話し言葉の整形作業における削除箇所の自動同定

尾嶋 憲治† 内元 清貴‡ 丸山 岳彦* 秋田 祐哉† 河原 達也†

†京都大学 情報学研究科 ‡独立行政法人 情報通信研究機構 *独立行政法人 国立国語研究所

1 はじめに

近年、音声認識技術の進展にともない、会議や講演などの話し言葉音声を対象とした筆記録の作成や整形・要約などの研究が進められている [1]。しかし、このような話し言葉音声の書き起こしや音声認識による結果は、書き言葉とは異なる点が多く、可読性がよいとは言えない。そのため、会議録や講演録といったテキストアーカイブとしての利用を念頭に置いた場合、音声の書き起こしをそのまま利用するのではなく、文の区切りで分割する、フィラーや冗長部分を取り除く、話し言葉特有の表現を適切に置き換える、話中で省略された助詞などの語を補う、などの整形処理を行う必要がある。しかしながら、現在このような整形処理は主に人手で行われているため、会議録や講演録の作成に大きなコストがかかっている。ここで自動的な処理が可能となれば、効率的な筆記録の作成に役立つと考えられる。

整形作業については従来より、整形を受ける言語構造の解析や機械的な処理の研究が進められている。話し言葉における繰り返しや言い直しなどの自己修復部の分析・検出については、RIM（Repair Interval Model）[2] が代表的なモデルとして知られており、自己修復部を被修飾部（RPD）、言いよどみ（DF）、修復部（RP）の連続した 3 つの区間に分割されると仮定して、その検出を行っている。この分析・検出は、英語だけでなく、日本語を対象としても行われている [3, 4]。また、自己修復部だけでなく、フィラーや言いよどみの検出 [5] も行われている。しかし、これまで分析や検出の研究にとどまっており、話し言葉の整形処理との関係は明らかではなかった。統計的機械翻訳の手法を応用して講演の書き起こしから整形を試みる研究 [6] も行われているが、整形処理を、句点の挿入、助詞の挿入、文体の統一に限定している。

本稿では、整形処理における削除箇所と自己修復部、フィラー、言いよどみ、定型表現との関係を分析し、講演録編集者が削除した箇所を自動的に推定する手法を提案する。本手法では、SVM（Support Vector Machine）を用いたテキストチャンキングにより削除箇所を自動推定する。素性としては、表現の繰り返しや係り受けなどの言語的情報や、自己修復部についての情報を採用する。

なお、本手法の分析および実験には『日本語話し言葉コーパス』（CSJ）[7] を利用している。CSJ は、学会講演や模擬講演などのモノローグを主な対象として収集・構築されたコーパスである。CSJ に収録されている講演のうち、コアと呼ばれる一部の講演に対しては、書き起こしテキストのほかに形態素・係り受け・節単位などの言語的情報やポーズ情報などの音響的情報が付与されている。

2 話し言葉に対する整形処理

会議や講演を収録した音声の書き起こしや認識結果から会議録や要約を作成する過程においては、主に以下のような順序で作業が行われる [6]。

1. 一次整形
2. 長文の分割、文法的チェック、ポリティカルチェック
3. 意味的チェック
4. 要約の作成

第一段階の一次整形では、フィラーの削除や書き言葉表現への変換、助詞の挿入などを行う。第二段階では、文法的チェックで不適切に用いられている助詞や接続詞を適切なものに修正し、ポリティカルチェックで不適切な表現の修正を行う。第三段階の意味的チェックでは、専門用語が正しく用いられているかの確認を行う。

以上の作業のうち、本研究では一次整形の処理を対象とする。一次整形として行われる処理としては、以下の 3 つが挙げられる。

- 削除：言いよどみや重複箇所など、冗長な部分を除く。
- 置換：語の誤りや話し言葉特有の表現などを適切な表現に置き換える。
- 挿入：発話中で省略された助詞などの語を補う。

一次整形を対象とするのは、これらの処理だけでもかなり読みやすいものに整形される反面、これ以上の処理については、内容の理解を含めた高度な処理が必要であると考えられるからである。

表 1 は CSJ を対象にして行った一次整形における 3 つの処理の内訳を示している。3 つの処理のうちでは削除箇所が最も多く、全体の約 6 割を占めている。また、図 1〜3 から分かるように、 書き起こしから削除箇所を除く作業だけでも、可読性が十分に向上する。以上より、本研究では話し言葉の一次整形における削除箇所の自動推定について検討を行う。

