地方議会会議録における発言文の推定

◯ [1] 桧森拓真　[1,2] 木村 泰知　[3] 坂地 泰紀

[1] 小樽商科大学　[2] 理研 AIP　[3] 東京大学

kimura@res.otaru-uc.ac.jp

1　はじめに

近年，公共データの活用に期待が高まっており，自治体のオープンデータ化が進んでいる．ウェブ上に公開されている公共データの一つに地方議会会議録がある．地方議会会議録は地方議会でのすべての発言，すなわち議会における議論の過程を書き起こしたテキストデータであり，いつ，どこで，だれが，なにを発言したのかが記録されている．

このため，地方自治体における政策の決定がどのように行われたのかを確認するために利用することができる．例えば，東京都議会の会議録を読むことで，築地市場の豊洲移転問題について，知事や議員がどの時点でどのような発言をしたのかを知ることができる．

我々は，全国の自治体のウェブサイトから地方議会会議録を収集し，地方議会会議録コーパスの構築を進めてきた [1]．本コーパスは，自然言語処理や社会言語学の分野において利用され，研究成果をだしている[1]．

しかしながら，本コーパスは，統計量を扱う計量経済学のような分野において，網羅性や正確性の観点から，利用されていない．そこで，我々は，政治学・経済学の分野における地方議会会議録コーパスの活用に向けて，会議録中のすべての発言に対して「自治体名」「発言日」「発言者」を正確に付与することを進めた [2]．その結果，発言者情報を正確に付与することにより，異なる表記（例えば,「梅沢佳一」や「梅澤よしかず」などと記載されている場合）でも同一人物として識別を行えることから，正確に話者を識別しながら，発言文字数を表示できるようになった．しかしながら，もう一つ大きな問題として，発言を正確に識別しできていない問題が残っている．ここで，発言とは，知事，議員，職員などの発言を書き起こした記述のことである．例えば，東京都議会の会議録には，下記のように「発言以外の文字列（これ以降，非発言と呼ぶ）」が含まれており，４割弱の「非発言」が含まれている．

議会会議録における「発言」「非発言」の例

発言：これにご異議ありませんか。
非発言：〔「異議なし」と呼ぶ者あり〕
発言：ご異議なしと認めます。
発言：よって、日程第 1 から第 24 までは、議案付託事項表のとおり、それぞれ所管の常任委員会に付託することに決定いたしました。
非発言：（別冊参照）
非発言：――――――――――
発言：これより追加日程に入ります。
発言：追加日程第 1、東京都副知事の選任の同意についてを議題といたします。
非発言：〔鈴木議事部長朗読〕
非発言：1、東京都副知事の選任の同意について 1 件
非発言：――――――――――
非発言：23 財主議第 133 号
非発言：平成 23 年 6 月 17 日
非発言：東京都知事石原慎太郎

上記のように，地方議会会議録は，状況説明・ト書き・添付資料・朗読[2]のような「非発言」が含まれているため，発言文字数を正確に数えらないという問題がある．

そこで，本研究では，正確な統計データの作成に向けて，地方議会会議録に含まれるテキストの発言箇所を自動で推定することを目的とする．本稿では，地方議会会議録の発言文のアノテーションについて説明するとともに，発言文の自動推定の性能評価実験の結果について述べる．

2　地方議会会議録における発言文

2.1　発言文へのアノテーション

我々は，平成 23 年 4 月から平成 27 年 3 月までの４年間を対象として，47 都道府県議会議事録を収集

[1] http://local-politics.jp/

[2]「朗読」は履歴書の読み上げなど通常の発言と異なることから，非発言としている．

している．また，それらを統一したデータ構造で管理しており，データベースに格納している．現在までに，47 都道府県の地方議会会議録の 4 年間分を漏れなく収集し，異なる表現の名前でも話者を識別することができるようにしている．これらの地方議会会議録コーパスから，4 つの自治体「青森県」「東京都」「大阪府」「福岡県」を対象に，発言文に発言と非発言のアノテーションを行う．

表 1 に 4 つの対象自治体の平成 23 年 4 月から平成 27 年 3 月までの 4 年間に含まれる行数を示す．

表 1: 4 年間の会議録に含まれる行数

都道府県名	青森県	東京都	大阪府	福岡県
行数	96,095	115,750	95,929	84,918

注釈者は大学生 4 人 (A,B,C and D) であり，アノテーションは一つの自治体（同一データ）に対して，2 人の注釈者が発言と非発言のラベルを一文単位で付与する．また，発言と非発言が混在する場合や判断が難しい場合には「その他」のラベルを付与する．アノテーションのラベルは，次の通りである．

