修 士 論 文

不要個所の削除と言い替えによる講演音声の要約

指導教官　奥村学 助教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

幅田 隆

2001 年 2 月 15 日

要旨

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。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。そこで、本研究では、人手によって講演音声の要約を行っている要約筆記の調査を行い、その調査結果を元にした文短縮型の要約システムの開発を目標としている。

調査の結果、間投詞、言い直し・繰り返し表現、挿入句表現、丁寧表現、「〜という〜」表現が削除または言い替えの対象となっている。さらに各表現の削除または言い替え処理が適用される条件についても調査を行い、その結果から各表現を個別に処理するモジュールをそれぞれ作成した。本要約システムは、各モジュールを組み合わせることによって構築しており、特定の表現だけを処理させることも可能である。

本要約システムの評価として削除率、および要約筆記を正解データとした場合の精度 (precision)、再現率 (recall) の測定を行った。その結果、削除率は 15% 〜 20% 程度、精度は 80% 程度、再現率は 50% 程度となった。精度の改善としては言い直し・繰り返し表現削除モジュールの改良、再現率の改善としては挿入句表現削除モジュールの改良と今回注目した表現以外の不要個所の調査の必要があると考えられる。

目 次

図 目 次

表 目 次

第 1 章

はじめに

1.1 研究の背景

要約筆記や字幕放送の様に、音声を文字化して提示する作業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、情報を欠落させずに適切な長さへと要約していく必要がある。この様な要約の場合、重要文抽出法などの様な文単位の要約手法では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。

例えば [白井 99] では、ニュース番組の電子化原稿を対象とし、重要文抽出法と形態素単位での文字数圧縮法を用いた要約手法を提案している。しかし、重要文抽出法では字幕作成には粗すぎるため、他の文字数圧縮法が必要と指摘している。

また [三上 99] では、ニュース音声の音声認識結果を対象とした要約について検討している。１文が長く１記事内の文数が少ないというニュース音声の特徴より、重要文抽出法では情報が大きく欠落すると指摘している。

これらニュース原稿・音声は比較的書き言葉に近いと考えられる。したがって、より話し言葉に近い音声やその書き起こしテキストに対して同様の要約を行なう場合、また違った問題が生じると考えられる。しかし、この様な話し言葉を対象とした要約手法に関する研究はあまりなされていない。

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。

話し言葉を対象として一文を短くまとめる文短縮形要約の実例として、要約筆記というものがある。これは、講演音声などの話し言葉を聞きながら、リアルタイムでその内容を要約し、その内容を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を開発する事もまた非常に有用だと考えられる。また、要約筆記は人間による話し言葉の文短縮形要約の良いモデルになると考えられる。

1.2 研究の目的

本研究は、自動要約筆記システム開発の第一歩として、より話し言葉に近いと考えられる講演音声の書き起こしテキストを対象とし、重要文抽出法ではなく、文中の不要個所を削除または言い替えすることによって要約を行なう文短縮形要約システムの開発を目的としている。

文中の不要個所として、本研究では話し言葉の冗長表現に注目している。話し言葉の特徴としては [?] などの先行研究が多く存在しており、その中から冗長表現と考えられ、不要個所として削除または言い替えが可能であると考えられる表現の検討を行う。

さらに、システムを実装するにあたり、実際の要約筆記データをモデルとして調査を行う。本研究で注目した表現が、実際の要約筆記ではどの様に処理されているか、そしてその処理をシステム化するにはどのような情報や条件が必要になるかについて調査を行う。

1.3 本論文の構成

本論文では、2 章において話し言葉要約について述べ、本研究の位置づけについて述べる。3 章では、既存の研究にて報告されている話し言葉の特徴の中から冗長表現と考えられる表現についての検討を行う。4 章では、本研究の要約システムにおいてモデルとしている要約筆記についての一般的な説明を行う。5 章では 3 章で検討した話し言葉の冗長表現について、実際の要約筆記データを用いて調査を行った結果について述べる。6 章では、調査結果を元にして構築した要約システムについて述べる。7 章では、要約システムに対して行った評価の結果と考察を述べる。8 章では本研究のまとめと今後の課題について述べる。

第 2 章

話し言葉要約

従来の自動要約研究は主に書き言葉を対象にしたものである。書き言葉と話し言葉ではそれぞれ異なった特徴を持っているため、本研究で目標としている話し言葉を対象とした要約システムの開発を考えた際に、従来の書き言葉要約とは異なる点がいくつか存在する。本章では、この様な書き言葉要約と話し言葉要約との違いについて述べ、本研究との比較を行う。

2.1 書き言葉要約と話し言葉要約の違い

従来の自動要約研究は、「大量のテキストに満ち溢れた情報洪水の中から、重要な情報を抽出して提示することにより、読み手の負担を軽減する事」が主な目的であると考えられる。この「大量のテキスト」というものが主に新聞記事などの書き言葉によるものであるため書き言葉を対象とした要約手法が主に研究されてきたと考えられる。また要約は、要約結果の利用目的によって、原文を参照する前の段階で用いる indicative な要約と、原文の代りとして用いる informative な要約に分ることができる [Hand97]。しかし、従来の研究では必ずしもこの分類を十分に考慮したものとは言えない [奥村 99]。

これに対して話し言葉要約の多くは、「音声によって伝えられた情報を、音声の代りに要約したテキストで伝える事」が目的である。例えば、[白井 99] などでは、ニュース番組の字幕作成に関連した要約手法を提案してる。字幕は聴覚障害者支援のために音声の代りに文字で情報を伝達するものでるが、読み手の読みやすさなどを考慮して 70% 程度の要約が必要であるとしている。この様な要約は明らかに informative な要約であると考えら

れる。この様な場合、重要文抽出法の様な文単位の要約手法では情報の欠落が多く、文短縮形の要約手法が必要となってくる。

2.2 話し言葉要約の関連研究

話し言葉要約の関連研究として、テレビ放送における字幕生成における要約がある。

[白井 99] では、ニュース番組の電子化原稿を対象とし、文字数にして 70% の要約を目標とした研究が行なわれている。要約手法としては、重要文抽出法と文末表現の言い替えなど形態素単位での文字数圧縮法を用いている。前者で約 80%の要約率を、後者で約 90%の要約率を、合わせて目標としている 70%の要約をほぼ達成している。しかし、重要文抽出法は文単位の要約であるため情報の欠落が多く、字幕作成には粗すぎると指摘している。しかし、文字数圧縮法だけでは 85% 程度の要約率が限界であり、他の文字数圧縮法の検討が必要としている。

[三上 99] では、ニュース音声の音声認識結果を対象とした要約について検討している。１文が長く１記事内の文数が少ないというニュース音声の特徴より、重要文抽出法では情報が大きく欠落すると指摘している。したがって要約手法としては、構文解析の結果をもとに文節単位で不要個所を削除する手法を用いている。しかし、音声認識の認識誤りが原因で構文解析に失敗し、不自然な要約結果や情報の欠落などが起こる可能性があると指摘している。

2.3 本研究の特徴

本研究は、講演音声の書き起しテキストを対象とした要約手法に関する研究を行う。関連研究において対象としていたのはニュース原稿・音声であったが、これらは比較的書き言葉に近いと考えられる。これに対して講演音声はより話し言葉に近いと考えられる。

また、本研究では、既存の研究で報告されている話し言葉の特徴の中から冗長表現と考えられる表現に注目し、その表現を不要個所として削除する手法を提案している。[白井 99] ではニュース原稿や字幕独特な表現を利用した要約手法であるのに対し、本研究はより一般的な話し言葉の特徴を利用した要約手法である。

さらに、本研究では、文字列や形態素情報などの比較的表層的な情報を使って要約する手法を検討している。これは、より話し言葉に近い講演音声を対象としているため、構文

解析などの処理がうまくいかないと予想されるためである。したがって、[三上 99] で提案されている係り受け情報などを用いた要約手法とは異なる。

第 3 章

話し言葉の冗長表現

話し言葉の特徴は [竹沢 94] などにおいてすでに多くの研究がなされている。本研究では、すでに明らかとなっている話し言葉の特徴の中から、冗長表現と考えられるものに注目した。本章では、[竹沢 94] にて調査されている話し言葉の特徴の内、本研究で注目した表現について説明する。

3.1 間投詞

間投詞は、内容計算や評価などの言い淀み系と、応答系や驚きなどの入出力制御系に分類されている。これらすべてが明らかに冗長表現であり削除可能であると考えられる。間投詞の例を図 3.1に示す。下線部が間投詞であり削除しても問題はない。

