2008－NL－187 (4)
2008／9／24

市議会会議録を対象とした概念体系構築へ向けた分析

長谷川 大[*1*5] 乙武 北斗[*1*5] 木村 泰知[*2]
渋木 英潔[*3*5] 高丸 圭一[*4*5] 荒木 健治[*1]

[*1] 北海道大学 大学院 情報科学研究科 [*2] 小樽商科大学 商学部 社会情報学科
[*3] 横浜国立大学 大学院 環境情報研究院 [*4] 宇都宮共和大学 シティライフ学部
[*5] ディクティオ

E-mail: {hasegawadai, hokuto, araki}@media.eng.hokudai.ac.jp
kimura@res.otaru-uc.ac.jp, shib@forest.eis.ynu.ac.jp, takamaru@kyowa-u.ac.jp

本研究は地方政治に特化した概念体系の自動構築に向けた，小樽市と帯広市の市議会会議録の分析を目的としている．分析は，会議録の段落毎に政治問題を示すキーワードを人手により抽出し，帯広市市議会の委員会体系を基に作成された概念体系から，キーワードに対して適切な上位概念を選択するという方法で行われた．2名の分析者による分析の結果，抽出したキーワードが一致した割合は0.68であり，概念体系の自動構築の可能性を示唆するものであった．また，キーワードの前後4単語から成るフレーズを用いて，既知の会議録に対してキーワード自動抽出を行った結果，精度は9割を示した．

An Analysis of Minutes of City Council Meetings for the Construction of a Province Politics Thesaurus

Dai HASEGAWA[*1*5] Hokuto OTOTAKE[*1*5] Yasutomo KIMURA[*2]
Hideyuki SHIBUKI[*3*5] Keiichi TAKAMARU[*4*5] Kenji ARAKI[*1]

[*1]Graduate School of Information Science and Technology, Hokkaido University
[*2]Department of Information and Management Science, Otaru University of Commerce
[*3]Graduate School of Environment and Information Sciences, Yokohama National University
[*4]Department of City Life Studies, Utsunomiya Kyowa University
[*5]Dictio LLC.

E-mail: {hasegawadai, hokuto, araki}@media.eng.hokudai.ac.jp
kimura@res.otaru-uc.ac.jp, shib@forest.eis.ynu.ac.jp, takamaru@kyowa-u.ac.jp

In this paper we analyze and annotate city council meeting minutes with the goal of automatically constructing a thesaurus of words related to provincial politics. Our target cities are Otaru and Obihiro. Two analyzers each extract a keyword that describes a politics topic for each paragraph in the minutes. They also select a hypernym for each keyword. The hypernyms are based on ontology from the Obihiro council. The agreement between the two analyzers on the extracted keywords was 0.68, which indicates that automatic construction of a thesaurus will be possible.

1 まえがき

国政と比較して地方政治の活動は，国民の認知度が低い傾向にあるが，その重要性が劣るものではない．また，地方政治への関心は個々の住民によって異なっており，それらに対応する政治情報も多様に存在する．それゆえ，住民が自分の関心に合った情報を探すことは容易ではなく，一人ひとりの住民にマッチした政治情報を提示するシステムの構築が望まれている．このような背景から，我々は住民本位型政治情報システムの開発を目指している [1, 2].

これまでの研究において，議員活動表現となるフレーズを会議録における発言から抽出した後[1]，地域住民

[1]議会における発言は，所属する政党や会派などの意向により，必ずしも発言した議員自身の主張であるとは限らない．しかしながら，その議員が発言することにより地方政治に影響を与えたという事実は住民が知るべき重要な事実であるため，会議録における発言

の政治的関心や要望をブログから抽出することで，両者のマッチングを試みている．このマッチングの精度を高めるために，地方政治の分野に対応した概念体系を利用することの必要性が確認されている．しかしながら，地方政治は国政と比べてその地域性が強く反映されている部分があるため，分類語彙表などの一般的な概念体系の政治部分を流用するだけでは不十分であり，地方政治に特化した概念体系を独自に構築する必要がある．

文献 [1, 2] では，小樽市に限定して会議録やブログを用いた研究を行ってきたが，我々の目的は，一つの都市だけを対象とすることではなく，多様な市町村を対象としながらも各市町村の政治的特徴を捉えることにある．それゆえ，複数の都市を比較することで，地域差がある政治的特徴とは何か，また，それらの特徴を同一の枠組みの中で扱うにはどうしたらよいか，といった点を調査する必要がある．