ラベル ”1” 発言
　　3 文字以内の「記号」は切り分けない
　　「――」のような記号が含まれる場合も入れる
ラベル ”2” 非発言
ラベル ”9” その他
　　発言と非発言が混在する場合
　　あるいは，判断が難しい場合

議長	和田宗春	この際、開議に先立ちまして、このたびの東日本大震災	1
議会局長	白石弥生子	全員ご起立願います。	1
議会局長	白石弥生子	〔全員起立〕	2
議会局長	白石弥生子	黙祷をお願いいたします。	1
議会局長	白石弥生子	〔黙　　祷〕	2
議会局長	白石弥生子	黙祷を終わります。	1
議会局長	白石弥生子	ご着席願います。	1
議会局長	白石弥生子	――――――	2
議長	和田宗春	これより本日の会議を開きます。	1
議長	和田宗春	――――――	2
議長	和田宗春	まず、議席の変更を行います。	1
議長	和田宗春	議席変更の申し出がありますので、会議規則第二条第三	1
議長	和田宗春	（別冊参照）	2
議長	和田宗春	――――――	2

図 1: 東京都議会会議録のアノテーションの例

図 1 は東京都議会会議録の各行に「発言」「非発言」を付与している例である．表 2 はアノテーションの結果であり，二人の注釈者が一致した発言数，非発言数，合計，二人の注釈者の不一致数,「その他」の行数，合計，κ 値[3]を自治体ごとに記載している．

[3]R の library(irr) の kappa2 関数を用いて計算した．

3 章では，アノテーションしたデータ用いて，評価実験を行う．

3　発言文の推定実験 1

3.1　実験の目的

本実験では，訓練データと評価データに「同一自治体のデータ」を利用して，地方議会会議録に含まれるテキストの発言箇所を自動で, 発言か非発言かを推定することを目的とする．

3.2　実験方法

地方議会会議録のテキストから特徴語辞書を作成し, さらに特徴ベクトルを作成する．特徴語辞書を作成するために, テキストを 1 文づつ Mecab を使用して形態素解析を行い, 特徴語辞書を作成する．作成した特徴語辞書を,Bag-of-words モデルを用いて特徴語ベクトルを作成する．作成した特徴語ベクトルを LSI(Latent Semantic Indexing) を用いて 100 次元のベクトルに圧縮する．作成した特徴語ベクトルを判別器を使用して発言か非発言かを判別する．学習データには, 推定する発言文と同一自治体の発言文を使用する. 例を挙げると, 東京都の発言文推定には, 東京都の発言文を学習データに使用する．判別の精度は十分割交差検証によって評価する．

本実験では，推定結果を評価するために，表 3 のように，True Positive(TP), False Positive(FP), False Negative(FN) and True Negative(TN) に分ける．正解数及び正解率の計算方法は，次の通りである．

$$\text{正解数} = TP + TN$$

$$\text{正解率} = \frac{TP + TN}{TP + FP + FN + TN}$$

また, 各自治体の正解数を足したものを, 各自治体の総文数の総和で割ったものを平均正解率とする．以上の予測結果の評価方法に基づいて実験結果を評価する．

3.3　実験データ

実験データには平成 23 年 4 月から平成 27 年 3 月までの定例会を対象とした会議録を使用する．対象となる自治体は, 青森県, 東京都, 大阪府, 福岡県の 4 つである．本実験では，注釈者の二人が一致したデータだけを用いることとした．つまり，各自治体の発言および非発言のデータ数は，表 2 における「一致」と記載されている「発言」「非発言」の値である．

表 2: 4 つの自治体に対するアノテーションの結果

自治体	注釈者	一致			不一致	その他	合計	κ
		発言	非発言	合計				
青森県	C,D	91,714	2,875	94,588	882	625	96,095	0.807
東京都	A,B	72,604	42,909	115,512	215	23	115,750	0.996
大阪府	C,D	74,592	18,320	92,912	167	2,850	95,929	0.914
福岡県	A,B	78,930	5,638	84,568	276	74	84,918	0.992

