2006－SLP－62　(12)
2006／7／8

会議音声の自動話題分割による単語辞書と言語モデルの適応

根本 雄介　　秋田 祐哉　　河原 達也

京都大学 情報学研究科 知能情報学専攻
〒606-8501 京都市左京区吉田二本松町
e-mail: nemoto@ar.media.kyoto-u.ac.jp

あらまし　広範な話題からなる会議音声を話題単位に自動分割し，得られた話題ごとに単語辞書と言語モデルの適応を行う手法を提案する．音声認識結果に対して PLSA（Probabilistic Latent Semantic Analysis）を適用して，話題を表す特徴ベクトルに変換し，その類似度に基づいて話題分割を行う．そして，話題ごとに類似したテキストを収集して，単語辞書を更新するとともに N-gram 言語モデルの適応を行う．衆議院予算委員会の音声で評価を行った結果，提案手法により単語辞書・言語モデルの適応を行うことで，ベースラインから未知語率を約 25%，テストセットパープレキシティを約 9%削減することができた．

Vocabulary and Language Model Adaptation based on Automatic Topic Segmentation of Meetings

Yusuke Nemoto　　Yuya Akita　　Tatsuya Kawahara

School of Informatics, Kyoto University, Kyoto 606-8501, Japan
email: nemoto@ar.media.kyoto-u.ac.jp

Abstract　We address a vocabulary and language model adaptation method based on topic segmentation of meetings that include various topics. The ASR result is segmented based on the similarity among the feature vectors that were extracted with PLSA (Probabilistic Latent Semantic Analysis). The relevant texts (newspaper articles) for each topic segment are retrieved. The vocabulary and N-gram language model are updated with this retrieved texts. Experimental evaluation on a meeting of the Lower House Budget Committee showed that the proposed model adaptation based on topic segmentation reduced the test-set OOV rate and perplexity.

1　はじめに

情報通信技術の発展により，大量の音声や動画をディジタルデータとして蓄積し，ネットワークを通じて配信するディジタルアーカイブの実現が可能になりつつある．大量のデータを蓄積，配信するアーカイブの利便性を向上させるために，インデックスや要約といった高次の情報を付加してユーザに提示することが考えられるが，これらの作業を人手で行うのは非常にコストが大きい．そこで，音声認識技術を活用した自動インデキシングなどが検討されている．我々も，会議を対象とした音声認識と自動インデキシングの研究を行っている [1].

会議では様々な話題・議題が次々と議論されるので，固有名詞や時事用語といった話題に特有な単語が未登録となる問題が生じる．また，これらの話題に関連する単語の出現が十分に予測できているとは限らない．これらの固有名詞や話題語は会議の話題に密接に関連する単語であるので，それを含む文を正しく認識できないことは，後のインデキシングなどの性能の低下につながる．

この問題に対処するために，音声認識結果を用いて言語モデルや単語辞書を話題に適応させる手法が提案されている [2][3][4]．これらの手法では，例えば放送番組における個々のニュースのように，明確な話題境界が存在することを仮定している．会議の場合には，長いポーズで区切られた発話や，一人の発言者が発話を続けた区間（ターン）といった単位で適応を行うことが考えられる [2]．しかし，これらは必ずしも話題適応を行うのに適切な単位であるとは限らない．したがって，適切な話題のまとまりの単位に音声を分割する必要があると考えられる．入力音声を話題単位に分割することは，後にインデックスや要約の作成を行う上でも非常に有用な処理である．

これまでにテキストや音声の話題分割に関して数多く研究されている [5][6]．米国では NIST によって TDT（Topic Detection and Tracking）といったコンテストも実施されている．これらの研究では，新聞記事やニュース音声などの話題の境界が比較的明確に定められる文書を主な対象としている．しかし，話題が非常に広範かつ境界も曖昧で，非定型な発話が繰り返される会議のようなタスクについては十分な検討は行われていない．さらに，音声に対して話題分割を行った場合に，その結果が話題適応に利用可能か否かといった検討は行われていない．

そこで本研究では，会議音声に対して話題適応を指向して話題分割を行う手法，及びこれにより得られた話題セグメントごとに単語辞書と言語モデルの適応を行う方法について検討する．

2　タスクとデータ

本研究では衆議院予算委員会の音声を対象とする．衆議院予算委員会は，国政全般に関する議論を行うため，一回の会議の中で外交問題や年金問題といったさまざまな話題が出現する．

