ビデオ音声認識テキストからの文認定

中澤 聡　　佐藤 研治　　奥村 明俊
NEC　マルチメディア研究所
{nakazawa, satoh, okumura}@ccm.cl.nec.co.jp

1. はじめに

近年、ニュース番組や講演映像などのビデオデータに、テキスト情報をリンクさせて元のビデオデータの言語インデックスとし、検索や動画要約に利用するシステムが研究・開発されている[1][2]。この言語インデックスとして利用できるテキスト情報としては、クローズド・キャプションやビデオの音声認識テキスト、台本等を人手でタグ付けしたテキストなどが挙げられるが、字幕付き番組の割合がまだ少ない日本の放送コンテンツや一般家庭で作成されるホームビデオなどを対象としたときには、ビデオから自動的に作成できる音声認識テキストを言語インデックスの主対象として利用することになる。

米国ではほとんどのＴＶ放送にクローズド・キャプションが付与されており、通常の書き言葉と同様にピリオドなどで文の区切りがつけられている。よって、クローズド・キャプション中で予め区切られている文を抽出単位として、重要文抽出を施すことにより、比較的良質のビデオ要約が得られる[3]。

一方、ビデオの音声認識テキストは、基本的にはビデオの長さ分の連続した認識単語列であり、言語処理の基本単位である「文」に区切られていない。そのため、得られた音声認識テキストを従来の重要文抽出処理の入力として取り扱うには、連続して得られる音声認識テキスト中の文末を認定し、１文ごとに区切っていく処理が必須となる。

これまでも、連続した音声認識テキストを文などの言語処理単位に分割する手法が研究されてきた[4][5][6]。しかし、ビデオの音声認識はその性質上、読み上げ文の音声認識に比べて、認識性能が大きく悪化する傾向にあり、これらの既存の手法をそのまま用いることができない。

そこで本稿では、まず音声認識時に統計的言語モデルを用いて認識結果に句読点を挿入し、ついでポーズ長とパターン・マッチング・ルールから、最終的に文の区切りとして採用する箇所を判定する２段階の文認定手法を提案する。本稿では、２節で本手法が前提とするビデオ要約について簡単に説明し、３節で提案手法についての説明および評価結果について述べ、４節でまとめる。

2. 文認定を用いたビデオ要約

前節で述べたように、ビデオのオーディオ・トラックに音声認識をかけて得られたテキストを、言語インデックスとして用いることで、文書データに対する場合と同様に、ビデオの要約処理が可能となる。本稿で提案するビデオ音声認識テキストからの文認定手法は、このようなビデオ要約で使用されることを前提としている。図１はビデオ要約全体の大まかな処理手順である。

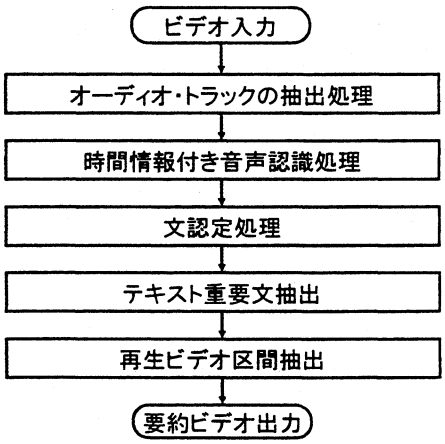

図１．ビデオ要約の処理手順

この手順では、ビデオデータが入力されると、データ形式変換等の前処理の後、ビデオのオーディオ・トラックを分離して、音声認識処理に渡す。

音声認識処理では、途中で話者が替わることを想定し男女共通不特定話者認識を行う。ただし、通常の音声認識処理と異なり、認識単語毎に各語が認識された時間情報を出力する。時間情報は元のビデオの先頭から計る。この時間情報により、ある単語が認識された時間が分かり、ビデオでその語が発話されたシーンにいきなり頭出しするといったことが可能となる。また、文認定処理により各文ごとのビデ