表 1: 整形要素の要素数

整形要素	要素数	
削除	19621	(57.3%)
置換	11640	(34.0%)
挿入	2946	(8.6%)

3 コーパスに付与された情報

本研究に際して、CSJ のコア (188 講演）およびコアに含まれないテストセット（11 講演）に対して、一次整形の箇所を開始タグと終了タグで記述したデータを構築した。このうち、削除箇所については、短単位を単位として IOB ラベリングスキームでラベリングを行い、自動推定の実験に用いた。また、今回利用したデータには他に、フィラー情報（F タグ）、言いよどみ（D タグ）、言い直し表現（R タグ）の各情報が、一次整形の各要素とは独立に付与されている。このうち F タグおよび D タグについては、CSJ に元来付与されているが、 F タグについては削除が容易であるので本研究の対象としない。

一方 R タグは、自己修復に関する代表的なモデルである RIM [2] を参考に、丸山らが CSJ の独話から収集した言い直し表現の体系 [4] をもとに人手で付与した情報である。この R タグは、言い直し表現の機能的な分類に基づき、R1〜R5 の 5 種類に分けられている。

でこれは私だけじゃなくて周り私の友人もやはり幼馴染みなんかはみんな青山の辺り住んでおりますからみんな同じことでこの前そのキディーランドに行って何を買ってもらったって言うともう小学校で自慢し合うようなそういう感じだったんですけども

図 1: 講演の書き起こしの例

これは私だけじゃなくて私の友人もやはり幼馴染みはみんな青山の辺り住んでおりますからみんな同じことでこの前そのキディーランドに行って何を買ってもらったって言うともう小学校で自慢し合うような感じだったんです

図 2: 削除箇所を除いた例

これは私だけではなくて私の友人もやはり幼馴染みは皆青山の辺りに住んでおりますから皆同じことでこの前そのキディーランドに行って何を買ってもらったと言うともう小学校で自慢し合うような感じだったのです

図 3: 削除・置換・挿入を行った例

(F ま) {R5 そんな風に思える |(F あの)| ペットボトルの使い方もそういう {R3 風な ‖ 風に } 考えられる } んじゃないかなと思います

図 4: 入れ子状に存在する R タグの例

- R1：発音エラーに伴う言い直し
  e.g. コンテキスト {R1 (D いぞ)‖ 依存 } モデルを使う
- R2：単純な繰り返し
  e.g. その下の {R2 波線 ‖ 波線 } を付けました
- R3：語彙的な誤りに伴う言い直し
  e.g. 独特の {R3 旋律の |(F えー)| 旋律を } 形作っている
- R4：情報不足に伴う言い直し
  e.g. 海外に {R4 興味 ‖ 元々興味 } がありまして
- R5：別表現への言い換え
  e.g. {R5 こちら側のグラフ ‖ 右側のグラフ } は何を表しているかと言いますと

なお、これらの R タグは、文の構造により、図 4 のように複数の R タグが入れ子状に存在する場合も見受けられる。

上で述べた削除箇所および R タグは、それぞれ異なる基準で独立に付与されたラベルである。削除箇所のラベルは言語的特徴を問わず整形作業をする上で削除すべき箇所に広く付与されている。一方、R タグは言い直し表現と認定された箇所について、整形作業に関係なく付与されている。削除箇所を R タグとの関係で分類すると、5 種類の R タグの被言い直し部（RPD）と言い直し部（RP）との品詞や意味の関係も考慮することにより、表 2 に示すような 4 つのタイプに分類することができる。各タイプ内では、比較的性質が似ているといえる。このうち R タグに関係する削除箇所には上で述べたような繰り返しの表現が多く見られ、「その他」つまりタイプ 4 に含まれるものには、以下のような種類の定形表現が多く見られた。

- 終助詞　e.g. 富士山がよく見えるんです よ
- 文中・文末の冗長表現　e.g. その前日 ですね
- 文頭の接続詞　e.g. で 今ちょうど冬の時期で

表 2: 各 R タグと削除箇所の関係

		総数	RPD を削除	RP を削除
タイプ 1	R1	1426	1421	113
タイプ 2	R2	1253	1183	69
	R3	1174	1108	57
	R4	961	624	60
タイプ 3	R5	595	250	116
タイプ 4	R なし	14679	—	—

