

話し言葉コーパスにおける文の切り分けと重要文抽出

野畑 周[†] 関根 聡^{††} 内元 清貴[†] 井佐原 均[†]

[†] 独立行政法人 通信総合研究所 けいはんな情報通信融合研究センター
〒 619-0289 京都府相楽郡精華町光台 2-2-2

^{††} Computer Science Department, New York University
715 Broadway, 7th floor, New York, NY 10003, USA

E-mail: [†]{nova,uchimoto,isahara}@crl.go.jp, ^{††}sekine@cs.nyu.edu

あらまし 本研究では、講演録コーパスにおいて文境界の検出と重要文抽出という二つの課題について実験を行った結果を報告する。文境界の検出課題では、ポーズ情報と人手で作成したパターンを用いた実験を行い、ポーズ情報と人手で作成したパターンを併用することによってパターンによる検出精度を F 値で 92%程度にまで向上させることができた。重要文抽出課題では、被験者による結果の比較から、話し言葉コーパスにおいて複数の被験者の結果から統一された正解を作成するのは容易でないことを示し、また重要文抽出システムの出力の評価から、正しい文境界が与えられれば、小さい要約率(10%)においては話し言葉コーパスにおける重要文抽出の評価が書き言葉における評価に匹敵することを示した。

キーワード 講演録、文境界検出、重要文抽出、自動要約

Sentence Segmentation and Sentence Extraction for Lecture Speech Corpus

Chikashi NOBATA[†], Satoshi SEKINE^{††}, Kiyotaka UCHIMOTO[†], and Hitoshi ISAHARA[†]

[†] Keihanna Human Info-Communication Research Center
Communications Research Laboratory, Japan

2-2-2 Hikaridai Seika-Cho Soraku-Gun Kyoto 619-0289 Japan

^{††} Computer Science Department, New York University
715 Broadway, 7th floor, New York, NY 10003, USA

E-mail: [†]{nova,uchimoto,isahara}@crl.go.jp, ^{††}sekine@cs.nyu.edu

Abstract We present the experiments of two tasks for a speech corpus: sentence segmentation and sentence extraction. In the sentence segmentation task for the corpora of lecture speech, our experiments showed that pause information alone was not sufficient for sentence segmentation, but useful for improving the performance of manually created lexical patterns to about 92% in F-measure. In the sentence extraction task, we showed that a unique key data was difficult to create because the agreements among the annotators were not high, and that if the results of sentence segmentation has no errors, the performance of our sentence extraction system for speech corpus was comparable to that for written documents in a small compression ratio (10%).

Key words Lecture Speech, Sentence Segmentation, Sentence Extraction, Automatic Summarization

1. はじめに

本研究では、書き言葉のための要約システムを話し言葉コーパスに適用し、その結果を人手で作成した結果と比較して評価を行なった。我々はこれまで、重要文抽出を基にした、新聞記事などのような書き言葉を対象とした要約システムを開発してきた。重要文抽出とは、文章中の各文について、いくつかの評価尺度を用いて重要度を求め、その結果に基づいて重要と思われる文を抽出する手法であり、書き言葉の要約において用いられる主な手法の一つである [1] ~ [5]。一方、話し言葉の要約を行う研究としては、堀・古井 [6] が、言語モデルや確率的 CFG に基づく重要度などを用いて、コーパスから重要な単語の連鎖を取り出すことによって要約を作成している。彼らは、人手による書き起し結果だけでなく、自動音声認識の結果に対しても手法を適用し、結果を評価している。また、伊藤ら [7] は、重要文抽出の手法をテレビ番組の講演録に適用して 5 文からなる短い要約と 20 文からなる長い要約を生成し、長い要約では手がかり表現を用いた手法が統計的な手法よりも有効であることを示している。本研究では、書き言葉の要約と話し言葉の要約の双方について評価結果を比較したこと、文境界の検出も含めた実験を行ったという点が彼らの研究と異なっている。

