トピックモデルを用いた映像コンテンツの理解支援

岡本 昌直 [1)]　祖父江 美香 [2)]　祖父江 翔 [1)]　中村 明 [3)]　田村 哲嗣 [2)]　速水 悟 [2)]

1)　岐阜大学大学院　工学部研究科
2)　岐阜大学　工学部
3)　三洋電機（株）エコロジー技術研究所

1. はじめに

現在，インターネットの発達に伴い，TV 番組のインターネット配信や動画投稿サイトなど，さまざまな形で映像コンテンツに触れる機会が増加している．しかし，膨大な映像コンテンツの中から，ユーザごとに求める情報のみをユーザ自身が発見するのは困難であり，また，閲覧にも多大な時間を消費する．そのため，映像コンテンツをリアルタイムで理解することを支援するシステムの開発が望まれている．その代表例として字幕が挙げられ，近年の音声認識技術の進展により，リアルタイムでの放送音声への字幕付与を目指した研究が行われている[1,2]．しかし，音声認識結果をそのまま字幕に用いた場合，冗長な箇所が多いという問題があるため，音声要約を行う必要があると考えられる．また，字幕提示方式もユーザの理解に影響を与える．これまでに，会議議事録のように複数話者を対象とした際の字幕提示方式の検討がなされている[3]．一方で，書き起こし文からキーワードを抽出する手法も考えられる．キーワードは映像コンテンツの内容を端的に表しているため，効果的にユーザに内容を伝えることが可能となる．これまでに我々は，TF-IDF など複数特徴量を用いた，線形回帰によるキーワード自動抽出，ジャンルにおける重要度ベクトルの調査・最適な字幕提示方式の検討を行った[4]．

トピックごとにキーワードを提示することで，ユーザはより深い理解を得ることができると考えられる．そこで本研究では，大語彙音声認識 エンジン Julius を用いた字幕自動生成へ向けた取り組みとして、音声区間検出(VAD)と条件付き確率場(CRF)を用いて，文境界推定を行った．また，LDA トピックモデルを用いて，トピック境界を推定し，キーワード抽出を行い，吹き出し型字幕として提示する．

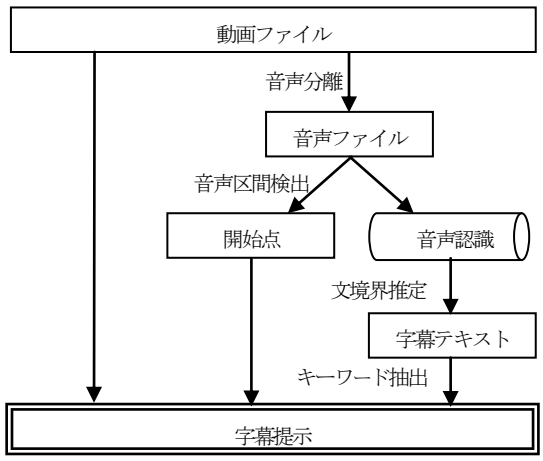

図 1　　システム概略図

2. 字幕自動生成システム

2.1 システムの概要

本研究が提案する字幕自動生成システムの概要を図 1 に示す．まず，動画ファイルより音声ファイルを取り出し，音声認識を行う．認識結果より，動画ファイル，音声区間開始点，字幕テキストを統合し，Adobe Flash を用いて吹き出し型字幕としてユーザに提示する．

音声区間検出・文境界推定の実験には 2008 年 11 月 18 日放送分 NHK 時論公論の動画ファイルを使用した．

2.2 VAD (Voice Activity Detection)

音声ファイルに対しフレーム分割を行い，各フレームにおいて得られた特徴を基に，音声/非音声の識別を行う．音声区間の特徴としては，音声/非音声区間の間にあるポーズ情報を利用したものが挙げられる[5]．

今回使用した音声ファイルには雑音が少ないため，ポーズ情報のみを特徴量に使用し，閾値処理を行うことにより音声区間の検出を行った．また，ハングオーバー処理を行った．ハングオーバー処理とは，音声区間の抜けている部分を補う処理のことである．最適フレーム数を検討するために，表 1 に示す条件で音声区間検出実験を行った．

