論文 / 著書情報
Article / Book Information

論題(和文)	講演音声のインデキシングを目的としたセグメンテーション手法の検討
Title(English)	
著者(和文)	新中庸介, 岩野公司, 古井貞熙
Authors(English)	Koji Iwano, SADAOKI FURUI
出典(和文)	日本音響学会2005年春季講演論文集, Vol. , No. 1-5-4, pp. 7-8
Citation(English)	, Vol. , No. 1-5-4, pp. 7-8
発行日 / Pub. date	2005, 3

1—5—4

講演音声のインデキシングを目的とした セグメンテーション手法の検討 *

◎新中 庸介　岩野 公司　古井 貞熙 (東工大)

1 はじめに

これまで講演音声のインデキシングを目的としたセグメンテーション手法が検討されている [1]．本稿では，ポーズ長や談話標識の確率モデルを利用した講演音声のセグメンテーション手法の提案を行う．ターゲットとして，講演時に用いられたスライドを基に，1スライドで発話される内容を1つのセグメントと定義する．提案手法におけるスライドの切り替わる位置 (セグメント境界) の検出性能について評価を行う．

2 セグメンテーション手法

講演音声の認識結果について，文献 [2] で用いられる手法により句点の挿入を行い，各句点位置について，セグメント境界であるかどうかの判定を行う．判定は，ポーズ長・出現単語に対する，境界・非境界の尤度に基づいて行われる．尤度算出のため，境界・非境界モデルを構築する．

モデルの学習データには，正しいセグメント境界位置をラベル付けしておく必要がある．そこで，学習データの形態素解析を行った講演書き起こしについて，句点の挿入を行い，人手によってスライドの切り替わりに対応する句点位置に境界のラベルを付与する．このとき，形態素解析には文献 [3] のシステムを用い，句点の挿入には，文献 [4] で用いられる手法を使用する．スライドの切り替わりが，自動挿入された句点位置に対応しない場合があるが，その場合には，書き起こしの該当位置に人手によって強制的に句点の挿入を行い，その句点位置に境界ラベルを付与する．また，ポーズについても人手によって時間ラベルが付与されている．

2.1 ポーズ長を利用した手法 (手法 P)

講演では，スライドの切り替わりの際に，通常のポーズよりも長いポーズが挿入されると考えられるため，ポーズ長の情報は，セグメンテーションに有効であると考えられる．以下にポーズ長を利用した手法の手順を示す．

(1) 学習データにおける，セグメント境界位置のポーズ長，境界以外の句点位置のポーズ長について，それぞれの確率モデルを作成する．具体的には，ポーズ長の逆数を正規分布に近似することでモデル化を行う．

(2) 評価データの各句点位置について，その位置のポーズ長を l としたときに，以下の式 (1) により，境界である尤度 $L_{pau}(l)$ を計算する．

$$L_{pau}(l) = \log P(l|C) - \log P(l|N) \qquad (1)$$

このとき，C は境界位置である事象，N は非境界位置である事象を表す．

(3) $L_{pau}(l)$ が大きい句点位置から順に，目標とする数まで境界を検出する．

2.2 談話標識を利用した手法 (手法 D)

講演のスライドが切り替わる位置の前後の発話と，それ以外の句点位置の前後の発話では，談話標識の有無などにより出現する単語に違いがあると考えられるため，これを利用した手法を提案する．以下に手法の手順を示す．フィラー，言い直しは学習データと評価データから除かれているものとする．

(1) 学習データにおける，境界位置の前 M_f 単語と，境界以外の句点位置の前 M_f 単語について，それぞれの言語モデルを作成する．これを後ろ M_r 単語についても同様に行う．

