音声と映像の一貫性を考慮したニュース動画の要約

News Movie Summarization in Consideration of the Consistency of Voice and Video Image

伊藤 一成♥ 酒井 康旭♦
斎藤 博昭♠

Kazunari ITO Yasuaki SAKAI
Hiroaki SAITO

自然言語処理を核とした新たな動画要約手法を提案する．動画内容はすべてメタデータを用いて表現できると仮定すると，音声と映像を分離して要約することが可能となる．すなわち，ユーザが指定する任意の要約率で音声テキストを要約した後に，対応する映像の重要区間を決定する．要約結果の提示の際には映像の重要区間を再生するのと同時に，日本語スピーチエンジンを利用して要約テキストを音声に変換することで，音声と映像の一貫性を考慮した要約生成が実現できる．ニュース報道番組の動画要約システムを試作し，提案手法の有効性を確認した．

This paper proposes a novel movie summarization method based on meta data analysis and text processing. Since all the contents of a movie can be described in a meta data format, it becomes possible to summarize the movie in two layers: voice and video image. Namely, the speech contents are firstly abridged at an arbitrary condense rate using natural language techniques. Then important video sections are determined corresponding to the selected speech parts. When the summarized result is presented, the video sections are reproduced along with the synthesized speech of the abridged text. This summarization method assures the consistency of sound and video. We have implemented a news summarization system and confirmed effectiveness of our approach.

1. はじめに

記憶装置の大容量化やネットワーク技術の発展などによって，音声や映像といったマルチメディアデータを容易に取り扱うことができるようになった．また，ディジタル放送やホームサーバの標準化並びに実用化の動向を見ると，番組コンテンツとして大量の映像データが一般視聴者に提供されることが近い将来に現実となることが考えられる．しかしその一方で，一般視聴者には，膨大な量の映像データの中から得たい情報を選択することを強いる結果となり，限られた時間の中で視聴者が必要とする情報だけを選択的に視聴することが困難な状況になることが予想される．これを解決するために，これらコンテンツを何らかの形で整理し，ユーザの要求を適切な形で提供できる，より高度なハンドリング手法が望まれており，そのアプローチの一つとして動画要約技術が挙げられる．これまでの動画要約の研究は，スポーツのような動画を扱う場合では映像の要約だけを行えば良かったが，ニュース報道番組の動画などは，音声と映像の両方の一貫性を考慮する必要があり，これを同時に満たすのは困難であった．そこで，本論文ではニュース報道番組の動画に焦点を当て，動画内容を予め記述したメタデータを用いて，そこに自然言語処理の手法を適用する．これにより，動画中に含まれる映像と音声を分離してそれぞれを要約し，再合成することにより両方の一貫性を考慮した新たな要約手法を提案する．

♥ 学生会員 慶應義塾大学大学院理工学研究科 後期博士課程 k_ito@nak.ics.keio.ac.jp
♦ 非会員 NTT コミュニケーションズ yasu@nak.ics.keio.ac.jp
♠ 正会員 慶應義塾大学理工学部 hxs@nak.ics.keio.ac.jp