オ区間が分かる。後の文認定処理で用いるため、認識単語の品詞情報も併せて出力する。表1は、ビデオからの時間情報付き音声認識テキストの例である。

表1．ビデオ音声認識テキストの例

開始時間(秒)	終了時間(秒)	認識単語	品詞	正解テキスト
27.17	28.14	円満	形容動詞	えー
28.14	28.20	を	格助詞	まず
28.20	28.65	最初	名詞	最初
28.65	28.74	に	格助詞	に
28.74	28.99	する	動詞	ですね
28.99	29.14	。	記号	、
29.14	29.39	で	格助詞	えー
29.39	29.98	暗号	名詞	暗号
29.98	30.03	を	格助詞	
30.03	30.27	という	引用助詞	という
30.27	30.38	の	形式名詞	もの
30.38	30.73	が	格助詞	が
30.73	31.05	悪い	形容詞	一体
31.05	31.25	という	引用助詞	
31.25	31.51	どんな	連体詞	どんな
31.51	31.69	もの	形式名詞	もの
31.69	31.81	な	形容動詞語尾	な
31.81	32.09	のか	終助詞	のか
32.09	32.10	、	記号	
32.10	32.32	の	格助詞	という
32.32	32.56	こと	形式名詞	こと
32.56	32.63	を	格助詞	を
32.63	33.05	簡単	形容動詞	簡単
33.05	33.14	に	格助詞	に
33.14	33.26	説明	サ変	説明
33.26	33.48	し	サ変語尾	し
33.48	33.76	たい	助動詞	たい
33.76	33.91	。	記号	
33.91	34.08	と	接続助詞	と
34.08	34.30	思い	動詞	思い
34.30	34.51	ます	助動詞	ます
34.51	34.90	。	記号	。

音声認識処理の次は、得られた認識テキストのどの単語からどの単語までを1文とするか判定し、文末と判定された箇所に文末情報を挿入する文認定処理を行う。文認定の手法は3節で詳しく説明する。

最後に、文単位に区切られた音声認識テキストに対して重要文抽出を行う。その後、重要と判断された文に対するビデオ区間を求め、そのビデオ区間を結合するとビデオ要約が得られる。

以上が、ビデオ要約全体の処理手順である。

3. ビデオ音声認識テキストからの文認定

3.1. ビデオ音声認識

日本語の放送コンテンツやホームビデオなどのビデオデータに音声認識をかける際の問題点として、次のようなことが考えられる。

- 音量など発話品質が一定でない
- BGM など人間の音声以外の音が含まれている場合がある
- 複数の話者が同時に発話する場合がある
- 口語調の発話が含まれる
- 予め話者学習をしておくことが難しい

この結果、口述筆記などで音声認識を使用した場合に比べて音声認識率が大きく悪化する傾向にある。

一方、音声認識テキストを文分割する従来手法として以下の3手法が提案されている。

1. 音声認識テキストを構文解析し、得られた部分構造から区切りを判定
2. 発話間のポーズ長(無音区間)を調べ、一定時間以上のポーズ長を区切りとする
3. 音声認識時に統計的言語モデルによって、区切りを認定

1の構文解析を用いる手法については、入力テキストの分野が限定されておらず、充分な構文解析規則を用意することが困難であることや、誤認識や口語表現のため、多くの間違った部分構造ができてしまう可能性が高いことから、ビデオの音声認識テキストに使用することは難しい。

発話間のポーズ長は重要な手がかりである。しかし、必ずしも文間のポーズ長が、文内でのポーズ長より長いとは限らない。実際にある講義映像で人間が書き起こしたテキストを対象として調査したところ、文内に文間より長いポーズが含まれている文が、3から5割含まれていた。このため、ポーズ長だけを用いる文認定手法は精度が不十分であろう。

統計的言語モデルは、句点を推定するために有効な手法であることが報告されている[6]。しかし、ビデオの音声認識テキストには、元の発話の音量が充分になかった場合や、効果音と発話が重なった場合など、局所的に音声認識結果が極端に悪くなる現象がよく見られる。そのような箇所では、句点推定がうまく働かない。

