文セット単位による評価を用いたCSJの講演からの重要文抽出

下岡 和也[†]　内元 清貴[‡]　河原 達也[†]　井佐原 均[‡]

[†]京都大学 情報学研究科

[‡]情報通信研究機構

e-mail: shitaoka@ar.media.kyoto-u.ac.jp

1　はじめに

近年，計算機性能の向上や音声認識技術の進展に伴い，講演や会議の自動書き起こしなどの研究・開発が行われてきている．音声アーカイブはそのままでは，手早く目的の箇所を検索したり，全体の内容を把握したりすることが困難であり，効率よく利用するためには，単純な書き起こしだけでなく，要約などの二次情報をアーカイブに付与することが必要となる．

要約自動作成のアプローチの 1 つとして,「重要文抽出結果は一種の要約とみなすことが可能である」という考えに基づき，新聞記事や「日本語話し言葉コーパス」(CSJ)[1] を対象とした重要文抽出の研究が数多く行なわれている [2, 3, 4, 5, 6]．これらの研究では，主に，テキスト中の文を一つの単位として，それらに何らかの情報を基に重要度を付与し，その上位のものから文を選択するという枠組みが用いられている．しかし，抽出結果が全体として妥当なものになるためには，文脈のつながりを考えたり，冗長な文を抽出しないようにするなど，文全体としての整合性を考慮した評価を行う必要がある．

そこで本稿では，従来研究のように，テキスト中から独立に文を選択するのではなく，文セットを単位とした評価も併せて行い，最適な重要文セットを選択する手法を検討し，CSJ の講演を用いて行った重要文抽出の評価結果を示す．

2　文セット単位による評価を用いた重要文抽出

本稿では，文セット単位による評価を行うために，2 つの重要度尺度を考える．

まず，従来のように，文単位での評価により各文に重要度を付与し，任意の文セットについて，各文に付与されている重要度 (文単位重要度) の平均を求める．次に，文セットそのものの評価を行い，本稿で提案する文脈整合度を求める．最後に，これら 2 つのスコアを統合することで，文セットに対する新たな重要度を求める．以下に各々について説明する．

2.1　文単位重要度に基づく重要度尺度

文単位重要度の計算法として，これまでに提案されている 2 つの手法を統合して用いることとした．

(1) 北出らの手法 [4]

この手法では，2 つの重要度尺度が利用されている．

(a) 話題語に基づく重要度尺度

話題を特徴づける単語 (=話題語) は重要であると考えて，文中の単語の tf*idf 値の合計を文の重要度としている．idf 値を求める際の文書集合には CSJ の講演が用いられている．

(b) 談話標識に基づく重要度尺度

話題の転換点の冒頭には,「次に，‥」や「最後に‥」のように重要な発話がなされることが多い．したがって，話題の転換点を示すような単語 (=談話標識) を多く含む文は重要であると考えて，文中の単語の談話標識らしさの統計量 (tf*idf 値と同様に定義) の合計を文の重要度としている．

以上 2 つのスコアを統合することにより，重要度尺度を定義する．具体的には，両方の重みつき幾何平均を求める．

(2) 野畑らの手法 [5]

この手法では，4 つの重要度尺度が提案されているが，そのうちの見出し情報については，本稿では用いることができないため，以下の 3 つの重要度尺度を用いる．

(a) 文の位置情報

先頭か末尾に近いほど重要であるという仮定に基づき，各文のスコア付けを行なう．具

体的には，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，値が大きい方を求める．

(b) 文の長さ

極端に短い文は重要文として選択されにくいという観測事実に基づき，各文のスコア付けを行なう．具体的には，長さ L が 20 文字以下の文には負の値 (L-20) をペナルティとして与える．

(c)tf*idf

北出らの手法における話題語に基づく重要度と同様である．なお，ここでは，idf 値を求める際の文書集合に新聞記事が用いられている．

以上 3 つのスコアを統合することにより，重要度尺度を定義する．具体的には，それぞれの重みつき線形和を求める．

本研究では，上記の 2 つの手法により得られた重要度尺度を統合したものを，文単位重要度とする．具体的には，両方の手法により得られたスコアの合計を文単位重要度とした．なお，それぞれの手法により得られたスコアは，各講演ごとに最大値で割ることによって正規化されているため，文単位重要度は 0〜2 の値となる．

2.2 文脈整合度を導入した重要度尺度

文脈整合度の計算のために，最大エントロピー法 (ME) に基づくモデルを採用する．各講演ごとに作成されうる文セット集合から，重要文セットであるかどうかと，そこで観測される素性との依存関係を学習することによって，各文セットに対して重要文セットであるかどうかの確率を計算する．重要文セットであると出力された場合はその確率値を文脈整合度とし，重要文セットではないと出力された場合は文脈整合度は 0 とする．

このようにして求めた文脈整合度と，前節で述べた手法により求めた各文の文単位重要度の平均を統合することにより，文セットの新たな重要度とする．具体的には，それぞれの重み付き線形和を求める．

