特集論文 「実践 AI システム」

地方議員マッチングシステムにおける能動的質問のための質問生成手法

An Automatic Question Generation Method for a Local Councilor Search System

木村 泰知 小樽商科大学
Yasutomo KIMURA Otaru University of Commerce
kimura@res.otaru-uc.ac.jp, http://minna.ih.otaru-uc.ac.jp

渋木 英潔 横浜国立大学
Hideyuki SHIBUKI Yokohama National University
shib@forest.eis.ynu.ac.jp

高丸 圭一 宇都宮共和大学
Keiichi TAKAMARU Utsunomiya Kyowa University
takamaru@kyowa-u.ac.jp

乙武 北斗 福岡大学
Hokuto Ototake Fukuoka University
ototake@fukuoka-u.ac.jp

小林 哲郎 国立情報学研究所
Tetsuro KOBAYASHI National Institute of Informatics
k-tetsu@nii.ac.jp

森 辰則 横浜国立大学
Tatsunori MORI Yokohama National University
mori@forest.eis.ynu.ac.jp

keywords: local politics, question generation, information extraction

Summary

This paper presents an automatic question generation method for a local councilor search system. Our purpose is to provide residents with information about local council activities in an easy-to-understand manner. Our designed system creates a decision tree with leaves that correspond to local councilors in order to clarify the differences in the activities of local councilors using local council minutes as the source. Moreover, our system generates questions for selecting the next branch at each condition in the decision tree. We confirmed experimentally that these questions are appropriate for the selection of branches in the decision tree.

1. はじめに

TV や新聞のように時間や紙面に限りがあるメディアで取り上げられる政治情報は，国政に関する内容が中心的であり，これに比べて地方政治に関する情報は少ない．議員活動についても同様で，地方議会議員は国会議員と同様に住民による選挙によって選ばれ，かつ，国政よりも身近な存在であるべきであるにもかかわらず，その活動に関する認知度は国会議員よりも低い．また，平成の大合併により，各自治体の広域化が進み，各議員の活動を把握することが困難となっている．そこで，住民に提供される地方政治の情報，特に地方議会議員に関する情報量の不足を解決するための方法の一つとして，ウェブ上の情報を有効に利用することが考えられる．議員側の情報には，議員や政党のホームページ，ニュースサイトの政治ニュース，議員のブログ，マニフェスト，議会の会議録などがある．このうち会議録には，議員からの一方的な情報発信ではなく，議論や反対意見などのやりとりが含まれ，公の場における各議員の活動や考え方を知ることができる．

国会の場合，国立国会図書館により会議録サイトが整備されており，第1回国会（昭和22年）以降のすべての会議録がテキストデータとして公開され，検索システムによって検索を行うことができる．しかしながら，地方議会会議録については，未だウェブ公開自体がなされていない自治体も多い．ウェブ公開されている会議録も自治体により公開方法が異なっており，国会会議録のように整備されているものはほとんどない．

会議録は，定例会のものだけでも膨大な量となる．例えば，北海道小樽市の市議会会議録の場合，定例会1回分の会議録は A4 版で 200 ページを超えている．このよ

図 1 議員マッチングシステムの外観

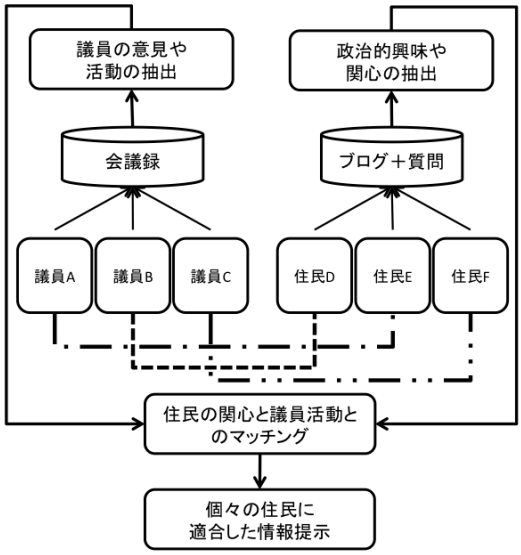

図 2 議員マッチングシステムの処理概要

うな大量のテキストデータを単純にウェブ公開しただけでは，能動的にアクセスしてこれを読もうと考える住民はほとんど存在しないことが予想されるため，公開方法や情報提供形態を工夫し，地方議会会議録を有効に利用することが望ましい．以上の背景から，我々は，ウェブ上に存在する政治情報を利用して，メディアで取り上げられる機会の少ない地方議会議員の政治の情報を提供する方法について研究を進めている [長谷川 08, 乙武 09, 木村 09a, Takamaru 09, 渋木 09, 木村 09b, 木村 10, 乙武 10].

本システムの利用者となる住民は日常生活における不満や要望に政治的問題が含まれているとは捉えていない場合が多く，また，住民の関心の対象はそれ自体が多様である．住民の関心に合う情報を探すためには，まず，住民の潜在的な関心を明確化して，それぞれの住民にマッチした情報を抽出・整理して提示するシステムが必要であると考えられる．このため，我々は，ウェブ上の情報から住民の関心にあわせた地方議会議員の情報を提示するシステム (議員マッチングシステム) の開発を行っている．議員マッチングシステムでは，利用者に対して政治的問題の関心を明確化するための質問を行うことで，その利用者の考えに近いと思われる n 名の議員の組み合わせを提示する．本論文では，これまでに我々が開発した議員マッチングシステムを紹介するとともに，本システムにおける住民の潜在的な関心を明確化するための能動的質問生成手法に関して論ずる．

本論文の構成は以下の通りである．2 章では，議員マッチングシステムの概要を述べ，本システムにおける能動的質問生成手法の位置付けを説明する．3 章では，能動的質問生成の基本的な考え方について述べる．4 章では，利用者の興味や関心を尋ねるのに適した質問表現の調査を行い，調査から得られた知見を反映した能動的質問生成手法を提案する．5 章では，能動的質問生成手法の評価実験について述べる．6 章では，まとめと今後の課題について述べる．

2. 議員マッチングシステムの概要

我々が開発した議員マッチングシステムの外観を図 1 に，処理の概要を図 2 にそれぞれ示す．まず，我々は，住民の潜在的な関心を明確化しマッチングを容易にするという目的から，会議録中の議員の意見や活動，および，利用者である住民の興味や関心を政治的カテゴリに分類した上でマッチングを行うことを考えた．そのために，議員の意見や活動と住民の興味や関心を分類するための政治的カテゴリ体系を構築する必要があり，2·1 節に記述する政治的カテゴリ体系の構築を行った．また，政治的カテゴリへの分類は決定木や SVM などの機械学習を利用することを想定しており，学習や評価を行うために，会議録に政治的カテゴリを付与した地方議会会議録コーパスを 2·2 節に記述する方法で構築した．