話し言葉のコーパスに対して重要文抽出に基づいた要約を行なうためには、コーパスが文に区切られていることが必要である。しかし、話し言葉のコーパスには、必ずしも文の境界が与えられているとは限らない。本研究では書き起しコーパスのみを入力データとしたため、韻律情報は用いていないが、基本周波数 (F0) などの韻律情報に基づいて、音声データから文境界を検出する研究が行われている [8], [9]。話し言葉のコーパスのみを用いて文の境界を与える研究としては、Stevenson and Gaizauskas [10] が英語の音声認識結果に対して、文境界を検出する実験を行ない、人間による結果とメモリーベースのシステムとの結果を比較し、両者の結果において、英語では字種情報の有無 (大文字小文字の区別) が性能に大きく関係することを示した。日本語の書き起しコーパスについては、長谷川ら [11] が、講演に談話ラベルを自動的に付与するための前処理として、単語 3-gram とポーズ情報から文境界の検出を行っており、4 講演録に対して再現率 98.1%、適合率 75.2%の

検出結果を得ている。彼らは自動インデキシングのために再現率を向上させることを重視しており、我々は再現率と適合率の双方を向上させようとしている点で立場が異なる。

本研究では、日本語書き起しコーパスについて文境界の検出を行い、さらにその結果を用いて重要文抽出を行なった結果を報告する。以下では、まず実験に用いたコーパスについて説明し (2. 節)、次に文境界の検出について、ポーズ情報単独での実験と人手で作成したパターンによる実験結果の双方について述べ (3. 節)、続いて重要文抽出の手法とその実験結果について述べる (4. 節)。

2. 話し言葉コーパス

本研究の実験に用いたコーパスは、国立国語研究所、東京工業大学、通信総合研究所の 3 団体が共同で構築作業をすすめている、CSJ コーパス (Corpus of Spontaneous Japanese) [12] から得たものである。CSJ コーパスは、学会講演など、モノログを対象として収集・構築されているコーパスである。本研究では、この CSJ コーパスのうち、1999 年日本音響学会秋季研究発表会 (AS99) の講演から 35 講演、2000 年言語処理学会年次大会 (NL00) の講演から 25 講演の計 60 講演を取り出して用いた。文境界の検出、重要文抽出の双方の実験ではともに、60 講演のうち 50 講演 (AS99 から 30 講演、NL00 から 20 講演) をトレーニングデータとし、10 講演 (AS99 から 5 講演、NL00 から 5 講演) をテストデータとして用いた。システムの出力を評価するために、三人の被験者に 60 講演全てについて文境界の検出と重要文抽出のデータ作成をしていただいた。被験者が作業を行ったときには、書き起こしのみを対象とし、ポーズ情報や音声は示されていない。文境界においては、各被験者の結果を言語学の専門家が統合して単一の正解データを作成した。正解データにおける、各講演の平均文数は 68.7 (4123/60) 文であった。一方重要文抽出においては、正解データを作成することは行わず、被験者の個々の抽出結果を実験に用いた。

3. 文境界の検出

本節では、話し言葉コーパスに対して、まずポーズ情報単独では文境界の検出がどの程度行えるかを示す。次に、人手で作成したパターンによって文境界を検出した結果の評価を行う。

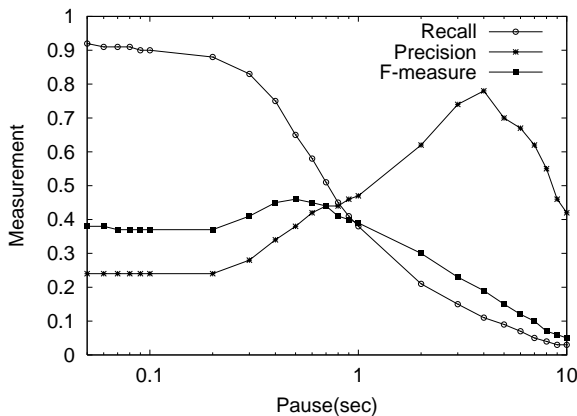


図1 ポーズ長に対するしきい値と文境界検出の評価との関係

3.1 ポーズ情報による文境界の検出

ポーズ情報による文境界の検出は、ポーズ長にしきい値を設定し、しきい値以上の長さのポーズがある所は文境界と見なすことによって行った。しきい値の値域は、10 ミリ秒から 10 秒までの範囲に設定した。図1のグラフは、しきい値の変化と、それに伴う文境界検出の評価結果を示したものである。x 軸はポーズ長に対するしきい値であり、y 軸は各評価基準の値(百分率)を示す。ここでは、再現率 (Recall)、適合率 (Precision)、F 値 (F-measure) の3種類の評価基準を用いている^(注1)。