2. 本手法の流れ

図1に本手法における処理の流れを示す．

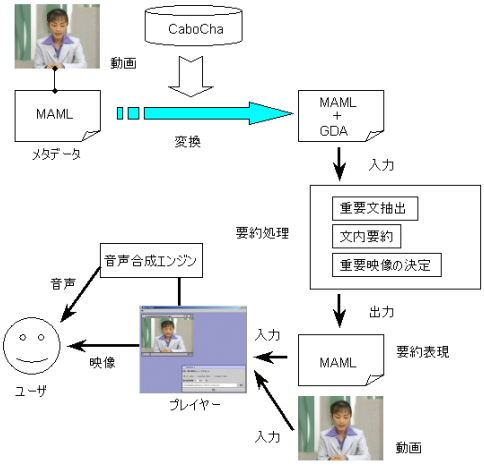

図1 処理の流れ図
Fig.1 Flowchart of Processing

今回，メタデータの記述形式として，我々が提唱しているMAML[1]と，橋田が提案するGDA[2]を利用し，ニュース動画の内容を記述する．MAMLは人間が理解及び記述しやすい表現構造を念頭においた，XML形式の汎用アノテーション記述言語である．メディアの種類やフォーマットに非依存な統一記述仕様であり，タグによるデータの構造化は最小限で，自然文章中心の構造となっている．はじめに，MAMLにより動画の転記情報を記述する．一方GDAは橋田が提案する，多言語間に共通の統語・意味などに関するXMLタグの標準を作って普及させようというプロジェクトである．GDAでは，文法機能（主語，目的語，間接目的語），主題役割（動作主，非動作主，受益者など），修辞関係（理由，結果など）や照応関係を表すことができ，既に検索，要約，翻訳，対話処理，質問応答システムをはじめとした自然言語処理の分野でGDAを活用した研究報告がなされている．今回は，発話転記テキストに日本語の係り受け解析器CaboCha[3]の出力結果から得られる形態素情報と構文情報から自動的にGDAタグを付与した．図2はMAML及びGDAの記述例である．図2の例では動画中の6.7秒から9.2秒まで“都内の家屋が倒壊”という発話があることを示している．

次に，このメタデータに対してテキスト要約処理及び映像の重要区間決定処理を行うことで音声と映像の重要部分をそれぞれ決定する．詳細については3章，4章で述べる．この際ユーザは任意の要約率を指定することが可能である．

その結果，要約された情報を新たなMAMLファイルとして出力する．要約されたMAMLファイルには，各トピック内で音声として読み上げる文章と，再生する映像区間情報が含まれる．要約されたMAMLファイルに基づいて専用のプレイヤーで元の動画から映像の重要区間だけを間引いて順再生する．この時，元の動画に含まれる音声は消音にして使用せず，代わりに日本語のスピーチエンジンを利用し，要約されたテキストから音声を合成し，映像と共に再生する．

```
<element id="1" begin="6.7" end="9.2">
  <audio>
    <utterance>
     <su>
      <adp>
       <adp>
        <np bfm="都内" prn="トナイ">都内</np>
        <ad bfm="の" prn="ノ" sem="連体化">の</ad>
       </adp>
       <ad>
        <np bfm="家屋" prn="カオク">家屋</np>
        <ad bfm="が" prn="ガ" sem="格助詞">が</ad>
       </ad>
      </adp>
      <n>
       <np bfm="倒壊" prn="トウカイ">倒壊</np>
       <x>。</x>
      </n>
     </su>
    </utterance>
  </audio>
</element>
```

図 2　MAML 及び GDA の記述例
Fig.2　An Example of MAML and GDA

3. テキスト要約処理

本手法では，メタデータ中に含まれるニュースキャスターの発話転記テキストについて要約処理を施す．テキスト要約処理の主なアプローチには大別して 1)重要文抽出による要約，2) 抽象化，言い換えによる要約，3) 文中の不要箇所削除による要約が挙げられる．

このうち本手法では，一段階目の要約処理として重要文抽出による要約を行い，二段階目の要約処理として文内の不要箇所削除による要約を行う．以下に本稿で用いるテキスト要約手法について説明する．

3.1 重要文抽出による要約

重要文抽出の手法としては吉見らの手法[4]を採用している．以下に吉見らの重要文抽出手法について説明する．

吉見らの手法は，文の重要度に関して次の 2 つの仮定に基づいている．

・表題はテキスト中で最も重要な文である．
・重要な文とのつながりが強ければ強いほど，その文は重要である．

最初に表題文 S_1 中に含まれる重要語を抽出して重みを付与し，それぞれの語の重要度を求める．ここで重要語とは形態素解析の結果から得られる品詞が，名詞，人称代名詞，動詞，形容詞，副詞のいずれかである辞書見出し語を指す．次に，表題文 S_1 の重要度を次式で求める．

