大学の講義におけるスライド情報を用いた発話の話題検出

北出 祐　　河原 達也

京都大学 情報学研究科 知能情報学専攻

kitade@ar.media.kyoto-u.ac.jp

1　はじめに

近年，高速なネットワーク環境が整備され，音声や映像などの大規模なディジタルコンテンツの閲覧が容易になった．教育分野においても，講義をアーカイブとして蓄積し，これらを教材として復習や遠隔学習に利用する環境が整いつつある．このようなアーカイブから目的の情報を検索したり，内容を容易に把握できるようにするためには，インデックスなどの二次情報の付与が極めて重要である．その反面，これらの情報の付与には膨大な人的・時間的コストがかかるため，自動的に作成する技術が望まれている．

我々は以前，学会講演を対象として談話標識に基づいてセクションに分割する方法を提案し，この情報が重要文抽出においても有用であることを示した [1][2]．これに対して近年，講演や講義においてスライドが用いられることが一般的になってきたので，このスライドを基にインデックスを作成することが考えられる [3]．本研究でもこのアプローチを採用し，スライドに記述されたキーワードを用いてスライドを単位として発話を対応付ける．ただし，スライドに記述されるキーワードは全般的に少ないので，複数のスライドから構成されるトピックを用意し，スライドまたはトピックの情報を用いて発話を対応付ける．

本稿では，スライドの提示順序や切り替えのタイミングの情報が利用可能な場合と，より汎用的な，講義の音声（の書き起こし）とスライドのみ（構成順）を利用した場合の二通りを想定して対応付けを検討する．

また，本研究で対象とする大学の講義では，スライドを用いた説明以外に，ビデオやホワイトボードを用いた説明や学生とのインタラクションなど，スライドと直接対応しない部分もかなりの割合を占めている．そこで，汎用的なトピック外モデルを構成し，スライドに直接対応しない発話の検出も試みる．

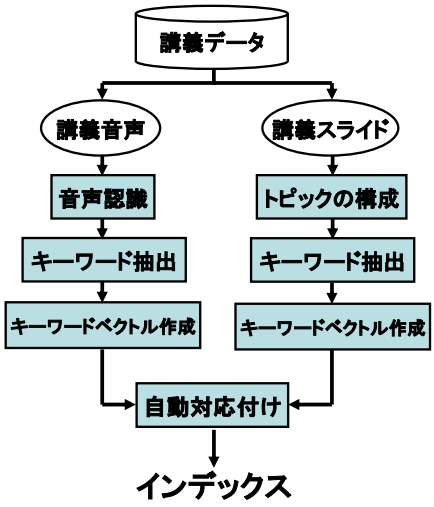

図 1: 講義のインデキシングの処理の流れ

2　スライドの言語情報に基づく発話との対応付け

処理の全体の流れを図 1 に示す．

あらかじめ人手により類似した内容のスライドをまとめあげて，トピックを作成する．トピック系列およびスライド系列は構成順に並べた系列である．次にスライドおよび発話からキーワードを抽出する．抽出するキーワードは，品詞が名詞，数詞，記号の単語である．その際，トピックについては，すべてのスライドのキーワードを用いる．これらのキーワードをもとに，講義の発話をスライドもしくはトピックに対応付ける（図 2）．

2.1　スライド /トピックとのマッチング尤度

講義の大半はスライドに基づいて行われているので，スライドに記されたキーワードを多く含む発話は，そのスライドの内容を説明している可能性が高い．そこで，スライドと各発話との関連度を測る尺度