表 3: 予測結果の評価

		正解 (注釈結果)	
		正 (発言)	負 (非発言)
予測結果	正 (発言)	TP	FP
	負 (非発言)	FN	TN

番号	発言文	正解ラベル	注釈者の注釈	予測ラベル
1	七十番　服部ゆくお君	1		1
2	百二番　大津　浩子君	2	朗読及び資料	1
3	以上、地方自治法第九十九条の規定により意見書を提出する。	1		1
4	一度失われた農地を取り戻すことは極めて困難であり、一刻も早い対応が必要である。	2	朗読及び資料	1
5	知事石原慎太郎君。	1		1
6	〔知事石原慎太郎君登壇〕	2		2

図 2: 実験結果

3.4 実験結果

実験結果を表 4, 表 6 に示す. ここで，RF はランダムフォレスト,LR はロジスティック回帰を示す. 東京都では SVM の正解率が 0.952 と一番高く, 決定木は 0.862 と一番低い正解率となった. 青森県, 大阪府及び福岡県では SVM,RF の正解率が 0.999 と非常に高い正解率であった. 表 6 の平均正解率をみると,SVM が 0.987 と一番正解率が高いことがわかる.

3.5 考察

図 2 に実験結果の一部を抜粋したものを示す. 1 番と 2 番は一文だけ見ると番号と名前を示している. しかしながら,1 番のものは実際に名前を読んでおり,2 番は資料に書かれている文である. 3 番と 4 番も同様で, 文だけみると誰かが発言している文に思われるが,3 番は実際の発言で,4 番は資料中の記述である. 以上のように一文だけ見て判別するのに難しいことがわかる. 一方，5 番と 6 番のように一見同じような文章でも鉤括弧の有無で 5 番を発言,6 番を非発言と判別できている例もあった. 以上から, 鉤括弧など記号のついている非発言部分は上手く判別できるていることがわかったが,2 番や 4 番のように発言文だけみると非発言と区別できないような文に対しては上手く判別できていないことがわかった.

4 発言文の推定実験 2

4.1 実験の目的

本実験では，学習データと評価データに「異なる自治体のデータ」を利用して，地方議会会議録に含まれるテキストの発言箇所を自動で, 発言か非発言かを推定することを目的とする.

4.2 実験方法

前実験では, 推定したい自治体の発言文を学習データに使用していた. 本実験では, 東京都の発言文を学習データとし, 青森県, 大阪府及び福岡県の発言文を判別器を用いて推定する. 学習データを東京都の発言文, テストデータには他の 3 つの自治体の発言文を使用するため, 交差検定は行わない. それ以外の実験の方法, 及び評価方法は前実験と同様である.

4.3 実験データ

実験データは前実験に用いたものと同様のものを使用する.

4.4 実験結果

実験結果を表 5 と表 6 に示す. 表 5 からは青森県と福岡県では,RF が 0.953 と一番正解率が高く, 大阪府では SVM が 0.968 と一番正解率が高くなったことがわかる. 表 6 からは, 平均正答率では,SVM が平均正解率 0.952 と他の手法よりも高くなることがわかる.

4.5 考察

前実験と同様に, 本実験では, 学習データを推定したい自治体と別の自治体の発言文に変更しても, 大きく正解率が下がらないことがわかった. また, 手法においては,SVM が一番平均正解率を高くすることがわかった. しかし, 前実験と同様に発言文だけを見ると, 発言文か, 資料の記述などの非発言文かを判別できないような文については, 解決できていない.

表 4: 同一自治体の会議録を学習データとした 10 分割交差検定の実験結果

評価データ	手法	パラメータ	TP	FP	FN	TN	総文数	適合率	再現率	F 値	正解率
東京都	SVM	kernel=rbf	71,693	4,613	911	38,295	115,512	0.940	0.987	0.963	0.952
東京都	決定木	深さ=3	69,039	12,351	3,565	30,557	115,512	0.848	0.951	0.897	0.862
東京都	RF	標本数=10	72,045	6,005	559	36,903	115,512	0.923	0.992	0.956	0.943
東京都	LR	C=1	70,501	5,607	2,103	37,301	115,512	0.926	0.971	0.948	0.933
青森県	SVM	kernel=rbf	91,647	47	66	2,828	94,588	0.999	0.999	0.999	0.999
青森県	決定木	深さ=3	91,449	100	264	2,775	94,588	0.999	0.997	0.998	0.996
青森県	RF	標本数=10	91,704	22	9	2,583	94,588	0.999	0.999	0.998	0.999
青森県	LR	C=1	91,643	75	70	2,800	94,588	0.999	0.999	0.999	0.998
大阪府	SVM	kernel=rbf	74,562	59	30	18,261	92,912	0.999	0.999	0.999	0.999
大阪府	決定木	深さ=3	73,851	408	741	17,912	92,912	0.995	0.996	0.995	0.988
大阪府	RF	標本数=10	74,588	57	4	18,263	92,912	0.999	0.999	0.999	0.999
大阪府	LR	C=1	74,506	58	86	18,262	92,912	0.998	0.999	0.999	0.998
福岡県	SVM	kernel=rbf	78,907	66	23	5,572	84,568	0.999	0.936	0.967	0.999
福岡県	決定木	深さ=3	78,595	398	335	5,240	84,568	0.999	0.913	0.954	0.991
福岡県	RF	標本数=10	78,927	44	3	5,594	84,568	0.999	0.996	0.998	0.999
福岡県	LR	C=1	78,873	107	57	5,531	84,568	0.996	0.938	0.966	0.998

