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

下岡 和也[†] 南條 浩輝[†] 河原 達也[†]

講演音声のような話し言葉の書き起こしや音声認識結果には、話し言葉特有の表現が数多く含まれており講演録などのアーカイブとして二次利用しにくいいため、文章として適した形態に整形する必要がある。本稿では、統計的機械翻訳の考え方に基づいて講演の書き起こしを整形された文章に自動的に変換する方法を提案する。本研究で扱う処理は、フィラーの削除、句点の挿入、助詞の挿入、書き言葉表現への変換、文体の統一である。これらの処理を統合的に行うようにビームサーチを導入した。実際の講演の書き起こしを用いた定量的な評価により統計的な手法の有効性が示され、句点と助詞の挿入に関して高い精度を得ることができた。

キーワード: 話し言葉, 音声認識, 統計的機械翻訳, 自動整形, 自動要約

Automatic Transformation of Lecture Transcription into Document Style using Statistical Framework

KAZUYA SHITAOKA[†], HIROAKI NANJO[†] and TATSUYA KAWAHARA[†]

Transcriptions and speech recognition results of lectures include many expressions peculiar to spoken language. Thus, it is necessary to transform them into document style for practical use of them. We apply the statistical approach used by machine translation to automatic transformation of the spoken language into document style sentences. We deal with deletion of fillers, insertion of periods, insertion of particles, conversion to written expressions and unification of the end-of-sentence style. A beam search is introduced to apply these processings in an integrated manner. Experimental evaluation using real lecture transcriptions confirms that the statistical transformation framework works well and we achieved high recall and precision rates of period and particle insertion.

KeyWords: *spontaneous speech, speech recognition, machine translation, text processing, summarization*

1 緒論

音声認識研究の対象は、読み上げ音声から講演や会議などの話し言葉に移行している。このような話し言葉は日本語では特に、文章に用いる書き言葉と大きく異なり可読性がよくない。そのため、書き起こしや音声認識結果を講演録や議事録などのアーカイブとして二次利用する際には、文章として適切な形態に整形する必要がある。実際に講演録や議事録の作成の際には、人手によりそのような整形が行われている。

これまでに、放送ニュースなどを対象とした自動要約の研究が行われている (加藤 1998; 若

[†] 京都大学情報学研究科, School of Informatics, Kyoto University

尾, 江原, 白井 1998; 中沢, 遠藤, 古川, 豊浦, 岡 1996; 堀, 古井 1999, 2000). これらは主に, 頻出区間や重要語句の抽出といった処理, つまり発話された表現をそのまま用いることによって要約を作成している. しかし, 話し言葉表現が多く含まれる場合には, 要約を作成する際にまず話し言葉から書き言葉へ変換する必要がある. 実際に人間が要約を作成する際には, このような書き言葉表現への変換に加えて, 不必要な部分の削除や必要な語の挿入, さらに 1 つの文書内での「ですます」調/「である」調などの文体の統一といった処理も行っている.

本研究では講演の書き起こしに対してこのような整形を自動的に行うことを考える. 現在, 文章を整形するソフトウェアも存在しているが, これらはパターンマッチング的に規則ベースで変換を行っており, 言語的な妥当性や前後との整合性はあまり考慮されていない. また, 基本的に 1 対 1 の変換を行っているので, 変換の候補が複数ある場合への対処が容易ではない. 学会講演とその予稿集との差分をとることで書き言葉と話し言葉の変換規則を自動抽出する研究が村田らにより行われている (村田 2002; 村田, 井佐原 2001) が, 変換の際の枠組みは本質的に同じと考えられ, また実際に変換を行い文章を整形する処理は実現されていない. これに対して本研究では, 規則に基づいて 1 対 1 の変換を行うのではなく, 話し言葉と書き言葉を別の言語とみなした上で統計的な機械翻訳の手法を適用し, 確率モデルによりもっともらしい表現に変換し実際に文章を整形することをめざす.

2 整形作業における処理

講演録編集者は, 書き起こしから講演録や要約を作成する際に, 通常次の 4 段階の作業を行う.

- (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. 第二段階の文法的チェックでは, 言語的に正しくない助詞や接続詞を適切なものに修正する. ポリティカルチェックでは, 差別用語などの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.

実際に講演録編集者が, 「日本語話し言葉コーパス」(CSJ) (Furui, Maekawa, and Isahara 2000) のいくつかの講演について整形・要約したものを参考にして, (a) から (i) で第一段階, (j) と (k) で第二段階の作業についてそれぞれ例を挙げながら説明する.

(a) フィラーの削除

「あの一」や「えっと」といった間投語はすべて削除する.