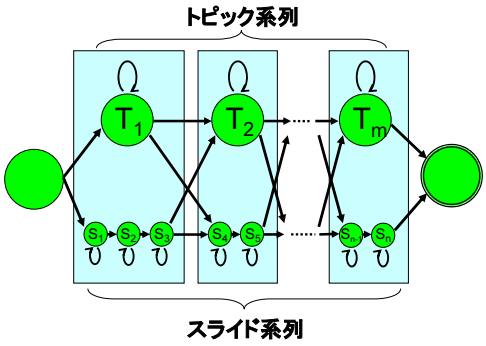

図 2: 対応付けのためのマルコフモデル

として，コサイン距離（式（1））を用いる．

$$dist(X,Y) = \frac{W(X)\cdot W(Y)}{|W(X)||W(Y)|} \quad (1)$$

ここで $W(X)$ は，X のキーワードの頻度ベクトルで，キーワード数の次元からなり，各要素に X 中のキーワードの出現頻度が代入されている．

スライドには講義内容を端的に表すキーワードが記されているが，記述されたキーワードは極めて少なく，発話の中に表れるキーワードはさらにその一部であるために，スライドと発話が関連していても全く対応付けられない場合が多数存在する．そこで，以下に述べる 2 パスの方法により状態出力尤度を計算する．

第 1 パス

スライドの状態出力尤度は，スライドのテキストと各発話のコサイン距離とする．よって，発話 U_i のスライド S_j に対する尤度 $O(U_i,S_j)$ は以下のようになる．

$$O(U_i,S_j) = dist(U_i,S_j) \quad (2)$$

トピック系列については，トピック T_k に含まれるすべてのスライド S_j と発話 U_i とのコサイン距離（から一定値 λ を引いた値），またはトピック T_k に含まれるすべてのスライドのテキストと発話 U_i のコサイン距離の最大値を状態出力尤度とする（式（3））．

$$O(U_i,T_k) = \max\{dist(U_i,S_j)-\lambda, dist(U_i,T_k)\} \quad (3)$$

第 2 パス

第 1 パスにおいて，スライドに対応付けられた発話集合も含めたテキストを構成してコサイン距離を求める．具体的には、スライド S_j（またはトピック T_k），および第 1 パスにおいてスライド S_j（またはトピック T_k）に対応付けられた発話 U_i の集合 S'_j（トピックについても同様に T'_k）とのコサイン距離を同様に計算する．

さらにここで得られた結果を用いて，トピック外の発話を検出する．具体的には，前後 2 発話 $\{U_{i-2},\cdots,U_{i+2}\}$ の尤度に基づく平滑化を行い，その値 $O'(U_i,S_j)$ がしきい値 δ 以下の発話をトピック外の発話とする．

2.2 状態遷移尤度

次に，ある発話において状態が遷移する尤度を定義する．スライドやトピックの転換点では，スライドを切り替えたり，少し間をとったりするために，他の箇所よりも比較的長めのポーズが挿入されると考えられる．そこで発話の直前のポーズ長の z-score による尺度 $\overline{pause}(U_i)$ を定義する．

これに加えて談話標識に基づく統計量も導入した．ここで談話標識は話題の転換点に用いられる特徴的な表現である．談話標識は『日本語話し言葉コーパス』の学会講演 889 講演から（教師なしで）学習した [1]．

この談話標識らしさを示す統計量を利用して，スライドやトピックの切り替えの尺度 $DM(U_i)$ を定義する．ただし，これもポーズ長と同様に z-score による正規化を行う（$\overline{DM}(U_i)$）．ポーズ長に基づく尺度と談話標識に基づく尺度はともに話題の境界らしさを表す尺度であるので，重み付き和によりスライド S_j（またはトピック T_k）から次のスライド S_{j+1}（またはトピック T_{k+1}）に遷移する尤度 $a_{j,j+1}(U_i)$ を定義する．同様に，同一のスライド /トピックにとどまる尤度 $a_{j,j}(U_i)$ も定義する．

$$\begin{cases} a_{j,j+1}(U_i) &= \frac{\beta}{1+\alpha}(\overline{pause}(U_i)+\alpha * \overline{DM}(U_i)) \\ a_{j,j}(U_i) &= -\frac{\beta}{1+\alpha}(\overline{pause}(U_i)+\alpha * \overline{DM}(U_i)) \end{cases} \quad (4)$$