グラフから、ポーズ長に対するしきい値を変えることで再現率と適合率のどちらか一方のみを上げることは可能だが、その両者、すなわち F 値は最大でも 50%より向上させることはできないことが分かる。

3.2 パタンによる文境界の検出

ポーズ情報の他に文境界の検出を行うための情報として、本研究ではパタンを人手で作成した。パタンは、単語の連続と、文境界であるか否かを示すフラグから成る。例えば、以下に示すパタンは、「~です/それで~」「~しますように~」の表現について

(注1): 各評価基準は、それぞれ以下のように定義される。再現率は、正解データ中の文境界の数 (G) のうち、正しく認識された文境界の数 (C) がどれだけであったかを示す。適合率は、文境界とみなされた個所の数 (S) のうち、正しく認識された文境界の数 (C) がどれだけであったかを示す。F 値は、両者の調和平均である。それぞれの評価基準を式で示せば以下ようになる。

$$\begin{aligned} \text{再現率 (R)} &= C / G \\ \text{適合率 (P)} &= C / S \\ \text{F 値} &= 2PR / (P + R) \end{aligned}$$

個々の講演録について各評価の値を求めた後、全講演録について平均値の百分率をとったものがグラフに示された値である。

文境界を調べ、前者には「~です/それで~」と文境界 (/) を挿入し、後者には文境界がないとみなして文境界を挿入せず、次の表現に進む。アスタリスク (*) は、その部分に当たる単語が任意で良いことを示す：

T	*	です	それで
F		します	ように

パタンの一般形を記述すると以下ようになる：

T/F	w_{eos-1}	w_{eos}	w_{eos+1}
-----	-------------	-----------	-------------

パタンは、文末の候補 (w_{eos}) の直後に文境界を挿入するかどうかを判定する。文末の候補とその前後の単語の表現がマッチしたとき、フラグが T ならば、文境界が挿入される。フラグが F ならば、文境界でないとマッチした表現は無視される。本システムは、形態素解析の前の段階で文境界を検出することを目的としているので、品詞や単語区切りの情報を前提としていない。従って、 $w_{eos(\pm 1)}$ は、単語単位であることを意図しているが、形態素解析によって得られる語の単位とは必ずしも一致しない。 $w_{eos(\pm 1)}$ は単語だけでなく、単語の集合を示す名前を表すこともできる。単語の集合も、パタンと同様に人手で記述したものである：

w_i	→	さて、しかし、しかも、したがいまして したがって、じゃ、そいから そいでは、そして、それから、...
-------	---	--

単語の集合名を含むパタンは、文境界の検出を行う前に、その集合に含まれる各単語に展開される。展開後のパタンの数は 3776 個であった。

我々は、さらにポーズ情報を考慮したパタンを用意した。これは、字面からは判断しにくい文境界を検出することを意図したものである。例えば「~ですけれども」という表現は、逆接を示す場合(「~です/けれども」)、次の話題への展開を示す場合(「~ですけれども/~」)、また単に挿入的な表現(表現中に文境界は存在しない)の場合の三通りに解釈され、文の境界をどのように認定すべきかはそれぞれ異なる。このような場合を区別するためにはポーズ情報が有効であると考え、ポーズ情報を併用するパタンを導入した。新たに加えられたパタンは、文末の候補とその前後の単語連続がマッチし、かつ文末の候補の直後にポーズ情報がある場合に適用される。

T/F	w_{eos-1}	w_{eos}	[pause]	w_{eos+1}
-----	-------------	-----------	---------	-------------

新たに加えられたパタンの展開後の数は 171 個であ

表 1 パタンによる文境界検出の評価結果

手法	再現率	適合率	F 値
パタンのみ	90.6	89.3	89.7
パタン+ポーズ	91.8	92.9	92.0

表 2 被験者とシステムの文境界検出結果

コーパス	被験者	システム
テスト (AS99)	93.6	87.0
テスト (NL00)	98.9	97.0
テスト (全部)	96.2	92.0

り、パタンの総数は 3947 個となった。

パタンによる文境界検出の結果を表 1 に示す。数値は各講演録の評価結果の平均である。トレーニングデータはパタン作成に用いたデータ 50 記事であり、テストデータは評価用に残した 10 記事である。パタン単独での文境界検出結果の評価は、F 値で 89.7% であり、パタンとポーズ情報を併用することで、文境界の検出結果をさらに約 2.0% 向上させることができた。すなわち、ポーズ情報を加えることによって、約 22% の文境界誤りを改善したことになる。