図 2 に示す会議録からの議員の意見や活動の抽出は，地方議会会議録コーパスを学習データとした SVM による分類器を用いて行われる*1が，利用者である住民の政治的興味や関心の抽出に関しては，以下の 2 通りの方法で行うことを想定している．第 1 の方法は，利用者が自らの日常生活などを綴ったブログを対象として，日常生活における不満や要望などから政治的興味や関心を自動的に推測する方法である．この方法は，政治に全く興味や関心がなかった利用者に対して，本人も意識していない政治的興味や関心を引き出せる可能性があるという点で優れているが，精度などの点で技術的に未解決な課題も多い．そこで，第 2 の方法として，本システムから政治的問題に関する質問を能動的に行うことで，利用者の政治的興味や関心を明確化する方法を併用することとした．第 2 の方法では，第 1 の方法よりも，システムから

*1 紙面の都合により，本論文では会議録中の政治的カテゴリの推定に関しては割愛する．これに関しては [乙武 10] を参照されたい．

の質問に回答するという点で利用者に負担がかかることが予想されるが，質問の回数を適正な数に抑えることで利用者の負担を軽減できると考えている．

利用者の政治的興味や関心を特定するために質問を行うシステムとして，[上神 09] などのボートマッチシステムが挙げられる．一般的なボートマッチシステムでは 20 問程度の質問が行われており，この回数は利用者の負担として軽いものではない．我々は，ある時点までの利用者の回答に基づいて，次の質問を，議員を特定するために最適な質問とすることで質問回数を最小限に抑えることができると考えた．この考え方は，議員集合を葉ノード，質問を内部ノードとする決定木の考え方で表すことができる．しかしながら，地方議会ごとに内部ノードとなる質問を人手で設定することは多大な労力が必要となるため，自動的に質問を生成する手法の開発を行うこととした．本論文では，この質問生成手法を能動的質問生成手法と定義し，3 章以降では能動的質問生成手法に関する議論を行う．

2·1 政治的カテゴリ体系

政治に関する既存のカテゴリ体系は地方議会に焦点を当てたものではないため，議会での発言内容を表現するのに必ずしも適しているとはいえない．それゆえ，我々は小樽市，帯広市，函館市，釧路市の 4 市を対象とした調査を行い，議題を区分するために存在する委員会体系に基づいて基本となる概念体系を構築した．その後，実際の会議録の内容と整合するよう概念体系の調整を行い，[長谷川 08] において 96 の政治的カテゴリを構築した．表 1 に政治的カテゴリの一部を示す．また，平成 19 年度小樽市市議会本会議定例会に政治的カテゴリを付与した頻度を表 1 の段落数に示す．平成 19 年度小樽市市議会本会議定例会は 7,084 段落あり，表 1 の割合は 7,084 段落を分母とした割合である*2．

現在の政治的カテゴリが政治情報システムにおいて最適な分類を可能とするものであるかは今後の課題だが，本論文で扱う範囲を超えるため，会議録中の発言は現在のカテゴリに分類されるものとする．

2·2 地方議会会議録コーパス

本論文で対象とする会議録データには，我々が以下の手順で構築した地方議会会議録コーパスを用いている．まず，北海道の 59 市町村*3を対象に電子化された会議録を自動収集した [乙武 09]．収集された会議録の内，小樽市と札幌市の本会議定例会を対象に，学習・評価用のタグ付けを行った．タグは XML 形式で付与され，段落区切りを示す <Paragraph> と，タグ付け作業者が政治的カテゴリを判断する際に手掛かりとなった語句を示す <Keyword> がある．<Keyword> は属性として，発言者を示す Member と，判断された政治的カテゴリを示す Category を持ち，<Paragraph> は Member のみを持つように記述されている．平成 19 年度小樽市市議会本会議定例会の会議録に対してタグ付けされた一部を表 2 に示す．タグ付け作業は，専用のツールを開発，利用して，1 会議録あたり 2 名の大学生により行った．なお，本論文では，マッチングを幅広く行いたいという動機から，2 名の作業者により付与されたカテゴリの積集合ではなく和集合を学習・評価用の正解カテゴリとしている．

表 1 政治的カテゴリの例

コード	ラベル	段落数	割合
1000	総務文教		
1010	財務	4,236	59.8%
1011	地方税	565	8.0%
1012	予算	821	11.6%
1020	総合的な行政の推進	859	12.1%
1021	条例	543	7.7%
1030	職員	1,112	15.7%
1050	情報	704	9.9%
1060	地域社会	479	6.8%
1061	住民活動	585	8.3%
1062	まちづくり	471	6.6%
1100	医療	1,500	21.2%
1101	病院事業	1,739	24.5%
1120	教育	1,306	18.4%
1121	学校	1,268	17.9%
1160	施設	880	12.4%
1162	スポーツ	498	7.0%
2000	厚生		
2013	廃棄物	427	6.0%
2065	介護保険	396	5.6%
3000	産業経済		
3030	労働行政	1,112	5.6%
3040	観光	988	13.9%
3060	海港	548	7.7%
4000	建築		
4020	道路	646	9.1%
4110	住宅	415	5.9%
5000	その他		
5030	どのカテゴリにも属さない	517	7.3%

3. 能動的質問生成の基本的な考え方

議員と住民のマッチングは，種々の政治課題に対して、同じ考え方（意思決定）をする相手を見つけるタスクだと考えられる．そこで，我々は，経営学などで意思決定

*2 1 段落に複数のカテゴリが付与される場合があるため，全てのカテゴリの合計率は 100%を超える．

*3 平成 20 年の時点で，北海道の 180 市町村のうち Web 上で会議録を公開しているのが 63 市町村あり，その中の 4 市町村がアクセス制限やリンク切れなどにより収集できなかった．

表 2 タグ付き会議録コーパスの例

<Paragraph Member="37 山田勝麿">
次に，<Keyword Member="山田勝麿" Category="4110"> 空き住宅 </Keyword> が発生する理由と戸数の市民周知でありますが，先ほどお答えしたとおり，空き家には募集停止している政策空き家と募集中のものがあります．これらが空き家になっている理由と <Keyword Member="山田勝麿" Category="4110"> 空き戸数 </Keyword> などについて，今後，<Keyword Member="山田勝麿" Category="1050"> 広報 </Keyword> おたるや <Keyword Member="山田勝麿" Category="1050"> ホームペ ージ </Keyword> などでお知らせをしてまいりたいと考えております．</Paragraph>
<Paragraph Member="見楚谷登志">
（「議長，１０番」と呼ぶ者あり）○議長（見楚谷登志）　１０番，斉藤陽一良議員．</Paragraph>
<Paragraph Member="斉藤陽一良">
○１０番（斉藤陽一良議員）　１点だけ再質問させていただきます．</Paragraph>
<Paragraph Member="斉藤陽一良">
不良債務解消の件ですが，<Keyword Member="斉藤陽一良" Category="1010;1101"> 病院事業会計 </Keyword> における不良債務の解消で <Keyword Member="斉藤陽一良" Category="1101"> 病院 </Keyword> が経営努力で解消をするという部分について，給与費で８億４，０００万円の減を見込んでいるという答弁だったのですが，給与表自体の見直し，あるいは <Keyword Member="斉藤陽一良" Category="1030"> 職員定数 </Keyword> の管理についての考え方，また，統合・新築までに至る <Keyword Member="斉藤陽一良" Category="1030"> 職員数 </Keyword> の削減のスケジュール等について，ある程度今の段階でお答えいただけることがありましたら，お答えいただきたいと思います．</Paragraph>
<Paragraph Member="見楚谷登志">
○議長（見楚谷登志）　理事者の答弁を求めます．</Paragraph>