2.3 最尤状態系列の導出による対応付け

以上により定義されたマルコフモデルに基づいて発話系列に対して最尤出力系列をビタビアルゴリズムを適用することによりスライドとの対応付けを行う．

3 自動対応付けの評価実験

3.1 実験データ

評価実験に用いたのは，京都大学で行われた 5 講義である．そのデータの概要を表 1 に示す．人手によっ

表 1: 実験データ

ID	MM031126	PA040707	IP040706	PR041110	PR041117
時間 (min)	78.8	75.3	81.4	80.7	83.4
発話数 （トピック外の発話）	437 (28)	648 (258)	643 (172)	293 (68)	340 (90)
総単語数	14519	14913	14845	11947	11171
スライド数 （異なりスライド数）	99 (43)	35 (23)	 (21)	95 (23)	 (9)
トピック数	17	7	13	11	5
キーワード総数 （異なりキーワード数）	2189 (241)	1204 (183)	3967 (447)	1498 (299)	599 (164)

てまとめられたトピックの数は，スライドの数のおよそ 30〜50%程度である．

本実験においては，状態出力尤度での式（3）における λ を 0.1，式（4）におけるパラメータ α は 1，β は 0.001 に，事後的に決定している．

講義の（人手による）書き起こしを対象に発話単位にあらかじめ人手によりスライド番号/トピック番号またはトピック外の発話のいずれかにタグ付けしておき，正解として用いる．

評価には，その正解数に基づく再現率（recall），適合率（precision），F 値（F-measure）を算出しその平均で評価を行う．

3.2 実験結果

スライド/トピックと発話の自動対応付けの結果を表 2 に示す.（図 2 において）スライド系列とトピック系列との間の遷移を許さず，それぞれスライド系列，トピック系列のみで行った結果もあわせて示す．正解のスライドとの一致精度（適合率）は，およそ 65%であった（表 3）．トピック系列のみの場合は，スライド系列のみの場合よりも対応付けが容易になり，同等もしくは高い精度が得られた．また，スライド/トピック系列間の遷移を許した方が，スライド系列のみで行うよりも精度が高くなった．

最後にトピック外の発話の検出精度を表 4 に示す．トピック外の発話の割合が，講義によってかなりの差がある．トピック外の発話の多い講義（PA040707）においては，その検出精度がスライド/トピックへの対応付けの精度（適合率）に大きな影響を及ぼすことがわかる（表 2）．

表 2: 自動対応付け結果

		再現率	適合率	F 値
MM031126	スライド系列のみ	0.611	0.661	0.635
	トピック系列のみ	0.665	0.743	0.702
	系列間の遷移あり	0.738	0.716	0.727
PA040707	スライド系列のみ	0.454	0.441	0.448
	トピック系列のみ	0.464	0.492	0.478
	系列間の遷移あり	0.405	0.386	0.395
IP040706	スライド系列のみ	0.773	0.735	0.754
	トピック系列のみ	0.762	0.754	0.758
	系列間の遷移あり	0.775	0.736	0.755
PR041110	スライド系列のみ	0.329	0.548	0.411
	トピック系列のみ	0.707	0.652	0.678
	系列間の遷移あり	0.742	0.690	0.715
PR041117	スライド系列のみ	0.600	0.679	0.637
	トピック系列のみ	0.680	0.806	0.738
	系列間の遷移あり	0.656	0.659	0.657
全講演平均	スライド系列のみ	0.582	0.623	0.601
	トピック系列のみ	0.654	0.685	0.669
	系列間の遷移あり	0.663	0.636	0.649

表 3: スライドに対応付けられた発話数

ID	適合率 (正解一致数/スライド判別数)
MM031126	0.699 (230/329)
PA040707	0.412 (122/296)
IP040706	0.736 (365/496)
PR041110	0.665 (147/221)
PR041117	0.659 (164/249)
全講演平均	0.646 (1028/1591)

4 スライドの切り替え時間情報を用いた対応付け（トピック外発話の検出）

最後にスライドの提示順序ならびに切り替え時間の情報を利用できる前提で発話との対応付けを行う．この場合，スライド S_j を提示している時間中に含まれる発話 U_i がスライド S_j（もしくはスライド S_j が属するトピック T_k）の内容を話しているか，トピック外の発話（OOT）を話しているかの二値判別の問題