被験者とシステムとの文境界検出結果の比較を表 2 に示す。被験者の評価値は、三人の被験者のうち一人のデータを正解とみなしたときの他二人のデータの評価 (F 値) を、正解とみなす被験者を変えてそれぞれ求め、その平均をとったものである。被験者とシステムの結果を比較すると、テストデータでは、AS99 の講演録において両者の結果の間に大きな開きがあることが分かる。トレーニングデータでは AS99・NL00 のどちらにおいても被験者との差は 2% 前後におさまっていた。

テストデータの各講演録において、パタンによる文境界検出の結果を示したものが表 3 である。上位 5 個の結果は、再現率・適合率ともに 95% を超えている。トレーニングデータでも、再現率・適合率の中央値はともに 95% を超えており (再現率: 96.4%、適合率: 96.2%)、半分以上の講演録においては、文境界の検出精度は 95% を超えている。

うまく文境界が見つからなかった例を見てみると、特定の講演録中に、パタン作成時に想定していない表現が集中して見られた。例えば、AS99-3・AS99-4 では、パタンとして記述した文末候補の数が不足していたために、文境界として検出すべき表現が検出できていないものが目立った。

が間違っていたと ではないだろうか	/	いわゆる それから
----------------------	---	--------------

表 3 テストコーパス中の各講演録に対する文境界検出の評価結果

講演録	再現率	適合率	F 値
NL00-1	100.0 (82/82)	100.0 (82/82)	100.0
NL00-2	100.0 (75/75)	98.7 (75/76)	99.3
AS99-1	100.0 (53/53)	98.1 (53/54)	99.1
NL00-3	97.4 (75/77)	98.7 (75/76)	98.0
NL00-4	95.6 (86/90)	95.6 (86/90)	95.6
NL00-5	88.0 (81/92)	96.4 (81/84)	92.0
AS99-2	86.4 (38/44)	88.4 (38/43)	87.4
AS99-3	81.8 (45/55)	86.5 (45/52)	84.1
AS99-4	73.6 (39/53)	97.5 (39/40)	83.9
AS99-5	95.3 (61/64)	69.3 (61/88)	80.3

一方、AS99-5 では、パタンの記述では文末表現と体言に係る表現との区別が不十分であったために、文境界でない個所を文境界とみなしたものが目立った。

ボウスによります 今申し上げました	(/)	MRI による研究ですと 三種類の
----------------------	-----	----------------------

これらの誤りは、パタンの数・記述が充分でないこととともに、話者の話し方によって文境界前後の表現が異なることから生じている。これを解消するためには、パタンの改良とともに、話者のタイプに応じて適用するパタンを変更することが有効と考えられる。また、書き言葉向けの形態素解析システムを暫定的に用いて、パタンの記述に用いる形態素解析結果を利用可能にする、また音声データから韻律情報を得てパタンと併用するなど、文境界を検出するための情報を追加することも今後の課題として検討したい。

4. 重要文抽出

前節では、文境界を検出する課題について、正解データと機械的な手法による結果との比較を行った。本節では、まず、重要文抽出に用いたシステムが用いた評価関数と、それらの関数の値に対する重み付けの方法について述べ、次いで実験結果について述べる。

4.1 重要文の評価尺度

重要文抽出とは、与えられた文書が重要だと考えられる文を文章から抜き出して、要約を作成する手法であり、書き言葉の自動要約に用いられている主な手法の一つである。重要文抽出システムは、文書中のどの文が重要であるかを判断するために、文の長さや文章中での位置、文中の単語の頻度などの評価尺度から文の重要度を見積もる。そして得られた重要度の高い文を与えられた長さになるまで出力し、

要約を生成する。

本システムでは、文の位置や文長などの各評価尺度について、さらに複数の関数を定義し、トレーニングデータを用いてそのうちの一つを選択する。各評価尺度の用い方を予め複数定義しておくことで、対象とする文書の性質に合うように評価尺度の用い方を変更できることを意図している。システムがもつ評価関数については[13]に述べられているので、ここでは説明を簡潔にするために、各評価尺度において選択された関数についてのみ述べる。