(b) 書き言葉表現への変換

次に挙げるような話し言葉表現の書き言葉表現への変換を行う。

(例) 「行ってるんですが」 → 「行っているの んですが」

「規則合成 っていう 方式」 → 「規則合成 という 方式」

上記の例では話し言葉と書き言葉が1対1に対応しているが、そうでない場合を以下に示す。

(例) 「このようにいたしております」 → 「このように しています」「このように している」
「このように しております」

このような場合は、全体として「ですます調」か「である調」に統一されるように留意する。

(c) 助詞の挿入

話し言葉では、しばしば助詞が脱落して発話されるので、適切な助詞を挿入する。

(例) 「観測されたらそれ適当な分布で」 → 「観測されたらそれ を 適当な分布で」

「先程も話しありましたが」 → 「先程も話ししがありましたが」

(d) 句読点の挿入

書き起こしに句読点がない場合は、適当な箇所句読点を挿入する必要がある。

(e) 倒置部の修正

話し言葉では、倒置を用いて発話されることがあるので、通常の語順に修正する。

(例) 「日本語は前寄りでも後寄りでも一応‘あ’なんですね 音韻的には。」 → 「日本語は前寄りでも後寄りでも 音韻的には 一応‘あ’なんですね。」

(f) 言い淀みをしている部分の削除

話者が言い淀みをした部分は全て削除する。

(例) 「このようなスペクトルの ドッ、ギャップがですね」

(g) 説明を付け加えている部分の削除

言葉の説明を付け加えている部分を削除する。

(例) 「モデル化を行い、そのモデル自体 HMM ですが、そのモデルから」

このような表現は必ずしも削除されるわけではなく、別の表現に書き換えられることもある。

(例) 「こういう HMM、我々は MSD-HMM と呼んでいる」

→ 「こういう HMM(MSD-HMM)」

(h) 独り言の部分の削除

本論と関係ないことを発話している部分を削除する。

(例) 「えー、そろそろ時間なんですが、」

続いてえ結果方見ていきたいと思います< pause-647-msec >でえーまず
 えーとーこちらあの英語話者アメリカ人のえー英語話者によるえ語頭に
 R Lを含む単語えーとRのライトとLのライトの調音について
 < pause-333-msec > えーとと一名の例を示しました< pause-513-msec >で縦
 軸とこのグラフでいう縦軸と横軸というのはえーこの< pause-304-msec >
 調音地図というのの元になっている距離マトリックス< pause-385-msec >を
 < pause-255-msec >より反映するようにそのMDSというまー< pause-654-msec >
 え手法がえ抽出したえー次元ーと二というふうになりますで一つ一つの
 < pause-326-msec >点グラフ内の一つ一つの点が< pause-306-msec >え発話の一回
 一回< pause-470-msec >をえー示していて< pause-389-msec >え同じ単語のえー
 < pause-203-msec >点をえー見易いようにちよっと< pause-311-msec >見易くえーと
 丸で囲んでグルーピングしております< pause-1298-msec >

図 1 書き起しの例

(i) 言い直し部分の削除

言い直しをしている部分は削除する。

(例)「これは、テキスト テキスト からの音声合成を」

「上唇の、ごめんなさい これ間違いです」

(j) 長文の分割

一文が非常に長い場合は、適切な接続詞を用いることでいくつかの単文に分割する。

(例)「～日本人としてある意味で嬉しいところ あるんですけども どうして こういう
 技術が考えられたかというのと、～」→「～日本人としてある意味で嬉しいところ
があります。どうして こういう技術が考えられたかというのと、～」

(k) 助詞・接続詞の修正

より適切な助詞や接続詞がある場合は、それに訂正する。

(例)「どのようにして音声の合成を するかということが」→「どのようにして音声の合
 成を するかが」

本研究では、これらの過程のうち、(a) から (d) の一次整形の処理について取り扱う。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。参考のために図 1 に書き起
 こし、図 2 に整形された文章の例を示す。

3 統計的手法による文体の整形

3.1 統計的機械翻訳のアプローチ

現在、音声認識や機械翻訳の研究において統計的な手法が広く用いられている。入力系列を
 X 、出力系列を Y とすると、これらは X を観測した際の Y の事後確率 $P(Y|X)$ を最大にする

続いて、結果の方を見ていきたいと思います。まず、こちらはアメリカ人の英語話者による語頭に R、L を含む単語、R のライトと、L のライトの調音について 1 名の例を示しました。このグラフで縦軸と横軸というのは、この調音地図のもとになっている距離マトリックスをより反映するように、MDS という手法が抽出した次元 1 と 2 というようになります。グラフ内の一つ一つの点が発話の一回一回を示していて、同じ単語の点を見やすく丸で囲んでグルーピングしております。

図 2 人手により整形された文章の例

Y を求めるという枠組みで捉えられ、ベイズ規則により次の式 (1) のように定式化される。ここで、 $P(Y)$ は系列 Y が生起する事前確率、 $P(X|Y)$ は系列 Y から系列 X が生起する条件付き確率である。右辺の分母 $P(X)$ は、 Y の決定に影響しないので無視できる。

$$\max_Y P(Y|X) = \max_Y \frac{P(Y)P(X|Y)}{P(X)} \quad (1)$$

音声認識 (鹿野, 伊藤, 河原, 武田, 山本 2001) の場合は、 X は入力音声、 Y は出力単語列となる。この場合は、音響モデルにより $P(X|Y)$ を、言語モデルにより $P(Y)$ を求めている。