- 間投的に用いられる表現　e.g. 何ですかね

表 2 には、学習データにおける各 R タグと削除箇所との関係も示した。R1,R2,R3 が付与された箇所では、その大多数で被言い直し部（RPD）が削除箇所に該当している。またこの 3 つに加えて R4 でも、言い直し部（RP）を削除する場合が、RPD を削除する場合の 1 割に満たないのに対し、R5 が付与された箇所では RP を削除する場合が、RPD の場合の約半数と、その比率が大きく異なっている。

4　削除箇所推定のアプローチ

本研究では、文章中の削除箇所の推定をテキストチャンキングの問題として扱い、SVM に基づくテキストチャンカである YamCha [8] を利用して、形態素列からの推定実験を行う。なお、チャンクラベルとして、IOB ラベリングスキームに基づいたラベルを形態素ごとに付与する。YamCha における多項式カーネルの次数は 3、解析方向は Right to Left とする。

SVM に与える素性としては、前後 3 形態素の単語情報（表層表現、読み、品詞情報、活用の種類、活用形）やポーズの有無、フィラー情報、言い直し情報（R タグ）および、表現の重複（繰り返し）の有無を用いた。また上記に加えて、書き起こしデータに対する実験では文節境界と係り受け情報 も素性として利用している。このうち、ポーズは CSJ の転記基本単位の定義に従う。

また表現の重複（繰り返し）の有無は、図 5 のように、ある形態素列に対しその前後 5 形態素および 10 形態素以内に、単語情報（表層、読み、品詞）が全く同じ形態素列および、単語情報が同じ自立語でかつその後に続く品詞が同じ形態素が含まれる形態素列を機械的に抽出し、5 形態素と 10 形態素の両方について抽出された繰り返しのうち文頭側の表現と文末側の表現にそれぞれ F ラベルと B ラベルを独立に与えた。

書き起こしを対象とした実験においては、品詞情報や文節、係り受けの情報はコーパスに付与された正解を用いており、フィラーは形態素単位に F タグが付与された感動詞としている。一方、音声認識結果での実験における削除箇所および R タグのラベルは、書き起こしデータに付与されたラベルを音声認識結果へ自動的に移行したものをもとに、その範囲が適切になるよう人手で修正したものである。また、認識結果において品詞が「感動詞」と認定されたものをフィラーとしている。

5　評価実験

5.1　書き起こしデータに対する実験

本節では、形態素解析や係り受け解析、R タグの情報が削除箇所の推定に及ぼす影響を調べる。本実験では、書き起こしを入力とし、形態素解析や係り受け解析、R タグ認定の精度がそれぞれ十分に高いと仮定した場合の、削除箇

検査	名詞	-	F	O	F	O
音	名詞	-	F	O	F	O
に	助詞	格助詞	O	O	F	O
対する	動詞	-	O	O	O	O
反応	名詞	-	F	O	F	O
検査	名詞	-	O	B	O	B
音	名詞	-	O	B	O	B
へ	助詞	格助詞	O	O	O	B
の	助詞	格助詞	O	O	O	O
反応	名詞	-	O	B	O	B
と	助詞	格助詞	O	O	O	O

右から 3 列目と 4 列目は全く同じ形態素列の場合に繰り返しと認定
右から 1 列目と 2 列目は自立語が同じでかつ付属語の品詞が同じ形態素列の場合に繰り返しと認定

図 5: 表現の重複（繰り返し）の例

所の推定精度の評価を行った。CSJ 公開版の音声認識テストセットのうちコアに含まれる 19 講演を評価に用い、これを除くコアのデータ（169 講演）を学習セットとした。形態素や係り受け、R タグについては、人手で付与あるいは修正したものを用いた。

削除箇所の推定結果を表 3 に示す。繰り返し素性を加えることにより、再現率に向上が見られた。また、文節や係り受け情報などの言語的情報を合わせて利用することで、より再現率が向上した。更に R タグを利用した場合と利用しない場合の性能の差は F 値で 0.03 程度であるが、人手で付与されている R タグの自動推定の困難さを考慮すると、比較的頑健に自動推定できる素性のみでもよい性能が期待できるといえる。