4.1.1 文の位置

文の位置情報に基づく関数としては、この関数は、先頭か末尾に近い文ほど重要である、という仮定に基づいた以下の関数が選択された。この関数は、先頭からの文の位置の逆数と末尾からの文の位置の逆数のうち、値が大きい方を与える：

$$Score_{pst}(S_i) = \max\left(\frac{1}{i}, \frac{1}{n-i+1}\right)$$

ここで n は講演録中の文の数を示す。システム中には、他に講演録の先頭の文に大きい値を与える関数も用意しているが、今回は、トレーニングデータを用いた実験では選択されなかった。

4.1.2 文の長さ

各文の文字数に基づく関数については、極端に短い文は重要文として選択されることが非常に稀であるという観測事実に基づいて、以下のものが用いられた。

$$Score_{len}(S_i) = 0 \quad (L_i \geq C \text{ のとき}) \\ = L_i - C \quad (\text{それ以外})$$

この関数は、長さ (L_i) が一定の値 (C) より短い文にはペナルティとして負の値を与えるものである。他に、長い文に大きな値を与える関数も用意されていたが、選択されなかった。評価の際には、トレーニングデータを用いた実験から一定値 C を 20 (文字) とした。

4.1.3 tf*idf

この関数は、講演録中の単語の頻度 (tf) と、その単語がある文書集合の中で現れた文書の数 (df) を用いて $tf*idf$ 値を計算し、文のスコア付けを行う。この関数を用いる意図は、「講演録に特有な単語をより多く含む文は、その講演録においてより重要だ」とみなす仮定に基づく。本システムでは、単語の切り分けには JUMAN [14] を用い、 $tf*idf$ 値を与える単語を、時相名詞や副詞的名詞を除いた名詞に限定し

た。 $tf*idf$ 値の定義としては、情報検索課題において効果を挙げている [15] の定義に基づき、以下の式を用いた：

$$tf*idf(w) = \frac{tf(w)}{1 + tf(w)} \log \frac{DN}{df(w)}$$

ここで DN は与えられた文書集合中の文書数である。ここでは、1994年と1995年の毎日新聞の記事を文書集合として用いた。また、文のスコアは、以下の式のようにそれらの和によって与えられる：

$$Score_{tf*idf}(S_i) = \sum_{w \in S_i} tf*idf(w)$$

4.1.4 見出し

本実験で用いた各講演録は、予稿との対応がとれており、予稿の情報を用いることが可能である。予稿の情報を用いて、講演録中の重要な文を抽出する精度を上げることができると考えられる。この関数は、予稿の情報を用いる方法の一つとして、各講演録に対応する予稿の見出しに含まれる単語に対する $tf*idf$ 値を用いて、文のスコア付けを行なう。これは「見出しと類似している文は重要である」という仮定に基いている。注目する単語は、前節の $tf*idf$ を用いた関数と同様に、時相名詞や副詞的名詞を除いた名詞に限定している。文 (S_i) 中の全名詞について、その名詞が見出し (H) に含まれていれば、その $tf*idf$ 値を文のスコアに加算する。文のスコアを与える式を以下に示す：

$$Score_{hl}(S_i) = \frac{\sum_{w \in H \cap S_i} tf*idf(w)}{\sum_{w \in H} tf*idf(w)}$$

4.2 重み付け

本システムでは、先に述べた各評価関数の値の和を文の重要度とする。各評価関数 ($Score_j$) の値には重み付け (α_j) を与え、それらの和が各文 (S_i) の重要度となる。

$$Total-Score(S_i) = \sum_j \alpha_j Score_j(S_i)$$

重みの値は、トレーニングデータから求めた。各重みの値域を予め定めておき、その値域内でトレーニングデータに対する重要文抽出の評価を繰り返し行い、最適な結果を与える重みの値を記録した。各被験者データごとに重み付けの値を求めたが、値は被験者間でほとんど変化がなかったため、以下に述べ

表 4 トレーニングデータを用いて得られた重み付けの値

評価尺度	重みの値
文の位置	1000
文長	100
tf*idf	5
見出し	100

る実験では、一種類の重み付けのみを用いた。トレーニングデータから得られた重みの値を表 4 に示す。

4.3 実験と考察

本節では、重要文抽出の結果について述べる。まず、被験者間の重要文抽出結果について比較し、次いで被験者の結果とシステムによる結果とを比較する。

4.3.1 被験者間の比較

三人の被験者（以下各々 A, B, C とする）が作成した、60 講演録に対する重要文抽出の結果を Kappa 値を用いて比較した。被験者の重要文抽出データを比較する際には、文境界は全て正解データのものに統一されている。表 5 は、二人の被験者の各組合せについて、重要文抽出データの Kappa 値を求めた結果を示している。表中の数値は 60 講演録についての Kappa 値を平均したものである。Kappa (κ) 値とは、二つのデータの間で、偶然に一致する割合を除いた一致度を示す指標であり、以下の式で定義される [16] :