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。

図 3.1: 間投詞の例

3.2 言い直し・繰り返し表現

言い直し表現は言い間違えを訂正する表現、繰り返し表現は訂正などの目的ではなく同様の表現が繰り返し使用される表現である。訂正された個所と言い直した個所、または繰

り返された個所と繰り返した個所はそれぞれ同一の内容を表す個所であり、どちらかが冗長表現であると考えられる。例えば図 3.2の言い直しの例では、二重下線部の個所を下線部の個所が言い直しており、二重下線部の「今後の課題として」を削除しても問題ないと考えられる。同様に、図 3.3の繰り返しの例では、下線部の個所と同様のな表現が二重下線部の個所で繰り返しており、二重下線部の「テレビ放送の」が削除されても問題ないと考えられる。

あの音声を文字にするという部分ですが、その部分がやはりあの自動化されるということは非常にこの字幕制作の効率化に与えるインパクトは大きいわけで、えーその部分についても 今後の課題として 大きな課題として 考えていきたいというふうに考えています。

図 3.2: 言い直しの例

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。

図 3.3: 繰り返しの例

3.3 挿入句表現

挿入句表現は、文の途中にあって、その文の内容とは関係なく別の次元から挿入された表現である。文の内容とは関係ないため冗長表現と考えられ削除が可能であると考えられる。挿入句の例を図 3.4に示す。下線部が挿入句であり削除しても問題はない。

3.4 丁寧表現

丁寧表現は主に述語部に現れ、普通体の表現と比べると冗長表現と考えられる。丁寧表現を普通体に言い替えても文の内容に変化は無いため、丁寧表現を言い替える事が可能で

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 3.4: 挿入句の例

で、えー字幕付きのテレビ放送と 申しましても→ 言っても いろいろなテレビ放送のジャンルが ございます→ ある。

図 3.5: 丁寧表現の例

あると考えられる。例えば、図 3.5 の下線部を → ボックス の様に言い替えても問題はないと考えられる。

本研究では、以上の表現に注目した。これら各表現が人手による話し言葉要約例である要約筆記においてどの様に処理されているのか、また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う。

第 4 章

要約筆記

本研究では、要約筆記をシステムのモデルとして調査する。本章では、まず要約筆記について一般的な説明をする。

4.1 要約筆記

要約筆記は、手話通訳やテレビ字幕放送など「情報保障」活動の一種であり、聴覚障害者支援の観点から非常に重要な活動である。「情報保障」の手段としては手話通訳と文字伝達の 2 種類があり、文字伝達はさらに、速記型文字伝達と要約型文字伝達に分類される。要約筆記は要約型文字伝達に分類される「情報保障」活動である。

一般に要約筆記と言った場合、OHP などへの手書きによって行われる要約型文字伝達の事を意味する。この手書きの部分をパソコンのキーボードによる入力に置き換えたものが、パソコン要約筆記と呼ばれている。本論文においては、特に断りのない限りパソコン要約筆記の事をさして要約筆記と表記することにする。

専用機器を用いている速記型文字伝達とは異なり、要約筆記では音声をすべて文字化することは不可能である。一般的に、ニュースのアナウンサーが話す速度は 350 字/分 〜 400 字/分、ゆっくりとした講演でも 250 字/分 〜350 字/分、早口の漫才などになると 700 字/分 程度となる。これに対して、タッチタイプに習熟した人が一般のワープロやパソコンを活用して入力した場合、入力速度が 100 字/分 〜 200 字/分 程度となっている。したがって、要約筆記の場合は、音声を聞きながらリアルタイムで要約を行い、その要約結果を入力する事になる。

4.2 要約筆記の注意点

要約筆記は聴覚障害者支援の観点から非常に重要な活動であると同時に、「人間による話し言葉の要約事例」として考えることができ、本研究においても要約システムのモデルとして、またシステム評価の際の正解データとして用いている。

しかし、要約筆記はたしかに「人間による話し言葉の要約事例」であるが、要約をしようと思って要約しているのではない。人間がリアルタイムで処理をしているため、キーボード入力が間に合わず結果的に要約になっているというものである [太田 99][太田 98]。そのため以下のような点に注意する必要がある。

- 入力速度が発話速度に間に合う場合は要約せずにそのまま入力する
- 入力速度が発話速度に間に合わない場合は入力をあきらめる
  (極端な要約処理になる)
- 聞き間違えなどによる誤入力がある。

要約筆記データを調査する場合、上記の点に十分注意しながら調査を行う必要があると考えられる。ただし、要約筆記データが全くの間違えであるわけではない。要約システムのモデルにする際も、システムの評価における正解データとして使用する際も非常に参考になるデータである。

第 5 章

要約筆記における要約事例の調査

要約筆記データ　本研究において用いる要約筆記データは、1999 年に 通信・放送機構 (以下 TAO と表記する) によって行われたワークショップ「聴覚障害者のためのテレビ用字幕制作に関する国際ワークショップ」において発表された講演の要約筆記データである。このデータにはそれぞれ元音声の書き起しデータも付属していてペアコーパスとなっている。以下ではこの書き起し・要約筆記のペアコーパスの事を TAO コーパスと表記する事にする。

TAO コーパスは全部で 8 講演分あり、その内容は表 5.1の通りである。

表 5.1より、TAO コーパスには以下のような特徴があると考えられる。

- 講演の発話速度が講演者によって開きがある
  200[char/m] から 350[char/m] までと発表者によってばらつく
- 要約筆記の表示速度はどの講演でも差が少ない
  150[char/m] から 200[char/m] と安定
  キーボード入力の限界がこのあたりであるため [太田 99][太田 98]
- 講演の発話速度が速ければ高い要約率になっている (図 5.1)

講演の発話速度が速いほど高い要約率となっている。これは要約筆記というものが結果的に要約となっているためと考えられる。入力速度がほぼ一定であるならば、発話速度が速くなるほど要約率も上がるのである。したがって、発話速度が速い講演の要約筆記デー

表 5.1: TAO コーパス

講演 No.	書き起しデータ				要約筆記データ			
	文数 [sent]	文字数 [char]	時間 [sec]	速度 [char/m]	文数 [sent]	文字数 [char]	速度 [char/m]	要約率 [%]
1-3	245	7854	1662	283.5	193	4974	179.6	63.3
1-4	152	8032	1422	338.9	143	3857	162.7	48.0
2-3	84	4596	1230	224.2	85	3398	165.8	73.9
2-4	137	7101	1702	250.3	123	4325	152.5	60.9
3-1	159	8357	1940	258.5	148	5375	166.2	64.3
3-2	158	5963	1096	326.4	129	3516	192.5	59.0
3-3	82	7372	1410	313.7	78	4554	193.8	61.8
3-4	76	5857	1455	241.3	93	3947	162.8	67.4

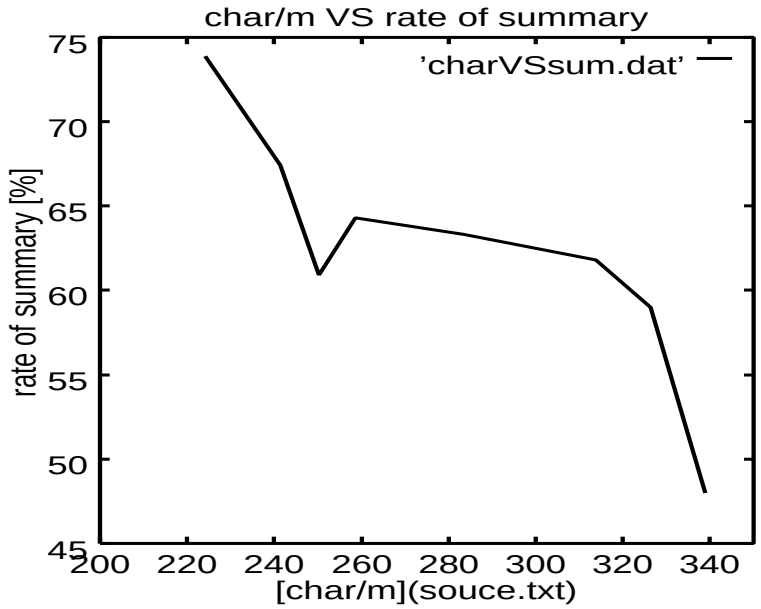

図 5.1: 発話速度と要約率

タはより多くの要約事例を含んでいると考えられ、その代わりとして、文単位や節単位での削除や言い替えなどの極端な要約事例が多くなってくると考えられる。

調査データ　今回調査対象としたのは表 5.1 における 講演番号 1-4 のデータセットである。このデータセットは、元音声の発話スピードが早く要約筆記データにおける要約率がもっとも高くなっている。そのため要約可能であるにも関わらず、入力速度が間に合うために要約しないという事が少なく、要約事例が多く観測できると考えられる。その一方で、極端な要約処理や誤入力などが行われている可能性が高いので注意が必要となる。

このデータセットを調査対象データとし、3章にて紹介した以下の各冗長表現に対する要約事例について調査を行った。

- 間投詞の削除
- 言い直し・繰り返し表現の削除
- 挿入句表現の削除
- 丁寧表現の言い替え

さらに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.

