地方議会会議録コーパスの学際的応用を目的とした n-gramデータの構築およびウェブUIの試作

乙武北斗[1]　渋木英潔[2]　高丸圭一[3]　木村泰知[4]　森辰則[5]
[*1]福岡大学　[*2*5]横浜国立大学　[*3]宇都宮共和大学　[*4]小樽商科大学
[*1]ototake@fukuoka-u.ac.jp　[*2]shib@forest.eis.ynu.ac.jp　[*3]takamaru@kyowa-u.ac.jp
[*4]kimura@res.otaru-uc.ac.jp　[*5]mori@forest.eis.ynu.ac.jp

1.　はじめに

近年，国会や多くの地方議会の会議録が電子データとしてウェブ上に公開されている．会議録は首長や議員の議論が書き起こされた話し言葉のデータであるとともに，長い年月の議論が記録された通時的データでもあることから，政治学，経済学，言語学，情報工学などの分野において研究対象のデータとして利用されている．しかしながら，地方議会会議録を利用した研究においては，自治体によってウェブ上に公開されているデータの書式が異なるため，収集作業や整形作業に労力がかかっている現状がある．

このような背景から，我々は多くの研究者に地方議会会議録データを利用してもらうことを目的として，地方議会会議録コーパスの構築を行っている[1][2][3]．コーパスは，ウェブ上に公開されている全国の地方議会会議録を対象として，「いつ」「どの会議で」「誰が」「何を発言したのか」等を検索可能な形式で収録している．しかしながら，コーパスに収録されている発言内容は会議録に記載されているそのままの生テキストであるため，単語単位でデータを利用したい場合には何らかの方法で形態素解析を行う必要がある．コーパスに収録されているデータは膨大な量であるため，利用者が各々で解析を行うことは効率が悪い．また，そのような解析ツールに詳しくない利用者にとっては，解析に多大な労力を要すると考えられる．

そこで我々は，現在までに整備した地方議会会議録コーパスの収録データを対象として，発言の単語n-gramデータの構築を行った．また，構築したn-gramデータを直観的な操作で検索できるウェブユーザインターフェイス（以降，ウェブUIと呼称する）の試作を行った．

以降，**2.**では単語n-gramデータの構築について詳細を述べる．**3.**では構築した単語n-gramデータを直観的な操作で検索することができるウェブUIを試作した結果について述べる．**4.**ではまとめと今後の課題について述べる．

2.　単語n-gramデータの構築

2.1.　対象データ

現在，我々が整備している地方議会会議録コーパスには，表1で示す項目で書式化されたデータが含まれている．今回，我々はコーパスに含まれるデータのうち，2003年度から2010年度の議会における発言内容を対象に，単語n-gramデータを構築した．しかしながら表1で示されるように，コーパスには年度情

表 1　地方議会会議録コーパスの書式

項目	備考
発言ID	自動採番
市町村コード	全国地方公共団体コード（6桁）
議会種別コード	独自のコード
年	西暦
回	開催数
月	開催月
議会名	定例会，臨時会，委員会など
号	会議が何日目なのか
日付	開催日4桁
表題	議会名の情報を含む文字列
段落番号	発言の段落番号
役職名	発言者の役職
議員フラグ	議員ならば1，それ以外は0
発言者名	会議録より抽出されたもの
発言者名表層	通常，発言者と役職のペア
議員ID	議員リストに対応がない場合，-1
発言	文単位の発言文字列
その他	発言以外の文字列

報は含まれていないため,「年」,「月」の情報から推定した年度を基準に，データの抽出を行った．具体的には,「月」が4未満の場合,「年」に1を引いた数値を年度とし，それ以外の場合は「年」の値を年度とした．その結果，対象データに含まれる434の自治体のうち，単語n-gramデータ構築対象の自治体の数は419（うち，19道県，27政令指定都市）となった．

2.2. 構築方法

単語n-gramデータを構築するためには，すべての発言文字列に対して形態素解析を行う必要がある．本研究では形態素解析ツールMeCab[4]を用いて解析を行った．また，定義が明確な短単位を用いることで形態素の境界誤りを防ぐことを狙って，解析辞書としてUniDic[5]を用いた．

次に，この結果を利用して単語n-gramデータを構築した．構築には，Googleが作成したWeb日本語Nグラム第1版[6]のデータ形式を参考にした．以下では，構築した単語3-gramと7-gramデータについて述べる．

単語3-gramデータの例を表2に示す.表2において，“〈S〉”は文頭記号，“〈/S〉”は文末記号を表す．表2で示されるように,単語3-gramデータは3つの単語もしくは文頭・文末記号の並びと，その並びの出現頻度で構成される．