2.2.1 学習に用いた素性

本稿では，任意の文セットを表す特徴として，以下に挙げる素性を用いた．なお，得られる素性値が実数の場合は，離散値に変換して用いる．

表 1: 接続助詞の種類

から	が	けど	けれど	し
つつ	とも	ながら	なり	ば

表 2: 接続表現の種類の分類

接続の種類	接続表現の例
並列	あるいは，および，かつ
順接	から，して，そして
逆接	が，けれど，しかし
条件	ただし，されば，もし
添加	しかも，じゃ，なお
転換	さて
強調	ただ，さらに，無論
選択	もしくは
対比	いっぽう
換言	すなわち，いわば，つまり
結論	結局，とにかく，やはり
相反	むしろ
例示	例えば
説明	そもそも

(1) 文の位置情報

(2) 文の長さ

前節，野畑らの手法により得られる値を用いる．

(3) 話題語に基づく重要度

(4) 談話標識に基づく重要度

前節，北出らの手法により得られる値を用いる．

(5) 出現する助詞の種類

各文において，格助詞，係助詞，終助詞，準体助詞，接続助詞，副助詞，の 6 種類の助詞が，それぞれ出現するかどうかを考える．

なお，接続助詞については，次に述べるようにより細かく考える．

(6) 出現する接続助詞の種類

各文において，表 1 に示す接続助詞が，それぞれ出現するかどうかを考える．

(7) 出現する接続表現の種類

各文の文頭に出現する接続表現を表 2 のように分類し [7]，各文について，どの種類の接続表現が出現したかを考える．ただし，フィラー的な振舞いをする「て」「で」「と」については無視する．

(8) 照応・接続関係の保存

各文において，前後の文との照応・接続関係が保

表 3: 指示語の種類

これ	それ	ここ	そこ	この
その	こんな	そんな	こう	そう

存されているかどうかを考える．具体的には，以下の 3 種類を考える．

- 文頭に表 2 に示す接続表現が出現した場合，直前の文が原文においても直前の文であったかどうか
- 文中に表 3 に示す指示語が出現した場合，直前の文が原文においても直前の文であったかどうか
- 文末が接続助詞，あるいは，接続助詞+副助詞の順で終わっている場合，直後の文が原文においても直後の文であったかどうか

(9) 単語連鎖
各文において，直前の文に出現した単語（名詞に限定）が何種類表れたかを考える．

以上に挙げた素性は，各文単位で得られる素性である．したがって，これらを文セット単位の素性に変換する必要がある．具体的には，それぞれの素性値を持つ文の割合を新たな素性値とする．

さらに，文セット単位で得られる以下の 2 つの素性を考える．

(10) 単語連鎖の長さ
各文セットにおいて，前記の単語連鎖が何文続いたか，つまり，単語連鎖に関する素性値が 0 以外のものが何文続いたかを考える．

(11) 接続表現の出現順
表 2 で分類したそれぞれの接続表現が，各文セットにおいて，どういった順で出現したかを考える．

2.2.2 擬似的な正解セットの作成

CSJ では，人手による抽出結果 (正解セット) が各講演 3 通りしかないため，これらのみを学習の際の正例とすると，正例の数がごくわずかとなってしまう．そのため，擬似的な正解セットを作成して，正例の数を増やすことを考える．具体的には，生成された文セットをある指標を用いて評価し，その評価値が人手による抽出結果に対する評価値を上回るような文セットを全て正解とみなすこととした．ここでは，その際の評価の指標として，機械翻訳の評価に用いられている BLEU スコア [8] を用いる．BLEU スコアは,「正解文に出現する N-gram が数多く出現するほど正解文に近い」という仮定に基づいて，N-gram 系列の適合率を求めて，各 N-gram の相乗平均により計算される．なお，本研究では，各文セットにおける文のつながり度合を考慮するため，N-gram を形成する単位を文とし，N=3 として評価する．

まず，人手による 3 通りの抽出結果のうち，1 つをテスト文，残り 2 つを正解文として BLEU スコアを計算する．これを 3 通り行い，そのうち最大のものを正解の閾値とする．次に，3 つの人手による抽出結果を正解文として，作成された全ての文セットに対して BLEU スコアを計算し，閾値を超えた文セットは全て擬似的な正解セットとする．

2.2.3 探索空間の限定

1 講演において，生成されうる文セットの総数は膨大な数となり，全探索を行なうことは困難である．したがって，学習時・テスト時ともに，探索空間を限定する必要がある．

(1) 学習時
あらかじめ原文を文単位重要度の高い文の集合とそうでない文の集合に分別し，重要度の低い文を順次入れ替えることにより，人手による抽出結果と類似した文セットを多く作成する．