(2) 評価データの各句点位置について，前 M_f 単語を w，w の単語数を $|w|$ としたときに，以下の式 (2) により，w が境界位置の前 M_f 単語である尤度 $L_f(w)$ を計算する．このとき $P(w|C)$，$P(w|N)$ は単語 N-gram によってモデル化する．後ろ M_r 単語についても同様に尤度 $L_r(w)$ を計算する．ただし，学習データ (評価データ) において，境界位置 (句点位置) の前の文の総単語数 M_f' が，M_f よりも小さい場合には，前 M_f' 単語を用いる．境界位置の後ろ M_r 単語についても同様である．

$$L_f(w) \text{ or } L_r(w) = \frac{1}{|w|}\{\log P(w|C) - \log P(w|N)\} \qquad (2)$$

(3) 式 (3) により，$L_f(w)$ と $L_r(w)$ を合わせた尤度 $L_d(w)$ を計算する．λ_r は重み係数である．

$$L_d(w) = L_f(w) + \lambda_r L_r(w) \qquad (3)$$

(4) $L_d(w)$ が大きい句点位置から順に，目標とする数まで境界を検出する．

2.3 手法 P と手法 D を組み合わせた手法 (手法 C)

評価データの各句点位置について，式 (4) により，手法 **P** による $L_{pau}(l)$ と，手法 **D** による $L_d(w)$ を合わせた尤度 $L_c(l,w)$ を計算する．$L_c(l,w)$ が大きい位置から順に，目標とする数まで境界を検出する．λ_{pau} は重み係数である．

$$L_c(l,w) = L_d(w) + \lambda_{pau} L_{pau}(l) \qquad (4)$$

3 評価実験

3.1 実験条件

評価データは，話し言葉コーパス (CSJ) の 17 名による学会講演音声とした．これについて，文献 [5] の音声認識システムを用いて認識を行い，認識結果について自動的に句点を挿入し，セグメンテーションを行った．音声認識結果の単語正解精度は約 69%であった．CSJ の人手による重要文抽出結果における句点位置を正解とした場合，自動挿入した句点位置の適合率/再現率は約 86%/85%であった．句点は 17 講演で 1211 箇所であり，セグメント境界は 237 箇所であった．なお，セグメント境界のうち，句点位置に対応しないものは，21 箇所であった．学習デー

* A study on automatic lecture segmentation for indexing purposes
By Yosuke Shinnaka, Koji Iwano and Sadaoki Furui (Tokyo Institute of Technology)

表 1. N-gram 別の境界検出率 (%)

	unigram	bigram	trigram
D	38.4	36.7	38.0
C(est1)	61.2	60.3	60.8
C(est2)	62.5	61.6	62.5

タには，評価データと独立な CSJ の 80 名による学会講演を用いた，

セグメンテーションを行う際のパラメータ (手法 **D** では M_f，M_r，λ_r，手法 **C** ではそれらに λ_{pau} を加えたもの) は，評価データや学習データとは独立な CSJ の 20 名による学会講演をデベロップメントセットとして推定した．デベロップメントセットは評価データと同様に，音声認識を行い，自動的に句点挿入を行った．手法 **C** についてパラメータを推定する際に，手法 **D** の 3 つのパラメータを推定した後，手法 **P** と組み合わせる際のパラメータ λ_{pau} を推定する場合 (est1) と，全てのパラメータを同時に推定する場合 (est2) の 2 つの場合について実験を行った．

なお，実験は，講演ごとの境界数が既知として行った．

3.2 評価結果

3.2.1 最適な N-gram の選択

手法 **D**，**C** では言語モデルとして単語 N-gram を用いており，unigram，bigram，trigram の 3 つの場合についてそれぞれ実験を行った．bigram，trigram については，Good-Turing 法に基づく back-off 平滑化を行った．表 1 に N-gram 別の境界検出率を示す．いずれも 17 講演の平均である．unigram を用いた場合にどの手法においても境界検出率が最大となった．bigram，trigram の境界検出率が unigram よりも良くならなかったのは，学習データの不足により，評価データに現れる談話標識となる表現をカバーできなかったことが原因であると考えられる．なお，手法 **D** において unigram を適用した場合に推定された M_f，M_r は，それぞれ 25，15 であった．