5.2 各 R タグ箇所に限定した場合

本節では、削除箇所タイプと削除箇所推定精度との関係を示す。本実験では、前節で用いた学習セット 169 講演およびテストセット 19 講演の書き起こしデータから、各 R タグについて、削除箇所のうち、R タグの付与された箇所と重なる部分のみを残し、同様に YamCha を用いて削除箇所の自動推定実験を行った。また、R タグの付与された範囲と重ならない削除箇所についても、その削除箇所のみを残して同様に実験を行った。ただし、R1 タグに該当する箇所を対象にした実験では、品詞情報による影響が顕著であるため、品詞情報および活用の種類・活用形を素性に用いなかった。これは、被言い直し部（RPD）が語断片など特殊な品詞情報が付与されているためである。

削除箇所の推定結果を表 4 に示す。3 章で示したタイプ 1,2 については、繰り返し等の素性による効果が見られたほか、R タグなど他の素性を加えても性能の向上が見られた。特に R2 については、繰り返しの素性の効果が大きく見られた。一方、R5 については、繰り返しの素性による効果は見られず、また R タグの利用に関わらず性能が非常に悪い。これは、他のタイプの R タグでは被言い直し部（RPD）が削除されることが多く、言い直し部（RP）が削除される場合がほとんどないのに対し、R5 タグに該当する箇所では、RPD が削除箇所となる場合だけでなく、RP が削除箇所となる場合も多く、削除すべきかどうかの判断には意味的な包含関係を考慮する必要があることが原因である。

5.3 音声認識結果に対する実験

本節では、実際の音声認識結果に対する本手法の性能を調べる。本実験では、CSJ 公開版の音声認識テストセット（30 講演）の音声認識結果に対し、6 分割の交差検定を行うことにより評価を行った。音声認識精度は 76.5%であった。なお、音声認識結果における削除箇所は、書き起こしとのマッチングにより削除箇所に相当する部分を自動推定して付与した。削除箇所の前後に認識誤りが含まれるものについては、認識誤りの範囲を考慮した上で削除箇所の範囲を広げた。

削除箇所の推定結果を表 5 に示す。書き起こしに対する実験と同様に、繰り返し素性や R タグ情報の導入により再現率の向上が見られた。このうち繰り返し素性については、表層表現の単語情報に基づいて定められていることから、音声認識結果に対しても頑健で効果的な素性であるといえる。

音声認識結果に対する実験でも、単語情報から下記 1 の定形表現を、また繰り返し素性の導入により 2,3 の繰り返し表現を推定することができたが、2 の短い繰り返し表現は多数を精度よく推定できたものの、3 のような長い繰り返し表現では、表現中に音声認識誤りが現われることが多く、その場合には繰り返し素性がうまく機能しないため、推定が難しかった。

1. 四メートル以上の距離を取る という風に
2. 同じ 場で (F えー) 場を共有して
3. 親しさの違い (F あ) 親しさの違い

5.4 自動推定結果に基づく削除箇所タイプの絞り込み

本節では、前節の推定結果における正誤を踏まえた上でのより効果的な方法を調べる。音声認識結果に付与された削除箇所の情報について、以下に示す基準で人手による削除箇所タイプの絞り込みを行った。

- 削除箇所に該当しているが、認識誤りが多く発生しているため削除すべき本来の内容が認められない場合、その削除箇所全体を外す。
- 言い直し部に認識誤りが見られるために本来の繰り返しが認められない表現について、被言い直し部が削除箇所に該当している場合は、それを外す。
- 本来は削除箇所に該当していないが実験で削除箇所と誤判定された部分について、周辺の文脈を参照した上で削除しても構わないと判断した場合は、削除箇所とみなす。

以上の絞り込みを行った結果、3 章で示した削除箇所のうちタイプ 4 の定形表現が多く残った。ただ、文頭の接続詞「で」についてはフィラーとの混同も見られたため、その場合には削除箇所から外した。一方、繰り返し表現については、表現全体が短いものについては比較的残ったが、長い表現や低頻度語については認識誤りが多く見られたため、削除箇所から外した。

