, vol. 109, no. 234, NLC2009-11, pp. 25-30, 2009.
- [3] 木村泰知, 洪木英潔, 高丸圭一, "地方議員と住民間の協働支援に向けたウェブの利用", 選挙研究 25 卷 1 号, pp. 100-118, 2009.
- [4] Takamaru, K., Shibuki, H., Kimura, Y., Hasegawa, D., Ototake, H., and Araki, K., "Extraction of Political Activity of Assemblyman from Minutes of Municipal Assemblies Using the Political Category", Pacling2009, B11, 2009.
- [5] 上神貴佳, 堤英敬, "投票支援のためのインターネット・ツール—日本版ボートマッチの作成プロセスについて—", 選挙学会紀要第 10 号, pp. 27-48, 2008.
- [6] 黒橋禎夫, 酒井康行, "国語辞典を用いた名詞句「A の B」の意味解析", 自然言語処理研究会報告, pp. 109-116, 1999.
- [7] 片岡明, 増山繁, 山本和英, "動詞型連体修飾表現の N1 の N2 への言い換え", 自然言語処理, Vol. 7, No. 4, pp. 79-98, 2000.
- [8] Web 日本語 N グラム第 1 版 by Google. GSK カタログ GSK2007-C.
- [9] 工藤 拓, 松本 裕治, "チャンキングの段階適用による日本語係り受け解析", 情報処理学会論文誌, Vol. 43, No. 6, 2002.