具体的な作成手順は以下の通りである．

（1）2.1 節で述べた手法による文単位重要度により原文をソートし，抽出率に応じて，文単位重要度の高い文の集合（B) とそうではない文の集合（R) とに分別する

（2）人手による抽出結果（A) と上記（B) とに共通して含まれる文を共通文，それ以外の文を入れ替え対象文として（B) を分別する．ただし，入れ替え対象文数が 10 を超えた場合，それ以降の文は共通文とする

（3）共通文は固定し，入れ替え対象文が全て上記（A) に含まれる文と入れ替わるまで文集合（R) に含まれる文と入れ替え対象文とを一文ずつ入れ替える

(2) テスト時
学習時との整合性を高くするため，以下の手順によりテストを行なう．

(1) 文単位重要度により原文をソートし，ベースライン抽出結果 (B) とそれ以外の文集合

表 4: 文脈整合度を用いた重要文抽出の精度

	再現率	適合率	F 値
文脈整合度利用	69.7%	52.6%	0.598
ベースライン	69.3%	52.3%	0.595

(R) とに分別する

(2) 以下を (R) が空集合になるまで繰り返す

(3)R の上位から 1 文 (r) を取り出す

(4)B の各文を一つずつ r と入れ替え，新たな文セット集合を作成する

(5) 作成された各文セットを文セット単位で評価し，重要度が B に対する重要度を越えるもので最大の文セットを新たに B とする

(6) 3に戻る

3 評価実験

提案手法による重要文抽出の評価を行った．実験には CSJ の 187 講演の書き起こしを用い，157 講演を学習データ，10 講演をチューニングデータ，20 講演をテストデータとした．なお，今回は抽出率が 50%の場合についてのみ評価し，文境界は人手により与えられているものとした．評価の際の正解セットには，任意の 2 名が共に重要文として抽出した文セット (計 3 通り) を用いる．この全文に対する割合は 37.1%である．それぞれの正解セットに対する再現率・適合率・F 値の平均で評価を行う．

まず，チューニングデータに対して，文脈整合度を用いた場合で精度が最大となったパラメータを求めた結果，(文単位重要度の平均, 文脈整合度)=(1,0.1) となり，F 値は 0.603 であった．

テストデータに対し，上記のパラメータを用いて重要文抽出を行なった．表 4 に結果を示す．わずかながら精度が上昇しているが，有意な差ではない．

テストデータ 20 講演のうち，文脈整合度を利用することでベースライン結果と異なる抽出結果となった各講演について，入れ替わった文数を調査したところ，20 講演中 8 講演は全く入れ替わっておらず，多くても抽出結果の文数の 1 割程度しか入れ替わっていないことが判明した．これは，文脈整合度による評価の精度が低く，文脈整合度による影響が小さくなるような重みが選ばれたためである．

参考のため，学習により得られた確率モデルの精度を調べるために，テストデータ 20 講演それぞれにおける人手による抽出結果 (60 セット) に対して文脈整合度を求めた．その結果，文脈整合度が 0 ではない，つまり重要文セットであると出力されたのは 27 セットであった．すなわち，現在の確率モデルの正解率は 45%程度であり，確率モデルの洗練が必要であると考えられる．確率モデルの洗練のためには，素性の洗練，疑似的な正解作成の際の指標の洗練，モデル化の際のカテゴリの洗練などが挙げられる．

4 まとめ

本稿では，重要文抽出において，文単位での評価によって独立に文を選択するのではなく，文セット全体の整合性なども考慮した評価を行うことによって，最適な文セットを選択する手法を検討した．CSJ の講演を用いた評価実験において，現段階では，本手法の有効性を明確に示す結果は得られていないが，今後，確率モデルを洗練することが課題である．

参考文献

[1] S.Furui, K.Maekawa, and H.Isahara. Toward the realization of spontaneous speech recognition - introducing of a Japanese priority program and preliminary results -. In *Proc. ICSLP*, Vol. 3, 2000.

[2] 平尾努, 磯崎秀樹, 前田英作, 松本裕治. Support Vector Machine を用いた重要文抽出. 情報処理学会論文誌, pp. 2230–2243, 2003.

[3] 菊池智紀, 古井貞熙, 堀智織. 重要文抽出と文圧縮による音声自動要約. 信学技報, SP2002-158, 2002.

[4] T.Kitade, H.Nanjo, and T.Kawahara. Automatic extraction of key sentences from oral presentations using statistical measure based on discourse markers. In *Proceedings of ICSLP*, pp. 2169–2172, 2004.

[5] 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 回年次大会発表論文集, pp. 301–304, 2001.

[6] 奥村学, 難波英嗣. テキスト自動要約に関する研究動向 (巻頭言に代えて). 自然言語処理学会誌, Vol.6, No.6, pp. 1–26, 1999.

[7] 田村直良, 和田啓二. セグメントの分割と統合による文章の構造解析. 自然言語処理学会誌, Vol.5, No.1, 1998.

[8] K.Papineni, S.Roukos, T.Ward, and W-J.Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of thr 40th Annual Meeting of ACL*, pp. 311–318, 2002.