表 4: トピック外の発話の検出結果

ID	再現率	適合率	F 値
MM031126	0.429	0.414	0.421
PA040707	0.519	0.561	0.539
IP040706	0.448	0.524	0.483
PR041110	0.309	0.412	0.353
PR041117	0.722	0.714	0.718
全講演平均	0.518	0.555	0.536

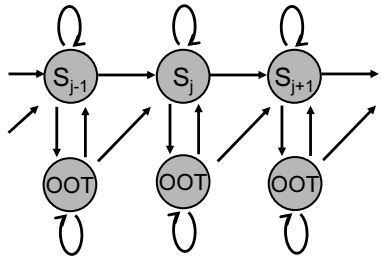

図 3: スライドの切替え情報を利用した講義の自動対応付け

として定式化する（図 3）.

4.1 トピック外モデルの導入

トピック外の発話の検出のために,『日本語話し言葉コーパス（CSJ）』の講演データベースから汎用的なトピック外モデルを構成する．具体的には，TFIDF 値を求めて，しきい値以上の単語 w を抽出する.

その上で講演コーパスから単語 w を含まない文を抽出し，文集合 S_{OOT} を得る．そして，再度文集合 S_{OOT} に出現する単語 w_{OOT} を抽出し，これらをトピック外モデル（OOT）のキーワードとした．この頻度ベクトルに基づいて，各発話 U_i のトピック外モデルに対する尤度 $O(U_i, OOT)$ を求める.

4.2 継続時間長に基づく遷移尤度

スライド系列とトピック外モデルの間で状態が遷移する尤度も導入する．ここでは，同一状態の発話の継続時間長に基づいて尤度を定義する（式（5））.

$$a'(U_i) = \begin{cases} \beta' * \exp(-\alpha' \log(t_{S_j})) \\ \beta' * (1 - \exp(-\alpha' \log(t_{OOT})) \end{cases} \quad (5)$$

ここで t_{S_j}，t_{OOT} は，それぞれスライド S_j，トピック外モデル OOT の継続時間長を表す.

表 5: スライドの切り替え時間情報を用いたトピック外の発話の検出結果

ID	再現率	適合率	F 値
MM031126	0.750	0.304	0.432
PA040707	0.583	0.791	0.671
PR041110	0.184	0.696	0.291
全講演平均	0.503	0.648	0.566

4.3 実験結果

3 章での評価実験に用いた講義のうち，スライドの切り替え時間が記録された MM031126，PA040707，PR041117 の計 3 講義を用いて評価実験を行った．式（5）におけるパラメータ α' は 0.001 に，β' は 0.01 に事後的に決定した.

トピック外の発話検出結果を表 5 に示す．スライド切り替え時間情報を用いない場合に比べ，トピック外の発話検出精度が高くなり，全発話の判別精度も改善された.

5 結論

講義におけるスライド情報を用いた発話の対応付けの方法を提案した．スライドのキーワードをもとに，発話とのコサイン距離を計算し，両者の関連度を定義した．また，遷移尤度としてポーズ長と談話標識に基づく尺度も導入した．これらの尺度に基づいたマルコフモデルによって発話との対応付けを行った．より柔軟な対応付けを行うために，複数のスライドから構成されるトピックを用意し，またトピック外の発話の検出も試みた．今後は，実験データを増やし，音声認識結果に対しても評価を行う予定である.

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. 情処学論, Vol. 43, No. 7, pp. 2222–2229, 2002.

[2] T. Kawahara, M. Hasegawa, K. Shitaoka, T. Kitade, and H. Nanjo. Automatic indexing of lecture presentations using unsupervised learning of presumed discourse markers. *IEEE Trans. Speech & Audio Process.*, Vol. 12, No. 4, pp. 409–419, 2004.

[3] 河原達也, 石塚健太郎, 堂下修司. 発話検証に基づく音声操作プロジェクタとそれによる講演の自動ハイパーテキスト化. 情報処理学会論文誌, Vol. 40, No. 4, pp. 1491–1498, 1999.