に利用されている決定木を用いて，「**決定木において同じ経路を選択する相手は同じ考え方をする相手と見なすことができる**」という仮説を立てた．このような考え方で、議員と住民という二つの異なる種類のものをマッチングさせるために決定木を利用する方法は，これまでに存在しない．下記では，この仮説に基づいた能動的質問生成について述べる．

能動的質問生成とは，利用者の政治的興味や関心を特定するための質問を生成することである．システムが自動的に質問を生成するには，質問内容を何にするかという点と，質問表現をどのようにするかという点の 2 点を決定する必要がある．図 3 は，質問内容と質問表現の関係を示す処理概要である．能動的質問生成では，注釈付けされた地方議会会議録コーパスを入力として，質問文付き決定木の状態遷移表を出力する．注釈付けされた地方議会会議録コーパスには，2·2 節で述べた通り，発言者である議員名と政治的カテゴリの情報が付与されている．我々は，質問内容に関しては，政治的カテゴリで近似することとし，議員集合を葉ノード，政治的カテゴリを内部ノードとする決定木を地方議会会議録コーパスから学習することとした．質問文生成処理では，政治的なカテゴリに対応する，利用者にわかりやすい質問表現を生成する．その生成された質問表現は，決定木学習で作成された状態遷移表に追加される．議員マッチングシステムでは，予め作成された質問付きの状態遷移表を利用することで，利用者に対して質問を行い，利用者に適した議員を決定する．

次に，質問内容の考え方について説明する．本研究における決定木の利用方法は，議員を分類することに加えて，最短経路探索問題の観点から利用している．最短経路探索問題の観点とは，分類結果に到達するまでの分岐数を少なくすることである．本研究における決定木の利用目的は，分岐数を少なくすることであり，利用者に対しての質問回数を減らすことによって，利用者の負担を軽減させることにある．分岐数を少なくする方法としては，情報利得の大きい属性を分岐の条件として選択する方法がある．情報利得を利用した決定木には ID3 や C4.5 などの手法があり，本研究では C4.5 を実装している Weka J48[*4]を利用することとした．他にも，決定木を利用する利点としては，分岐の条件が明確になっている点がある．決定木は，分類の過程を把握することが容易にできるため，分類結果に対する説明も可能となる．決定木の分類過程を利用する方法は，利用者への信頼を高めるために，荒井らの研究で利用されている [荒井 03]．政治的カテゴリに基づいた決定木の学習に関しては，3·1 節で説明する．

次に，質問表現の考え方について説明する．質問表現に関しては，ある政治的カテゴリに対応する適切な表現を会議録から抽出し，テンプレートを用いて生成することとした．このテンプレートを利用した質問文は，Weka J48 により作成された決定木の遷移表に追加される．政

∗4 http://www.cs.waikato.ac.nz/ml/weka/

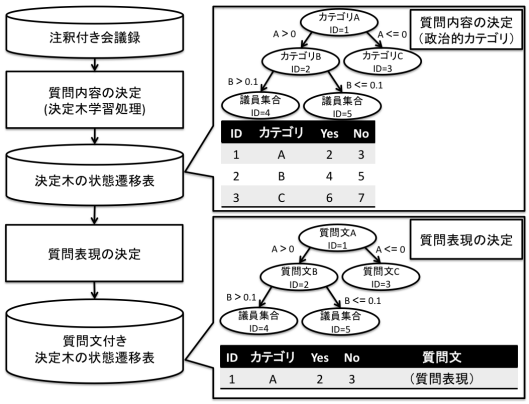

図 3 能動的質問生成の処理概要

治的カテゴリに基づいた質問文の生成に関しては，3·2 節で説明する．

3·1 政治的カテゴリに基づいた決定木学習

質問内容となる政治的カテゴリを求めるための決定木は，地方議会会議録コーパスから以下の手順により学習された．まず，各議員の発言内容を政治的カテゴリに基づいて分類し，各議員の政治的カテゴリごとの発言頻度を求める．利用者の政治的関心に最も適合した議員を 1 名だけを選ぶ場合には，上記の政治的カテゴリにより特徴付けられた議員データを用いて議員を葉ノードとする決定木を作成する．しかしながら，本システムが議員情報を提示する目的は，利用者の考えに近い複数の議員を提示することであり，選挙のように最適な議員 1 名を選択することではない．したがって，本システムでは 1 名の議員に絞り込むのではなく，利用者の考えに近いと思われる n 名の議員の組み合わせを提示することとした．提示する議員の組み合わせは，必ずしも議員間の意見の近さを意味しない．例えば，利用者の政治的関心が政治的問題 P1 と P2 にある場合，P1 に関して積極的に取り組んでいるが P2 に関して無関心な議員 M1 と，P2 に関して積極的に取り組んでいるが P1 に関して無関心な議員 M2 の組み合わせが提示されることは十分に考えられることである．この問題を解決するために，我々は，n 名の議員の組み合わせを葉ノードとする決定木を作成することで対処する．それゆえ，n 名の議員の組み合わせを全て求め，各組み合わせを 1 つの議員集合として，各議員集合の政治的カテゴリごとの発言頻度に基づいて学習データを作成した．本システムでは，提示する議員数を暫定的に 3 名とした．表 3 は，1 つの議員集合を 3 名とした場合の学習データ作成例である．学習データの属性値は，各カテゴリにおける 3 名の議員の発言を合計してから，各カテゴリの合計を発言総数で割ることによって，正規化をしている．この正規化された属性値は，決定木の分岐条件に利用される．我々は，この分岐条件となる数値を質問表現に反映させることが困難と考え，分岐条件に不等号情報のみを利用することとし，分岐条件の値よりも大きければ，条件のカテゴリに興味があると判断している．この分岐条件を利用した質問文の生成については、次節で説明する．

表 3 決定木の学習データ作成の例

	カテゴリ A	カテゴリ B	カテゴリ C	合計
議員 1	5	10	10	25
議員 2	10	10	20	40
議員 3	5	10	20	35
合計	20	30	50	100
正規化	0.2	0.3	0.5	1.0