以上の背景から，本稿では，地方政治に特化した概念体系の自動構築に向けた市議会会議録の分析を行う．具体的には，地域・年代による概念体系の相違の分析，会議録中の政治問題を指すキーワードの周辺に特徴的に出現する表層表現の発見，及び，未知の会議録における表層表現の有効性の調査を行う．

本稿の構成は以下の通りである．2 章で会議録を扱った従来研究に関して述べる．3 章で分析対象とするデータの詳細を述べ，4 章で分析手順について詳述する．5 章で分析結果を示し，自動化の可能性を考察する．6 章は結論である．

2 関連研究

議事録を対象とした研究には以下のものがある．川端ら [3] や山本ら [4] は，特徴的な表層表現を手がかりに国会議事録を対象とした自動要約を行っている．我々が目的とする市議会会議録においても特徴的な表現が存在すると考えられるが，それらの表現が全ての市町村に共通のものであるかは確認されていない．それゆえ，複数の市議会会議録を対象とした分析を行う必要がある．

友部ら [5] や本村ら [6] では，議事録の半自動生成や知識源としての再利用を目的としたディスカッションマイニングの研究を行っている．将来的にこれらの技術が市議会会議録に導入されれば効率的な会議録の情報利用が可能になると考えられるが，現状では一般公開されている会議録を対象として研究を進める必要がある．

表 1: 分析対象データ

	段落数	段落当りの文字数
平成 14 年度小樽	598	107
平成 19 年度小樽	695	173
平成 14 年度帯広	1,246	61
平成 19 年度帯広	1,314	73

3 対象データ

分析の対象としたデータは，平成 14 年度小樽市議会本会議における第 1 回定例会の会議録，平成 19 年度小樽市議会本会議における第 1 回定例会の会議録，平成 14 年度帯広市議会定例会における第 1 回 3 月定例会の会議録，平成 19 年度帯広市議会定例会における第 1 回 3 月定例会の会議録の 4 つである．小樽市と帯広市を選択した理由としては，都市の規模が近く，地域性が異なるためである．小樽市の人口は平成 19 年の時点で 137,456 人，帯広市は 169,156 人であり，小樽市は海に面している一方で，帯広市は内陸に位置しているため，地域性による政治問題の相違が現れやすいと考えられる．さらに，年代による相違も考察するため，平成 19 年度と平成 14 年度を選択した．分析に使用したデータの詳細を表 1 に示す．

4 分析方法

我々は，会議録における議員の発言と，ブログから抽出した地域住民の政治的関心や要望のマッチングを試みているが，ブログで使用される表現は会議録で使用される表現とは異なるため，表層的な対応付けは困難である．しかし，地方政治に特化した概念体系を利用することで，意味的な面からのマッチングが可能になると考えられる．そこで，市議会会議録から政治問題を表わすキーワードを抽出し，その概念を体系化することが目下の課題であり，本稿ではそのための分析を行う．

分析に際し，まず，会議録とブログの表現をマッチングさせる共通概念をどのように設定するかが問題となる．各市町村の会議録とブログに含まれる表現をそれぞれボトムアップにクラスタリングしていくことで共通概念を探す方法も考えられるが，労力や網羅性の

を議員活動の一つとみなしている．

点で問題がある．そこで，我々は，各自治体に存在する政治問題には共通の上位概念体系が存在すると仮定して，まずはじめに，基本となる概念体系を作成した．その後，人手により政治問題を指すキーワードを会議録から抽出し，その上位概念を選択することで，政治問題の自動抽出方法の検討と自動体系化の可能性の考察を行う．分析は，大学院生2名により行われた．以下に分析の流れを示す．

1. 基本となる概念体系の作成
2. 会議録の段落による分割
3. 政治問題を含む段落の判断
4. 政治問題を含む段落におけるキーワード抽出
5. キーワードが属する上位概念の選択
6. 判断の一致率の算出

以下に手順の詳細を示す．

4.1 基本となる概念体系の作成

小樽市，帯広市，函館市，釧路市の4市における予備調査では，議題を区分するために存在する委員会体系が非常に類似していた．そこで，本稿では委員会の体系から地方政治における基本の概念体系を設定できると仮定した．そして，最も細目化されている帯広市の市議会における常任委員会とその所管事項の名称[2]をもとに基本となる概念体系を作成した．帯広市の常任委員会は総務文教委員会，厚生委員会，産業経済委員会，建設委員会の4つが存在し，上位概念となっている．それらに加えて，それぞれの委員会に属する事項としていくつかの下位概念が存在している．概念体系作成の手順を以下に示す．

1. 名称末尾の「委員会」を削除
2. 名称末尾の「に関する事項」を削除
3. 名称末尾の「に属する事項」を削除
4. それぞれの上位概念に属する概念として「その他」を追加
5. 上位概念に並列な概念として「その他」を追加