表 1　　音声区間検出実験条件

動画	NHK 時論公論（約 10 分）
フレーム数	1~7
時間誤差	5ms
字幕提示	TV 型字幕
字幕文字数	制限なし
話者	1 人

フレーム数 1~7 で音声区間検出実験を行った結果，フレーム数 4,5,6 のときに，精度が最も高く，85.8%であった．

2.3 CRF (Conditional Random Fields)による文境界推定

VAD による音声区間検出結果に句点を付与し，「文」を入力単位と仮定する自然言語処理を行うため，文境界の推定を行う必要がある．

文境界の推定はラベリング問題として考えることができる．そこで，対象テキストの形態素解析結果列にラベルを付与する．また，ポーズの部分は，書き起こしと比較した際，句読点部分である可能性が高いと考えられるため，これら 2 つを特徴量とし，識別モデルとして CRF を用いた．CRF は，入力例 x に対する各出力ラベルの列 y の条件付き確率$P_\theta(y|x)$を表現する．θは学習により求められるモデルのパラメータで，それらをベクトルにし

たものが Θ である．位置 i の素性ベクトルを $f(y,x,i)$，それに基づく大域素性ベクトルを $F(y,x)=\sum_i f(y,x,i)$ とすると，$P_\theta(y|x)$ は次式で求められる．

$$P_\theta(y|x)=\frac{\exp\left(\Theta\cdot F(y,x)\right)}{Z_\theta(x)} \quad (1)$$

$$Z_\theta(x)=\sum_y \exp\left(\Theta\cdot F(y,x)\right) \quad (2)$$

音声認識などで使われる隠れマルコフモデル(HMM)は，特徴が互いに独立である必要がある．これに対し，CRF はその必要がなく，HMM より細かい特徴の指定が可能である．また，条件付き確率により確率が直接推定できるという特徴がある．

本研究では，句点があらかじめ挿入された新聞記事と話し言葉のテキストデータより，モデルを作成する．音声認識のテキストデータに，テキスト情報のみで作成した識別モデルを適用して文境界推定を行う．学習に用いる適切な素性を決定するため，前後の形態素数，使用する単語情報について，CRF による学習で予備実験を行った．予備実験により，学習素性は，前後 2 形態素を用いて単語情報に表層形と品詞を用いる場合が適切という結果になった[6].

今回使用した動画ファイルにおける文境界推定結果の精度，再現率，F 値を表 2 に示す．

$$\text{精度}=\frac{\text{正解の句点と本手法で挿入した句点との一致数}}{\text{本手法で挿入した句点数}} \quad (3)$$

$$\text{再現率}=\frac{\text{正解の句点と本手法で挿入した句点との一致数}}{\text{人手で挿入した句点数}} \quad (4)$$

$$\text{F 値}=\frac{2\times\text{精度}\times\text{再現率}}{\text{精度}+\text{再現率}} \quad (5)$$

表 2　　文境界推定結果

精度	再現率	F 値
92.6	98.4	95.5

2.4 吹き出し型字幕提示方式

映画などでみられる，一般的な TV 型字幕は，発話内容が話者の下に表示されることが多い．これに対し，吹き出し型字幕提示方式とは，図 2（右）のように，話者の顔付近に字幕を表示する方式である．この字幕表示方式をとることによって，話者が複数存在する場合においても，誰が・いつ・何を話したのか認識しやすくなると考えられる．図 2 に TV 型字幕と吹き出し型字幕の例を示す．

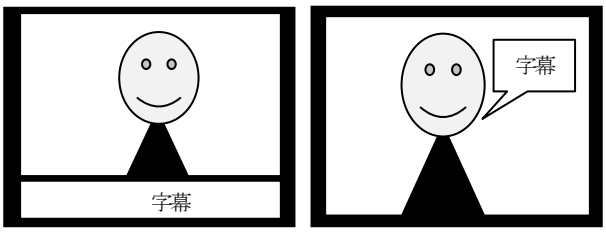

図 2　字幕提示方式（左：TV 型，右：吹き出し型）

2.5 システム評価実験

本研究の提案手法によって作成されたコンテンツによるユーザの理解支援への有効性を示すため，5 段階評価にて被験者実験を行った．以下に評価項目を示し，実験条件を表 3 に示す．また，実験結果を図 3 に示す．