機械翻訳 (Brown, Pietra, Pietra, and Mercer 1993; Garcia-Varea, Casacuberta, and Ney 1998; Wang and Waibel 1998) の場合は、 X を入力言語、 Y を出力言語として $P(Y|X)$ を最大にする Y を求めることで、入力言語 X を出力言語 Y に変換する。この場合、 $P(Y)$ は出力 Y の言語的な自然性を評価するもので、音声認識と同様に言語モデルにより求める。 $P(X|Y)$ の計算には変換モデルを仮定し、その確率を求める。変換モデルとは、入力単語はある出力単語系列 (null を含む) に対応づけられるという仮定の下で、どの単語に対応するかを文全体における相対的な位置も考慮して確率で表したものである。

2 章で述べたように、書き起こしの文体と整形した文章の文体はかなり異なっており、書き起こしの単語列と整形された文章の単語列を異なる言語とみなすことができる。そこで、本研究では書き起こしの文体を整形する際に、書き起こしを整形した文章に翻訳すると考えて、機械翻訳と同様に統計的手法を適用することを検討する。ただし、一般の機械翻訳と異なり、本研究で扱う処理では単語の順序の入れ替わりは考慮しない。

本研究では上記の変換確率と言語モデル確率を 1 対 1 で組み合わせるのではなく、言語的妥当性を重視するために言語モデル確率に重みを乗じることにする。また、言語モデル確率の値として対数尤度を用いるが、対数尤度は単語の数が多くなるほど値が小さくなるので、単語数に応じた補正を行う。これら是对数スケールで行われ、以下の式 (2) で示される仮説スコアを定義する。 (a, b) のパラメータは音声認識のデコーディングにおいて通常用いられ、 a は言語重み、 b は挿入ペナルティと呼ばれている。

$$\max_Y \{\log(P(X|Y)) + a * \log(P(Y)) + b * (Y \text{ の単語数})\} \quad (2)$$

本研究では、 Y の言語モデルとして単語 3-gram を用いる。この場合、 $Y=(y_1 \cdots y_N)$ につ

表 1 書き言葉への変換規則・確率の一部

書き言葉 (y)	話し言葉 (x)	$P(x y)$
という	っていう	0.14
	という	0.86
が	けども	0.036
	けど	0.042
	が	0.922
どのよう	どういう風	0.46
	どのよう	0.54
(～し) ている	(～し) てる	0.12
	(～し) ている	0.88

いて, $P(Y) = \prod_{i=1}^N P(y_i|y_{i-1}, y_{i-2})$ として求められる. 変換モデル確率 $P(X|Y)$ の計算においても, 同様に $X=(x_1 \cdots x_M)$ と $Y=(y_1 \cdots y_N)$ の部分列に対する確率を規定して (デフォルトは $P(x_i|y_j) = 1$ for $x_i = y_j$) その積を求めるが, この単位は可変長 (1 から数単語) である.

以下では, フィラーの削除, 書き言葉表現への変換, 助詞の挿入, 句点の挿入, 文体の統一のそれぞれの処理において, この枠組みをどのように実現するかについて説明する.

なお本研究では, 書き起こしを形態素解析した結果を入力データとして用いている. 形態素解析には, ChaSen ver2.02 を用いている. また, 話者がポーズをおいた箇所にはその情報がポーズ長とともに記録されている.

3.2 フィラーの削除

ここでは, $P(X|Y)$ は整形文 Y に対応する話し言葉 X においてフィラーが挿入される確率とする. フィラーは, 発話のどの部分にも出現する可能性があるが, 特に句読点の後によく出現する傾向がある.

ただし, 整形された文章にフィラーは一切出現しない. つまり, Y をフィラーを含む単語列であると仮定すると, $P(Y)$ の値は 0 となり, $P(Y|X)$ の値も 0 となる. これは $P(X|Y)$ の値にかかわらず書き起こしの単語列 X から全てのフィラーが削除されることを意味している.

3.3 書き言葉表現への変換

ここでは, $P(x|y)$ (x, y は 1 つ以上の単語からなる句) は書き言葉表現の話し言葉表現への変換確率と解釈される. 講演録編集者が一次整形を行う際, 文章の順序を入れ替える操作は行わないので, 元の書き起こしと講演録編集者により一次整形された文章を照合し, 句単位で対応づけを行うことで, $P(x|y)$ の値を学習・推定することができる. ただし本研究では, 正解の講演録の数が完全な学習には不十分であったため, あらかじめ人手により書き言葉から話し言葉への変換規則を作成しておき, それらに対してのみ確率を推定することとした. 作成した変換

表 2 助詞の脱落パターンと確率の一部

パターン y	脱落確率 $P(x y)$
名詞 は 名詞	0.073
名詞 を 名詞	0.032
名詞 は 動詞	0.056
名詞 を 動詞	0.042
名詞 は 形容詞	0.20
名詞 が 形容詞	0.024
名詞 は 接続詞	0.16

規則数は 64 個あり, その一部を表 1 に示す.