結果として5つの上位概念と54個の下位概念を作成した．作成した概念体系の一部を表1に示す．

[2]帯広市の委員会の所管事項 http://www.city.obihiro.hokkaido.jp/sigikai/inkaisyokan.jsp

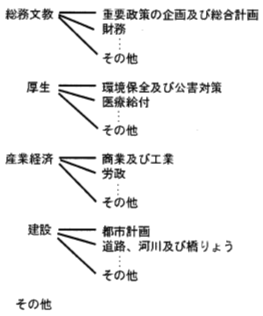

図 1: 基本となる概念体系

4.2 会議録の段落による分割

1つの会議録には，複数の政治問題が含まれている．政治問題を抽出する前処理として，会議録を議題となっている政治問題毎に分割する必要がある．さらに，複数の分析者による分析結果を比較することを考慮すると，分析の単位が同一のほうが好ましい．平成19年度の小樽市議会会議録における予備調査により，1つの段落に政治問題が1つ以上含まれることは稀であった．そこで，1つの段落には政治問題は1つ以上含まれないと仮定して，会議録を段落毎に分割し，分析対象とした．

4.3 政治問題を含む段落の判断

各々の段落に対して，政治問題に関する記述を含んでいるか，または含んでいないかの2値判定を行う．その際，政治問題を含んでいるが適切なキーワードが存在しないと判断した場合は，その段落は政治問題に関わる記述を含んでいないと判定した．

4.4 政治問題を含む段落におけるキーワード抽出

政治問題に関わる記述が含まれると判断された段落に対して，その段落中の文章から内容を反映する適切なキーワードを抽出し，政治問題概念とした．キーワードは概念のラベルに相応しいものを選択するため，「助詞を含まない名詞句」に限定した．

4.5 キーワードが属する上位概念の選択

抽出されたキーワードが属する上位概念を，基本となる概念体系から選択する．我々は1つのキーワード

は1つの上位概念にしか属さないと仮定したが，1つの上位概念に絞れない場合は，2つ以上の上位概念を選択することを許した．

4.6 判断の一致率の算出

2人の分析者における分析結果の類似性を，政治問題が含まれる段落判定の一致率 (1)，抽出されたキーワードの一致率 (2)，唯一の上位概念に絞り込めたかどうかの一致率 (3)，同一の上位概念を選択したかどうかの一致率 (4) を計量することで評価する．ここで分析の対象とした全段落数を D，政治問題が含まれているという判断が一致した段落数を D_1，政治問題が含まれていないという判断が一致した段落数を D_2，キーワードが一致した段落数を A，一致したキーワードに対して上位概念を1つ選択するか複数選択するかの判断が一致した段落数を B，唯一の上位概念を選択し，かつ，同一の上位概念の選択が一致した段落数を C，とする．なおキーワードの一致に関しては完全一致に加えて，包含関係も一致と見なしている．

$$\text{政治問題が含まれる段落の判定の一致率} = \frac{D_1 + D_2}{D} \quad (1)$$

$$\text{抽出されたキーワードの一致率} = \frac{A}{D_1} \quad (2)$$

$$\text{唯一のカテゴリーに絞り込めたかどうかの一致率} = \frac{B}{A} \quad (3)$$

$$\text{同一のカテゴリを選択したかどうかの一致率} = \frac{C}{B} \quad (4)$$

5 結果と考察

各判断の一致率を表2に，各段落数を表3に示す．抽出されたキーワード・選択された上位概念が共に一致した段落の中から出現回数上位5位までのキーワードを表4と表5に示す．以下では，地域・年代による相違，段落毎の分割の妥当性，キーワードの自動抽出，上位概念の選択についての考察を述べる．

5.1 地域・年代によるキーワードの相違

まず，地域・年代による概念体系の相違について考察する．平成14年度の小樽では，学校におけるティームティーチングの制度に関する不正が問題となってたことから，ティーム・ティーチングに関わるキーワードが抽出されているが，平成19年度には自治体の財政赤字が問題となっていることから財政に関するキーワードが頻繁に抽出された (表4)．さらに平成14年度の帯広では乳幼児医療など福祉制度の問題が主に議論されている一方で，平成19年度では「ばんえい競馬」の存続をめぐる議論が多数を占めていた (表5)．このことから，年度による相違があることは明らかである．さらに，「ばんえい競馬」(表5) に象徴されるように，帯広市に存在し，小樽市には存在しないような地域性から生じる政治問題の相違も見られた．このように，地域や年代によって必要となる概念体系は多様であるため，人手により，それぞれの地域・年代ごとに適切な概念体系を構築するには，多大なコストが必要となる．

