不要個所削除による講演音声の要約

幅田隆†　　奥村学‡†
†北陸先端科学技術大学院大学 情報科学研究科　‡東京工業大学 精密工学研究所

1　はじめに

字幕放送などの様に，音声を文字化して提示する事は聴覚障害者支援の観点から重要視されている．その際，音声をそのまま文字化するのではなく，適切な長さへと要約していく必要がある．しかしこの場合，要約の結果は音声の代りとして情報を伝えるものであるため，要約による情報の欠落は極力避ける必要がある．これに対して，重要文抽出法などの要約手法は，文単位の要約手法であるため情報が多く欠落する可能性があり不適切であると指摘されている [1][2].

その一方で，講演音声の様な話し言葉の場合，一文中に多くの冗長表現が含まれている．この冗長表現は情報伝達という観点から考えると明らかに不要個所であると考えられる．したがって，この様な表現を不要個所として削除する事により，情報を欠落させずに要約が行えるものと考えられる．

この様な話し言葉を対象とした人手による要約事例として要約筆記がある．要約筆記とは，講演音声などの話し言葉を聞き，その内容を要約しながら手書き文字やキーボード入力などによって伝える活動の事である．この活動は，聴覚障害者支援の観点から非常に重要な活動であり，自動要約筆記システムの開発は有用であると考えられる．

そこで，実際の講演音声の要約筆記データをモデルとして調査し，自動要約筆記システム開発の第一歩として文短縮型の要約システムを開発した．本稿では，その調査結果と要約システムについて述べる．

2　話し言葉の冗長表現

話し言葉の特徴についてはすでに研究されており，多くの特徴的な表現が報告されている [3]. 今回はその特徴的な表現の中から冗長表現として考えられる以下の表現について注目した．各表現の事例を図1に示す．

で、えー字幕付きの テレビ放送と 申しましてもいろいろな テレビ放送の ジャンルがございます。
で、現状で申し上げますと、ここに書いてありますように、報道番組への字幕付与の希望が多いというレポートが出されております。

図 1: 話し言葉の冗長表現の例

- 間投詞 (例中の Box)
- 言い直し・繰り返し表現 (例中の下線部)
- 挿入句表現 (例中の二重下線部)

各表現について，実際の要約筆記ではどの様に処理されているのか，そしてその処理をシステム化する際にはどの様な情報や条件を用いれば良いのかについて調査を行った．

3　要約筆記の調査

今回調査を行ったデータは，通信・放送機構 (TAO) によって行われたワークショップにおける1発表分の書き起しデータとキーボード入力による要約筆記データであり，表1に示す様なデータである．

表 1: 調査データ

	文数	文字数
書き起し	152 文	8032 文字
要約筆記	143 文	3857 文字
要約率	3857/8032=0.480	

このデータを用いて，前説で述べた各表現の調査を行った．

3.1　間投詞

間投詞が削除されている事例は 159 事例存在した．これらは形態素解析の結果「フィラー」または「感動詞」から削除処理が可能であると考えられる．ただし，「まあ」「あの」「その」「あのう」などが正しく解析されない事がある．

「あの」「その」は連体詞として残される場合と間投詞として削除される場合が混在しているが,「まあ」と「あのう」はすべて間投詞として削除されいるため,形態素解析の結果とは関係なく削除処理が可能であると考えられる.

3.2 言い直し・繰り返し表現

言い直し表現は直前の発話を訂正する目的で類似した発話が再び出てくる表現であり,繰り返し表現は訂正の目的とは関係なく類似した発話が再び現れる表現である.これらは近くに類似した発話が存在しているため,形態素単位や文節単位の類似度などをもとに削除処理が行えると考えられる.

しかし,類似形態素を削除してしまうと格助詞など文の構成上重要な形態素が頻繁に削除されてしまうため,ここでは文節単位の類似度をもとに処理を行うことを考える.

また,文節単位で削除する場合,述語となる文節,必須格になる文節,被修飾文節などは削除せずに残されるべきと考えられる.しかし,言い直し表現の場合と繰り返し表現の場合とでこれらの処理は異なってくる.

言い直し表現 この表現が文節単位で削除されている事例は全部で6事例存在した.その特徴は以下の通りである.

- 類似発話が必ず1文中に存在している
- 類似発話間に述語は存在しない
- 類似発話の内先に出現した発話が削除されている
- 述語となる発話でも削除されている
- 被修飾発話でも削除されている

繰り返し表現 この表現が文節単位で削除されている事例は全部で17事例存在した.その特徴は以下の通りである.

- 類似発話が数文はなれて存在する事もある
- 一文内に類似発話が存在する場合はその間に述語存在する
- 類似発話の内後に出現した発話が削除されている
- 述語となる発話は基本的に削除されない
- 被修飾発話は基本的に削除されない

3.3 挿入句表現

挿入句表現が削除されている事例は13事例存在した.これらを処理する有効な手段は発見できていないが,唯一「～ように,」という句末表現に関してはすべて削除されていた.したがって,現在のところはこの句末表現のみを対象とする.