表 2　単語3-gramデータの例

〈S〉 報告 事項	42
報告 事項 は	19
事項 は 、	19
は 、 特に	28
、 特に ござい	20
ん 。 〈/S〉	351

表 3　単語7-gramデータの例

〈S〉 報告 事項 は 、 特に ござい	42
報告 事項 は 、 特に ござい ませ	19
事項 は 、 特に ござい ませ ん	19
は 、 特に ござい ませ ん 。	28
、 特に ござい ませ ん 。 〈/S〉	20

次に，単語7-gramデータの例を表3に示す．単語7-gramデータも，単語3-gramデータと同様，7つの単語もしくは記号の並びと，その頻度で構成される．ただし7-gramデータの構築の際は，低頻度単語列を除外して構築処理の効率化を図るため，先頭3-gramの頻度による閾値を設け，閾値未満の頻度を持つ7-gramは処理の対象としない．今回，閾値は3に設定した．

単語3-gram，単語7-gramデータは2.1で述べた対象データのすべての発言文字列から，自治体と年度別に構築した．対象データの発言文字列における総単語数は，419自治体合わせて約30億語となった．主な政令市の総単語数，異なり3-gram数，異なり7-gram数を表4，5，6に示す．基本的には総単語数が多ければ多いほど，構築されるn-gramデータの規模も大きくなることがわかる．また，表4の札幌市と京都市との比較で見られるように，地域によって総単語数の差が大きいことがわかる．これは，地方議会会議録コーパスに含まれるデータの議会種別が主な理由として挙げられる．例えば札幌市の場合，コーパスに含まれるデータは定例会と臨時会のほかに，各種委員会や特別委員会のものがあるが，京都市は定例会と臨時会のみとなっている．各自治体によって会議録の公開方法が異なるため，このように自治体によって収集量に差が生じる結果となっている．

3. 検索用ウェブUIの試作

地方議会会議録コーパスのデータは，自治体や会議の数が多いため，データサイズが非常に大きくなる．そのような巨大なデータから構築されるn-gramデータのサイズも非常に巨大なものになる.さらに,巨大なデータから検索を行ったり，集計を行ったりする場合，プログラミング等の専門知識が必要になる．そこで，本研究では，本稿で述べたn-gramデータのサイズを意識せずに容易に情報にアクセスできることを目的として,検索用のウェブUIを試作した.

本システムでは大量のn-gramデータから指定した

表 4　主な政令指定都市における各年度の発言の総単語数

自治体名	2010	2009	2008	2007	2006	2005	2004	2003
札幌市	703,206	2,965,669	3,379,468	3,162,158	2,596,116	2,924,745	3,291,320	3,418,811
新宿区	3,496,492	3,756,860	3,883,815	3,766,527	3,564,597	3,497,193	4,097,073	2,916,119
横浜市	2,709,129	5,018,437	5,211,368	4,435,905	5,238,646	5,009,796	5,357,124	5,434,996
名古屋市	1,076,304	1,852,574	1,643,036	1,326,941	641,199	720,823	672,282	644,214
京都市	324,071	452,085	465,036	407,963	420,556	438,831	396,507	402,727
大阪市	1,345,342	3,990,299	4,217,031	3,495,335	4,161,278	4,107,615	2,751,902	3,140,550
北九州市	739,464	1,140,095	1,120,191	1,381,088	882,710	1,111,522	1,134,943	1,337,305

表 5　主な政令指定都市における各年度の発言の単語3-gram異なり数

自治体名	2010	2009	2008	2007	2006	2005	2004	2003
札幌市	254,728	786,449	877,047	831,457	705,018	774,554	891,976	882,770
新宿区	881,649	937,573	974,531	912,197	881,721	876,223	967,556	756,674
横浜市	757,284	1,238,925	1,242,019	1,111,433	1,231,442	1,197,223	1,269,608	1,257,452
名古屋市	371,092	572,131	528,606	441,045	247,928	273,532	258,557	244,540
京都市	152,245	195,160	204,644	180,390	188,477	199,035	182,502	187,297
大阪市	451,075	1,018,051	1,067,198	935,047	1,047,622	1,020,468	784,289	862,892
北九州市	288,795	403,704	392,315	459,631	329,679	386,713	394,233	390,927