3.4 助詞の挿入

ここでは, $P(x|y)$ は整形された文章のある単語列パターン y に含まれる助詞が話し言葉 x において脱落する確率と解釈される. 書き起こしと講演録編集者により整形された文章を比較したところ, 次のような「品詞」「助詞」「品詞」のパターン y において助詞が脱落していた.

- 「名詞」(助詞)「名詞」
(例)「このお話し を 幹事の方から」
- 「名詞」(助詞)「動詞」
(例)「我々 は 作ってきたわけです」
- 「名詞」(助詞)「形容詞」
(例)「非常に能力 が 高くなって」
- 「名詞」(助詞)「接続詞」
(例)「これは は つまりサンプルごとの」

そこで, これらのパターンの助詞が脱落する規則をあらかじめ作成しておき, それが生起する確率 $P(x|y)$ を書き起こしと一次整形の対応づけにより推定する. 用意した変換パターン数は 13 個あり, その一部を表 2 に示す. そして, それらの助詞を挿入する場合としない場合との尤度を比較することによって助詞を挿入するかしないか, どの助詞を挿入するのか判定を行う.

3.5 句読点の挿入

「日本語話し言葉コーパス」(CSJ) の書き起こしには句読点がなく, その代わりに話者のポーズ情報が記録されている. ここでは, 整形された文章の単語列 Y に含まれる句読点が, 音声(書き起こしの単語列 X) においてポーズに変換される確率 $P(x = \text{ポーズ} | y = \text{句読点})$ を考える. なお, 本研究では句点のみを扱う. なぜなら, 句点を挿入する位置は人によらずほぼ一定であるが, 読点を挿入する位置は様々であり定量的な評価を行うのが難しいためである. 音声

認識における句点の挿入に関する研究は、これまでにも行われているが(中嶋, 山本 2001; 竹沢他 1999; Chen 1999), これらが扱っているのは旅行会話などの短い話し言葉であり, 講演のように長い話し言葉は扱っておらず, また, 主に言語モデル ($P(Y)$ に相当) による情報しか用いられていない.

文末には, 「です」「ます」などのような典型的な表現が多い反面, 話し言葉においては独特の表現が文の区切りになることがある. 文末での「〜と」と文頭での「で〜」である. その例を以下に挙げる.

(例) 「単位が使われていたと. あるいは」

「大きく違う. でそのままでは」

そこで, 本研究では $P(x_i = \text{ポーズ長} | y_j = \text{句点})$ の確率において, 以下の3通りのモデル化を考える. x_i においてはポーズの長さの情報も考慮する.

- (1) すべてのポーズを対象 ($x_i > 0$): 整形した文章 Y における句点があらゆる長さのポーズに変換されうるとする. この場合は, 書き起こし X におけるポーズがすべて句点に変換されうることになり, その判定を言語モデル確率 $P(Y)$ を用いて行う. 著者らが以前報告した講演音声の自動インデキシングの研究(長谷川, 秋田, 河原 2001)では, この考え方が用いられている.
- (2) 平均以上の長さのポーズを対象 ($x_i > \theta$): 句点の挿入箇所には, ある程度長いポーズがおかれると考えられるので, 変換確率 $P(x_i = \text{ポーズ長} | y_j = \text{句点})$ において, 句点がある閾値 θ 以上の長さのポーズに変換されうるとする. ここで, 講演では発話速度が話者によってまちまちであり, それに応じてポーズのおかれる長さも大きく異なるため, すべての講演者に対して同一の値を閾値として用いるのは適切でない. そこで, 閾値として各話者ごとの平均ポーズ長を用いることにする. 予備実験においても, 一定の値を閾値として用いる場合には, 平均ポーズ長が最良であった. この場合, 平均ポーズ長以上のポーズに対してのみ $P(Y)$ を考慮して句点に変換するか否かの判定を行う.
- (3) 表現に依存してポーズ長が変わると仮定 ($x_i > \theta$, θ が y_{j-1}, y_{j+1} に依存): 「です」「ます」などの典型的な文末表現に付随する句点はあらゆる時間長のポーズになりうるが, 書き言葉では通常文の区切り表現とならない「〜と」「で〜」, あるいは文中にも頻繁に使われる「〜た」の部分にある句点は平均ポーズ長以上の長さのポーズになると仮定する. この場合, 「ます」「です」などの後のポーズは長さに関係なく句点に変換されうるが, 「〜と」「で〜」「〜た」の部分のポーズは平均長以上のポーズに限り, 句点に変換されうることになる. この場合も, 最終的には $P(Y)$ を考慮して判定を行う.