3.4 「～という～」表現

実際の講演音声書き起しテキストと要約筆記テキストを調査していく中で「～という～」という表現も不要個所として削除されていた.その例を図2に示す.例中の下線部が実際に削除されている個所を示している。

えー別の手段としまして、文節単位での圧縮<u>ということ</u>を考えております。

図2:「～という～」表現の例

この表現は97個所存在している.この中で文・節単位で要約処理がされている39事例を調査の対象外とし,残りの58について調査を行った.その結果,「という」前後の品詞関係から以下のような削除処理が適用できる.

「という」が単独で削除 この削除処理が適用されるの前後の品詞関係が以下の場合である.

- 動詞 という 名詞
- 形容詞 という 名詞
- 助動詞 という 名詞 の一部
  (助動詞が「ない」の場合のみ)

「という」と後の形態素が削除 この削除処理が適用されるの前後の品詞関係が以下の場合である.

- 名詞 という 名詞
- 助詞 という 名詞

例外処理 上記5種類の前後品詞パターンであっても,「～というふうに」という表現の場合,「と」を残して「いうふうに」を削除する例外処理が適用される.

4 要約システム

前説までに述べてきた各表現を処理するモジュールをそれぞれ個別に実装し,各モジュールを組み合わせることで要約システムを構築している.また,各モジュールは処理を行うために形態素情報必要としているため,必要に応じて形態素解析システム (茶筅) も用いている.

各モジュールの組み合わせ方は以下の通りである.

1 間投詞削除モジュール
2 「という」表現削除モジュール
3 丁寧表現の言い替えモジュール
4 挿入句削除モジュール
5 言い直し・繰り返し削除モジュール

間投詞削除モジュール 形態素情報から品詞が「フィラー」「感動詞」である形態素, および「まあ」「あのう」という文字列を削除する.

「という」表現削除モジュール 形態素情報から文字列「という」の前後品詞関係を抽出し, 各品詞関係に当てはまる削除処理を行う.

丁寧表現の言い替えモジュール 参考文献[5][6]を参考にし, 助動詞「です」「ます」および「ござる」「申す」など特定の謙譲語を言い替えるルールを自前で作成した.

挿入句削除モジュール 句は読点「,」にて区切られているものとし, 形態素情報から名詞「よう」助詞「に」記号「,」で終わる句を削除する.

言い直し・繰り返し削除モジュール 文節および隣接2文節列を処理単位として, 前後2文内に含まれる各文節 (列) との類似度をもとに削除をする. 文節 (列) 間類似度の計算は各文節 (列) 構成形態素列のマッチングから式 (1) の様に計算する.

$$\text{文節 (列)}AB\text{ 間類似度} = \frac{\text{文節 (列)}AB\text{ 間類似スコア}}{max\{\text{ 文節 (列)}AB\text{ 構成スコア }\}} \quad (1)$$

$$\text{文節 (列) 間類似スコア} = 2 \times \text{自立語の一致数} + 1 \times \text{非自立語の一致数} \quad (2)$$

$$\text{文節 (列) 構成スコア} = 2 \times \text{文節 (列) 構成自立語の数} + 1 \times \text{文節 (列) 構成非自立語の数} \quad (3)$$

形態素の一致によって加算される類似スコアを, 比較文節 (列) の構成スコアによって正規化している. したがって, 文節 (列) 間類似度は最低値が0, 最高値が1となる. この類似度計算の結果0.5以上となる文節 (列) 間を類似文節 (列) とし, さらに以下の条件にあてはまる文節 (列) を削除する.

- 類似文節 (列) が同一文に存在して以下を満たす場合は言い直しとして文頭側の文節 (列) を削除
  - 類似文節 (列) の間に動詞が存在しない
- 類似文節 (列) が同一文に存在して以下を満たす場合は繰り返しとして文末側の文節を削除
  - 類似文節 (列) 間に動詞が存在する
  - 各文節 (列) が動詞を含まない
  - 文末側の文節 (列) が「ガヲニ格」ではない
  - 文末側の文節 (列) の一つ前の文節が助詞「の」によって連体化されていない
- 類似文節 (列) が同一文に存在せずに隣接文 (前後2文まで) に存在して以下を満たす場合は繰り返しとして文章末側の文節を削除
  - 類似文節 (列) が動詞を含まない
  - 文章末側の文節 (列) が「ガヲニ格」ではない
  - 文章末側の文節 (列) の一つ前の文節が助詞「の」によって連体化されていない

5 システム評価

5.1 削除率

本システムに講演音声書き起しテキスト (原文) を入力し, システムが要約した結果 (出力) の削除率を式 (4) にしたがって計算した. その結果を表2に示す. また, 各モジュール別の削除率を表3に示す. なお入力した講演データは調査には用いていないデータである.

$$\text{削除率} = \frac{\text{原文文字数} - \text{出力文字数}}{\text{原文文字数}} \times 100 \quad (4)$$