表 4 決定木の条件分岐数

	組合せ	平均	最大値	最小値	中央値
小樽市	210	8.5	13	5	9
札幌市	367	9.4	14	6	10

次に，決定木の葉ノードの数について述べる．決定木の葉ノードの数は，葉ノードが議員の組み合わせにより決められることから，各自治体の議員定数に影響する．議員定数は，自治体の規模や人口によって決められており，選挙の時期に変更されることもあるため，対象年度と対象都市によって異なる．我々は，決定木の学習データを作成するために，小樽市と札幌市の平成 19 年の市議会会議録を対象とした．しかしながら，小樽市と札幌市の議会は，平成 19 年の 4 月に統一地方選挙が実施されたため，平成 19 年第 1 回の定例会と平成 19 年 2-4 回の定例会を構成している議員が異なっている．そこで，本論文では，選挙前と選挙後の両方の議会を構成する議員，つまり，2 期連続で当選している議員を対象とすることとした．この条件を満たす議員数は，小樽市が 20 名，札幌市が 49 名おり，この中から，3 名一組の組み合わせを作成する．単純に議員の組み合わせを考えた場合，葉ノードの数は，小樽市と札幌市，それぞれ，1,140 と 18,424 存在することになる．

- 小樽市の場合 $\quad {}_{20}C_3 = 1,140$
- 札幌市の場合 $\quad {}_{49}C_3 = 18,424$

しかしながら，上記の組み合わせでは，3 名の議員が一つの組としてまとめても良い議員であるかを考慮していない．そこで，本システムでは，同じような内容を発言していることを考慮するために，25 以上のカテゴリにおいて 3 名が共通して発言していることを条件とした．上記の条件を満たす小樽市と札幌市における組み合わせは，それぞれ，210 と 367 であった．表 4 は，この組み合わせ結果を用いて作成した決定木の条件分岐の平均，最大値，最小値，中央値を示したものである．この結果から，平均 9 回程度の質問で，利用者の考えに近いと思われる 3 名の議員の組み合わせを提示できることがわかる．

ここで，政治的カテゴリに基づいた決定木学習の評価

について議論する．決定木の分類精度を評価する方法としては，次の 2 つの評価方法が考えられる．

(1) 住民を被験者とした場合の評価
(2) 地方議員を被験者とした場合の評価

まず，1 点目の住民を被験者とした場合の評価について説明する．住民を被験者とする場合には，決定木の葉として提示されるマッチング議員が正しいことを確認するために，住民の考えに近い議員を決める必要がある．住民の考えに近い議員を決めるためには，被験者となる住民に，地方議会会議録を読んでもらい，他の議員の意見も考慮しながら，決定してもらう必要がある．小樽市の地方議会会議録を利用する場合，正解となる議員を決めてもらうために，平成 19 年の会議録 1 年分の会議録，約 800 ページを読んでもらうことになる．この正解作成については，被験者に負荷がかかり過ぎることに加え，地方議会会議録を全て読んでいるのかを確認することも困難であるという問題がある．

次に，2 点目の地方議員を被験者とした場合の評価について説明する．本研究で利用している決定木は，地方議員会議録に記録されている地方議員の発言を利用して作成しているため，地方議員を被験者として評価する方法がある．つまり，決定木を評価するもう一つの方法としては，地方議員本人に本システムを利用してもらい，マッチングする議員として提示された 3 名の中に本人が含まれているのか，確認する方法が考えられる．しかしながら，この評価方法は，被験者が地方議員に限定されるため，有効な回答数を得ることが難しいという問題がある．

我々は，住民からの評価と地方議員からの評価方法について検討したが，評価方法が困難であることから，評価を行っていない．本システムの評価については，今後の課題として，検討していかなければならない．

3·2 政治的カテゴリに基づいた質問文生成

本システムは，3·1 節で作成された決定木に基づいて，現在までの利用者の回答から次の質問の内容となる政治的カテゴリを決定する．その後，決定した政治的カテゴリに基づいて，その政治的カテゴリに興味や関心があるかを利用者に尋ねる質問文を生成する．

表 1 に示すように，政治的カテゴリには「医療」,「観光」,「建築」といったラベルが付与されており，これらのラベルを利用して質問文を生成することが考えられる．しかしながら，これらの抽象化されたラベルをそのまま用いることは「医療に興味がありますか？」といった漠然とした質問となりやすい．それゆえ，会議録における各議員の発言記述から，政治的カテゴリに対応する具体的な記述を抜粋して利用することで，この問題を解決することを試みる．政治的カテゴリに対応する会議録中の記述を抜粋するために，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する必要があり，その調査を行った結果を 4 章に記述する．

本システムが生成する質問文は二者択一式の質問であり，会議録から抜粋された記述を「○○について詳しく知りたい.」というテンプレートに適用することで生成される．利用者の興味や関心を尋ねるテンプレートとして,「○○に興味がありますか？」といった表現を用いることも考えられるが，一般に，どのような政治的カテゴリであれ，全く興味がないということは稀であると考えられる．したがって,「○○に興味がありますか？」といったテンプレートを用いた場合に，利用者によっては「ないとはいえない」といった意味で「ある」と回答することが予想されるため，より積極的に興味や関心があることを利用者に認識させるように「○○について詳しく知りたい.」というテンプレートを用いることとした．また，二者択一の質問とは，会議録中に含まれる地方政治に関する問題を 2 つ提示し，利用者に興味ある内容を 1 つ選択してもらうための質問である．

二者択一の質問は，決定木の結果とテンプレートを利用することで，生成される．本システムでは，二者択一の質問を生成するために，現在の分岐条件のカテゴリと現在の分岐条件において興味ないと選択された場合に進む下位にある分岐条件のカテゴリを利用している．決定木が図 3 のように作成されている場合，カテゴリ A が分岐条件となっている ID=1 では，次のような質問をする．

次の内容のどちらに興味がありますか？
- 「(カテゴリ A の内容)」について詳しく知りたい
- 「(カテゴリ C の内容)」について詳しく知りたい

上記の「(カテゴリ A の内容)」と「(カテゴリ C の内容)」の質問表現は，カテゴリが付与された発言を対象として，名詞句 A の B を含む表現を収集した集合から，ランダムに選択する．各カテゴリに適した質問表現を選択する場合には，カテゴリが付与されている発言の条件として，政治的カテゴリが一つだけ付与されていることとしている．これは，複数のカテゴリが付与されている場合に，対象としていないカテゴリの内容を抽出してしまうことを防ぐためである．

また，カテゴリ B が分岐条件となっている ID=2 では，次のような質問をする．

次の内容のどちらに興味がありますか？
- 「(カテゴリ B の内容)」について詳しく知りたい
- 「(議員集合 ID=5 に含まれる議員の発言)」について詳しく知りたい

上記の「(議員集合 ID=5 に含まれる議員の発言)」の質問表現は，議員集合に含まれる議員の発言集合から，名詞句 A の B を含む表現をランダムに選択する．この質問文生成処理において,「観光情報の発信について詳しく知