表 3: 各段落の総数

	小樽 (H14)	小樽 (H19)	帯広 (H14)	帯広 (H19)
D	598	695	1,246	1,314
D_1	303	466	191	249
D_2	137	152	1,017	1,034
A	201	307	135	176
B	197	244	98	163
C	152	142	26	55

表 4: 小樽市会議録から抽出されたキーワード

平成14年度小樽		平成19年度小樽	
キーワード	頻度	キーワード	頻度
加配	12	財政再建	8
TT	11	財政	8
生徒指導	8	予算	7
TT 加配	7	協働	6
ティーム・ティーチング	7	病院	4

5.2 段落毎の分割の妥当性

会議録の分析において，本稿では，1つの段落には政治問題は1つ以上含まれないという仮説に立脚し，分析を行った．表2をみると，2人の分析者による，各段落に政治問題が含まれているか否かの判断の一致率は全体で0.89(κ=0.55) であり，唯一のカテゴリーに絞り込めたかどうかの一致率は0.86であった．このことから，自動化の際にも先の仮説を前提とすることが可能であると考えられる．

表 2: 分析結果

	小樽 (H14)	小樽 (H19)	帯広 (H14)	帯広 (H19)	全体
政治問題が含まれる段落の判定の一致率	0.74	0.89	0.97	0.98	0.89
抽出されたキーワードの一致率	0.66	0.66	0.71	0.71	0.68
唯一のカテゴリーに絞り込めたかどうかの一致率	0.98	0.80	0.73	0.93	0.86
同一のカテゴリを選択したかどうかの一致率	0.77	0.58	0.27	0.34	0.53

表 5: 帯広市会議録から抽出されたキーワード

平成 14 年度帯広		平成 19 年度帯広	
キーワード	頻度	キーワード	頻度
乳幼児医療	4	ばんえい競馬	16
介護保険	4	後期高齢者医療制度	2
児童扶養手当	3	後期高齢者	2
学童保育	3	北海道市営競馬組合	2
予算	3	事故	2

表 6: キーワードの前後 3 単語のフレーズ

前 3 単語	後 3 単語
次に、	であります
次に、	について御説明
また、	につきましては
なお、	におきましては、
次に、	の充実について

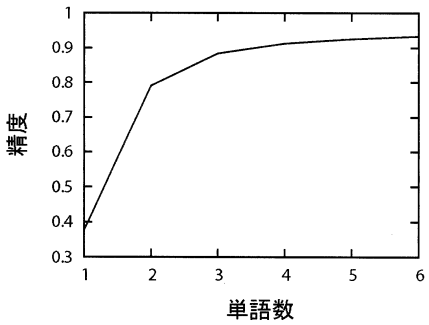

図 2: キーワード抽出精度

5.3 キーワードの自動抽出

表 2 において，2 人の分析者が抽出したキーワードが一致する段落の割合は全体で 0.68 であった．この数値は，政治問題が含まれている段落の中で，政治問題の概念ラベルとして利用可能なキーワードが存在する段落の割合を示している．今後は，人手でさえキーワードの抽出が困難な段落は除外し，キーワードが一意に抽出可能な段落を対象に自動化を進めていくべきであると考えられる．

キーワード抽出の自動化手法として，キーワードを含む文に特徴的な表層表現を利用する方法が考えられる．そこで，分析によって得られたキーワードの前後 n 単語から成るフレーズを抽出し，キーワード自動抽出の精度を以下の手順で算出する．まず，キーワードの前後 n 単語のフレーズを抽出する．次に，全てのフレーズを頻度により尤度を付加し，キーワード抽出ルールとする．このルールを分析に使用した会議録に対して使用し，自動的にキーワードを抽出する．人手による分析と同様に会議録は段落毎に分割され，各段落から，キーワードとして「助詞を含まない名詞句」を 1 つ抽出する．対象とする段落は人手でのキーワード抽出結果が一致した段落に限定し，正解の判定は対象段落について人手で抽出されたキーワードと一致または包含関係であることを基準とした．キーワードの前後 3 単語のフレーズの例を表 6 に示し，前後 1 単語から前後 6 単語までのフレーズについてのキーワード抽出精度を図 2 に示す．キーワードには 2 人の分析者の判断が一致したものを使用している．

図 2 では，前後 3 単語で抽出精度が 8 割を超え，前後 4 単語では精度は 9 割を超えた．このことは，会議録から得られた表層表現は，クローズドデータに対して高精度にキーワードを抽出することが可能であることを示している．