$$S_1\text{の重要度} = \frac{S_1\text{の重要語の重み和}}{S_1\text{の重要語の数}}$$

この表題文の重要度を元に，後に続く文との関連度を算出してそれぞれの文の重要度を決定する．文 S_j の先行文 S_i へのつながりの強さ（関連度）を求める式を以下に示す．

$$S_i\text{と}S_j\text{の関連度} = \frac{S_j\text{中の重要語のうち}S_i\text{の題述中の重要語につながるものの重みの和}}{S_i\text{の題述中の重要語の数}}$$

ここで，文 S_j の主題は，S_j 中の重要語のうち S_j の関連文中の重要語につながるものから構成され，文 S_j の題述は，つながらない重要語から構成される．ただし，関連文を持たない標題文 S_1 では，それに含まれる重要語すべてが題述を構成する．重要文選択の手順を図 3 にまとめる．本論文で扱うメタデータ内には，文章に対する表題が存在しないため，映像中に含まれるクローズドキャプション文字列を表題文 S_1 に代用した．

ステップ1　GDA文書を入力とする.

ステップ2　表題への重み付け処理を行う.

ステップ3　表題文S_1の重要度を次式で求める.

$$S_i\text{の重要度} = \frac{S_1\text{中の重要語の重み和}}{S_i\text{の重要語の数}}$$

ステップ4　各文 S_j (j=2,3,・・・,n)について，S_jから五文前までの先行文 S_iの範囲(j-5 ≦i<j)で重要度を求める.

ステップ5　あらかじめ定められた数だけ文を重要度の順に選択し，それらをテキストでの出現順に出力する.

図 3　重要文選択手順
Fig.3　Procedure of Selecting Important Sentences

3.2 文内の不要箇所削除による要約

重要文抽出による要約処理により，テキスト全体からユーザが指定する要約率を指定して重要な文の集合を得ることができる．その結果から得られた重要文の各文をさらに要約することで要約率を高めるために，我々が既に提案した手法[5]を基本にして，各文内で重要度が低いと思われる節を削除する．以下に基本的な概念を示す．この手法では，係り受け情報に加えて，照応・代用・省略といった詳しい情報があらかじめ手動によって付与された GDA 文書を処理対象としている．基本的には GDA タグの文法機能（主語，目的語，間接目的語），主題役割（動作主，非動作主，受益者など），修辞関係（理由，結果など）の情報を利用して得られる文のテキスト構造から各節に非重要度のスコアを付け，スコアの小さい語のみを抽出することで行う．まず，文の必須語（主辞，主語，目的語）を抽出する．次に，各節に対して GDA タグの文法の種類と係り受けに応じて非重要度を付与し，ある閾値以上の語を抽出する．この際，閾値は要約率と各文の非重要度の最大値から決定する．本手法では，機械的に付与された係り受け構造のみから非重要度を決定する．各処理単位内において，最後の節の非重要度を 0 とし，次に，その他の節の非重要度を決定する．その他の節の非重要度については非重要度が 0 の節に係るまでの距離をその節の非重要度とする．

図 4 に非重要度付与の例を示す．以上の処理により決定した各節の非重要度が 2 以上の場合，その節を削除の候補とし，経験的に定めた独自ルールにより削除の可否を決定する．

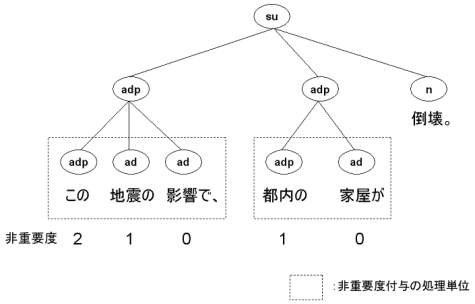

図 4　非重要度付与の例
Fig.4　An Example of Setting Non-Importance Degree

以上のテキスト要約処理により，ユーザが指定した要約率に応じて適切な長さの文章を返すことができる．この時，一秒間に発話可能な文字数を決定しておき，元となる動画の総時間と要約率より求められる要約結果の時間長制約に応じて，重要文抽出の結果に含める文字数を決定するようにする．