表 5 名詞句 A の B

「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

りたい.」と「雪に関する条例の制定について詳しく知りたい.」のような 2 つの政治問題を挿入することが可能である．この二者択一の質問は，興味や関心のより強い方を選択させることで，利用者に自らの興味や関心を認識させることも意図している．

4. 能動的質問生成手法

4·1 質問表現の定義

3·2 節において，本システムが生成する質問文は，会議録から抜粋された記述を「○○について詳しく知りたい.」というテンプレートに適用することを述べた．本節では，このテンプレートに埋め込まれる表現として，相応しい表現について定義する．

我々は，[渋木 09] において，テンプレートに埋め込まれる適切な表現を明らかにするためのアンケート調査を行った．6 名の大学生を被験者として調査した結果，テンプレートに埋め込まれる表現としては,「名詞」よりも「名詞句 A の B」が適していることを確認した．つまり,「ごみ」という名詞だけを埋め込むよりも,「ごみの削減」,「ごみの処理」,「ごみのポイ捨て」のような表現が適している．しかしながら,「名詞句 A の B」は名詞 A と名詞 B の関係によって様々な意味になることが知られており,「名詞句 A の B」の表現を全て利用することはできない．例えば,「名詞句 A の B」の 2 つの名詞を結んでいる「の」の関係には，所有（私の車），範疇（野球の選手），道具（トランプの手品）などが存在する [黒橋 99]．そこで，[渋木 09] の調査において回答された質問表現の関係を確認したところ，2 つの名詞の関係は「対象-述語」になっていることがわかった．例えば,「医師の配置」,「生活習慣病の早期発見 」,「中心商店街の活性化」などが,「対象-述語」の関係である．ここで，対象とは目的語になる名詞であり，述語とは動詞化できるサ変名詞のことである．

次に，質問表現の具体性について考える.「対象-述語」の関係を持つ「名詞句 A の B」は，テンプレートに埋め込んだ場合，文法的に問題はないが，曖昧な表現になっている場合がある．例えば,「費用の試算」は,「費用」の用途がわからないため，テンプレートに埋め込まれても，質問表現としては相応しくない質問文となる．このような曖昧性を解消するための一つに,「名詞句 A の B」を前方に拡張する方法がある.「費用の試算」の場合には，前方に拡張することで,「プールに係る費用の試算」という表現になり，質問表現として利用できる．

上記の内容をまとめて，適切な質問表現を定義する．適切な質問表現とは「の」の関係が「対象-述語」となる「名詞句 A の B」とする．また,「名詞句 A の B」の表現が抽象的な場合には,「名詞句 A の B」を前方に拡張することで，具体的な表現とする．さらに，前方に拡張した表現は，利用者にとって，具体的であり，冗長性がなく，理解しやすい表現になっていることとする．

次節では，名詞句 A の B に関する調査について述べる．

4·2 名詞句 A の B に関する調査

本節では，平成 19 年小樽市議会会議録を対象に実施した,「対象-述語」の関係を持つ「名詞句 A の B」の調査について述べる．まず，平成 19 年小樽市議会会議録に含まれる「名詞句 A の B」を収集したところ，14,336 の表現が存在した．表 5 に「名詞句 A の B」の例を示す．4·1 節で記述したように，表 5 の「A の B」には「市民の皆さん」,「質問の概要」などの「対象-述語」以外の関係になっている表現が存在する．そこで，我々は,「対象-述語」の関係をもつ「名詞句 A の B」だけを選択するために，言語資源の利用を検討した．まず，我々は IPADIC 辞書[*5]に登録されている「名詞-サ変接続」を利用することを検討した.「対象-述語」の関係の「述語」は「サ変名詞」になることから，IPADIC 辞書の「名詞-サ変接続」として登録されている単語であるかを確認することで判断できると考えた．しかしながら，IPADIC 辞書の「名詞-サ変接続」に登録されていることを確認するだけでは,「市民の生活」,「４月の選挙」,「ほかの施設」のように「対象-述語」以外の関係を取り除くことができなかった．

そこで,「対象-述語」の関係の「対象」も考慮するために,「対象」が目的語になることを「言い換え」を利用することで確認することとした．言い換えに関する従来研究には，片岡らの言い換え可能表現の絞り込みがある．片岡らは，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 [片岡 00]．我々は，片岡らの考え方を参考にして「対象」が目的語となることを確認する，ここでは,「A の B」から「A を B する」へ言い換え可能な表現であることを確認する．例えば,「ごみの削減」は「ごみを削減する」と言い換えることができるため，質問表現の候補とする.「A の B」を「A を B する」と言い換えられるか確認するために，GoogleN-gram を利用した [Google 07]．Google N-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．ウェブ上に 20 回以上出現する表現は，言い換えできると判断して，質問表現として利用する．その結果，14,336 の中から 805 の表現を得ることができた．表 6 にフィルタリングの結果を示す．

∗5 http://chasen.aist-nara.ac.jp/chasen/doc/ipadic-2.6.3-j.pdf

次節では 曖昧な名詞句 A の B に対して,「A の B」の「A」を前方に拡張する方法について検討する.

4·3 適切な質問表現の決定方法

本研究において適切な質問表現とは，4·1 節で述べた通り，具体的であり，冗長性がなく，理解しやすい表現である．本システムでは，曖昧な「名詞句 A の B」を前方に拡張することで，適切な質問表現を生成する．本節では，会議録に含まれる表現の中から，適切な表現を被験者の評価によって決定する方法について述べる.

まず，質問表現の評価データの作成方法について説明する．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，本論文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 N-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 を用いる [Kudo 03]. 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 7 に示す. 表 7 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プ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.

次に，質問表現の評価基準について説明する.

作成された質問表現候補は，質問表現を評価する前に，質問文として成立することを確認する必要がある．質問文として成立しない場合には，質問表現の評価を行わないこととした．質問文として成立する質問表現に対しての評価基準は，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．被験者からの回答結果は，上記の 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.

最後に，上記の評価基準により調査した結果について述べる．我々は男性 2 名，女性 4 名の大学生に対して,「A」の拡張に関する調査を行った．表 8 に最も評価の高い質問表現の例を示す[*6]．我々は，この結果を適切な質問表現の正解とする.

*6 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.

4·4 適切な質問表現の特徴

本節では，4·3 節の「A」の拡張に関する調査結果を用いて，適切な表現と判断された特徴を明らかにすることで，能動的質問生成手法を検討する.

まず，名詞句 A の B の拡張について考察する．表 8 の結果から，20 表現のうち 7 つの表現は，名詞句 A の B を拡張していないことがわかる[*7]．これは，拡張する必要がない，あるいは，拡張できない「名詞句 A の B」があるということである．次に,「質問文として成り立たない」と判断された表現候補と適切な表現との違いについて述べる．適切な質問表現を決定するために利用した表現候補数は，拡張なし「名詞句 A の B」の 20 表現を含め，90 表現である．この 90 表現の中で，被験者 6 名のうち 2 名以上に「質問文として成り立たない」と判断された表現は 79%=(71/90) であった.