【評価項目】
(1) 字幕の切り替えのタイミング
(2) 表示文の句切れ
(3) 吹き出し型字幕の見やすさ
(4) 表示文字数
(5) 全体の評価

表 3　　システム評価実験条件

話者数	1
字幕文字数	制限なし
コンテンツの長さ	約 3 分
被験者数	14
字幕提示方式	吹き出し型

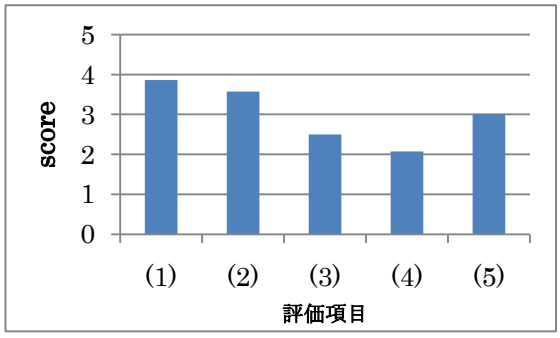

図 3　　システム評価実験結果

評価項目(1)，(2)より，VAD による音声区間検出と CRF を用いた文境界推定を用いた字幕提示は有効であることが分かる．しかし，1 つ 1 つの音声区間内において，文字数にばらつきがみられた．これにより，1 度の吹き出しに表示される文字数が増加し，複数行にまたがって字幕が表示される，被験者が 1 秒当たりに読みとる文字数が増えるといった問題のために，(3)，(4)の評価項目を下げたと考えられる．そのために，改行の挿入，1 度に表示する字幕文字数制限，キーワード抽出などの検討を行う必要があると考えられる．また，吹き出し型字幕は，本実験のように話者 1 人における状況下では，TV 型字幕提示方式よりも理解を損なう可能性があるため，複数話者での吹き出し型字幕の有効性を検討する必要がある．

3. キーワード抽出

キーワードはトピックを端的に表すという仮定に基づき，複数のトピックが存在する場合においても，トピックごとにキーワードを提示することにより，ユーザは現在述べられているトピックを容易に理解することができる．トピックごとにキーワードを提示するために，トピック境界の推定，キーワード抽出を行う必要がある．本

研究では LDA を用いて，トピック境界推定，キーワードの抽出を行った．形態素解析には形態素解析エンジン MeCab を使用した．また，学習テキストには 2008 年の毎日新聞 1 年分の記事を用いた．

3.1 LDA(Latent Dirichlet Allocation)

確率・統計的自然言語処理や音声認識の分野では，単語の生起確率を直前の(N-1)単語を用いてモデル化した N-gram モデルや，単語間の大域的な依存関係を単語対の関係でモデル化したトリガーモデルやキャッシュモデルが多用されている．これらに対し，単語間の大域的な依存関係を話題としてモデル化したものにトピックモデルがあり PLSI（Probabilistic Latent Semantic Indexing）[7] や，DM（Dirichlet Mixtures）[8] などが挙げられ，LDA もこれに該当する．トピックモデルは，現在の話題に応じて単語の生起確率を動的に推定でき，言語モデルの高精度化が期待できる．LDA は，各潜在トピック$(z_1, z_2 \dots z_C)$(C : 潜在トピック数) の生成確率$\theta = (\theta_1, \theta_2 \dots \theta_C)$ が多項分布の共役事前分布であるディリクレ分布 $Dir(\theta|\alpha)$ に従うと仮定したモデルである．文書 $d = (w_1, w_2 \dots w_{|d|})$ の出現確率は次式で表される（$|d|$ は文書 d の総単語数を表す）

$$P(d \mid \alpha, \beta) = \int Dir(\theta \mid \alpha) \left(\prod_{n=1}^{|d|} \sum_{k=1}^{C} P(w_n \mid z_k, \beta) P(z_k \mid \theta) \right) d\theta \tag{6}$$

α，β が LDA のモデルパラメータであり，β_{kj} はトピック z_k における語 w_j の uni-gram 確率 $P(w_j|z_k)$ を表す$(1 \leqq j \leqq V)$(V: 語彙数)．$\alpha = (\alpha_1, \alpha_2 \dots \alpha_C)$はディリクレ分布のパラメータである．パラメータ α，βの学習には変分ベイズ法による近似計算が用いられる[9]．本研究では対象となる文書にフレーム化処理を行う．未知のフレーム化された文書 fに対するトピック適応は，学習時と同様の変分近似により計算される．即ち，fに対する変分パラメータγ_kおよびϕ_{kj} を導入し，学習済みの α，βを用いて以下の手順を収束するまで繰り返す．

$$VB - Estep : \phi_{kj} \propto \beta_{kj} \exp(\Psi(\gamma_k) - \Psi(\sum_{k'=1}^{C} \gamma_{k'})) \tag{7}$$

$$VB - Mstep : \gamma_k = \alpha_k + \sum_{j=1}^{V} n(h, w_j) \phi_{kj} \tag{8}$$

$\Psi(\gamma)$ は digamma 関数であり，$n(h, w_j)$ は h における語 w_j の出現回数を表す．得られたγ_k をフレーム化された文書 f の元での各潜在トピックの混合比とする．したがって，フレーム化された文書 f の元での語w_j の生起確率は次式により与えられる．

$$P(w_{j'} \mid h) = \frac{\sum_{k=1}^{C} \gamma_k \beta_{kj'}}{\sum_{k=1}^{C} \gamma_k} \tag{9}$$