3.2.2 各手法ごとの検出性能の比較

図 1 に各手法ごとの境界検出率を示す．**R** は，比較のための，句点位置に対してランダムに境界を選択したときのセグメンテーション結果である．手法 **D**，**C**(est1)，**C**(est2) については，unigram を用いた場合の結果を示した．提案手法 **P**，**D**，**C**(est1)，**C**(est2) は，ランダムの手法 **R** に比べて，それぞれ，約 42%，20%，43%，44%の改善が得られ，提案手法の有効性が確認された．また，手法 **C**(est1)，**C**(est2) はともに手法 **P**，**D** をそれぞれ単独で用いた場合よりも良好なことから，手法 **P** と手法 **D** を組み合わせることが有効であることが示された．また，手法 **C**(est1) と手法 **C**(est2) では，手法 **C**(est2) の結果の方が良好なことから，全てのパラメータを同時に推定する場合の方が有効であることが確認された．

3.2.3 書き起こしに対するセグメンテーション

手法 **P**，**D** については，人手により句点を挿入した評価データの書き起こしについても実験を行った．学習データはこれまでの実験と同様のものを用いた．デベロップメントセットは，評価データと同様に書き起こしを用い，句点の挿入は自動的に行った．これら全てのデータのポーズには，人手によって時間ラベルが付与されている．表 2 に，手法 **P**，**D** について，認識結果と書き起こしを用いた場合のそれぞれの境界検出率を示す．手法 **D** については，unigram を用いた場合の結果を示した．手法 **P**，**D** ともに，書き起こしを用いることで，境界検出率がさらに高くなった．書き起こしを用いる場合，認識結果を用いる場合と比べて，句点の挿入精度，ポーズ長の推定精度，単語の認識精度が改善されるため，境界検出率が向上したと考えられる．

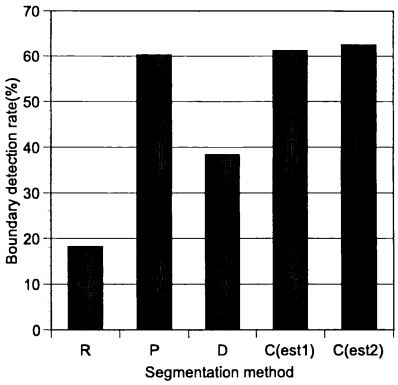

図 1. 各手法ごとの境界検出率

表 2. 認識結果と書き起こしに対する境界検出率 (%)

	認識結果	書き起こし
P	60.3	64.6
D	38.4	43.5

4 まとめ

本稿では，講演音声のインデキシングを目的として，スライドを単位としたセグメンテーションについての検討を行った．提案手法についてセグメント境界の検出性能を評価したところ，その有効性が確認された．

今後の課題としては，学習データ，デベロップメントセットの増加，境界数の自動決定などがある．

謝辞 本研究の一部は，文部科学省 21 世紀 COE プログラム「大規模知識資源の体系化と活用基盤構築」の支援を受けて行われた．

参考文献

[1] 長谷川将宏，秋田祐哉，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,” 情報処理学会研究報告, NL-143-11 / SLP-36-6, pp.35-42, 2001.

[2] 北出祐，南條浩輝，河原達也: “談話標識と話題語を用いた重要文抽出手法の CSJ の学会講演における評価,” 話し言葉の科学と工学ワークショップ 講演予稿集，pp.111-118, 2004.

[3] K.Uchimoto, C.Nobata, A.Yamada, S.Sekine and H.Isahara: “Morphological analysis of Corpus of Spontaneous Japanese,” Proc. SSPR2003, Tokyo, Japan, pp.159-162, 2003.

[4] 北出祐，南條浩輝，河原達也，奥乃博: “談話標識と話題語に基づく統計的尺度による講演からの重要文抽出,” 情報処理学会研究報告, NL-155-13 / SLP-46-2, pp.7-12, 2003.

[5] T.Kawahara, H.Nanjo, T.Shinozaki, and S.Furui: “Benchmark test for speech recognition using the corpus of spontaneous Japanese,” Proc. SSPR2003, pp.135-138, 2003.