ここで，使用されたフレーズが未知の会議録に対しても有効であるかを考察する．フレーズは年代・地域の異なる 4 つの会議録から得られたものであり，各会議録間のフレーズの共通率を調べることでフレーズの一般性を計量する．共通性として，全ての年代・地域に共通するフレーズ，小樽に共通するフレーズ，帯広に共通するフレーズ，平成 14 年度に共通するフレーズ，平成 19 年度に共通するフレーズの 5 つの項目について，全フレーズに対して占める割合を調査する．前後 4 単語までのそれぞれの割合を表 7 に示す．

表 7 では，全てに共通するフレーズの割合は前後 1

表 7: フレーズの共通率

	前後 1	前後 2	前後 3	前後 4
小樽に共通	0.76	0.17	0.05	0.03
帯広に共通	0.57	0.08	0.05	0.05
H14 年度に共通	0.62	0.10	0.02	0.01
H19 年度に共通	0.66	0.11	0.03	0.01
全てに共通	0.55	0.05	0.01	0.00

表 8: 上位概念の不一致の例

平成 14 年度小樽	
キーワード	上位概念
雇用問題	労政/他の委員会の所管に属しない事項
廃棄物処理施設	環境保全及び公害対策/清掃その他環境衛生
平成 14 年度帯広	
歳入	財務/会計管理者の所管
医療保険制度	医療給付/社会福祉

単語では 0.55 であったが，前後 2 単語では，0.05 であった．これは，各会議録において，キーワードの前後のフレーズには共通性がないことを示している．このフレーズをそのまま利用しても，他の会議録からキーワードを自動抽出することは望めない．フレーズに対する適切な一般化が必要であると考えられる．

5.4 概念の体系化

表 2 において，唯一の上位概念に絞り込めたかどうかの一致率は，全体で 0.86 であった．これは，会議録から抽出されるキーワードは 1 つの上位概念に属するという仮定が成り立つことを示している．今後の概念体系の自動構築においては，この仮説のもとにアルゴリズム開発を進めるべきである．

しかし，同一の上位概念を選択したかどうかの一致率は全体で $0.53(\kappa = 0.40)$ であり，人手による判断すら困難であったことを示している．表 8 にキーワードが一致しているにも拘らず上位概念が一致していない例を示す．例として「廃棄物処理施設」を取り上げてみると，2 人の分析者がそれぞれ「環境保全及び公害対策」と「清掃その他環境衛生」を選択している．自動体系化においては，各概念間の帰属関係の曖昧性を解決することが必要となる．

6 おわりに

本研究では，地方政治に特化した概念体系の構築に向けて，小樽市と帯広市の市議会会議録の分析を行った．分析は帯広市議会の委員会の体系をもとに基本となる概念体系を作成し，段落毎に分割された会議録から政治問題を指すキーワードを抽出し，その上位概念を選択するという方法で行なった．2 名による分析の結果，抽出したキーワードが一致した割合は 0.68 であり，自動化の可能性を示唆するものであった．また，キーワードの前後 4 単語のフレーズを用いて，キーワードの自動抽出を行った結果，クローズドデータに対して，9 割以上の精度を示した．しかし，これらのフレーズは，各会議録間での共通性が低く，未知の会議録に対しての有効性は低いと考えられる．今後，フレーズを適切に一般化することで，汎用的に使用可能なフレーズを生成し，キーワード自動抽出システムの実現を目指したいと考えている．

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．

参考文献

[1] 渋木英潔, 木村泰知, 山崎記敬, “議員発言録からの重要単語抽出システムの提案”, FIT2007 情報科学技術フォーラム 一般講演論文集 第 2 分冊, pp.275-276, 2007.

[2] 木村泰知, 渋木英潔, “ブログに潜在する政治的意見と議員活動とのマッチング手法”, 電子情報通信学会言語理解とコミュニケーション (NLC) 研究会, pp.19-23, 2008.

[3] 川端正法, 山本和英, “話題の継続に着目した国会会議録要約”, 言語処理学会第 13 回年次大会, pp.696–699, 2007.

[4] 山本和英, 安達康昭, “国会会議録を対象とする話し言葉要約”, 自然言語処理, Vol.12, No.1, pp.51–78, 2005.

[5] 友部博教, 長尾確, “ディスカッションマイニング: 議事録集合からの知識発見”, 情報処理学会第 67 回全国大会, 2005.

[6] 本村可奈子, 友部博教, 長尾確, “ディスカッションマイニングシステムにおける会議活性化支援”, 情報処理学会第 67 回全国大会, 2005.