この絞り込みを踏まえ、前節の実験と同様に 6 分割の交差検定により改めて学習と評価を行った。絞り込み後の実験結果を表 6 に示す。絞り込みに伴い、F 値で 0.2 程度の性能の向上が見られた。また精度は 8 割近い値となり、書き起こしに対する精度が R タグを利用しない場合で 83.2%〜84.9% であることを踏まえると、絞り込みにより音声認識結果に対しても効果的に削除箇所の検出を行えることがわかった。

6 おわりに

本稿では、一次整形における削除箇所を付与したコーパスを紹介し、そのコーパスを用いて削除箇所を自動推定する手法を提案した。本手法では SVM による機械学習により削除箇所を自動推定する。素性としては、係り受けなどの言語的情報や表現の繰り返しの情報、自己修復部の情報などを用いた。その結果、書き起こしに対して F 値で 0.8 前

表 3: 書き起こし（19 講演）に対する削除箇所推定精度

素性	範囲が一致			始端が一致		
	再現率	精度	F 値	再現率	精度	F 値
単語情報のみ	1239/1874 (66.1%)	1239/1460 (84.9%)	0.743	68.4%	87.8%	0.769
+ポーズ・フィラー	1270/1874 (67.8%)	1270/1495 (84.9%)	0.754	69.7%	87.4%	0.776
+ポーズ・フィラー・繰返し	1308/1874 (69.8%)	1308/1573 (83.2%)	0.759	72.1%	86.0%	0.784
+ポーズ・フィラー・繰返し・文節・係り受け	1412/1874 (75.3%)	1412/1657 (85.2%)	0.800	78.0%	88.2%	0.828
+ポーズ・フィラー・繰返し・文節・係り受け・R タグ	1495/1874 (79.8%)	1495/1726 (86.6%)	0.831	82.9%	90.0%	0.863

表 4: 各タイプごとの削除箇所推定精度

		素性	範囲が一致			始端が一致		
			再現率	精度	F 値	再現率	精度	F 値
タイプ 1	R1	表層・読み	22/ 136 (16.2%)	22/ 43 (51.2%)	0.246	16.9%	51.2%	0.257
		+ポーズ (P)・フィラー (F)・繰返し	105/ 136 (77.2%)	105/ 135 (77.8%)	0.775	79.4%	80.0%	0.797
		+P・F・繰返し・文節・係受け・R タグ	133/ 136 (97.8%)	133/ 140 (96.4%)	0.964	99.3%	96.4%	0.978
タイプ 2	R2	単語情報のみ	15/ 82 (18.3%)	15/ 46 (32.6%)	0.234	25.6%	45.7%	0.328
		+ポーズ・フィラー・繰返し	47/ 82 (57.3%)	47/ 69 (68.1%)	0.623	65.9%	78.3%	0.715
		+P・F・繰返し・文節・係受け・R タグ	73/ 82 (89.0%)	73/ 81 (90.1%)	0.896	92.7%	93.8%	0.933
	R3	単語情報のみ	35/ 181 (19.3%)	35/ 81 (43.2%)	0.267	22.1%	49.4%	0.305
		+ポーズ・フィラー・繰返し	74/ 181 (40.9%)	74/ 125 (59.2%)	0.484	49.2%	71.2%	0.582
		+P・F・繰返し・文節・係受け・R タグ	138/ 181 (76.2%)	138/ 166 (83.1%)	0.795	86.2%	94.0%	0.899
	R4	単語情報のみ	25/ 97 (25.8%)	25/ 37 (67.6%)	0.373	26.8%	70.3%	0.388
		+ポーズ・フィラー・繰返し	39/ 97 (40.2%)	39/ 59 (66.1%)	0.500	47.4%	78.0%	0.590
		+P・F・繰返し・文節・係受け・R タグ	63/ 97 (64.9%)	63/ 78 (80.8%)	0.720	70.1%	87.2%	0.777
タイプ 3	R5	単語情報のみ	22/ 98 (22.4%)	22/ 40 (55.0%)	0.319	26.5%	65.0%	0.377
		+ポーズ・フィラー・繰返し	24/ 98 (24.5%)	24/ 55 (43.6%)	0.314	32.7%	58.2%	0.418
		+P・F・繰返し・文節・係受け・R タグ	33/ 98 (33.7%)	33/ 66 (50.0%)	0.402	42.9%	63.6%	0.512
タイプ 4	R なし	単語情報のみ	1036/1320 (78.5%)	1036/1167 (88.8%)	0.833	80.2%	90.7%	0.851
		+ポーズ・フィラー・繰返し	1045/1320 (79.2%)	1045/1227 (85.2%)	0.821	80.5%	86.6%	0.835
		+P・F・繰返し・文節・係受け・R タグ	1076/1320 (81.5%)	1076/1209 (89.0%)	0.851	83.0%	90.7%	0.867