上記3通りについて, 4章で比較・評価する.

3.6 文体の統一

話し言葉を書き言葉に変換する際に、その変換候補が複数ある場合には、どれを採用するか
の判定を言語モデル確率 $P(Y)$ に基づいて行う。ここで、使用する言語モデルを異なるもの
にすると、変換結果も異なったものになる。

本研究では2種類の言語モデルを用いることによって、それぞれ文体が「ですます」調ある
いは「である」調に統一することができるか検討する。使用する言語モデルは、講演録のコー
パスから作成されたものと、新聞記事のコーパスから作成されたものを用いる。前者の言語モ
デルを用いると「ですます」調の文体に、後者の言語モデルを用いると「である」調の文体に
統一されると期待される。

ただし、例えば「ですます」調の文でも「～であると思われます」といった表現があるため、
単純に「である」という表現をすべて「です」や「であります」という表現に変えればよいわ
けではない。ここでは、変換モデル確率において、表3の各グループ(同一行)間で相互に変換
可能としたが、変換確率は等しいものとし、その選択は $P(Y)$ に基づいて行うものとした。な
お、変換の際に動詞の語幹が変化する表現、例えば「思います」→「思う」などの表現につ
いては変換規則を用意していない。

3.7 デコーディングアルゴリズム

これまでに、講演の書き起こしから整形された文章を生成する処理とモデルについて述べて
きた。本研究では、確率 $P(Y)$ の計算に単語 3-gram モデルを用いるので、これらの処理を逐次
的に行うのではなく、統合的に行うように実装する必要がある。なぜなら、これらの処理を個
別に行うと、すぐ前後に別の処理をする必要のある表現が存在する場合、式(2)の尤度の計算
に影響を与えるからである。

したがって、前後2単語に着目する必要がある表現が存在しなくなる範囲において、その
すべての変換パターンの尤度を比較して出力単語列を決定する。その様子を図3に示す。変換
で複数の候補が生成されうることと考慮すると、可能な仮説の数は組み合わせ的に爆発するの
で、探索アルゴリズムを導入する必要がある。本研究ではビームサーチを行う。具体的には、生
成したパターン数が100を越えた場合は、そこまでの範囲で尤度を計算し、上位100個のパ
ターンのみを選択することにした。

4 実験と評価

4.1 データと実験条件

以上の手法を用いて講演の書き起こしを整形し、その評価を行った。評価データとして、CSJ
に含まれる実際に学会で行われた4講演を用いた。評価データの概要および処理時間を表4に

表 3 「ですます」調・「である」調の変換パターン

です・であります・である・だ
でした・でありました・であった・だった
します・いたします・する
しました・いたしました・した
おります・います・いる
おりました・いました・いた
あります・ございます・ある
ありました・ございました・あった

図 3 変換の仮説生成

示す．なお使用したマシンの CPU は Intel Xeon 2.8 GHz，メモリは 2 GB である．また， $P(Y)$ の計算に用いる言語モデルは，毎日新聞記事データで学習されたもの (河原他 2000) と，Web 講演録で学習されたもの (加藤，南條，河原 2000) の 2 種類であるが，4.5 節以外では Web 講演録の方を使用している． $P(X|Y)$ 推定のための講演の書き起こしと整形文章のデータは評価データ以外の CSJ の 18 講演を用いている．次節以降，行った実験の結果と評価について述べる．なお，3.2 節のフィラーの削除は完全に機械的に行える．また，3.3 節の書き言葉表現への変換もおおむね行えた．

4.2 デコーディングパラメータについて

3.1 節で述べた式 (2) の 2 つのパラメータについてまず検討した．評価尺度として，句点の挿入および助詞の挿入についての F 値の平均を用いる．句点挿入のモデルは最もよいものを用いている．様々な言語重み・挿入ペナルティの値について行った実験結果を図 4 に示す．結果と

図 4 パラメータの種々の値に対する句点および助詞挿入の F 値の平均

表 4 評価データの概要および処理時間

	講演		整形文	処理時間 (sec)
	時間	サイズ	サイズ	
A01M0035	28 分	5557 語	5378 語	17.42
A01M0007	30 分	3899 語	3802 語	13.65
A01M0074	13 分	2509 語	2451 語	5.97
A05M0031	27 分	5371 語	4854 語	19.99

して (言語重み, 挿入ペナルティ)=(5,8) の時に F 値が最大となった。したがって, 以降の実験はこれらの値を用いて行う。表 4 の処理時間もこの場合の計測値である。