表 2: システム出力結果の削除率

講演番号	3-1	3-3	3-4
原文文字数	8357	7372	5857
要約結果文字数	6912	6272	4573
削除率	17.3%	14.9%	21.9%

表 3: 各モジュールによる個別削除率

講演番号	3-1	3-3	3-4
間投詞	1.9%	1.3%	1.9%
という	3.3%	3.7%	4.6%
丁寧語	3.7%	3.1%	5.3%
挿入句	1.1%	1.6%	0.2%
繰り返し	7.4%	5.3%	10.0%
合計	17.3%	14.9%	21.9%

5.2 精度と再現率

前節で入力とした講演データの要約筆記テキストを正解データとし, システムの出力の精度と再現率による評価を行った. 精度および再現率は以下の様に計算する.

$$\text{精度 } (Precision) = \frac{\text{システムと正解データの一致数}}{\text{システム削除個所数}} \times 100 \quad (5)$$

$$\text{再現率 } (Recall) = \frac{\text{システムと正解データの一致数}}{\text{正解データ削除個所除数}} \times 100 \quad (6)$$

ただし, システムの出力と正解データとが部分的に一致する場合がある. この様な事例は,「削除する個所は一致しているが, 削除の手法が異なる事例」と考えられるため, 完全不一致とは区別して考える. その結果を表 4 に示す.

表 4: 精度と再現率

	精度 (Precision)		再現率 (Recall)	
完全一致	$\frac{130}{296}$	43.9%	$\frac{130}{405}$	32.1%
部分一致	$\frac{170}{296}$	57.4%	$\frac{170}{405}$	42.0%

また, 要約筆記はその性質上, 要約が必要でない場合 (キーボード入力が十分間に合う場合など) は要約をせずそのまま音声を書き起す事がある. そのため, システムの出力が正解データとまったく一致していない場合 (完全不一致) でも, システムの出力が明らかに正しい場合がある,

この様な場合も正解と見なした場合, 精度 (precision) は 70.6% となっている. ただし, この場合の再現率 (recall) に関しては, 正解データが変化しているため, 再現率定義式 (6) の分母が定まらず測定不能である.

6 考察

本稿にて提案してきた各削除手法により 10%～15% 程度の削除率に, また既存の言い替え手法と会わせることにより 15%～20%程度の削除率となっている. しかし, 挿入句表現削除のための有効な手法が発見できてなく, 削除率が低くなっている. 実際の講演データの中にはより多くの挿入句表現が出現しており, これらを削除する有効な手法を考案する必要がある. また, 再現率が低いことを考えると, まだ他にも削除可能な表現が残っていると考えられ, そういった表現の再調査を行う必要があると考えられる.

精度については, 不一致内正解まで含めたものが本システムのもっともらしい評価と考えているが, その値が 70%程度となっている. 明らかに不正解である 30% は, 主に言い直し・繰り返しの削除モジュールによるものが多く, 文節間類似度の計算法や削除可能条件の再検討などが必要と考えられる.

7 おわりに

本稿では, 実際の要約筆記データの調査結果と, その調査結果を元にして開発した文短縮型の要約システムについて述べた. 評価の結果, 挿入句削除モジュールと言い直し・繰り返し削除モジュールについては改善の必要があると考えられる.

また, 自動要約筆記システムを考えた場合, システムへの入力は書き起しテキストではなく音声認識システムによる認識結果となる. したがって, 音声認識結果を扱う際の問題点も今後検討する必要があると考えられる.

謝辞

本研究を進めるにあたり, 講演音声データ, 書き起しデータ, 要約筆記データを提供して頂きました, 通信・放送機構 (TAO) に感謝致します.

参考文献

[1] 三上真, 石ざこ友子, 赤松裕隆, 増山繁, 中川聖一. ニュース音声の認識結果を用いた要約による字幕生成. 情報処理学会第 58 回全国大会論文集,1999.

[2] 白井克彦, 江原暉将, 沢村英治, 福島孝博, 丸山一郎, 門馬隆雄. 視聴覚障害者向け放送ソフト制作技術研究開発プロジェクトの研究状況.Proceedings of TAO WORKSHOP ON TV CLOSED CAPTIONS FOR THE HEARING IMPAIRED PEOPLE,1999.

[3] 竹沢寿幸, 田代敏久, 森元逞. 音声言語データベースを用いた自然発話の言語現象の調査. 人工知能学会研究会資料, SIG-SLUD-9403-3,pp.13-20,1994

[4] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 浅原正幸. 日本語形態素解析システム『茶筅』version2.0 使用説明書 第二版.

[5] 若尾孝博, 江原暉将, 白井克彦. テレビニュース番組の字幕に見られる要約の手法 . 情報処理学会自然言語処理研究会報告, pp.83-89,122-13,1997.

[6] 山崎邦子, 三上真, 増山繁, 中川聖一. 聴覚障害者用字幕生成のための言い替えによるニュース文要約 . 言語処理学会第 4 回年次大会発表論文集, pp.646-649, 1998.