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

$P(A)$ は実際に両データ間で一致した割合を、 $P(E)$ は偶然に両データ間で一致する割合を表す。重要文抽出のデータでは、文数 S_n の記事に対して、一方の被験者が抽出した重要文の数を S_a 、もう一方の被験者が抽出した重要文の数を S_b とし、両被験者がともに抽出した重要文の数 S_c とすれば、 $P(A) \cdot P(E)$ は以下のように定義される :

$$P(A) = \frac{S_c}{S_n}$$

$$P(E) = \frac{S_a S_b}{S_n} + \frac{(1 - S_a)(1 - S_b)}{S_n}$$

表 6 は、[17] によって示された Kappa 値を解釈するための分類表である。Kappa 値が高いほど、両データの一致度が高く、データの内容に信頼性があると考えられる。[18] における考察では、Kappa 値が 0.7 未満の場合には被験者同士の関連を示すことは困難な場合が多いと述べられている。表 5 の結果を表 6

表 5 3人の被験者に対する Kappa 値

A ↔ B	B ↔ C	C ↔ A	平均
0.602	0.527	0.560	0.563

表 6 Kappa 値の分類表

Kappa value	Reliability
< 0	Poor
.00 - .20	Slight
.21 - .40	Fair
.41 - .60	Moderate
.61 - .80	Substantial
.81 - 1.0	Near perfect

に基いて解釈すれば、被験者間の一致度は“Moderate”にあたり、それほど信頼度が高くないことが分かる。これは、講演録の重要文抽出という課題において、統一された正解を作ることが難しいことを示しているといえる。従って、重要文抽出システムの出力を評価する際には、統一した正解を作成してそれと比較するのではなく、個々の被験者によるデータと比較した値を用いた。

4.3.2 システムの評価

重要文抽出システムを評価する際には、文境界の検出実験と同様に、50 記事をトレーニングデータとし、10 記事をテストデータとして用いた。重要文抽出システムの出力と個々の被験者によるデータとを比較した値とその平均値を表 7 に示す^(注2)。この結果は、文境界検出システムの出力に対して重要文抽出を行なったものである。評価結果は、対象とする被験者による変化が大きく、また F 値の平均は 30% を超えない。

一方、正しい文境界を予め与えた場合の重要文抽出結果は表 8 のようになる。F 値で 36.8% という評価は、システムが行った文境界検出結果を用いた場合よりも平均で約 7% 上回っている。この結果を書き言葉である日本語新聞記事に対する重要文抽出の

(注2): 重要文抽出における各評価基準は、文境界の検出と同様に、それぞれ以下のように定義される :

$$\text{再現率 (R)} = C / G$$

$$\text{適合率 (P)} = C / S$$

$$F \text{ 値} = 2PR / (P + R)$$

すなわち、再現率は被験者データ中の重要文の数 (G) のうち、システムによって抽出されたものと一致した文の数 (C) がどれだけであったかを示す。適合率は、システムによって抽出された重要文の数 (S) のうち、被験者データと一致した文の数 (C) がどれだけであったかを示す。F 値は、両者の調和平均である。個々の講演録について各評価の値を求めた後、全講演録について平均値の百分率をとったものが表に示された値である。

表 7 重要文抽出の評価結果（文境界検出を行った場合）

	被験者			
	A	B	C	平均
再現率	35.1	27.6	25.4	29.37
適合率	35.6	32.4	24.9	30.97
F 値	35.2	29.7	25.0	29.97

表 8 重要文抽出の評価結果（正しい文境界を与えた場合）

	被験者			
	A	B	C	平均
再現率	40.7	33.1	35.4	36.40
適合率	41.6	39.7	32.2	37.83
F 値	41.1	35.9	33.4	36.80