4.3 句点の挿入

3.5 節で述べた 3 通りの句点挿入の実験結果を表 5 に示す。

ポーズ長制限なしで挿入する場合は適合率が低い。その主な原因は, 「～と」「～た」の後と「で～」の前のポーズが誤って句点に変換されることが多いためである。「ます」「です」などの後のポーズが誤って句点に変換されることはなかった。「～た」という表現は文末表現であるが, 同時に文中にもよく使われるため, 例えば「～使われてきた <pause> データベースが～」→「～使われてきた。データベースが～」のような誤りが起こる。また, 「～と」の場合は, 例えば, 「～しようと思うと <pause> このように～」→「～しようと思う。このように～」のような場合に誤ってポーズが句点に変換された。

一方, 平均ポーズ長以上に挿入する場合においては再現率が低い。その主な原因は, 「ます」「です」などの, 通常書き言葉で文末表現になるものの後で, ポーズ長が短い箇所が対象外になったためである。したがって, 「です」「ます」などの典型的な文末表現部分にある句点はあらゆる時間長のポーズになりうるが, 「～と」「で～」 「～た」の部分にある句点は平均ポーズ長以

表 5 句点挿入の実験結果

	再現率	適合率	F 値
ポーズ長制限なし	309/371 (83.2%)	309/410 (75.4%)	0.791
平均ポーズ長以上	239/371 (64.4%)	239/255 (93.7%)	0.763
表現に依存して変化	283/371 (76.3%)	283/306 (92.3%)	0.835

表 6 助詞挿入の実験結果

$P(x y)$ の値	再現率	適合率	F 値
統計的に推定	37/47(78.7%) (プロの講演録と一致)	68/123(55.3%) (意味的に妥当)	0.650
	42/47(89.4%) (意味的に妥当)	81/123(65.9%) (許容範囲)	0.759
あらかじめ 全て 1 に設定	33/47(70.2%) (プロの講演録と一致)	83/187(44.4%) (意味的に妥当)	0.544
	42/47(89.4%) (意味的に妥当)	109/187(58.3%) (許容範囲)	0.706

上の長さのポーズになるとするモデルを導入した。その結果、再現率・適合率とも高い精度を得ることができた。

4.4 助詞の挿入

次に、3.4 節で述べた助詞の挿入に関する評価を行った。助詞が脱落している箇所は 4 講演で計 47 箇所あった。それらに対して、

- ・プロの編集者が作成した講演録において挿入されている助詞
- ・上記に含まれないが、意味的に妥当な助詞

を正解としてそれぞれ再現率を評価した。

また、挿入されたすべての助詞に対して

- ・意味的に妥当である
- ・許容範囲である

の 2 段階で適合率を評価した。

これらの評価基準に対して、統計的手法の有効性を調べるために、3.4 節で述べたパターンの変換確率 $P(x|y)$ に関して次の 2 つの場合の比較を行った。

(1) $P(x|y)$ の値として、統計的に推定したものをを用いる

(2) あらかじめ $P(x|y)$ の値を全て 1 に設定する

結果を表 6 に示す。適合率で大きな差が見られ、許容範囲のものまで正解にした場合において $P(x|y)$ も用いた場合の方が 7.6% 向上している。さらに、挿入された絶対数も約 2/3 に減少しており、実際に生成されたテキストの読みやすさという観点において大きく改善されている。

表 7 言語モデルの違いの影響

$P(Y)$ のモデル	「ですます」調	「である」調
新聞記事モデル	52.3 %	47.7 %
講演録モデル	81.1 %	18.9 %

誤って助詞が挿入された箇所の多くは「名詞」「名詞」のパターンであった。学会講演において用いられる専門用語の多くが複合名詞であるため、形態素解析を行うと「名詞」「名詞」と分解され、助詞の挿入箇所の候補になる。そこで、「名詞」「名詞」と連続しているもののうち、3回以上出現すればその箇所は専門用語であると判断するようにした。これにより多くの専門用語は助詞の挿入箇所の候補にならなくなったが、1~2回しか使われていない専門用語や、人名・機関名などの固有名詞が複数の形態素に分解された結果、助詞が挿入されることがあった。例えば「データ ベース」や「東京 大学」などである。参考までに、誤り箇所からこれらの複合名詞の箇所を除いて集計すると、適合率は許容範囲のものまでを正解とした場合で 79.4%(81/102) となった。

4.5 言語モデルの使いわけ

次に、3.6 節で述べた言語モデルの違いによる文体の変化について評価を行った。結果を表 7 に示す。

入力データでは「ですます」調が 81.5%、「である」調が 18.5%であった。

講演録言語モデルを用いて整形した場合は 81.1%が「ですます」調に、新聞記事言語モデルを用いて整形した場合は 47.7%が「である」調になった。入力データが基本的に「ですます」調であるため、講演録モデルを用いた場合はほとんど変換されていない。一方、新聞記事モデルを用いた場合は、「である」調の割合が 30%程度増えている。しかし、全体の半分程度しか「である」調にならないのは、3.6 節でも述べたように、今回の変換モデルでは動詞の語幹が変化する表現に対応していないためである。