- 「〜という〜」表現の削除

これらを調査した結果を表 5.2 にまとめた。本章ではこの調査結果について述べる。

表 5.2: 要約筆記の調査結果

間投詞	形態素解析の結果から間投詞、感動詞、副詞「まあ」を削除 「あのう」「えーと」は文字列マッチングで削除
言い直し	類似文節が 1 文中に存在し、その間に同士が存在しない場合削除 削除対象となるのは先に発話された文節
繰り返し	類似文節が数文離れて存在する場合に削除候補 類似文節が同一文中に存在し、その間に同士が存在する場合も削除候補 削除候補になる文節はいづれも後から発話された文節 削除候補文節が動詞を含む、被修飾文節、必須格の場合は削除しない
挿入句表現	句末表現が「〜ように、」の場合のみ削除
丁寧表現	助動詞「です」「ます」を削除または言い替え 助動詞「です」「ます」の前後の形態素を適切な形に言い替え 謙譲語を通常の動詞に言い替え 接頭辞「お」「ご」を削除
「〜という〜」表現	「という」の後が名詞の場合は削除候補 「という」の前が動詞、形容詞、助動詞の場合「という」を削除 「という」の前が名詞、助詞の場合「という名詞」を削除 「という」の前が文頭、文頭+間投詞の場合「という文節」を削除 「というふうに」の場合は例外的に「と」に言い替え 「という」の後が特定の名詞の場合は例外的に削除しない

5.1 間投詞の削除

調査データにおいて間投詞が削除されていた事例を表 5.3 に示す。

表 5.3: 間投詞削除事例

	あ	あの	あのう	あー	え	えっと	えー	えーと	この	その	ま	まあ	合計
個数	4	32	25	1	6	1	63	1	1	2	14	9	159
文字数	4	64	75	2	6	3	126	3	2	4	14	18	321

これは書き起しデータに存在するすべての間投詞が削除されている事になる。

書き起しデータを形態素解析し、上記の間投詞が正しく解析されるか確認してみた。形態素解析には茶筅 [松本 99] を用いている。結果、「フィラー」または「感動詞」として解析されたものを正解出力とした場合、精度は、recall=68%,precision=96%という結果になった。

recall が低い値になった原因を探るため、正しく解析されなかった事例を調べてみたところ、表 5.4に示す様になった。それぞれを更に調査し削除可能か検討した。

表 5.4: 茶筅解析失敗事例

	あの	まあ	あのう	その	え	えーと	合計
解析結果	連体詞	副詞	感動詞+助詞+感動詞 連体詞+名詞 (の一部) 連体詞+形容詞	連体詞	動詞	接続詞	
個数	29	9	8	2	2	1	51

連体詞「あの」「その」　形態素解析の結果、連体詞として解析された「あの」「その」の中には、本当に連体詞であるものと、本当は間投詞であるものが混在している。したがって、形態素解析の結果からこれらを削除することは不可能であり、連体詞として解析された「あの」「その」は削除せずに残す必要がある。

動詞「え」　形態素解析の結果、動詞として解析された「え」の中には、本当に動詞であるものと、本当は間投詞であるものが混在している。したがって、形態素解析の結果からこれらを削除することは不可能であり、動詞として解析された「え」は削除せずに残す必要がある。

副詞「まあ」　書き起しデータ中に「まあ」という文字列は全部で 9 個所出現しており、茶筅の解析結果ではそのすべてが「副詞」として解析されている。一方要約筆記の際、これらすべてが間投詞として削除されている。したがって、副詞と解析される「まあ」は間投詞として削除する事が可能と考えられる。

感動詞+助詞+感動詞「あのう」他　書き起しデータ内に「あのう」という文字列は 25 個所出現している。それらすべてが要約筆記では削除されている。したがって、「あのう」という文字列は間投詞として削除することが可能と考えられる。

接続詞「えーと」　書き起しデータ内に「えーと」という文字列は 2 個所出現している。それらすべてが要約筆記では削除されいる。したがって、「えーと」という文字列は間投詞として削除することが可能と考えられる。

形態素解析において正しく解析されなかった事例のうち、間投詞として削除が可能と考えられる事例も正解出力とした場合、recall = 78.6 % となる。しかしこれでもまた十分とは言えず、出現数の多い連体詞「あの」をうまく解析する事が必要になると考えられる。

5.2　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内容の発話が存在しているため、形態素単位や文節単位での類似度をもとに削除処理が可能であると考えられる。

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削除処理を行うことを考える。

また、文節単位で削除する場合、動詞を含む文節、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、削除されない条件が異なっている。

5.2.1 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。

- 類似文節が必ず１文中に存在している
- 類似文節間に動詞は存在しない
- 類似文節のうち先に発話された文節が削除されている
- 動詞を含む文節でも削除されている (図 5.2)
- 被修飾文節でも削除されている (図 5.3)

また、削除例を図 5.2 図 5.3 に示す。なお、⇓ の上のボックスが処理前のテキスト、下のボックスが処理後のテキストを表す。以下、本論文で示す要約事例はすべて同様の意味を表す事とする。また、ここの例では、A は、A が言い直した個所である事を示し、(B) は言い直された個所として削除されていることを示す。

（いう）こういう フローで、えー字幕番組が現在作られているわけです。
⇓　「いう」が削除
こういう フローで字幕が作られているわけです。

図 5.2: 言い直し削除事例 1

言い直し表現の場合、文の構成上削除すべきでないと考えられていた、動詞を含む文節、被修飾文節も削除されている。これは、言い直し表現の場合、言い直された個所の代りとなる個所 (言い直した個所) がすぐ近くに存在するためと考えられる。また同様に考えると、実際の調査データの中には存在しなかったが、必須格となる文節もまた削除することが可能であると考えられる。

そういうこともあって、現在日本ではニュースに字幕が付いていないわけですが、そういうこともありましてあの報道番組への (期待は) あの 字幕付与の希望が 多いと、いうことが言えると思います。

⇓　「期待は」が削除

そうしたことから報道真番組への 字幕の期待が 高いわけです。

図 5.3: 言い直し削除事例 2

5.2.2 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。また、削除例を図 5.4 図 5.5に示す。A は A が繰り返しされた個所を表し、(B) は B が繰り返した個所として削除されている事を表す。

- 類似文節が数文はなれて存在する事もある (図 5.4)
- 一文内に類似文節が存在する場合はその間に動詞が存在する
- 類似文節のうち後に発話された文節が削除されている
- 動詞を含む文節は基本的に削除されない
- 被修飾文節は基本的に削除されない
- 必須格が削除されている場合がある (図 5.5)

繰り返し表現の場合、文の構成上削除すべきでないと考えられていた、動詞を含む文節、被修飾文節は削除されていない。これは、繰り返した個所と繰り返された個所が比較的離れて出現するため、代りとなることができないからと考えられる。

しかし、必須格となる文節は削除されている場合がある。この事例はいずれも、その文節を必須格とする述語も互いに一致しているか、類似していた。したがって、述語と必須格となる文節が会わせて繰り返されている場合は削除が可能であると考えられる。

これは 自然言語処理技術の応用として自動要約を考えております。
それから、自動同期ですね。
(これは) 同期を自動的にするという意味で、(これは) 音声処理技術の応用です。

⇓　「これは」が削除

これは、自然言語技術出の応用として、です。それから、自動同期です。同期を自動的に、音声処理技術。

図 5.4: 繰り返し削除事例

で、えーこれはあのその制作で作った字幕の画面の例でございますけども、これはあのう画面外に 字幕を 表示したものです。
こちらの方は画面内にあの (字幕を) 表示したものです。