そこで本稿では、第1段階として、音声認識時に統計的言語モデルを用いて句読点を挿入し、ついで第2段階として、ポーズ長の手がかりと、よくある句読点の挿入誤りに対応し、要約として適切な範囲が切り出せるよう調整したパターン・マッチング・ルールから最終的に文末として採用する箇所を取捨選択する、2段階の文認定手法を提案している。

3.2.文認定手法

文認定手法の手順は以下のようになる。

1. 統計的言語モデルを用いて、句読点を含む音声認識テキストを作成。
2. 音声認識テキストの時間情報から、単語間のポーズ長を求める。ただし、句読点が挟まれる場合は、句読点をとばして句読点の前の単語の終了時間と、句読点の次の単語の開始時間との差をポーズ長とする。
3. ポーズ長が0.1秒以上の箇所を、文末候補とする。各文末候補に対して、最初0点の得点を与える。
4. 各文末候補に対して、そのポーズ長に従って、得点を与える。
5. 各文末候補に対して、予め用意された文末強化ルールがその箇所に成立するか、1つずつ確認する。成立する場合には、各ルールに定められた得点を、その文末候補の得点に加算する。
6. 各文末候補に対して、予め用意された文末抑制ルールがその箇所に成立するか、1つずつ確認する。成立する場合には、各ルールに定められた得点を、その文末候補の得点から減算する。
7. 最終的に、一定値以上の得点となった文末候補を、文末と見なし、文末と文末の間を1つの文と認定する。

手順1では、通常の音声認識処理の結果を、表1の形式で出力する。句読点の挿入には、従来の統計的言語モデルによる手法を用いる。

手順4では、ポーズ長に比例した得点を与える。ただし、ポーズ長が1.5秒以上の場合には、そこが後の処理に関わらず文末と判定されるような高得点を与えている。これらの得点やポーズ長の閾値は、データのポーズ長の傾向から分析したものである。

手順5，6で各文末候補に適用するパターン・マッチング・ルールの一部を表2に示す。ただし、得点は手順7で認定する文末候補の閾値を100点としたときのものである。これらのルールでは、文末候補の前後の単語の表記、品詞、語形と、文末候補のポーズ長を調べて、各ルールの条件が成立したとき、成立した文末候補に得点を与える。

表2で挙げたルールの基本的な考え方は、まず音声認識時に挿入された句点「。」をとりあえず文末と見なし、ついで統計的言語モデルの局所的な条件では正しく推定できなかった句点を抑制する、というものである。音声認識時には読点「、」と推定された箇所や読点が挿入されなかった箇所も、ポーズ長や他の文末強化ルールによって、基準値100点を超えた箇所は最終的に文末と判断する。

表2．文末強化・抑制ルールの例

種類	ルール成立条件	得点
文末強化	文末候補の単語が句点「。」であること	100
文末強化	文末候補の前方:6単語以内に動詞「思う」の変化形	60
文末強化	文末候補の前方:4単語以内に「まさし	40
文末強化	文末候補の後方:直後の表記が「最初」「まず最初」「たとえば」	50
文末強化	文末候補の後方:直後の表記が「そうすると」「そうしますと」「そうしたら」	30
…		
文末抑制	文末候補の前方:動詞または形容詞の終止形 文末候補の位置:句点「。」 文末候補の後方:「様子」「わけ」「こと」「用」などの係り受けされやすい名詞類	-70
文末抑制	文末候補の前方:助動詞「た」の終止形 文末候補の位置:句点「。」でポーズ長0.2秒以上 文末候補の後方:「様子」「わけ」「こと」「用」などの係り受けされやすい名詞類	-70
文末抑制	文末候補の前方:動詞または形容詞の終止形 文末候補の位置:句点「。」 文末候補の後方:格助詞または接続助	-60
文末抑制	文末候補の前方:表記が「…けれども」 文末候補の後方:6単語以内に動詞が存在すること	-80
…		