4.6 規則ベースの手法との比較

最後に、完全な規則ベースによる手法を用いて実験を行い、本論文で提案した統計的手法との比較を行う。ここでは、句点の挿入および助詞の挿入について評価した。

句点の挿入については次の規則を用いた。3.5 節及び 4.3 節における考察に基づいて、ポーズの閾値についても最も妥当なものにした。

- ・「です」「ます」などの典型的な文末表現の後のポーズは全て句点に変換
- ・「～と」「～た」「で～」の部分のポーズは、平均ポーズ長以上なら句点に変換

また助詞の挿入については、コンテキストを考えずに言語モデルの学習テキストにおいて出現頻度が高いものを挿入規則として抽出した。例えば学習テキスト中において、「名詞 助詞 動

表 8 規則ベースと提案手法との比較 (句点挿入)

	再現率	適合率	F 値
規則ベース	301/371(81.1%)	301/392(76.8%)	0.789
提案手法	283/371(76.3%)	283/306(92.3%)	0.835

表 9 規則ベースと提案手法との比較 (助詞挿入)

	再現率	適合率	F 値
規則ベース	26/47(55.3%)	60/91(65.9%)	0.601
	(プロの講演録と一致)	(意味的に妥当)	
	31/47(66.0%)	67/91(73.6%)	0.696
	(意味的に妥当)	(許容範囲)	
提案手法	37/47(78.7%)	68/123(55.3%)	0.650
	(プロの講演録と一致)	(意味的に妥当)	
	42/47(89.4%)	81/123(65.9%)	0.759
	(意味的に妥当)	(許容範囲)	

続いて結果の方を見ていきたいと思います。まずこちら英語話者アメリカ人の英語話者による語頭に R L を含む単語 R のライトと L のライトの調音について一名の例を示しました。縦軸とこのグラフでいう縦軸と横軸というのはこの調音地図というのの元になっている距離マトリックスをより反映するように MDS という手法が抽出した次元一と二というようになりますで一つ一つの点のグラフ内の一つ一つの点が発話の一回一回を示していて同じ単語の点を見易いように見易く丸で囲んでグルーピングしています。

図 5 提案手法による整形結果

詞」の並びで出現頻度が最大となる助詞は「が」であったため、「名詞 動詞」となっている箇所には、その前後のコンテキストは考えずに、出現回数が最大の「が」を挿入するという規則を抽出した。

以上の手法により行った実験結果と提案手法との比較を表 8、表 9 に示す。句点の挿入に関しては、規則ベースの方が再現率は若干高いものの、適合率が大幅に低下しており、誤った挿入がおおよそ 4 倍に増えている。助詞の挿入に関しては、F 値が 0.759 から 0.696 に低下している。以上より、提案手法の有効性が確認された。

4.7 整形結果の具体例

ここでは、提案手法により図 1 で示した書き起こしを整形した具体例を図 5 に示す。なお、本来入力形態素解析結果であるが、図 1 では見やすくするために活用形や形態素番号などの情報は省いて示している。句点の挿入はおおむね正しく行えており、また、「結果の方見ていきたい」に「を」が挿入されているなどの整形が行えている。しかし、「なりますで1つ1つ」の所では、ポーズが本来の文末位置とは少しずれた所にあるため、句点が正しく挿入され

ていない。また、この場合、「調音 地図」というパターンが3回以上出現しているため、これを専門用語であるとみなして助詞の挿入箇所の候補にはしていない。ただし、図2と比較すると言い直し部分の削除などを行っていない。

5 結論

統計的な機械翻訳の考え方に基づいて文体の整形を自動的に行う手法を提案した。行った処理は、フィラーの削除、句点の挿入、助詞の挿入、書き言葉表現への変換及び文体の統一である。ビームサーチを導入してこれらを統合的に行い、実際の講演の書き起こしを整形された文章に変換した。正解の文章として講演録編集者によって一次整形されたものを用いて、句点の挿入と助詞の挿入に関して定量的な評価を行った。句点の挿入においてはF値で0.835、助詞の挿入においてはF値で0.759という高い精度が得られた。また、実験的評価により、規則ベースの手法に比べて統計的なアプローチが有効であること、及び変換モデル確率 $P(x|y)$ の効果が示された。

今後の課題としては、書き言葉表現への変換に関して人手により変換規則を作成するのではなく、大規模なコーパスから規則を抽出して変換確率 $P(x|y)$ を推定することや、文体の統一に関して不十分であった箇所に対応することが挙げられる。また、今回は正しい書き起こしを用いて評価を行ったが、今後は音声認識結果に適用していく予定である。

謝辞

本研究は、開放的融合研究『話し言葉工学』プロジェクトの一環として行われた。東京工業大学の古井貞熙教授をはじめとして、ご協力を頂いた関係各位に感謝いたします。

参考文献

- Brown, P., Pietra, S., Pietra, V., and Mercer, R. (1993). "The Mathematics of Statistical Machine Translation : Parameter Estimation." In *Proc. Computational Linguistics*, Vol. 19.
- Chen, C. (1999). "Speech Recognition with Automatic Punctuation." In *Proc. Eurospeech*.
- Furui, S., Maekawa, K., and Isahara, H. (2000). "Toward the realization of spontaneous speech recognition - introducing of a japanese priority program and preliminary results -." In *Proc. ICSLP*, Vol. 3.
- Garcia-Varea, I., Casacuberta, F., and Ney, H. (1998). "An Iterative, DP-Based Search Algorithm For Statistical Machine Translation." In *Proc. ICSLP*, Vol. 4.