LDA はトピックの事前分布にディリクレ分布を用いることにより，トピックの拡がりやトピック間の関係を表現できる点で PLSI より優れている．またベイズ推定に基づくため過適応の問題が少ないとされている．

3.2 トピック境界推定

LDA は，1 つの文書内に複数の潜在トピックが同時に混在していると考えたモデルである．そのため，トピックの混合比は，潜在トピック数を要素とし，トピック混合比ベクトル$(\gamma_1, \gamma_2 \dots \gamma_C)$(C:潜在トピック数)として表される．本研究では，対象となる文章を移動幅 1 文でフレーム化し，次式に表すコサイン尺度を算出する．

$$cos(t_1, t_2) = \frac{t_1 \cdot t_2}{\|t_1\| \|t_2\|} \tag{10}$$

トピック混合比ベクトルt_1，t_2間のなす角度を測り，閾値以下の場合にトピックの変化点と判定する．以下の図 4 にトピック混合比ベクトルの概略を示す．

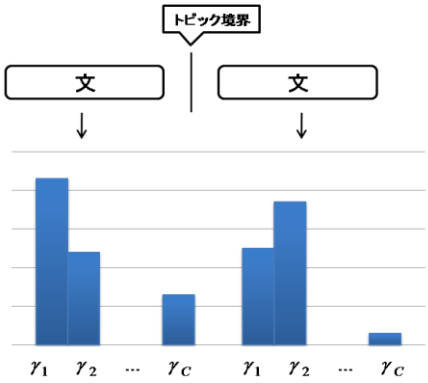

図 4　トピック混合比ベクトル概略図

3.3 複合語

複合語とは，2 つ以上の単語が組み合わさってなる語のことである．形態素解析において，MeCab 標準の IPA 辞書を用いた結果では，例として「日経平均株価」という語は，「日経」，「平均」，「株価」として解析される．キーワードとして提示される語としては，不十分であると考えられるため，複合語の処理を検討する必要がある．複合語を扱うために，形態素解析に使用する MeCab 辞書に,2009 年 9 月時点での Wikipedia に存在するページのタイトルとなる語，約 90 万語を素性「名詞・複合語」として登録した．それらに加え，学習テキスト内で，名詞と名詞が隣接している場合，それらを 1 つの語とする，約 56 万語を辞書に登録した．

3.4 キーワード抽出

3.2 より，同トピックと判定された範囲内で，キーワードの抽出を行う．LDA を用い，トピックz_k における語 w_j の uni-gram 確率β_{kj} をソートすることにより，単語ごとに出現しやすいトピックを判定し，そのトピックに属するものとする．その際，潜在トピック数は 100 とした．文書 d に対し，トピック境界を推定した後，同じトピックであると判定された範囲 R において，トピック混合比を求め，閾値 T 以上となる潜在トピックγ_k に属する単語を範囲 R におけるキーワードとする．

$$R_n = \{\gamma_k \mid \gamma_k \geq T\} \quad (1 \leq k \leq 100) \tag{11}$$

キーワードとされた語にはトピックごとに色を付け，ユーザに提示する．図 5 にトピックごとに抽出されたキーワード例を示す．

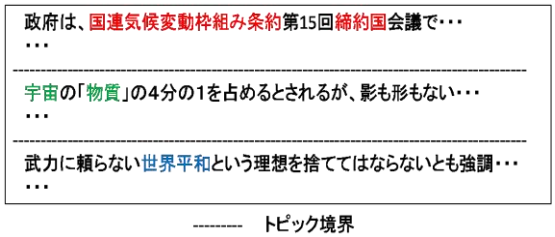

図 5　キーワード抽出例

4. 被験者実験

2 章の表 3 と同条件にて，抽出されたキーワードの適切さ・キーワード提示によるユーザの理解支援への有効性を示すために，キーワード提示に対する被験者実験を行った．以下に評価項目を示し，図 6 に作成したコンテンツの提示例，図 7 に被験者実験結果を示す．