会議は，質問者が自己の主張や見解を表明し，質問を行うのに対して，閣僚や官僚がそれに答弁する形を繰り返して進行する．数ターンにわたって同一の話題が続いたり，同一の質問者が異なる話題にふれることもあるため，ターンや話者により話題が定まるとは限らない．また，会議の中での話題の移り変わりが曖昧な部分も存在する．

テストデータとして用いるのは，2003 年 2 月 14 日の予算委員会の音声認識結果と人手による書き起こしである．

音声認識は，Julius 3.5 により行った．音声認識に用いた音響モデルは，「日本語話し言葉コーパス」（CSJ）に含まれる 274 時間の学会講演から SAT 学習した状態共有 triphone HMM に対して，教師なし MLLR 話者適応を行ったものである．ベースライン言語モデルは，1999 年から 2002 年の衆議院会議録を用いて構築したものと，CSJ の模擬講演のみを用いて構築したものを，相補的バックオフ [7] を用いた手法により重み付け混合したものである．混合重みは国会 0.5，講演 0.5 である．ベースラインの語彙サイズは 29731 である．ベースライン言語モデルによるテストセットパープレキシティは 61.94，未知語率は 0.47%（332 単語）であった．また，音声認識精度は，単語正解精度が 80.16%，単語正解率が 83.94%であった．

提案する処理の流れを図 1 に示す．まず，音声認識結果をターン毎に分割し，特徴抽出と類似度計算を行う．それに基づいて類似しているターンを連結していくことで話題の単位を構成する．次に，得られた話題ごとに，新聞記事データベースから類似度の高いテキストを収集する．そして，収集されたテキストに含まれる未登録語を単語辞書に追加する．さらに，得られたテキストを使用して言語モデルを構

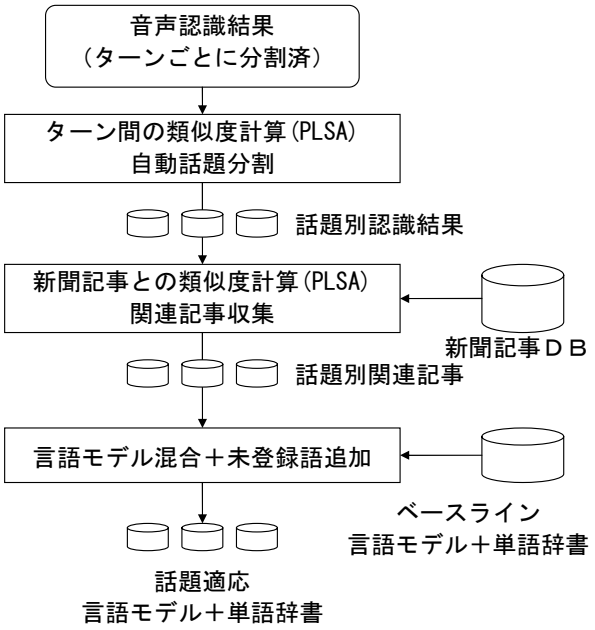

図 1: 処理の概要

築し，ベースライン言語モデルと混合することで適応を行う．

3 PLSA による話題分割

3.1 PLSA による特徴抽出

PLSA（Probabilistic Latent Semantic Analysis）[8] は，コーパスに含まれる文書ごとの単語頻度をもとに単語の生起確率を推定し，これを部分空間に射影することで各文書を特徴づける枠組みである．この部分空間に各ターンの発話を射影することにより，話題分割処理に用いる特徴量の抽出を行う．

PLSA における単語の生起確率は，コーパス中の文書 d，単語 w に対して (1) 式で定式化される．

$$P(w|d) = \sum_{j=1}^{N} P(w|t_j)P(t_j|d) \tag{1}$$

ここで，t_j は観測されない潜在変数であり，その総数 N は部分空間の基底数に相当する．

(1) 式における単語，潜在変数の生起確率は，(2) 式で表される対数尤度 L を最大化するように EM アルゴリズムにより推定される．ここで，$n(d,w)$ は文書 d において単語 w が出現する頻度である．

$$L = \sum_{d} \sum_{w} n(d,w) \log P(d,w) \tag{2}$$

学習コーパスに含まれる文書群 $\{d_i\}(i = 1,...,M)$ の各文書における単語 $w_k(k = 1,...,K)$ の頻度 $n(w_k,d_i)$ から $P(w_k|t_j)$ および $P(t_j|d_i)$ が推定される．これは，部分空間の基底と，各文書の部分空間における座標の推定に相当する．それから文書に非依存な $P(w|t_j)$ を固定し，特徴抽出を行う文書 d' (=ターンの発話）に対する $P(t_j|d')$ を同様に推定する．これにより，部分空間上で d' を表わすベクトル $(P(t_1|d'),...,P(t_N|d'))$ が得られる．これを，話題を表す特徴ベクトルとする．