この結果から，文節単位で前方に拡張する場合，ほとんどの表現は質問表現として利用できないことがわかる. 利用できない表現の特徴の 1 つは，テンプレートに埋め込むことを考慮せずに拡張しているため，非文になることである．例えば,「費用の試算」の拡張を考える場合,「係る費用の試算」や「場合のプールに係る費用の試算」をテンプレートに埋め込んでも，質問文として成り立たない．これらを解決するためには，係り受けや品詞情報を利用する方法が考えられる.

まず，係り受けの利用について考える.「名詞句 A の B」の「名詞 A」あるいは「名詞 B」に係っている文節に限定して，範囲を拡張することで，前述した問題は解決できると考えられる．下記の文を例として考える.
(例) なお、当面は市民からの要望のあるバス路線の新設や延長などの実現に向け、バス事業者に働きかけてまいりたいと考えております.
この文の「バス路線の新設」を拡張する場合,「バス路線」と「新設」を係り先とする文節を選択することで,「当面は」は対象外となる．上記の内容を考慮して,「名詞句 A の B」の名詞 A あるいは名詞 B に対する係り元を最も長い範囲まで選択する方法が良いと考えられる.

次に，品詞情報の利用について考える．質問表現として成り立たない表現は,「が」や「を」などの特定の助詞と副詞を含む傾向にあった．例えば,「企業の育成」を拡張する場合,「やはり企業の育成」という表現は副詞という情報から拡張しないということも考えられる．上記の内容を考慮して，文節単位の情報と品詞情報を合わせることが考えられる.

さらに，本研究において適切な表現として定義した「具体的であり」,「冗長性がない」 の観点から，単語の重み付けを利用した手法を検討する.「具体的であり」と「冗長性がない」については「特徴的な単語」から構成される表現と考えると，単語の重み付けとして考えることができる．政治的カテゴリの違いを考慮した質問表現にな

*7 「#」が付いている表現は，拡張していない.

表 6 Google N-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 N-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

表 7 質問表現の評価例

発言内容							
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

ることを考えた場合，IDF の考え方に基づく単語の重み付けが良いと考えられる．例えば,「受付窓口を 1 か所にすることがサービスの向上」に対して「サービスの向上」を拡張する場合,「する」や「こと」などの特徴のない単語が数多く含まれる表現は拡張しないことで問題を解決できる．

4·5 提 案 手 法

本節では，4·4 節の調査結果に基づいた手法について記述する．

(1) Baseline(拡張なし)

「対象-述語」の関係を抽出するために,「A の B」から「A を B する」へ言い換え可能な表現であることを Google N-gram を用いて確認した表現であり，拡張はしていない．拡張しない表現が適切な表現として選択されることも多いことから，拡張しない名詞句 A の B を Baseline とする．

(2) 係り受け

名詞句 A の B の「名詞 A」あるいは「名詞 B」に係る最大の範囲を抽出する．本研究における調査では，実験条件として，読点が出現する場所までを対象範囲としていることから，読点よりも前方にある係り元は対象外とする．

(3) 品詞による制約

特定の品詞が文節に含まれていれば，拡張を行わない．ここで，特定の品詞とは助詞，副詞であり，助詞については（「で」「が」「も」「は」「から」「における」「により」）としている．品詞による制約は助詞による制約，副詞による制約，助詞と副詞による制約の 3 つがある．

(4) 単語の重み

IDF(Inverse Document Frequency) の考え方を参考に ICF(Inverse Category Frequency) として利用する．ICF は IDF の Document を Category に置き換え，政治的カテゴリが付与された発言を収集した文集合を Document の代わりとして，扱うこととする．CF は単語 t が出現する政治的カテゴリ頻度である．ICF は単語 t が出現する政治的カテゴリ頻度の逆数の対

表 8 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）	74.63

“ # ”が付いている表現は拡張していない「A の B」である．

数を計算している．

$$ICF(w_i) = 1 + log_2 \frac{N}{CF(w_i)}$$

ここで，w_i は単語，N は政治的カテゴリ頻度の総数である．本節では平成 19 年の小樽市会議録を利用しており，段落数は 7,084 段落である．平成 19 年の小樽市会議録の場合，96 の政治的カテゴリを設定しており，政治的カテゴリ頻度の総数は 96 で計算する．また，96 のカテゴリ文書に含まれるか否か数えたものを CF(政治的カテゴリ頻度) としており，その政治的カテゴリ頻度の逆数の対数を ICF としている．

単語の重み ICF

拡張表現に含まれる各単語の ICF を計算し，その平均が最大となる表現を選択する．

$$\text{単語の重み } ICF = \frac{1}{n}\sum_{i=1}^{n} ICF(w_i)$$

上記の条件で，名詞のみを対象として各単語の ICF を計算し，その平均が最大となる表現を選択する手法を”単語の重み ICF(名詞)”とする．

単語の重み ICF + L

各単語の文字長 (L) と ICF を重みを掛け合わせて，文字長 × 単語の重み ICF が最大となる表現を選択する．

$$\text{単語の重み } ICF + L = \frac{1}{n}\sum_{i=1}^{n} ICF(w_i) \cdot logL(w_i)$$

ここで，$L(w_i)$ は w_i の文字長である．上記の条件で，名詞のみを対象として，文字長 × 単語の重み ICF が最大となる表現を選択する手法を”単語の重み ICF(名詞) + L”とする．

5. 評 価 実 験

評価実験の目的は，能動的質問生成手法の評価を行うことである．本実験のタスクは，質問内容（政治的カテゴリ）が決まっていることを前提としており，政治的カテゴリが付与された会議録中の発言から適切な質問表現を選択することである．

5·1 評 価 方 法

質問内容（政治的カテゴリ）に対応した名詞句 A の B が選択されている　最適な質問表現として利用するための評価である．

実験は，クローズドデータとオープンデータに分けて評価を行う．クローズドデータは，4·3 節の調査において述べた，小樽市議会会議録から収集した 20 の発言内

容であり，提案手法を検討するために利用したデータである．また，オープンデータは，平成 19 年札幌市議会会議録を対象として，小樽市議会会議録と同様の方法で作成した 37 の発言内容を含むデータである．

評価指標は，2 つの観点から評価を行う．最適な質問表現の評価は，4·3 節において最も調和平均の高かった表現を正解として，2 つの観点から行うこととした．

まず，正解率の評価方法は，最も調和平均の高かった表現を正解として，正解率を計算する評価である．

$$\text{正解率} = \frac{\text{システムが出力した正解数}}{\text{評価数}}$$

次に，調和平均による評価尺度の評価方法は，4·3 節の調査において，被験者から得られた調和平均の結果を利用する評価である．この評価は，それぞれの評価文の表現候補の中で，最も高い調和平均の合計を分母として，システムが出力した表現の調和平均を分子として，計算する．この評価式を次の通りである．

$$\text{調和平均による評価尺度} = \frac{\text{システムが出力した表現の調和平均の合計}}{\text{最も高い調和平均の合計}}$$

この評価は正解率と異なり，僅差で正解にならなかった表現を正しく評価できる．つまり，最も調和平均の高い表現以外が必ずしも不正解とは限らないため，最適な表現以外の利用可能な表現を考慮している．