- 長谷川将宏, 秋田祐哉, 河原達也 (2001). “談話標識の抽出に基づいた講演音声の自動インデキシング.” 情報処理学会研究報告, 01-SLP-36-6.
- 堀智織, 古井貞熙 (1999). “話題語と言語モデルを用いた音声自動要約法の検討.” 情報処理学会研究報告, 99-SLP-29-18.
- 堀智織, 古井貞熙 (2000). “係り受け SCFG に基づく音声自動要約法の改善.” 電子情報通信学会技術報告, SP2000-116.
- 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄 (2001). 「音声認識システム」. オーム社.
- 加藤一臣, 南條浩輝, 河原達也 (2000). “講演音声認識のための音響・言語モデルの検討.” 信学技報, SP2000-97, NLC2000-49 (SLP-34-23).
- 加藤直人 (1998). “ニュース文要約のための局所的要約知識獲得とその評価.” 情報処理学会研究報告, 98-NL-126-10.
- 河原達也他 (2000). “日本語ディクテーション基本ソフトウェア (99 年度版) の性能評価.” 情処学研報, SLP-31-2, NL-137-7.
- 村田真樹 (2002). “diff を用いた言語処理 — 便利な差分検出ツール mdiff の利用 —.” 言語処理学会誌, 9 巻 (2 号), pp. 91-110.
- 村田真樹, 井佐原均 (2001). “同義テキストの照合に基づくパラフレーズに関する知識の自動獲得.” 情報処理学会研究報告, 2001-FI-61, 2001-NL-142.
- 中嶋秀治, 山本博史 (2001). “音声認識過程での発話分割のための統計的言語モデル.” 情報処理学会論文誌, 42 巻 (11 号), pp. 2681-2688.
- 中沢正幸, 遠藤隆, 古川清, 豊浦潤, 岡隆一 (1996). “音声波形からの音素片記号系列を用いた音声要約と話題要約の検討.” 電子情報通信学会技術報告, SP96-28.
- 竹沢寿幸他 (1999). “発話単位の分割または接合による言語処理単位への変換手法.” 言語処理学会誌, 6 巻 (2 号), pp. 83-95.
- 若尾孝博, 江原暉将, 白井克彦 (1998). “短文分割を利用したテレビ字幕自動要約.” 情報処理学会研究報告, 98-NL-126-9.
- Wang, Y. and Waibel, A. (1998). “Fast Decoding For Statistical Machine Translation.” In *Proc. ICSLP*, Vol. 6.

略歴

- 下岡 和也: 2002 年 京都大学工学部情報学科卒業. 現在, 同大学院情報学研究
科知能情報学専攻修士課程在籍. 音声認識, 自然言語処理の研究に従事.
- 南條 浩輝: 1999 年 京都大学工学部情報学科卒業. 2001 年 同大学院情報学研
究科修士課程了. 現在, 同博士後期課程在学中. 音声認識・理解の研究に従
事. 情報処理学会, 日本音響学会, 電子情報通信学会各会員.
- 河原 達也: 1987 年 京都大学工学部情報工学科卒業. 1989 年 同大学院修士課

程修了. 1990 年 同博士後期課程退学. 同年 京都大学工学部助手. 1995 年 同助教授. 1998 年 同大学情報学研究科助教授. 2003 年 同大学学術情報メディアセンター教授. 現在に至る. この間, 1995 年から 96 年まで 米国ベル研究所客員研究員. 1998 年から ATR 客員研究員. 1999 年から 国立国語研究所非常勤研究員. 2001 年から 科学技術振興事業団さきがけ研究 21 研究者. 音声認識・理解の研究に従事.

京大博士 (工学). 1997 年度 日本音響学会粟屋賞受賞. 2000 年度 情報処理学会坂井記念特別賞受賞. 情報処理学会連続音声認識コンソーシアム代表. IEEE SPS Speech TC 委員. 情報処理学会, 電子情報通信学会, 日本音響学会, 人工知能学会, 言語処理学会, IEEE 各会

(2003 年 5 月 21 日 受付)

(2003 年 9 月 17 日 再受付)

(2003 年 12 月 1 日 再々受付)

(2004 年 1 月 27 日 採録)