表 5: 音声認識結果（30 講演）に対する削除箇所推定精度

素性	範囲が一致			始端が一致		
	再現率	精度	F 値	再現率	精度	F 値
単語情報のみ	589/2443 (24.1%)	589/ 904 (65.2%)	0.352	25.7%	69.6%	0.376
+ポーズ・フィラー	614/2443 (25.1%)	614/ 955 (64.3%)	0.361	26.9%	68.8%	0.387
+ポーズ・フィラー・繰返し	671/2443 (27.5%)	671/1106 (60.7%)	0.378	29.6%	65.3%	0.407
+ポーズ・フィラー・繰返し・R タグ	983/2443 (40.2%)	983/1551 (63.4%)	0.492	45.9%	72.3%	0.561

表 6: 絞り込み後の認識結果に対する削除箇所推定精度（音声認識結果（30 講演）を対象）

素性	範囲が一致			始端が一致		
	再現率	精度	F 値	再現率	精度	F 値
単語情報のみ	852/1937 (44.0%)	852/1070 (79.6%)	0.567	45.1%	81.7%	0.581
+ポーズ・フィラー	868/1937 (44.8%)	868/1101 (78.8%)	0.571	46.2%	81.2%	0.589
+ポーズ・フィラー・繰返し	927/1937 (47.9%)	927/1205 (76.9%)	0.590	49.4%	79.3%	0.609

後の性能が得られた。一方、音声認識結果に対しても、音声認識誤りを考慮した上で削除箇所の絞り込みを施すことにより、8 割近い精度が得られた。また、書き起こしおよび音声認識結果の両方に対して、表層表現の単語情報に基づいた頑健な素性である、表現の重複の有無（繰り返し）の素性が有効に働くことを確認した。

今後の課題として、音声認識時における信頼度などの情報の利用が挙げられる。単純に素性として利用した場合にその効果はほとんど見られなかったことから、他の情報と複合的に利用する必要があると考えられる。またこれ以外にも、音声認識結果に対する性能の向上のための新たな素性の追加などの検討を行う予定である。

謝辞

本研究の一部は、総務省 SCOPE の支援により行われた。

参考文献

[1] 河原達也. 筆記録作成のための話し言葉処理技術. 電子情報通信学会技術研究報告, SP2006-120, NLC2006-64 (SLP-64-36), pp. 209–214, 2006.

[2] C. Nakatani and J. Hirschberg. A speech-first model for repair identification and correction. In *Proc. ACL*, 1993.

[3] 下岡和也, 河原達也, 内元清貴, 井佐原均. 『日本語話し言葉コーパス』における自己修復部（D タグ）の自動検出および修正に関する検討. 情報処理学会研究報告, 2005-NL-167-14, 2005-SLP-56-14, pp. 95–100, 2005.

[4] 丸山岳彦, 佐野真一郎. 自発的な話し言葉に現れる言い直し表現の機能的分析. 言語処理学会第 13 回年次大会, pp. 1026–1029, 2007.

[5] 浅原正幸, 松本裕治. 形態素解析とチャンキングの組み合わせによるフィラー/言い直し検出. 言語処理学会第 9 回年次大会, pp. 651–654, 2003.

[6] 下岡和也, 河原達也, 奥乃博. 講演の書き起こしに対する統計的手法を用いた文体の整形. 情報処理学会研究報告, 自然言語処理 (149-12), 音声言語情報処理 (41-3), pp. 81–88, 2002.

[7] 古井貞煕, 前川喜久雄, 井佐原均. 科学技術振興調整費開放的融合研究制度：大規模コーパスに基づく『話し言葉工学』の構築. 日本音響学会誌, Vol. 56, No. 11, pp. 752–755, 2000.

[8] T. Kudo and Y. Matsumoto. Chunking with Support Vector Machines. In *Proc. NAACL*, 2001.