表 6　主な政令指定都市における各年度の発言の単語7-gram異なり数

自治体名	2010	2009	2008	2007	2006	2005	2004	2003
札幌市	321,486	1,550,188	1,811,109	1,675,496	1,365,350	1,541,045	1,755,357	1,819,885
新宿区	1,848,345	1,998,928	2,062,785	1,994,953	1,896,419	1,850,827	2,215,919	1,527,674
横浜市	1,341,126	2,670,556	2,800,106	2,322,692	2,793,028	2,656,761	2,872,487	2,939,894
名古屋市	495,298	899,522	765,862	605,782	264,078	298,146	274,160	259,936
京都市	110,588	168,341	171,623	149,706	152,661	152,521	139,330	138,553
大阪市	628,347	2,060,668	2,190,411	1,778,757	2,172,140	2,128,129	1,372,042	1,582,535
北九州市	315,978	517,472	508,195	652,849	380,736	505,043	511,097	530,074

単語の並びを検索する必要があるため，各自治体・年度のn-gramデータから全文検索システムのインデックスを作成し，それを利用することとした．全文検索システムには様々なものが存在するが，本システムはApache Lucene[7]を用いた．2.1で述べた対象データから構築されたLuceneインデックスサイズは137GBとなった．

本システムの検索インターフェイス画面を図1に示す．本システムでは3種類の条件を用いてn-gramデータの検索を行うことができる．図1の上から順に，1つ目はn-gram単語検索であり，このテキストボックスには検索したい単語を含む文字列を入力する．ただし，8単語以上から成る文字列を入力した場合は，検索対象データが最長でも7-gramまでの情報しか含まないため，検索結果は出力されない．入力された文字列はシステム内部で形態素解析され，その単語の並びを含むn-gramデータが検索される．2つ目は検索対象市町村で，3つ目は検索対象の年度である．市町村，年度ともに，何も指定しなかった場合はすべての自治体・年度から検索を行う．

本システムで検索を行った結果の画面例を図2に示す．図2は，n-gram単語検索条件に「です。</S>」（文末が“です。”となる単語列）と設定して出力された結果を表している．検索結果は，n-gramの頻度順に出力され，頻度，自治体名，年度，当該n-gramを含んだ情報が表示される．また，結果上部には総ヒット数と，出力までにかかった処理時間が表示される．一度に出力される結果は100件に設定しており，100件目の表示部分にそれ以降の項目を表示するためのボタンを配置している．そのボタンをクリック

することで, 101位以下の結果を次々と表示することが可能である．しかしながら，現在はすべての結果をまとめてCSVなどの他の形式でエクスポートすることはできない．

4. まとめと今後の課題

本稿では，我々が整備を行っている地方議会会議録コーパスの発言内容を対象とした単語n-gramデータの構築について述べた．また，構築されたn-gramデータを直観的な操作で検索できるよう試作したウェブUIを紹介した．

現在，単語n-gramデータは3-gramと7-gramのみ構築が完了している．引き続き，7-gram以下のn-gramデータの構築を行う予定である．

ウェブUIに関しては，現在，3種類の条件による検索結果を100件単位で閲覧できる状態であるため, データ全体を用いた集計を行うことが非常に困難である．今後の課題として，閲覧可能な項目数をカスタマイズできるようにするだけでなく，結果をCSVやXML，JSON形式でエクスポートするウェブAPIを構築することを予定している．

謝辞 本研究の一部は，科研費 No:22300086 による．

参考文献

[1] 木村泰知，渋木英潔，高丸圭一，乙武北斗，森辰則，“地方議会会議録コーパスの構築とその利用”，第26回人工知能学会全国大会，3B3-NFC-4-3，2012

[2] 菅原晃平，大城卓，齋藤誠，永井隆広，渋木英潔，木村泰知，森辰則，“地方議会会議録コーパスの拡充における問題点の分析と対処”，言語処理学会第18回年次大会発表論文集，pp. P1-15，2012

[3] 齋藤誠，大城卓，菅原晃平，永井隆広，渋木英潔，木村泰知，森辰則，“地方議会会議録の収集とコーパスの構築”，言語処理学会第17回年次大会発表論文集，pp. P2-21，2011

[4] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

[5] 伝康晴他(2007)「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」『日本語科学』22号 pp.101-122

[6] 工藤拓, 賀沢秀人著, “Web 日本語Nグラム第１版”，言語資源協会発行

[7] Apache Lucene，http://lucene.apache.org/

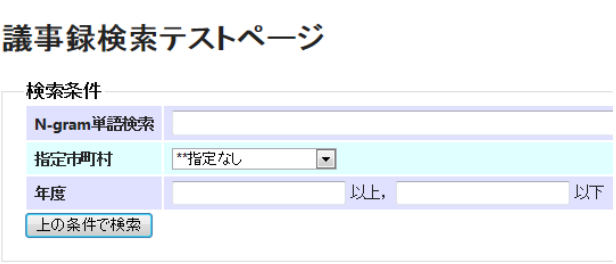

図 1 検索インターフェイスの画面

図 2 検索結果画面の例