⇓　「字幕を」が削除

画面の外に 字幕を 表示したものです。
これ（スクリーン）は、中に表示したものです。

図 5.5: 繰り返し削除事例 2

5.3 挿入句表現の削除

調査データにおいて挿入句表現として削除されている事例が 13 個所 281 文字存在した。それらを句末表現によって表 5.5に示すように分類した。

表 5.5: 挿入句表現の分類

～けども、	5 事例
～ように、	4 事例
～が、	3 事例
～ので	1 事例

この中で「～ように、」以外の表現については、削除されない句の句末表現としても頻繁に出現しているため、挿入句表現の検出に用いることはできず、他に何か情報が必要である。

しかし「～ように、」という表現に関しては、表 5.5の事例以外には 1 事例しか存在していない。さらに、その 1 事例は、調査データにおいては削除されていないが、明らかに挿入句として削除が可能であった。したがって、「～ように、」という表現に関しては句末表現のみを条件として削除する事が可能であると考えられる。

5.4 丁寧表現の言い替え

文末表現の言い替えに関しては、テレビニュース放送における字幕のための要約に関する先行研究の中ですでに研究されている [若尾 97][山崎 98]。ただし、これらの手法は、体言止めや動詞性接尾辞の削除など字幕特有な表現への様な言い替えルールとなっている。したがって、本研究ではこれら既存の手法を参考にし自前で言い替えルールを作成している。

丁寧表現の有無は述語の位置に現れ、その述語形式は以下の通りである。[益岡 92]

- 動詞連用形 + 「ます」
- イ形容詞 基本型 + 「です」

- ナ形容詞 語幹 + 「です」
- 判定詞「で」 + 「ござる」 + 「ます」

また、[益岡 92] には書かれていないが、実際の要約筆記データを調査したところ以下の形式の場合も存在する。

- 助詞 + 「ござる」 + 「ます」
- 感動詞 + 「ござる」 + 「ます」

これらの丁寧表現は、助動詞「です」「ます」を目印にして検出することができるが、「です」「ます」の語尾形によって言い替え処理が異なる。丁寧表現として使われる「です」「ます」の語尾形には以下のものがある。

- 基本形「です」「ます」
- タ形「でした」「ました」
- タ形条件形「でしたら」「ましたら」
- テ形「でして」「まして」
- タリ形「でしたり」「ましたり」
- 否定形「くないです」「ません」
- 否定のタ形「くなかったです」「ませんでした」

また「です」について、[益岡 92] には書かれていないが、実際の要約筆記データを調査したところ以下の形式の場合も存在する。

- 終助詞形 「ですね」

5 つの丁寧表現述語形式について、「です」「ます」の各語尾形ごとにどの様な言い替え処理が必要か調査した。

5.4.1 動詞連用形 + 「ます」

「ます」の語尾形が基本形である場合

『動詞連用形 + 「ます」』を『動詞基本形』に言い替える事ができる。

「ます」の語尾形が否定形または否定のタ形の場合

『動詞連用形 + 「ます」 + 「ん」』(出現形では『動詞連用形 + 「ません」』) を『動詞未然形 + 「ない」』に言い替えることができる。

また、『動詞連用形 + 「ます」 + 「ん」 + 「です」 + 「た」』(出現形では『動詞連用形 + 「ませんでした」』) を『動詞未然形 + 「ない」 + 「た」』(出現形では『動詞未然形 + 「なかった」』) に、それぞれ言い替えができる。

「ます」の語尾形が基本形・否定形・否定のタ形以外の場合

動詞の種類ごとに活用形変化のさせ方が以下のように異なってくる。[松本 99]

- 接続するときにイ音便になる動詞
  - 五段・カ行イ音便
  - 五段・ガ行
- 接続するときにウ音便になる動詞
  - 五段・ワ行ウ音便
- 接続するときに促音便になる動詞
  - 五段・カ行促音便
  - 五段・タ行
  - 五段・ラ行
  - 五段・ラ行特殊
  - 五段・ワ行促音便
- 接続するときにハツ音便になる動詞

- 五段・ナ行
- 五段・バ行
- 五段・マ行

- 接続するときに音便化しない動詞
  - 上記以外

これを参考にして、『動詞連用形 + 「ます」 + 「た」他』(出現形では『動詞連用形 + 「ました」(他)』 を『動詞の各種音便化 (もしくは音便化させずにそのまま) + 「た」(他)』に言い替えする事ができる。

5.4.2 イ形容詞基本形 + 「です」

この場合「です」の語尾形は基本形、否定形、否定のタ形しかとらない。それぞれ「です」を削除することができる。

5.4.3 ナ形容詞語幹 + 「です」

「です」の語尾形が基本形以外の場合、言い替えても文字数が同じであるかむしろ増えている。したがって、要約という事だけを考えた場合、「です」が基本形である場合のみ言い替えを行うべきであるが、要約文全体の表現を統一するために、「です」が基本型以外である場合も言い替えを行う事とする。また、否定形、否定のタ形は存在しない。

なお、茶筅ではナ形容詞を『名詞 + 助動詞「な」』と解析するので、『ナ形形容詞語幹 + 「です」』は『名詞 + 「です」』となる。

「です」の語尾形が基本型の場合

『名詞 + 「です」』 を 『名詞 + 「だ」』に言い替える事ができる。

「です」の語尾形が テ形の場合

『名詞 + 「です」 + 「て」』(出現形は『名詞でして』) を『名詞 + 「で」』に言い替える事ができる。

「です」の語尾形が終助詞形の場合

『名詞 + 「です」 + 「ね」』を『名詞 + 「だ」』に言い替えることができる。

5.4.4 判定詞「で」 + 「ござる」 + 「ます」

「ます」の語尾形が基本形の場合

『「で」+「ござる」 + 「ます」』(出現形では『でございます』) を『だ』に言い替える事ができる。

「ます」の語尾形が否定形の場合

『「で」 + 「ござる」 + 「ます」 + 「ん」』(出現形では『でございません』) を『「で」 + 「は」 + 「ない」』に言い替える事ができる。

「ます」の語尾形が否定形のタ形の場合

『「で」 + 「ござる」 + 「ます」 + 「ん」 + 「です」 + 「た」』(出現形では『でございませんでした』) を『「で」 + 「は」 + 「ない」 + 「た」』(出現形では『ではなかった』) に言い替える事ができる。

「ます」の語尾形が上記以外の場合

『「で」+「ござる」 + 「ます」 + 「た」他』(出現形では『「でございました」』) を『「で」 + 「ある」+「た」』(出現形では 『であった』) に言い替えることができる。

5.4.5 助詞 + 「ござる」 + 「ます」

「ます」の語尾形が基本形の場合

『助詞 + 「ござる」 + 「ます」』(出現形では『助詞 + ございます』) を『助詞 + ある』に言い替える事ができる。

「ます」の語尾形が否定形の場合

『助詞 + 「ござる」 + 「ます」 + 「ん」』(出現形では『助詞 + ございません』) を『助詞 + 「ない」』に言い替える事ができる。

「ます」の語尾形が否定形のタ形の場合

『助詞 + 「ござる」 + 「ます」 + 「ん」 + 「です」 + 「た」』(出現形では『助詞 + ございませんでした』) を『助詞 + 「ない」 + 「た」』(出現形では『助詞 + なかった』) に言い替える事ができる。

「ます」の語尾形が上記以外の場合

『助詞 +「ござる」 + 「ます」 + 「た」他』(出現形では『「助詞 + ございました」』) を『助詞 + 「ある」+「た」』(出現形では 『であった』) に言い替えることができる。

5.4.6 感動詞 + 「ござる」 + 「ます」

この表現は、それだけで文 (または節) となっており、その文 (または節) をそのまま削除する事ができる。また「ます」の語尾形としては基本形かタ形しか存在していない。

実際に要約処理を行う際には、すでに感動詞は削除されている。そこでこの表現は『文頭 (または「、」) + 「ござる」 + 「ます」 (または「ます」+ 「た」)』(出現形では『(〜、) ございます (まし た)』) をすべて削除する事にした。

5.4.7 「ます」に接続する特殊な動詞

「ます」に「おる」「いたす」などの尊敬語や謙譲語などの動詞が接続している場合に 5.4.1の処理を行うと、文法的には正しいがあまり一般的ではない表現となる。

したがって、要約目的ではないがこの様な特殊な動詞を適当な動詞へと言い替えた後に、5.4.1節の処理を行うこととする。今回、この言い替え処理の対象としたのは、実際の講演データの調査結果もとにして、表 5.6の様な言い替え処理を対象とした。