評価結果 [19] と比較する。要約率 10% のときに日本語新聞記事に対して、本システムによる重要文抽出の結果は F 値で 36.3% であり [13]、これは評価に参加したシステムの中でも上位に位置する結果であった。この評価で用いられた正解データに対する複数の被験者の重要文抽出結果は現在ないため、Kappa 値によるコーパスの比較はできないが、講演録・新聞記事の両者に対する重要文抽出の結果を比較すると、要約率が小さいときには、文境界が正しく検出できれば、講演録に対する重要文抽出は新聞記事に対する重要文抽出に匹敵しうることを示しているといえる。

表 9 に、評価関数の組み合わせを変えたときに講演録における重要文抽出の評価結果がどう変化したかを示す。表内の値は、各被験者と比較して求めた F 値の平均を示している。各評価関数を単独に用いた場合には、論文の見出しを用いた関数と文の位置情報を用いた関数の精度が良い。評価関数の組み合わせに対する結果を見ると、文の位置情報は他の評価関数と組み合わせることによって精度が向上するが、論文の見出しと他の評価関数との組み合わせは、位置情報との組み合わせほどには精度が向上しない。特に、全ての評価関数を用いた場合と見出し以外の全ての関数を用いた場合を比較すると、わずかではあるが全ての評価関数を用いた場合の方が精度が低くなっている。この理由としては、見出しと関連の強い重要文は、他の評価関数の組み合わせによっても得られるということ、つまり今回の実験で用いた評価関数が独立でないことが考えられる。見出し以外に予稿から得られる情報が重要文抽出にどの程度有効であるかを分析することは、今後の検討課題と

表 9 評価尺度の組合せごとの結果（講演録）

P---	-L--	--T-	---H		
23.2	17.0	12.1	26.3		
PL--	P-T-	P--H	-LT-	-L-H	--TH
28.3	23.3	24.8	12.1	26.3	18.8
PLT-	PL-H	P-TH	-LTH	PLTH	
30.7	27.9	24.1	19.6	30.0	

Key: P: 文の位置
L: 文の長さ
T: Tf*idf 値
H: 論文の見出し

表 10 評価尺度の組合せごとの結果（新聞記事）

P---	-L--	--T-	---H ₁	---H ₂
35.9	28.6	27.6	24.2	23.9
PLT-	PLTH ₁	PLTH ₂		
36.8	36.3	39.2		

Key: P: 文の位置
L: 文の長さ
T: Tf*idf 値
H₁: 論文の見出し（名詞）
H₂: 論文の見出し（固有表現）

したい。

新聞記事からの重要文抽出において、評価関数の組み合わせごとの評価結果を表 10 に示す。新聞記事に対する重要文抽出結果では、見出しを用いる評価関数として、見出し中の名詞に注目した評価関数（H₁）と、見出し中の固有表現^{注3}に注目した評価関数（H₂）の2種類を用意していた。見出しを用いた評価関数単独での結果は、他の関数よりも精度が低かった。他の評価関数と組み合わせたときの結果では、H₁ においては精度が向上せずむしろ低下を招くが、H₂ においては精度がさらに向上した。話し言葉における重要文抽出においても、見出しの情報を用いる際には、名詞からさらに限定して専門用語などに注目することが改良法の一つとして考えられる。

5. まとめ

本研究では、日本語書き起しコーパスに対して文

(注3): 固有表現 (Named Entity) とは、情報抽出の分野において定義された用語の一つであり [20], [21]、新聞記事中の人名や組織名などの固有名詞や、また日付や金額を示す表現などの数値的表現をさす。これらの表現は新聞記事からの情報抽出を行う際の基本的な構成要素とされている。

境界の検出と重要文抽出という二つの課題について実験結果を示した。文境界検出課題においては、

- ポーズ情報だけでは文境界の検出には不十分であること、

- パタンを用いた文境界の検出はF値で約90%の精度を得たこと、

- パタンとポーズ情報を併用することでさらに約2%精度が向上(約22%の文境界誤りが改善)したこと、

を示した。今後の課題としては、形態素解析システムを暫定的にポーズ区切りごとに用いて文境界検出のための情報を増やすこと、機械学習による文境界の検出精度を求めてパタンの精度と比較することなどを行い、また話者の話し方に応じて用いるパターンを変えるなどの手法を用いてさらに文境界検出の精度を高めることを考えている。また、本研究では文境界の検出に韻律情報を用いなかったが、韻律情報とパターンを併用することも今後検討したい。