4. 映像重要区間の決定

ニュースでは，キャスターの発話内容は同時刻の映像に何らかの関係があると考えられるため，3.1 節で概説した重要文抽出の結果から得られる各文の重要度を用いて，時間的に重なりのある映像区間に対し，各ショットごとに映像重要度を付与する．図 5 に，映像重要度の付与方法を示す．

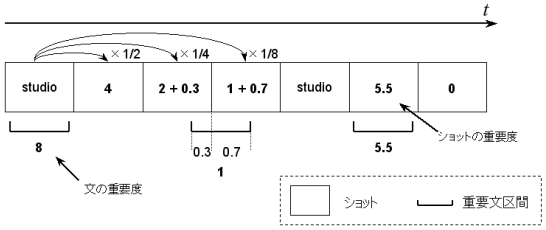

図 5　映像重要度の付与方法
Fig.5　Procedure of Setting Visual Importance Degree

ここでは時間的に重なりがあるショットに対して，重なる時間の割合に応じて文の重要度を割り当てる．この際，スタジオのショットは候補から除外する．もし重要文区間がスタジオのショットと重なった場合は，その重要度を後方のショットに分配する．この時，重用度を分配するのはスタジオ以降の 3 つのショットまでとし，図 5 のとおり，それぞれ 1/2，1/4，1/8 とする．ここで重要度を後方に分配する理由は，スタジオのショット中でなされるキャスターの発話内容を補完する内容が，それに続くショット中にあることが多いという経験則に基づいている．

以上のルールで各ショットの重要度を決定しておき，ユーザが指定する要約率に応じて 2 通りの方式で映像の要約候補を決定する．

方法 1：各ショットに割り当てられた重要度を用いずに，トピック中のスタジオを除く全ショットから，時間的に等しく映像重要区間を抽出する（図 6 参照）．

方法 2：元となる動画の総時間とユーザによって指定された要約率から算出された要約結果時間を，ショットの重要度に比例して割り当てる（図 7 参照）．

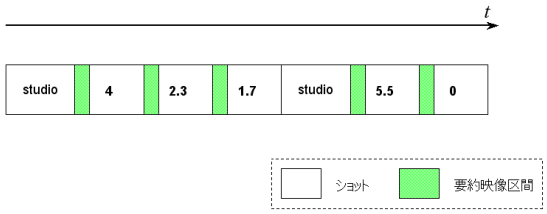

図 6　要約映像区間決定方法 1
Fig.6　Method 1 of Deciding Summary Video Sections

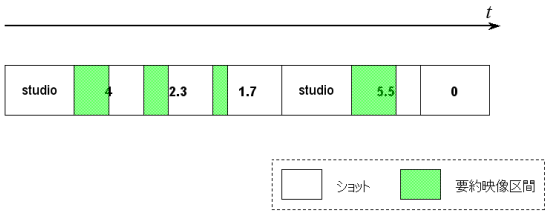

図 7　要約映像区間決定方法 2
Fig.7　Method 2 of Deciding Summary Video Sections

これら二つの方法を試みる理由として，要約結果時間の長さによって，ユーザに結果を提示した際の，内容の理解度に差が出ると考えられるためである．例えば，要約結果時間が非常に短い場合に，方法 1 の様に全てのショットを等間隔で抽出すると，一つのショットの再生時間が非常に短くなり，映像の内容を理解するのに不十分である考えられるからである．また逆に，要約結果時間が長めの場合には，方法 2 によりショットの選択をすると，情報が激しく欠落する恐れがある．よって評価実験を行う際には，これら二つの方法で要約映像を生成し，比較することにする．また，ニュースに含まれる映像は，スポーツなどの映像と違いカメラワークが少ないため，ショットのどの部分を抽出しても重要度に変化がないことが考えられる．よって本手法では，割り当てられた時間に応じて各ショットの先頭から抽出し，それらを先頭から繋ぎ合わせたものを要約結果の映像とする．