5.4.8 接頭辞の削除

丁寧表現を表す接頭辞「お」と「ご」を削除する。

表 5.6: 尊敬語・謙譲語の言い替え

おる	→	いる
いたす	→	する
ござる	→	ある
まいる	→	くる
申す	→	いう
申し上げる	→	述べる
いただく	→	もらう

5.5 「〜という〜」表現

調査データにおいて「〜という〜」表現は、71 文において出現し、合計 97 個所、291 文字が出現している。この表現の扱い方は 7 通りあり、それぞれの内訳は表 5.7の通りである。

表 5.7: 要約筆記における「という」表現の要約事例

A　単独での削除	6 個所
B　前後の単語を伴った削除	30 個所
C　文・節単位の削除に含む	35 個所
D　単独での言い替え	2 個所
E　前後の単語を伴った言い替え	11 個所
F　文・節単位の言い替えに含む	4 個所
G　そのまま残る	9 個所

これらのうち文・節単位で処理されている C と F は、「〜という〜」表現が直接関係しているとは考えられず、今回の調査対象からは除外し、A B D E G の 58 事例を今回の調査対象とした。

また、言い替えによる要約事例である D と E は、図 5.6図 5.7に示す様に削除処理によっても要約する事が可能である。したがって、削除処理による要約手法に統一して考

える。

なお、(A) は A が削除されている事を表し、[B] → [C] は B が C に言い替えられていることを表す。以下本節における要約事例では同様の意味を表すものとする。

番組を制作する [ということ] → [か] が、我々のプロジェクトの使命でございます。
⇓　「ということ」 が 「か」に言い替え
番組を制作するかが、我々のプロジェクトの使命でございます。

図 5.6: 言い替えによる要約事例

番組を制作する (という) ことが、我々のプロジェクトの使命でございます。
⇓　「という」 が 削除
番組を制作することが、我々のプロジェクトの使命でございます。

図 5.7: 言い替えによる要約事例を削除によって要約した例

削除による要約事例は以下に示す 3 通りの削除処理に分ることができる。

a　「という」が単独で削除

b　「という」とその後の形態素が削除

c　「という」とその後の文節が削除

d　a,b,c にあてはまらない例外処理

これらのうち、a,b,c の 3 種類の削除処理の分類は、「という」の前後に現れる形態素の品詞の関係 (以後、品詞関係と表記する) と、ほぼ一致している。また、c の処理は「という」の後に現れる単語の種類によって適用される。

以下では、これらの各処理と、その処理が適用される条件について説明する。

5.5.1 「という」が単独で削除

この削除処理が適用されるのは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「ない」 という 名詞

動詞 という 名詞 この品詞関係の場合、18 事例中 17 事例が削除可能であった。残りの 1 事例は例外処理が適用されている。削除例を図 5.8に示す。

えー字幕制作はこのフローに沿って進めているというわけではなくて、あの人間がいくつかのパートを同時に行う (という) やり方をしている場合もございますけども。
⇓ 「という」 が削除
字幕制作は、このフローにそってというわけではなく、人間がいくつかのパートを同時に行うやりかたをしている場合もあります。

図 5.8: 動詞 という 名詞 の削除事例

形容詞という名詞 この品詞関係の場合、5 事例中 4 事例が削除可能であった。残りの 1 事例は例外処理が適用されている。削除例を図 5.10に示す。

クローズドキャプションを画面の中にそのまま出すと、重なってしまって見にくい (という) ことがあのう考えられます。
⇓ 「という」 が削除
クローズドキャプションを画面に出すと重なり見にくいこともある。

図 5.9: 形容詞 という 名詞 の削除事例

助動詞 という 名詞　この品詞関係の場合、5 事例中 2 事例が削除可能であり、その 2 事例共に助動詞が「ない」であった。したがって、この品詞関係で助動詞が「ない」である場合のみ削除が可能であると考えられる。削除例を図 5.10に示す。

リアルタイムで字幕を付与しなければならない (という) 面がございますけども。
⇓　「という」 が削除
リアルタイムで字幕を付与しなければならない面がございますけども。

図 5.10: 助動詞 という 名詞 の削除事例

5.5.2 「という」とその後の形態素が削除

この削除処理が適用されるのは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

名詞 という 名詞　この品詞関係の場合、22 事例中 21 事例が削除可能であった。残りの 1 事例は例外処理が適用されている。削除例を図 5.11に示す。

ですから、あのう非常に荒っぽい要約になるといいますか、なってしまいますので、えー別の手段としまして、文節単位での圧縮 (ということ) を考えております。
⇓　「ということ」 が削除
あらっぽい要約になってしまいますので、別の手段として、文節単位での圧縮を考えています。

図 5.11: 名詞 という 名詞 の削除事例

つまり、聴覚障害者向けのテレビジョンサービスというところで、どのような技術的な支援ができるか (というところ) を主体にあの研究を進めております。

⇓　「というところ」 が削除

つまり、聴覚障害者史向けのテレビサービスに、どのような技術支援ができるかを主体に研究しています。

図 5.12: 助詞 という 名詞 の削除事例

助詞 という 名詞　この品詞関係の場合、3 事例中 2 事例が削除可能であった。残りの 1 事例は例外処理が適用されている。図 5.12

5.5.3　「という」とその後の文節が削除

この削除処理が適用されるのは品詞関係が「文頭 (間投詞) という 名詞」の場合のみである。この場合、すべてが削除可能であった。削除例を図 5.13に示す。

(ということで、) まいろいろな字幕提示方法をシミュレーションで作成いたしまして、聴覚障害者の方にえー評価していただいて、どの提示方法が一番望ましいかというようなことを調べております。

⇓　「ということで、」 が削除

いろいろな字幕提示方法を、シュミレーションで作成して、聴覚障害者に評価していただいて、一番合った表示方法を。

図 5.13: 文頭 (間投詞) という 名詞 の削除事例

5.5.4　例外処理

「〜 という ふうに」表現の例外処理

「〜という ふうに」表現の場合のみ、例外処理として「というふう」を削除して「に」を「と」に言い替える処理が可能である。処理例を図 5.14に示す。

ただそれでもだいたい 15 %の削減が、この現在やっている手法ではあのう限界と言いますか、そこぐらいにしかいかないのではないかな (というふう) に → と 考えております。
⇓　「というふう」 が削除　「に」 を 「と」 に言い替え
だいたい１５%くらいの削減が、現在の手法の限界ではないかと考えています。

図 5.14: 「というふうに」の例外処理事例

5.5.5 削除不可能な例外事例

削除不可能な事例として、図 5.15に示すような「多いという レポート」という表現がある。これは、「という」の前が他の形容詞であっても、動詞であっても削除不可能である (例えば図 5.16)。したがって、「という」の後に続く名詞「レポート」に削除不可能な原因があると考えられる。

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。
⇓　「という」 は削除されない
現状では、報道番組への希望が多いという報告がされています。

図 5.15: 削除不可能な事例 1

字幕生成のための要約は人手によって行う (という) レポートがある。
⇓　「という」 は削除すると不自然
字幕生成のための要約は人手によって行うレポートがある。

図 5.16: 削除不可能な事例 2

この様に「用言 という 名詞」の場合、特定の名詞が後にくると削除できないと考える事ができる。この様な名詞として、「報告」や「文書」「記録」といった名詞も同様な現象になる。これらの名詞は意味的に類似していると考えられ、角川類語国語辞典で調べてみ

たところ、「文書」と「記録」という小分類に属していた。したがって、この小分類に属する名詞が「という」の後に現れる場合は、例外処理として削除せずにそのまま残すべきと考えられる。

第 6 章

要約システム

6.1 要約システム全体の構成

最終的な要約システムとして、各モジュールを以下の順で処理させる。

1. 茶筅による形態素解析
2. 間投詞削除モジュール
3. 「〜という〜」表現削除モジュール
4. 丁寧表現言い替えモジュール
5. 茶筅による形態素解析
6. 「〜ように」表現 (挿入句) 削除のモジュール
7. 文節区切り処理
8. 繰り返し・言い直し表現削除モジュール

間投詞は文中のどのようなところにも出現するため、先に削除しておかないと他のモジュールの処理に悪影響を与える場合がある。したがって、一番最初に処理を行う。

丁寧表現の言い替えモジュール、および「〜という〜」表現削除モジュールの一部において、言い替え処理を行っているため、形態素情報が失われている恐れがある。したがっ

て、この二つのモジュールは残りのモジュールより先に行い、その再度形態素解析を行うものとする。

さらに、一番計算能力を必要とする繰り返し・言い直し表現削除のモジュールを最後に処理することで、他のモジュールで削除できるものはこのモジュールでは処理しないですむようにしている.