5·2 クローズドデータによる評価実験

前述したようにクローズドデータは小樽市議会会議録から収集した 20 の評価文であり，90 の表現候補がある．各評価文は 4 - 5 の質問表現候補から最適な質問表現を選択する．

クローズドデータによる評価実験の結果を表 9 に示す．正解率の評価では，最も高い結果は係り受けを利用した手法の 0.85(=17/20) である．また，調和平均による評価尺度の評価についても，係り受けを利用した手法が最も高い結果となり，0.9887 であった[*8]．品詞の制約による手法については，助詞と副詞の制約による手法が係り受けによる手法の次に高い評価となっている．また，単語の重みを利用した手法については，正解率および調和平均による評価尺度の評価に対して，全体的に低い結果となった．

§1 考 察

単語の重みを利用した誤りは，質問文としての構造を無視した拡張に原因があるといえる．例えば,「環境の変化」を拡張する場合,「めぐる環境の変化」と選択している．これは,「めぐる」という単語が，2 つの政治的カテゴリの文書にしか出現していないため，特徴的な単語として計算され，ICF の値が高くなったかったことが原因である．上記の結果から，単語の重みを工夫するだけでは，質問文を生成することが困難であることがわかった．

次に，係り受けの誤りの原因を明らかにするとともに，改善点について考える．下記に 3 つの誤りの例を示す．ここで，係り受けによる手法が出力した結果を【係】，調和平均の最も高い結果を【正】として記述する．

誤り例 1

【係】対象区域を広げた場合のプールに係る費用の試算

【正】プールに係る費用の試算

誤り例 2

【係】近年における市場をめぐる環境の変化

【正】市場をめぐる環境の変化

誤り例 3

【係】異常プリオンなどの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

【正】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

誤りの特徴の一つとして，係り受けによる手法が出力した表現は，正解よりも長いことが挙げられる．そこで，Mecab の品詞情報を利用することで，正解の一つ前にある文節の特徴をみつけ，拡張を制限する方法を検討する．誤り例 1 の特徴としては，正解の一つ前にある文節となる「場合の」に含まれる「場合」に着目すると，IPA 品詞辞書の名詞:副詞可能に分類されているという点が挙げられる．また，誤り例 2 の特徴も誤り例 1 と同様に，正解の一つ前にある文節をみると,「近年における」の「近年」は名詞:副詞可能に分類されている．誤り例 3 の特徴については，正解の一つ前にある文節となる「異常プリオンなど」の「など」が含まれているところであり，品詞は助詞:副助詞に分類されている．上記の品詞情報を利用して，係り受けと品詞による制約の組み合わせ手法の比較実験を行うこととした．上記の結果を利用した品詞による制約とは,「名詞:副詞可能」と「副助詞」が含まれている文節は拡張しないことである．そして，表 9 の実験で提案した品詞による制約と係り受けを組み合わせた手法についても比較を行う．

係り受け+品詞の制約による手法の追加実験の結果を表 10 に示す．表 10 からわかるように，係り受けと品詞の制約（名詞:副詞可能）を組み合わせる手法が正解率の評価で 0.95(=19/20)，調和平均による評価尺度の評価が最も高い結果となった．

次節のオープンデータでは，これらの手法も比較手法として，評価実験を行う．

5·3 オープンデータによる評価実験

本節では，前節で検討した評価手法を加えて，オープンデータによる評価を行う．評価データは，平成 19 年の札幌市議会会議録に対して，4·3 節と同様の方法で作成

*8 4·3 節の調査結果から，最も高い調和平均の合計は 74.63 であったため，調和平均による評価尺度の評価を計算するために利用している．

表 9 クローズドデータによる実験結果

手法	正解数/評価数	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(拡張なし)	7/20	0.35	61.84/74.63	0.8286
品詞による制約 (助詞)	15/20	0.75	66.62/74.63	0.8927
品詞による制約 (副詞)	5/20	0.25	46.74/74.63	0.6227
品詞による制約 (助詞と副詞)	16/20	0.80	70.24/74.63	0.9412
係り受け	**17/20**	**0.85**	**73.79/74.63**	**0.9887**
単語の重み ICF	3/20	0.15	44.54/74.63	0.5968
単語の重み ICF(名詞)	4/20	0.20	51.91/74.63	0.6956
単語の重み ICF + L	6/20	0.30	40.60/74.63	0.5440
単語の重み ICF(名詞) + L	5/20	0.25	45.10/74.63	0.6043

表 10 係り受けと品詞による制約を組み合わせた手法の実験結果

手法	正解数/評価数	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
係り受け	17/20	0.85	73.79/74.63	0.9887
係り受け+助詞	15/20	0.75	66.62/74.63	0.8927
係り受け+副詞	17/20	0.85	73.79/74.63	0.9887
係り受け+助詞+副詞	16/20	0.80	70.24/74.63	0.9412
係り受け+名詞:副詞可能	**19/20**	**0.95**	**74.51/74.63**	**0.9983**
係り受け+名詞:副詞可能+副助詞	**19/20**	**0.95**	71.60/74.63	0.9593

表 11 オープンデータによる実験結果

手法	正解数/評価数	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(拡張なし)	8/37	0.22	98.86/128.36	0.7702
品詞による制約 (助詞)	21/37	0.57	109.40/128.36	0.8523
品詞による制約 (副詞)	20/37	0.54	103.26/128.36	0.8045
品詞による制約 (助詞+副詞)	22/37	0.60	110.83/128.36	0.8634
係り受け	24/37	0.65	115.74/128.36	0.9016
係り受け+助詞	24/37	0.65	116.00/128.36	0.9037
係り受け+副詞	**25/37**	**0.68**	**117.17/128.36**	**0.9128**
係り受け+助詞+副詞	**25/37**	**0.68**	**117.43/128.36**	**0.9148**
係り受け+名詞:副詞可能	23/37	0.62	112.14/128.36	0.8736
係り受け+名詞:副詞可能+副助詞	22/37	0.60	110.11/128.36	0.8578

した 37 の評価文，133 の表現候補である．

オープンデータによる評価実験の結果を表 11 に示す．クローズドデータにおいて最も評価の高かった手法である「係り受け+名詞:副詞可能」と「係り受け+名詞:副詞可能+副詞」については,「係り受け」よりも低い結果となった．これは，クローズドデータに特化したことによることが原因であり，オープンデータに弱いことがわかった．