重要文抽出課題においては、

- 複数の被験者の重要文抽出結果をKappa値によって比較し、複数の被験者の結果から統一された正解を作成するのは困難であること、

- 文境界検出システムの出力を用いた重要文抽出結果と正しい文境界を与えたときの重要文抽出結果との比較から、要約率が小さいときには、文境界が正しく検出できれば、講演録に対する重要文抽出は新聞記事に対する重要文抽出に匹敵しうることを示した。今後の課題としては、まずより大きな要約率に対して重要文抽出の評価を行い、今回の実験によって得られた結果が要約率によって異なるかどうかを比較したい。さらに抽出結果の精度を向上させるために、予稿から得られる情報の有効性や、被験者による重要文抽出結果の分析を行いたいと考えている。また、本システムに文短縮の手法を組み合わせ、文を抽出するだけでなくフィルターや繰り返し表現などを除去しより読みやすい要約を生成する処理を行うことも今後検討したい。

文 献

- [1] H. Edmundson. New methods in automatic abstracting. *Journal of ACM*, Vol. 16, No. 2, pp. 264–285, 1969.
- [2] Julian Kupiec, Jan Pedersen, and Francine Chen. A Trainable Document Summarizer. In *Proc. of SIGIR '95*, pp. 68–73, 1995.
- [3] Hideo Watanabe. A Method for Abstracting Newspaper Articles by Using Surface Clues. In *Proc. of COLING '96*, pp. 974–979, 1996.

- [4] 野本忠司, 松本祐治. 人間の重要文判定に基づいた自動要約の試み. In *IPSJ-NL 120-11*, pp. 71–76, July 1997.
- [5] C. Aone, M. E. Okurowski, and J. Gorlinsky. Trainable, Scalable Summarization Using Robust NLP and Machine Learning. In *Proc. of COLING-ACL '98*, pp. 62–66, 1998.
- [6] Chiori Hori and Sadaoki Furui. Automatic Speech Summarization Based on Word Significance and Linguistic Likelihood. In *Proc. 2000 IEEE International Conf. On Acoustics, Speech, and Signal processing*, pp. 1579–1582, 2000.
- [7] 伊藤山彦, 松本賢司, 谷田泰郎, 柏岡秀紀, 田中秀輝. 講演文を対象にした重要文抽出実験. 話し言葉の科学と工学ワークショップ講演予稿集, pp. 157–164, March 2001.
- [8] 野村和弘, 河原達也, 堂下修司. 講義の自動アーカイブ化のための韻律情報を用いた講義音声の文境界の検出. 電子情報通信学会 信学技報 SP 1998-11, pp. 17–24, 11 1998.
- [9] 野村和弘, 河原達也, 堂下修司. F0パターンに基づく講義音声の文単位へのセグメンテーション. 電子情報通信学会 信学技報 SP 1999-05, pp. 31–38, 05 1999.
- [10] Mark Stevenson and Robert Gaizauskas. Experiments on Sentence Boundary Detection. In *Proceedings of ANLP-NAACL2000*, pp. 84–89, 2000.
- [11] 長谷川将宏, 秋田裕哉, 河原達也. 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. In *IPSJ-NL 143-11*, pp. 75–82, June 2001.
- [12] 古井貞熙, 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推進制度 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.
- [13] 野畑周, 関根聡, 村田真樹, 内元清貴, 内山将夫, 井佐原均. 複数の評価尺度を統合的に用いた重要文抽出システム. 言語処理学会 第7回年次大会, March 2000.
- [14] 黒橋禎夫, 長尾真. 日本語形態素解析システム JUMAN version 3.61. 京都大学, 1999.
- [15] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proc. of SIGIR '94*, 1994.
- [16] Jean Carletta. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, Vol. 23, No. 1, pp. 249–254, 1996.
- [17] Jean Carletta, Amy Isard, Stephen Isard, Jacqueline C. Kowtko, Gwyneth Doherty-Sneddon, and Anne H. Anderson. The Reliability of a Dialogue Structure Coding Scheme. *Computational Linguistics*, Vol. 23, No. 1, pp. 13–31, 1997.
- [18] Klaus Krippendorff. *Content Analysis: An introduction to its methodology*. Sage Publications, 1980.
- [19] TSC. <http://oku-gw.pi.titech.ac.jp/tsc/>, 2001. Text Summarization Challenge.
- [20] DARPA. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, Fairfax, VA, USA, May 1998.
- [21] IREX. <http://cs.nyu.edu/cs/projects/proteus/irex/>, 1999. Information Retrieval and Extraction Exercise.