各要約モジュールはそれぞれが形態素情報を受けて要約結果を出力できるようになっている。したがって、各モジュールはそれ単体による要約も可能である。

6.2 各要約モジュールの実装

6.2.1 間投詞削除モジュール

形態素解析の結果 (もしくはその情報を保持していると考えられる各モジュールの結果) を入力として以下に該当するものを言い替える。

- 品詞が「フィラー」である形態素
- 品詞が「感動詞」である形態素
- 見出し (出現形) が「えーと」である形態素
- 見出し (出現形) が「まあ」である形態素
- 見出し (出現形) が「あの」「う」の連続となる形態素列
- 見出し (出現形) が「あ」「の」「う」の連続となる形態素列

6.2.2 「〜という〜」表現の削除モジュール

形態素解析の結果 (もしくはその情報を保持していると考えられる各モジュールの結果) を入力として以下に該当するものを言い替える。

- 「という」が文頭の場合は「という文節」を削除
- 「という」が文頭ではない場合

- 「〜 という ふうに」の場合は「というふう」を削除して「に」を「と」に言い替える
- 上記以外の場合
  - 「名詞 という 名詞」の場合は「という 名詞」を削除
  - 「動詞 という 名詞」で名詞が例外名詞でない場合は「という」を削除
  - 「形容詞 という 名詞」で名詞が例外名詞でない場合は「という」を削除
  - 「ない という 名詞」で名詞が例外名詞でない場合は「という」を削除
  - 「助詞 という 名詞」は「という 名詞」を削除

なお、「という」は茶筅の結果においては、格助詞 (連語)「という」と、格助詞「と」＋動詞「いう」の二通りに解析されるが、ここではそのどちらも対象としている。

6.2.3 丁寧表現の言い替えモジュール

形態素解析の結果 (もしくはその情報を保持していると考えられる各モジュールの結果) を入力として以下に該当するものを言い替える。

- 動詞連用形 + 「ます」が存在する場合
  - 「ます」が基本形の場合 動詞を基本型に言い替えて「ます」を削除
  - 「ます」が否定形・否定のタ形の場合 動詞を未然形に言い替えて「ます」を「ない」又は「なかった」に言い替える
  - 「ます」が上記以外の場合 音便化動詞は音便化させてそれ以外の動詞はそのままにして「ます」を削除する
- イ形容詞 + 「です」が存在する場合「です」を削除する
- ナ形容詞語幹 + 「です」が存在する場合
  - 「です」が基本型の場合 「です」を「だ」に言い替える
  - 「です」がテ形の場合「です」を「で」に言い替えて「て」を削除
  - 「です」が終助詞形の場合 「です」を「だ」に言い替えて終助詞を削除

- 「です」がそれ以外の場合 「です」を「だっ」に言い替える

- 判定詞「で」+「ござる」+「ます」 が存在する場合
  - 「ます」が基本型の場合 「でございます」を「だ」に言い替える
  - 「ます」が否定形の場合 「でございません」を「ではない」に言い替える
  - 「ます」が否定のタ形の場合 「でございませんでした」を「ではなかった」に言い替える
  - 「ます」がそれ以外の場合「でございまし」を「であっ」に言い替える

- 助詞 + 「ござる」 +「ます」 が存在する場合
  - 「ます」が基本型の場合 「ございます」を「ある」に言い替える
  - 「ます」が否定形の場合 「ございません」を「ない」に言い替える
  - 「ます」が否定のタ形の場合 「ございませんでした」を「なかった」に言い替える
  - 「ます」が上記以外の場合 「ございまし」を「あっ」に言い替える

- 「ございます」が文頭にある場合
  - 「ます」が基本型の場合 「ございます」を削除
  - 「ます」がタ形の場合 「ございました」を削除

6.2.4 「〜ように」表現削除モジュール

形態素解析の結果 (もしくはその情報を保持していると考えられる各モジュールの結果) を入力として以下に該当する者を削除する。

- 「、〜ように、」が存在する場合「〜ように、」を削除
- 「文頭〜ように、」が存在する場合「〜ように、」を削除

6.2.5 言い直し・繰り返し表現の削除

形態素解析の結果 (もしくはその情報を保持していると考えられる各モジュールの結果) を 文節区切りプログラムにかけて、その結果を受けて以下のものを削除する。

- 同一文内における 2 文節について以下を満たす場合は言い直しとして文頭側の文節を削除
  - 2 文節が類似している
  - 2 文節間に動詞が存在しない
- 同一文内における 2 文節につい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
  - 2 文節それぞれが動詞を含まない
  - 文末側の文節が「ガヲニ格」ではない
  - 文末側の文節の一つ前の文節が助詞「の」によって連体化されていない
  - 2 文節が類似している
  - 2 文節間に動詞が存在する
- 隣接する 2 文 (前後 2 文まで) からそれぞれ一つづつ取ってきた 2 文節につい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
  - 2 文節それぞれが動詞を含まない
  - 文章末側の文節が「ガヲニ格」ではない
  - 文章末側の文節の一つ前の文節が助詞「の」によって連体化されていない
  - 2 文節が類似している

文節間類似度は、文節構成形態素をそれぞれ比較し、式 (6.1) の様に計算する。

$$\text{文節 (列)}AB\text{ 間類似度} \quad = \quad \frac{\text{文節 (列)}AB\text{ 間類似スコア}}{max\{\ \text{文節 (列)}AB\text{ 構成スコア}\ \}} \tag{6.1}$$

文節 (列) 間類似スコア (6.2)

$=$ 2 $\times$ 自立語の一致数 + 1 $\times$ 非自立語の一致数

文節 (列) 構成スコア (6.3)

$=$ 2 $\times$ 文節 (列) 構成自立語の数 + 1 $\times$ 文節 (列) 構成非自立語の数

ある文節 (列)A の構成スコアとは、その文節 (列)A と全く同じ形態素列によって構成される文節 (列) との類似スコアと考えることができ、文節 (列)A がとりうる類似スコアの最高値を表すことになる。したがって、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) の各構成スコアの高い方の値で割ること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(全く異なる形態素列の文節との類似度)、最高値が 1(全く同じ形態素列の文節との類似度) 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。

第 7 章

システムの評価

評価用データセットとして TAO コーパスの中から表 7.1のものを用いた。システムに入力すデータは書き起しデータであり、まずシステムの出力の削除率にて評価する。次に、入力したデータの要約筆記データを正解データとし、システムの削除個所の精度 (precision) と再現率 (recall) によって評価を行う。

表 7.1: 評価データセット

講演 No.	書き起しデータ (入力データ)				要約筆記データ (正解データ)			
	文数 [sent]	文字数 [char]	時間 [sec]	速度 [char/m]	文数 [sent]	文字数 [char]	速度 [char/m]	要約率 [%]
1-4	152	8032	1422	338.9	143	3857	162.7	48.0
3-1	159	8357	1940	258.5	148	5375	166.2	64.3
3-3	82	7372	1410	313.7	78	4554	193.8	61.8
3-4	76	5857	1455	241.3	93	3947	162.8	67.4

7.1 削除率

本システムに講演音声書き起しテキストを入力し、システムが要約した結果の削除率を式 (7.1) にしたがって計算した。その結果を表 7.2に示す。

なお使用した講演データのうち、1-4 は今回調査に用いたデータであり、それ以外は調査には用いていないデータである。

$$\text{削除率} \quad = \quad \frac{\text{書き起しテキスト文字数} - \text{システム出力文字数}}{\text{原文文字数}} \times 100 \qquad (7.1)$$

表 7.2: システム出力結果の削除率

講演番号	1-4	3-1	3-3	3-4
原文文字数	8032	8357	7372	5857
要約結果文字数	6237	6912	6272	4573
削除率	22.3%	17.3%	14.9%	21.9%