著者らはこれまでに PLSA に基づいた言語モデルの適応法について研究しており [2]，この部分空間に基づく特徴量を話題分割に用いることは本研究の言語モデル適応においても効果が期待できる．

3.2 自動分割処理

次に，PLSA により推定した各ターン d_i の N 次元特徴ベクトル $v_{d_i} = (P(t_1|d_i),...,P(t_N|d_i))$ を用いて，ターンを話題セグメントに統合する．

1. 先頭のターン d_1 を最初の話題セグメント s_1 に含める．また，$i = 2$，$j = 1$ とする．
2. 話題セグメント s_j の特徴ベクトル v_{s_j} を，次の式により求める．D_{s_j} は話題セグメント s_j 含まれるターンの総数である．
$$v_{s_j} = \frac{1}{D_{s_j}} \sum_{d_k \in s_j} v_{d_k} \tag{3}$$
3. ターン d_i と，その直前の話題セグメント s_j の類似度を次のコサイン距離により求める．
$$sim(d_i, s_j) = \frac{v_{d_i} \cdot v_{s_j}}{|v_{d_i}||v_{s_j}|} \tag{4}$$
4. $sim(d_i, s_j) > \theta_1$ ならば，d_i を s_j に追加する．それ以外の場合，j に 1 を加え，d_i を新たな話題セグメント s_j に含める．θ_1 は閾値である．
5. 末尾のターンならば終了．それ以外は i に 1 を加えて手続き 2 に戻る．

話題数	正解範囲	再現率	適合率	F 値
12	± 0	0.12	0.18	0.15
($\theta_1 = 0.12$)	± 2	0.12	0.18	0.15
25	± 0	0.43	0.29	0.35
($\theta_1 = 0.15$)	± 2	0.50	0.33	0.40
38	± 0	0.56	0.24	0.34
($\theta_1 = 0.20$)	± 2	0.62	0.27	0.38
65	± 0	0.68	0.17	0.28
($\theta_1 = 0.25$)	± 2	0.87	0.21	0.35

表 1: 書き起こしに対する話題分割結果

話題数	正解範囲	再現率	適合率	F 値
16	± 0	0	0	0
($\theta_1 = 0.12$)	± 2	0.06	0.06	0.06
26	± 0	0.06	0.04	0.05
($\theta_1 = 0.15$)	± 2	0.12	0.08	0.12
45	± 0	0.18	0.06	0.10
($\theta_1 = 0.20$)	± 2	0.31	0.11	0.17
69	± 0	0.56	0.13	0.21
($\theta_1 = 0.25$)	± 2	0.68	0.16	0.26

表 2: 音声認識結果に対する話題分割結果

3.3 自動話題分割実験

テストデータの人手による書き起こしと音声認識結果に対して類似度の閾値 θ_1 の値を変化させて自動話題分割を行った結果を表 1，2 に示す．会議全体で発話ターンは 296 個あり，人手で付与した正解話題境界は全部で 16 個であった．正解とした話題境界は，発言者がそれまでの議論の流れから離れて，別の事柄に関する議論を始めたと考えられるターンの直前に設定した．なお，議長の発言は議論内容とは関連がなく，議長席に設けられたマイクにより録音されるので評価対象から除外した．

また，PLSA により計算される特徴ベクトルの次元数は250である．これは，以前の言語モデル適応の研究 [2] において最適な結果が得られたものである．

正解範囲を± 2 ターンのずれまで許容した際に，書き起こしに対して話題数が 25 の時に F 値が 0.40 と最大になった．音声認識結果に対する最大の F 値は 0.26 であり，認識誤りの影響によって大きく低下した．特に，同じ話題について議論を続けているにもかかわらず，比較的短い発話の応酬が続き，話題に関連する語が認識されない部分で境界が誤挿入されることが多くみられた．

4 テキスト収集による話題適応

4.1 話題関連テキストの収集

それぞれの話題セグメントに関連するテキストを収集するために，得られた各話題セグメントをそれぞれ一つの文書とみなし，(3) 式の特徴ベクトルを求める．

次に，新聞記事データベースを対象に，会議音声の各ターンに対して行ったのと同様に，各記事を一つの文書単位として PLSA による確率 $P(t|d)$ の推定を行い，話題セグメントと同一の部分空間に写像された特徴ベクトルを得る．