【評価項目】
(1) キーワード提示数
(2) 提示キーワードの適切さ
(3) 理解支援に役立つか
(4) トピックの変化を把握できるか

図 6　コンテンツ提示例

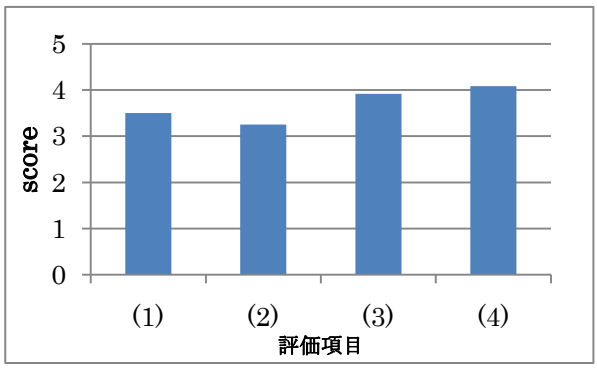

図 7　被験者実験結果

評価項目(4)より，提案手法はトピック変化の把握を支援することがわかる．しかし，本研究では，トピック混合比が閾値以上となる潜在トピックに属する単語すべてをキーワードとしたために，1 度に多数のキーワードを提示する場面がみられた．そのため，評価項目(1)，(2)を下げたと考えられる．また，キーワードを提示する際の色によって，被験者の理解度の変化がみられる可能性がある．そのため，字幕における色の影響について検討を行う必要があると考えられる．

5. まとめ

本研究では，ユーザの理解を支援することを目的とし，音声認識，VAD，CRF による文境界推定，LDA を用いて，吹き出し型字幕自動生成に取り組んだ．今後の課題として，完全な自動化へ向けて，吹き出し位置の決定を自動で行う必要があるため，動画内の話者の顔の位置を特定する必要があると考えられる．被験者実験では，話者 1 人の状況下における実験を行ったが，話者識別を行い，複数話者が存在する状況下での TV 型字幕，吹き出し型字幕の双方で検討する．また，トピック境界推定，キーワード抽出の精度を上げることで，よりユーザの理解を支援することができると考えられるので，検討する必要がある．

参考文献

[1] 堀智織,古井貞熙, “単語抽出による音声要約生成法とその評価”,電子通信情報通信学会誌,D-Ⅱ NO.2 ,pp.200-209 (2002)
[2] 大野誠寛,松原茂樹,柏岡秀紀,稲垣康善,“同時的な独話音声要約に基づくリアルタイム字幕生成”,情報処理学会研究報告 Vol.2006, No.73, 2006 - SLP–62 -(10),pp.51-56 (2006)
[3] 藤井絢子,南條浩輝,吉見毅彦, “会議の情報保障を目的とした吹き出し型字幕提示方式の検討”,情報処理学会研究報告,2009-SLP-75-14,pp.75-82 (2009)
[4] 岡本昌直,祖父江美香,山本けい子,田村哲嗣,速水悟,“映像コンテンツの理解支援のためのキーワード提示方式の検討”,第 8 回情報科学技術フォーラム, E-021,pp.299-300 (2009)
[5] 羽柴隆志,竹内伸一,田村哲嗣,速水悟, “マルチストリーム HMM を用いた音声と画像による音声区間検出”,日本音響学会 2009 年春季講演論文集,1-P-5,pp.131-132(2009)
[6] 祖父江翔,山本けい子, 田村哲嗣,速水悟, “音声認識結果の文境界推定における識別モデルの評価”,言語処理学会,第 15 回年次大会,P2-28,pp.582-585(2009)
[7] T.Hofman,“Probabilistic latent semantic indexing”, Proc.of 22nd Annual ACM Conference on Research and Development in Information Retrieval, pp.50-57(1999)
[8] 貞光九月,三品拓也,山本幹雄, “混合ディリクレ分布を用いたトピックに基づく言語モデル”,電子情報通信学会論文誌 D-Ⅱ Vol.J88-D-Ⅱ,NO.9, pp.1771-1779(2005)
[9] D.Blei,A.Y.Ng and M.Jordan, “Latent dirichlet allocation”, journal of Machine Learning Research, Vol.3,pp.993-1022(2003)
[10] 津田 裕亮, 中村 明, 速水 悟, 松本 忠博, 池田 尚志, “LDA トピックモデルに基づく話題変化点検出”,言語処理学会, 第 15 回年次大会, P2-25, pp.570-573 (2009)
[11] 門馬隆雄,江原暉将,白井克彦,沢村英治,三橋哲雄,“聴覚障害者向けニュースの字幕提示方法に関する主観評価”, 映像情報メディア学会誌 Vol.54, No.9, pp.1288–1297 (2000)
[12] J.Lafferty, M.Andrew, P.Fernando, “Conditional Random Fields:Probabilistic Models for Segmenting and Labeling Sequence Data”, In Proceeding of the 18th International Conference on Machine Learning(2001)