ルールでは、口語的な表現での典型的な認識単語出現パターンへの対処もされる。表1の認識文字列では、音声認識時に挿入された3つの句点のうち、最後のものだけが望ましい句点であり、パターン・マッチング・ルールにより、はじめの2つが抑制される。本ルールでは、成立条件として文末候補の前後最長8単語までチェックしている。このように統計的言語モデルでは対処できない長い範囲の性質も、ルールにより正しく処理される。

各ルールの得点は、ビデオを音声認識したテキストデータを分析して、誤りの頻度や副作用の度合いから調整したものである。

3.3.評価実験

本稿で提案した音声認識テキストからの文認定手法の評価として、ビデオの音声から人手で書き起こしたテキストに人手で文区切りをつけたものと、本手法での結果との比較実験を行った。

表3が2つのTV放送コンテンツに対して、比較実験を行ったときの結果である。ともに番組の最初から最後までを1つの映像ファイルとして録画し、処理したものであり、また文認定手法を開発するために利用していないオープン・コンテンツである。

表3．人手による文認定との比較

	ニュース番組	講義番組
文数(人手)	223	198
音声認識による句点数	193	171
句点再現率	0.655	0.475
句点適合率	0.756	0.550
文認定数	181	171
文認定再現	0.686	0.571
文認定適合	0.845	0.661

表3で、句点再現率、句点適合率は音声認識時に統計的言語モデルを用いて挿入された句点をそのまま文末と見なしたときの正解率であり、文認定再現率と文認定適合率は、統計的言語モデルとパターン・マッチング・ルールをともに用いる本手法で認定された文末の正解率である。

この2つのコンテンツでは、ニュース番組の方が、口語表現を多く含んだ講義番組よりも読み上げ発音に近く、比較的音声認識精度が高いが、どちらの場合も標準的な読み上げ音声認識に比べて大きく音声認識精度が下がっている。

表3から、どちらの番組に対しても、音声認識時の句点をそのまま文末と見なしたときに比べ、再現率、適合率ともに改善しており、改善の度合いは、もとの音声認識率が低い方が効果が高いことが判る。しかし、音声認識率が低くなるにつれて、最終的な文認定精度も下がっている。この傾向は、表3であげた2つのコンテンツ以外に対しても同様であった。

4. まとめと今後の課題

ビデオ音声に対する音声認識テキストを文単位に分割する手法を提案し、2種類のＴＶ番組を用いて評価することで有効性を確認した。音声認識テキストに対し文認定処理を行うことで、重要文抽出のような言語処理をビデオに対しても適用することが可能となるが、文認定精度はビデオの音声認識性能に大きく依存することが判った。

今後は、本手法を用いてビデオを要約した場合に、画像全体として、要約に適切な単位で区切られているか検討する予定である。またルールの拡充法も課題である。現在は数十程度のパターン・マッチング・ルールを人間が作成し、その得点付けデータの傾向分析から、実験的に調整しているが、これらのルールの成立条件や得点付けをコーパスから半自動的に調整することを検討する。入力ビデオの種類毎に適用するルールを切り替える手法も考えられる。

他に、入力がマルチメディアデータである特性をいかして、映像情報を利用することも検討中である。シーンチェンジのタイミングや、カラーレイアウト、人物・建物等のオブジェクト認識、テロップ文字認識などの画像情報を組み合わせて活用する手法についても研究する予定である。

参考文献

[1] Infomedia Project (CMU) http://www.informedia.cs.cmu.edu/

[2] Videologger (Virage 社) http://www.virage.com/products/videologger.html

[3] Yihong Gong and Xin Liu: Generic Text Summarization Using Relevanec Measure and Latent Semantic Analysis, SIGIR 2001

[4] Osamu FURUSE, Setsuo YAMADA, and Kazuhide YAMAMOTO: Splitting Long or Ill-formed Input for Robust Spoken-language Translation, COLING-ACL-98 (1998)

[5] Klaus Zechner and Alex Waibel : DIASUMM: Flexible Summarization of Spontaneous Dialogues in Unrestricted Domains, COLING-2000 (2000)

[6] 中嶋 秀治、山本 博史：音声認識過程での発話分割のための統計的言語モデル，情報処理学会論文誌 Vol42 No.11 (2001)