表 5: 学習データを東京都として，評価データを異なる自治体とした場合の比較実験

評価データ	学習データ	手法	パラメータ	TP	FP	FN	TN	総文数	適合率	再現率	F 値	正解率
青森県	東京都	SVM	kernel=rbf	85,840	1	5,873	2,874	94,588	0.999	0.936	0.967	0.938
青森県	東京都	決定木	深さ=3	83,698	34	8,015	2,841	94,588	0.999	0.913	0.954	0.915
青森県	東京都	RF	標本数=10	87,334	25	4,379	2,850	94,588	0.999	0.996	0.998	0.953
青森県	東京都	LR	C=1	86,035	384	5,678	2,491	94,588	0.996	0.938	0.966	0.936
大阪府	東京都	SVM	kernel=rbf	72,479	1,099	2,113	17,221	92,912	0.985	0.972	0.978	0.968
大阪府	東京都	決定木	深さ=3	59,584	125	15,008	18,195	92,912	0.998	0.799	0.887	0.837
大阪府	東京都	RF	標本数=10	71,311	7,789	3,281	10,531	92,912	0.902	0.956	0.928	0.881
大阪府	東京都	LR	C=1	70,634	1,464	3,958	16,856	92,912	0.980	0.947	0.963	0.942
福岡県	東京都	SVM	kernel=rbf	76,299	1,483	2,631	4,155	84,568	0.980	0.967	0.974	0.951
福岡県	東京都	決定木	深さ=3	71,689	1,514	7,241	4,124	84,568	0.979	0.908	0.942	0.896
福岡県	東京都	RF	標本数=10	76,814	1,449	2,116	4,189	84,568	0.981	0.951	0.966	0.958
福岡県	東京都	LR	C=1	74,882	1,781	4,048	3,857	84,568	0.977	0.949	0.963	0.931

表 6: 学習データごとの平均正解率

学習データ	手法	パラメータ	平均正解率
同一自治体	SVM	kernel=rbf	0.985
同一自治体	決定木	深さ=3	0.953
同一自治体	RF	標本数=10	0.983
同一自治体	LR	C=1	0.979
東京都	SVM	kernel=rbf	0.951
東京都	決定木	深さ=3	0.882
東京都	RF	標本数=10	0.930
東京都	LR	C=1	0.936

5 おわりに

本稿では，正確な統計データの作成に向けて，地方議会会議録に含まれる発言および非発言の調査を行うとともに，発言文の自動推定の性能評価実験について述べた．性能評価実験では，同一自治体の発言文の推定実験ではSVMの手法が平均正解率0.987となり，最も高いことを確認した．学習データと評価データを異なる自治体にした比較実験では,SVMの手法が平均正解率0.952と前実験と同様に最も高いことを確認した．今後は，一文だけで，発言か非発言かわからないような文に対して，前後の文脈を考慮する手法について検討する．

謝辞

本研究は，JSPS 科研費 JP16H02912 の助成，および，国立国語研究所プロジェクト「議会会議録を活用した日本語のスタイル変異研究」の助成を受けたものです．

参考文献

[1] 井原大将, 内田ゆず, 高丸圭一, 木村泰知, 江崎浩, 全地方議会会議録の横断検索に向けたデータ収集とデータ構造の検討, 第 233 回自然言語処理研究会, 2017.

[2] 木村泰知, 内田ゆず, 高丸圭一, 都道府県議会会議録のパネルデータ作成に向けた発言者情報の付与, 第 33 回ファジィシステムシンポジウム講演論文集, pp.701-706, 2017.