また、各モジュールによる削除率は表 7.3から表 7.6の通りとなった。なお単独削除率とは各モジュール単体での削除率であり、累積削除率は各モジュールで順次処理を行っていった際の途中経過を表している。それぞれの計算は、モジュール A・モジュール B の順で処理が行われた場合のモジュール B による単独削除率と累積削除率を式 (7.2) 式 (7.3) で計算している。

$$\begin{aligned} &\text{単独削除率} \\ &= \frac{\text{モジュール } A \text{ 処理後の文字数} - \text{モジュール } B \text{ 処理後の文字数}}{\text{原文の文字数}} \times 100 \end{aligned} \qquad (7.2)$$

$$\begin{aligned} &\text{累積削除率} \\ &= \frac{\text{原文の文字数} - \text{モジュール } B \text{ 処理後の文字数}}{\text{原文の文字数}} \times 100 \end{aligned} \qquad (7.3)$$

表 7.3: 1-4 データ

	文字数 [char]	単独削除 文字数 [char]	単独 削除率 [%]	累積削除 文字数 [char]	累積 削除率 [%]
原文	8032				
間投詞	7758	274	3.4%	274	3.4%
という	7447	311	3.9%	585	7.3%
丁寧	6808	644	8.0%	1229	15.3%
ように	6744	59	0.7%	1288	16.0%
繰り返し	6237	507	6.3%	1795	22.3%

表 7.4: 3-1 データ

	文字数 [char]	単独削除 文字数 [char]	単独 削除率 [%]	累積削除 文字数 [char]	累積 削除率 [%]
原文	8357				
間投詞	8201	156	1.9%	156	1.9%
という	7927	274	3.3%	430	5.1%
丁寧語	7621	306	3.7%	736	8.8%
ように	7533	88	1.1%	824	9.9%
繰り返し	6914	619	7.4%	1443	17.3%

表 7.5: 3-3 データ

	文字数 [char]	単独削除 文字数 [char]	単独 削除率 [%]	累積削除 文字数 [char]	累積 削除率 [%]
原文	7372				
間投詞	7278	94	1.3%	94	1.3%
という	7006	272	2.9%	366	5.0%
丁寧語	6788	218	3.0%	584	7.9%
ように	6673	115	1.6%	699	9.5%
繰り返し	6281	392	5.3%	1091	14.8%

表 7.6: 3-4 データ

	文字数 [char]	単独削除 文字数 [char]	単独 削除率 [%]	累積削除 文字数 [char]	累積 削除率 [%]
原文	5857				
間投詞	5746	111	1.9%	111	1.9%
という	5478	268	4.6%	379	6.5%
丁寧語	5176	302	5.2%	681	11.6%
ように	5166	10	0.2%	691	11.8%
繰り返し	4582	584	10.0%	1275	21.8%

7.2 精度と再現率

各データに付属している要約筆記データを正解データとみなし評価を行う。

評価尺度は、要約システムにおける削除個所がどれだけ要約筆記データにおける削除個所と一致したかを表す精度 (precision) と、要約筆記データにおける削除個所が要約システムにおいてどれだけ再現できていたかを表す再現率 (recall) にて評価を行う。それぞれの定義式を式 (7.4) 式 (7.5) し示す。

$$\begin{aligned} &\text{一致率}\,(Precision) \\ &= \frac{\text{要約システム・要約筆記ともに削除されている数}}{\text{要約システムにおいて削除された数}} \times 100 \end{aligned} \tag{7.4}$$

$$\begin{aligned} &\text{再現率}\,(Recall) \\ &= \frac{\text{要約システム・要約筆記ともに削除されている数}}{\text{要約筆記において削除された数}} \times 100 \end{aligned} \tag{7.5}$$

しかし、要約筆記データを正解データと見なす場合、いくつかの点において注意する必要がある。

7.2.1 部分一致

要約筆記データは要約システムとは異なり、必ずしも文節単位で処理されているとは限らない。したがって、システムの結果と要約筆記データとが部分的に一致するという事例が存在する。その例を図 7.1 に示す。システムが「約 10%」を削除しているのに対して、要約筆記では「10%、」が削除されている。

また、システムの出力では削除によって要約されているが、要約筆記では言い換えによって要約されている場合もある。その例を図 7.2に示す。システムが「通して」を削除しているのに対して、要約筆記では「を通して」を「で」に言い替えている。

これらの事例は、完全に一致しているわけではないが、「要約する個所は一致していて、要約の手法が異なるもの」と考えられる。したがって、完全な不一致と一緒に扱うべきではないと考えられる。そこで、一致度の評価の際には上記の事例を「部分一致」と定義して、「完全一致」「部分一致」「完全不一致」の三段階に別けて評価を行うことにする。

(注 例文中の 文字 1 → 文字 2 は「文字 1」が「文字 2」へと言い替えられた事を表す。また (文字 3) は文字 3 が削除された事を表す。以下、本節では同様に表記する。)

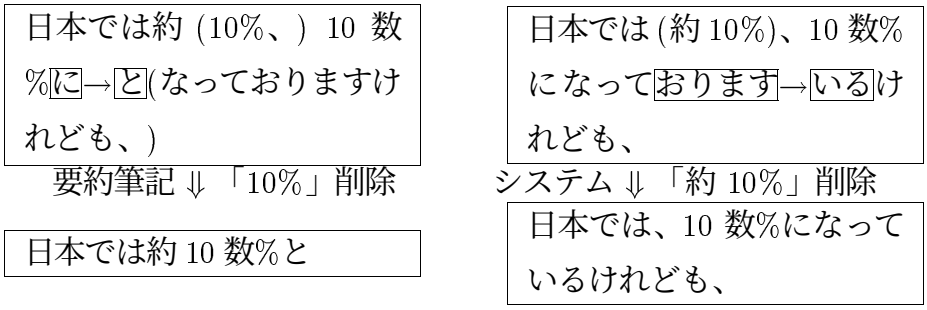

図 7.1: 部分一致の例

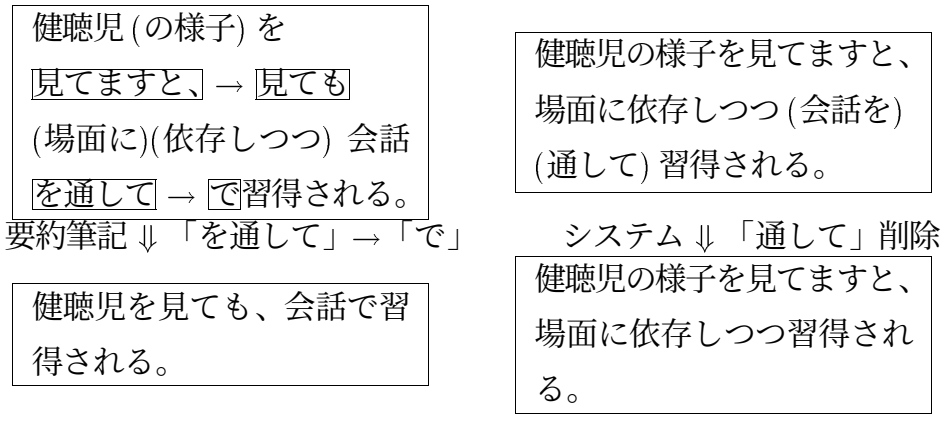

図 7.2: 部分一致の例 2

7.2.2 完全不一致内正解

また、要約筆記データは必ずしも正しいとか限らない。要約筆記の性質上、発話速度が入力速度より遅い場合は不要な個所でも削除せずそのまま残される事がある。したがって、「完全不一致」と判定された個所であっても実際には削除しても問題がない場合や、むしろ削除すべき個所である場合も存在する。その例を図 7.3に示す。システムでは「という」が削除されているのに対して、要約筆記では削除されていない。しかし、この「という」は明らかに削除可能である。

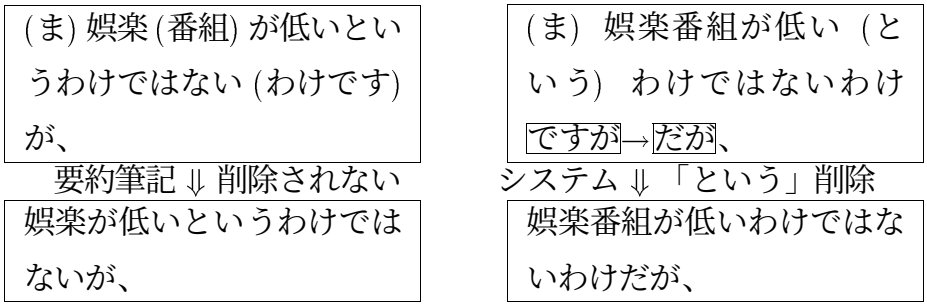

図 7.3: 完全不一致内正解の例

この様なものもすべて不正解としてしまうのは問題があると考えられる。したがって「完全不一致」と判定された個所については、「明らかに削除できる個所」「明らかに削除すべきでない個所」「どちらとも判断しにくいもの」の３つに別けて評価を行う必要がある。