そして，得られた話題セグメントと新聞記事の特徴ベクトル間の類似度に基づき，話題セグメントごとに関連記事を収集する．話題セグメント s_j，新聞記事 r に対して特徴ベクトル v_{s_j}，v_r が計算されているとき，(4) 式により与えられる類似度 $sim(s_j, r)$ が，

$$sim(s_j, r) > \theta_2 \tag{5}$$

を満たすとき，記事 r を話題セグメント s_j に関連するものとして収集する．θ_2 は，記事収集のための閾値であり，すべての話題セグメントで同一の値を用いた．

4.2 単語辞書と言語モデルの適応

各話題セグメントごとに収集したテキストに含まれる未登録語を単語辞書に追加して，各話題セグメントごとに単語辞書を更新する．

さらに，収集した新聞記事を用いて，N-gram 言語モデルを構築する．そして，この言語モデルと最初の音声認識に用いたベースライン言語モデルの N-gram 確率を重みつきで結合することで，話題セグメントごとに適応言語モデルを作成する．結合重みは，予備実験により，新聞から作成した言語モデルが 0.15，ベースラインが 0.85 と事前に定めた．

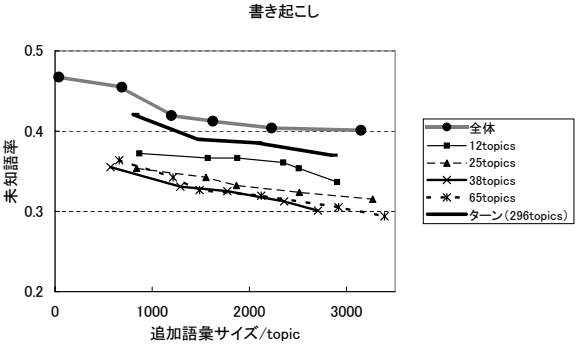

図 2: 書き起こしを用いた単語辞書適応結果

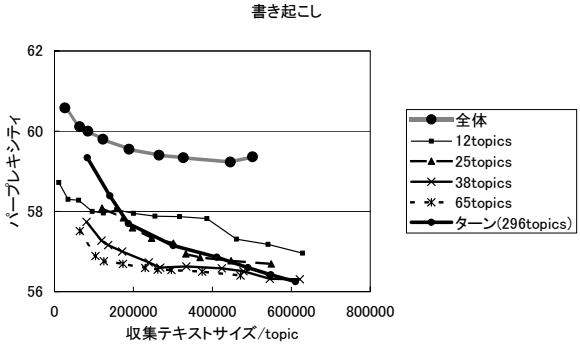

図 4: 書き起こしを用いた言語モデル適応結果

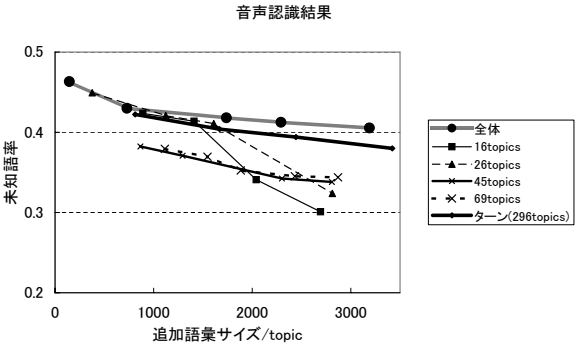

図 3: 音声認識結果を用いた単語辞書適応結果

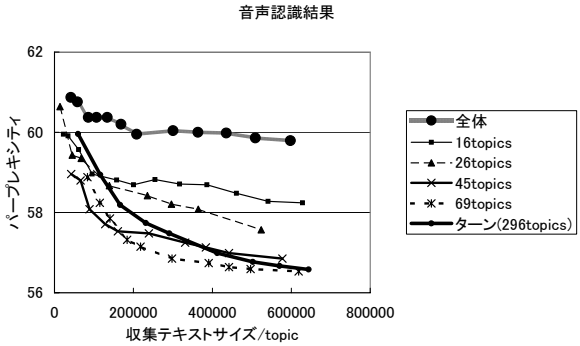

図 5: 音声認識結果を用いた言語モデル適応結果

4.3 実験結果

3 章の自動話題分割実験により得られたセグメントごとに単語辞書と言語モデルの適応を行った．単語辞書と言語モデルの適応に使用した新聞記事データベースは,「CD-毎日新聞」の 2002 年 7 月 1 日から 2003 年 2 月 13 日の 1, 2, 3, 国際，経済，社会，総合面に掲載された 41714 記事，合計単語数は 11M である．