次に，最も高い評価結果となった手法について述べる．正解率の評価では，最も良い結果は「係り受け+副詞」と「係り受け+助詞+副詞」を利用した手法が最も高い結果となり，0.6756(=25/37) であった．調和平均による評価尺度の評価については,「係り受け+助詞+副詞」を利用した手法が 0.9148 となり，最も良い結果となった．オープンデータを対象にした実験結果から判断すると,「係り受け+助詞+副詞」が最も良い事になるが，クローズドデータの結果では「助詞」を利用することで，低い結果になっている．一方,「係り受け+副詞」は，クローズドデータとオープンデータの両方に対して高い結果となっていることから，品詞の制約として副詞を利用することがよいといえる．

これらの評価実験を踏まえ，議員マッチングシステムでは，係り受けと品詞の制約を組み合わせた「**係り受け+副詞**」の手法 により能動的質問の質問生成を行っている．

6. ま と め

本論文では会議録を利用して住民の関心にあわせた地方議会議員の情報を提示する議員マッチングシステムの開発報告を行った．議員マッチングシステムは，カテゴリ推定，決定木，文生成技術を利用しているという特徴

がある．また，地方議員を対象としているため，地域特有の政治的問題を市町村ごとに抽出する点が国会議員を対象とした場合との違いである．

議員マッチングシステムの概要について説明し，政治的カテゴリをマッチングシステムの質問文生成に利用する方法を述べた．そして，住民が興味を持っている政治的問題を明らかにするために，システムから能動的に質問するための質問文生成手法について述べた．議員マッチングシステムの質問表現は，係り受け＋品詞による制約を用いて生成しており，http://www.hokkaido-politics.net で公開している．

今後は，全国の市町村議会を対象に本システムを適用することを考えている．そのために，市議会会議録のコーパスを作成することから始める予定である．また，2 章の議員マッチングシステムで記述したように，住民の政治的興味や関心の抽出する第 1 の方法であるブログを対象とする方法についても検討する．今後は，twitter(ミニブログ) を対象にすることで，潜在的な政治的意見の抽出を行うことを考えている．

謝　　辞

本研究の一部は科学研究費 22300086 の助成を受けたものである．

◇ 参　考　文　献 ◇

[荒井 03] 荒井 幸代, 村上 陽平, 杉本 悠樹, 田仲 正弘：Boosting を用いた評判の信頼性評価方法, 第 4 回セマンティックウェブとオントロジー研究会資料集 (2003)

[Google 07] Google 株式会社：Web 日本語 N グラム第 1 版 by Google, GSK カタログ GSK2007-C (2007)

[長谷川 08] 長谷川 大, 乙武 北斗, 木村 泰知, 渋木 英潔, 高丸 圭一, 荒木 健治：市議会会議録を対象とした概念体系構築へ向けた分析, 情報処理学会研究報告 2008-NL-187, pp. 23–28 (2008)

[片岡 00] 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)

[木村 09a] 木村 泰知, 渋木 英潔：市議会会議録と住民ブログとのマッチングのための共通カテゴリの付与, 人工知能学会全国大会 2009 (2009)

[木村 09b] 木村 泰知, 渋木 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 巻 1 号, pp. 100–118 (2009)

[木村 10] 木村 泰知, 渋木 英潔, 高丸 圭一, 森 辰則：具体性と記述長を考慮した質問文自動生成手法の提案, 言語処理学会第 16 回年次大会, pp. 563–566 (2010)

[Kudo 03] Kudo, T. and Matsumoto, Y.: Fast Methods for Kernel-Based Text Analysis, *ACL 2003* (2003)

[黒橋 99] 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, Vol. 129, No. 16, pp. 109–116 (1999)

[乙武 09] 乙武 北斗, 高丸 圭一, 渋木 英潔, 木村 泰知, 荒木 健治：地方議会会議録の自動収集に向けた公開パタンの分析, 言語処理学会第 15 回年次大会, pp. 298–301 (2009)

[乙武 10] 乙武 北斗, 渋木 英潔, 木村 泰知, 高丸 圭一, 森 辰則：地方議会会議録における政治的カテゴリの自動推定手法の提案, 信学技報 NLC2010-1, Vol. 110, No. 142, pp. 7–12 (2010)

[渋木 09] 渋木 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, pp. 25–30 (2009)

[Takamaru 09] 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, *Proc. 11th Conference of Pacific Association for Computational Linguistics (PACLING 2009)*, p. B11 (2009)

[上神 09] 上神 貴佳, 佐藤 哲也：政党や政治家の政策的な立場を推定する - コンピュータによる自動コーディングの試み -, 選挙研究, Vol. 25, No. 1, pp. 61–73 (2009)

〔担当委員：竹内 広宜〕

2010 年 8 月 1 日　受理

著 者 紹 介

木村　泰知（正会員）

2004 年北海道大学大学院工学研究科電子情報工学専攻博士後期課程修了．博士 (工学)．2005 年，小樽商科大学商学部助教授着任．2007 年，同准教授，現在に至る．この間，2010 年 10 月より 2011 年 9 月まで New York 大学客員研究員．自然言語処理に関する研究に従事．言語処理学会，情報処理学会，電子情報通信学会各会員．

渋木　英潔

2002 年北海道大学大学院工学研究科博士後期課程修了．博士 (工学)．2006 年北海学園大大学院経営学研究科博士後期課程修了．博士 (経営学)．2002 年北海学園大ハイテク・リサーチ・センター研究員，2008 年横浜国大環境情報研究院産学連携研究員．同特任教員を経て，現在，同科学研究費研究員．自然言語処理に関する研究に従事．言語処理学会，電子情報通信学会，情報処理学会，日本認知科学会各会員．

高丸　圭一

2004 年北海道大学大学院工学研究科電子情報工学専攻博士後期課程単位修得退学．修士（工学）．同年，那須大学都市経済学部専任講師．2006 年，宇都宮共和大学シティライフ学部専任講師（名称変更）．方言学，社会言語学，音声情報処理などの研究に従事．日本音声学会，社会言語科学会，電子情報通信学会各会員．

乙武　北斗

2010 年北海道大学大学院情報科学研究科博士後期課程修了．博士（情報科学）．同年，福岡大学工学部助教着任，現在に至る．自然言語処理，教育工学などの研究に従事．言語処理学会，電子情報通信学会各会員．

小林　哲郎

2007 年東京大学大学院人文社会系博士課程単位取得退学．社会心理学博士．2007 年，国立情報学研究所助教着任．この間，総合研究大学院大学助教，Rutgers 大学客員研究員，Stanford 大学客員研究員．社会心理学，メディア論，情報社会論などの研究に従事．日本社会心理学会，情報通信学会各会員．

森　　辰則（正会員）

1991 年横浜国立大学大学院工学研究科博士課程後期修了．工学博士．同年，同大学工学部助手着任．同講師，同助教授を経て，現在，同大学大学院環境情報研究院教授．この間，1998 年 2 月より 11 月まで Stanford 大学 CSLI 客員研究員．自然言語処理，情報検索，情報抽出などの研究に従事．言語処理学会，情報処理学会，ACM 各会員．