ただしこの場合、正解データが変化している事になる。したがって、再現率の定義式である式 (7.5) の分母が定まらないため、再現率は測定不能であると考えられる。

7.2.3 評価結果

7.2.1節および 7.2.2節に従った評価を行った。その結果が表 7.7表 7.8である。ただし、部分一致の累積一致数および完全不一致内正解の累積一致数はそれぞれ式 (7.6) 式 (7.7) の様に定義している。この累積一致数を式 (7.4)(7.5) の分子として精度 (precision) および再現率 (recall) を計算している。

部分一致の累積一致数

$$= \text{完全一致数} + \text{部分一致数} \tag{7.6}$$

完全不一致内正解の累積一致数

$$= \text{完全一致数} + \text{部分一致数} + \text{完全不一致内正解一致数} \tag{7.7}$$

表 7.7: 精度と再現率

	一致数	累積一致数	精度 (Precision)	再現率 (Recall)
完全一致	359	359	359/748　48.0%	359/969　37.0%
部分一致	107	466	466/748　62.3%	466/969　48.1%

表 7.8: 完全不一致内正解の精度

	一致数	累積一致数	精度 (Precision)
完全不一致内正解	131	597	597/748　79.8%

7.3　考察

評価の結果、本要約システムよる削除率が 15%〜20%程度となっている。モデルとして調査した講演データ 1-4 の書き起しテキストを本システムで要約させた場合、 22.3% の削除率となっているが、それと比べるとやや削除率が低くなっている。これは、講演発表者の個人差によるものと考えられる。

また、挿入句表現削除のための有効な手法が発見できていないため、挿入句表現削除モジュールによる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事で削除率が向上すると考えられる。

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 80%程度となっている。明らかに不正解である 20% 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などに再検討の余地があると考えられる。

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事で削除率が向上すると考えられる。また講演発表者ごとに個人差があることを考えると、より多数の講演データの調査を行う必要があり、それによりさらに削除率の向上が期待できる。

第 8 章

おわりに

8.1 まとめ

本研究は、話し言葉の特徴のうち冗長表現と考えられる表現に注目し、その冗長表現を不要個所として削除または言い替えをする事による要約システムを提案した。

要約システムの実装にあたり、要約筆記をシステムのモデルと考え、実際の要約筆記データの調査を行った。調査の対象とした表現は「間投詞」「言い直し・繰り返し表現」「挿入句表現」「丁寧表現」を対象とし、さらに調査を進めて行く中で「〜という〜表現」も不要個所として削除されることがわかった。また、調査の結果、「間投詞」形態素解析の結果から削除が可能であった。「言い直し・繰り返し表現」は文節間類似度と文節構成形態素などからの条件を元に削除する事が可能であった。「挿入句表現」は有効な削除手法が発見できなかったが、「〜ように、」という句末表現となっている句に関しては削除が可能であった。「丁寧表現」は字幕生成に関する要約手法の研究において言い替え手法が提案されており、その手法を参考にしている。ただし、先行研究では字幕特有の表現へと言い替え手法であるので、言い替え手法を自前で作成した。「〜という〜表現」は「という」の前後の形態素の品詞関係より削除手法が異なっており、さらに一部例外処理が存在する事がわかった。

この調査結果を元に、各表現を削除または言い替えするモジュールの実装を個別に行った。この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。

システムの評価としては、実際の講演の書き起しデータを入力とした際の削除率、および、その講演の要約筆記データを正解とした精度 (precision) と再現率 (recall) によって

評価を行っている。削除率は、調査に用いた講演データの書き起しデータを入力した際の削除率と比べてやや低い削除率になっており、発表講演者の個人差による影響があった。また、精度に関しては、完全不一致内正解がもっともらしい評価と考えられ、その結果が80% 程度になっていた。不正解であった削除個所の主なものが言い直し・繰り返し削除モジュールによるものであり、改善が必要と思われる。また、再現率が低いことから、他にも削除可能な表現が存在していると考えらる。

8.2 今後の課題

削除率に関して 本来削除率に大きく貢献すると考えられる、挿入句表現削除モジュールが低い削除率になっていた。このモジュールが理想的に動作すれば、さらに 5% 程度の削除率上昇が期待できる。このモジュールの現在の問題点は、句末表現のみを条件にして特定の表現のみしか扱っていない点である。改善方法としては、挿入句に含まれる動詞や自立語などの傾向を調べる事で新たな条件を発見できるのではないかと考える。たとえば、挿入句表現として多かったものとして「先ほど述べたように、」や「ここに示してありますが、」という表現がある。これらの表現には、方向や時間を指し示す単語 (「先ほど」「ここ」) や、物事を述べたり提示したりする単語 (「述べる」「示す」) などが存在している。この様に、挿入句に含まれる単語にはある傾向が存在するのではないかと考えられる。

また、「〜という〜」表現削除モジュールでは、「〜と申す〜」「〜といった〜」「〜といわれている〜」など色々なバリエーションが考えられる。実際に、システムの評価の際にシステムが出力した結果を確認すると、「〜という〜」とほぼ同じ用途で使われている「〜といった〜」という表現が頻繁に出現している。こういったバリエーションに対応する必要もあると考えられる。

精度に関して 精度を下げているのは主に言い直し・繰り返し表現削除モジュールによるものであった。このモジュールの改良点として、類似度計算の改良が考えられる。現在は形態素単位の文字列マッチングによる類似度計算を行っている。自立語の一致と付属語の一致の重み付けを変えて工夫はしているが、まだ改善の余地はあると考えられれる。例えば自立語の意味的類似性を考慮した類似度計算などが考えられる。

再現率に関して　再現率を上げるためには、より多くの要約筆記データの調査を行う必要がある。本研究では１講演の要約筆記データしか調査を行っていない。しかし、話し言葉は個人差が多いと考えられるため、他の要約筆記データを調査することで、さらに削除可能な表現を発見することができると考えられる。ただし、本研究は informative な要約を想定しているため、精度を犠牲にしてまで再現率を上げる必要はないと考えられる。

正解データに関して　本研究ではシステム評価の際に要約筆記データを正解データと見なしている。しかし、要約筆記は人間がリアルタイムで処理しているものであるため、間違えが含まれる事が考えられる。本研究では、その対策として「完全不一致内正解」というものを定義して対応している。しかし、より正確な評価を行うためには、正解データを別の方法で作成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、終始熱心な御指導を賜わりました奥村学助教授に心から感謝致します。また、数多くの御教授を頂きました島津明教授に厚く御礼申し上げます。さらに、多大な助言をして頂きました望月源助手に厚く御礼申し上げます。

中間審査などの折には、諸先生方から貴重な御意見を頂きました。深く感謝致します。

本研究において調査データとして使用した、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) の方々に感謝致します。

自然言語処理学講座の皆様には、貴重な御意見、討論をして頂きました事を感謝致します。

最後に、多くの方々の御援助によって本研究を行うことができましたことを厚く御礼申し上げます。

参考文献

[Hand97] Hand,T. "A Proposal for Task-based Evaluation of Text Summarization Systems." In Proc. of the ACL Workshop on Intelligent Scalable Text Summarization,pp.9-16. 1997.

[太田 98] 太田晴康. " パソコン要約筆記入門 – 「聞こえ」を支えるボランティア". 人間社1998.

[太田 99] 太田晴康. " 要約筆記への招待 – 活動現場の視点から". 言語,Vol.28,No.9,pp.73-79. 1999.

[奥村 99] 奥村学, 難波英嗣. " テキスト自動要約に関する研究動向". 自然言語処理,Vol.6,No.6,pp.1-26. 1999.

[益岡 92] 益岡隆志, 田窪行則. " 基礎日本語文法 — 改訂版 —". くろしお出版1992.

[松本 99] 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". 1999

[三上 99] 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. 1999.

[白井 99] 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.

[竹沢 94] 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.

[若尾 97] 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,122-13,pp.83-89. 1997.

[山崎 98] 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.