まず，話題分割に基づいて単語辞書の更新を書き起こしと音声認識結果に対して行った結果を，それぞれ図 2, 3 に示す．これらは閾値 θ_2 の値を変化させて得られた結果をプロットしたものである．書き起こしに対しては，テストデータ中の未知語数は，各セグメントに平均で 1000 語追加時に最大約 10%，3000 語追加時に最大約 30%削減された．各ターンごとに単語辞書適応を行った場合には，3000 語追加時に約 10%しか未知語数が削減されなかったことに比べて大きく改善された．

音声認識結果に対して，話題分割と単語辞書適応を行った場合は，3000 語追加時に約 25%未知語数を削減することができたが，書き起こしを用いた場合の削減率には及ばなかった．特に，26 セグメントに分割された場合は，1500 語追加時まではターンごとに行われた単語辞書適応と大きな差は見られなかった．これは，話題分割の精度の低さが悪影響を及ぼしたものと考えられる．

次に，話題セグメントごとの言語モデル適応を，類似度の閾値 θ_2 の値を変化させて実行した．書き起こしに対する結果を図 4 に，音声認識結果に対する結果を図 5 に示す．書き起こしに対して 60 万語のテキストを用いて適応を行った場合に，テストセットパープレキシティがベースラインに対して最大約 9%削減された．これは，ターンごとに適応を行った場合とほぼ同程度の削減率であった．しかし，適応に用いるテキストが少ない場合のパープレキシティの削減率は，ターン毎に適応を行う場合に比べて大きかった．音声認識結果に対しても同様にテストセットパープレキシティが削減された．次に，3 章で行った自動

話題分割の結果と比較・考察を行う．書き起こしに対する話題分割で F 値が最大となったのは話題数が 25 のときである．これに対して，テストセットパープレキシティの削減率が最大となるのは話題数が 65 のときであった．話題数が 38, 65 のときも F 値は話題数 25 の場合に近い値であり，話題分割の精度を高めることで言語モデル適応の効果が上げられたといえる．音声認識結果に対しても同様に，話題分割の精度が高かった話題数 45, 69 のときにパープレキシティの削減が大きくなった．

また，話題分割の精度が高い場合には，少量のテキスト収集によりパープレキシティが大きく削減され，収集テキストを増加させていくと収束する傾向が見られた．これにより，最適な閾値の設定も容易であると考えられる．

5　おわりに

本稿では，会議音声を対象に，PLSA により抽出した特徴量に基づき自動話題分割を行い，得られた話題ごとに新聞記事データベースから関連記事を収集し，単語辞書への未登録語の追加と言語モデルの適応を行う手法を提案した．衆議院予算委員会の書き起こしと音声により評価を行ったところ，未知語率が書き起こしに対して約 30%，音声認識結果に対して約 25%，テストセットパープレキシティが書き起こし，音声認識結果のそれぞれに対して約 9%削減された．

参考文献

[1] Y.Akita, M.Hasegawa, and T.Kawahara, Automatic audio archiving system for panel discussions. In Proc. ICME, 2004.

[2] Y.Akita and T.Kawahara, Language model adaptation based on PLSA of topics and speakers for automatic transcription of panel discussions, IEICE Trans., Vol.E88-D, No.3, pp.439–445, 2005.

[3] 伊藤友裕，西崎博光，関口芳廣，WEB 上の類似記事を利用した音声文書の認識性能の改善，情報処理学会研究報告，2005-SLP-59-9, 2005.

[4] K.Ohtsuki, N.Hiroshima, M.Oku, and A.Imamura, Unsupervised Vocabulary Expansion for Automatic Transcription of Broadcast News, In Proc. ICASSP, 2005.

[5] 別所克人，クラスター内変動最小アルゴリズムに基づくトピックセグメンテーション，情報処理学会研究報告，2002-NL-154-25, 2003.

[6] 越中孝文，磯健一，奥村明俊，HMM の変分ベイズ学習によるテキスト文書の話題分割法，情報処理学会研究報告，2004-SLP-51-9, 2004.

[7] 長友健太郎，西村竜一，小松久美子，黒田由香，李晃伸，猿渡洋，鹿野清宏，相補的バックオフを用いた言語モデル融合ツールの構築，情報処理学会論文誌， Vol. 43, No9, pp. 2884-2893, 2002.

[8] T. Hoffman, Probabilistic Latent Semantic Indexing, In Proc. SIG-IR, 1999.