5. 評価実験

テストデータとしては国立情報学研究所が配布している評価用映像メディア DB[6]を用いた．これに含まれる約 15 分間のニュース動画 2 本の中から 8 つのトピックを選択し，これらに対して本手法を適用した．ここでは実験用動画に対して人間が予め作成した要約動画と，本手法で生成した要約動画を再生し，被験者に視聴・比較してもらうという主観評価形式を採用する．まず，人間の要約者が作成した要約動画は以下の 2 種類である．

動画 1.映像情報を重要視し区間を間引いた要約
動画 2.音声情報を重要視し区間を間引いた要約

ここで 2 種類の人手による要約を用意する理由としては，まず，人間が実際に動画を要約する際，重要と思われる区間を決定するにあたっては，要約者によって様々な決定方法が考えられるからである．これにより，既存の手法で正解とされてきた要約結果との比較を行うことができる．本手法により

生成する要約は，前章で述べたとおり，2 通りの映像区間決定方法を試す．これを以下に示す．

動画 3.スタジオ以外の全ショットから等しく重要区間を抽出する要約

動画 4.ショットの重要度に応じて重要区間を抽出する要約

また，要約結果の再生時に本手法で実装した専用プレイヤーを使用することで，音声情報と映像情報を完全に分離した要約と再生が可能となる．つまりこれは，全く新しい結果提示方法の提案であり，こうした結果提示の良否についても検討する必要がある．よって，更にここで人間の要約者に映像情報と音声情報のそれぞれを独立に要約し，再合成してもらい，これとその他の要約を比較する．

動画 5.映像情報と音声情報のそれぞれを要約

以上の 1〜5 の要約方法で作成した 10 秒と 30 秒の動画をそれぞれ用意し，これらを 14 人の被験者に視聴してもらい，動画要約としての良否ついてそれぞれのアンケートにより 5 段階（1：適切とは言えない〜5：適切である）で良否を評価してもらった．評価結果を表 1 に示す．

	動画 1	動画 2	動画 3	動画 4	動画 5
10 秒要約の平均結果	1.75	2.43	2.41	2.30	3.81
30 秒要約の平均結果	2.72	3.60	3.06	3.02	4.07

表 1　各要約動画の評価の平均値
Table 1　An Average of Evaluation of Each Summrized Movie

まずこの結果から，従来手法における正解とされてきた動画（動画 1，動画 2）と，本手法により生成した動画（動画 3，動画 4）を比較すると，10 秒に要約した場合では動画 3，動画 4 の結果は動画 1 の結果を大きく上回っており，動画 2 と比べてもほぼ同等の結果であることがわかる．また，30 秒に要約した場合では，動画 2 との比較ではやや劣るものの，動画 1 よりは良い結果を得ることができた．さらに，概して動画 1 よりも動画 2 の方が良い結果であることから，少なくともニュース番組の動画については，人間は映像情報よりも音声情報を重視するということが言える．つまり，本手法のように音声情報を重要視し，その処理結果に基づいて動画作成を行なうアプローチは正しいと考えられる．また，本手法の結果を下げている要因としては，言語解析時の形態素解析誤りや係り受け解析誤りなどから生じる重要な節の削除によって不可解な要約テキストが生成されることや，スピーチエンジンのイントネーションの不自然さなどが考えられるが，これらは今後各研究分野において精度が向上していくことで解決され，本手法を用いた場合でもさらに良い結果が得られると思われる．

次に，本手法で実装した専用プレイヤーを用いて要約結果を再生した場合，従来のように単に動画区間を再生するのではなく，動画中の映像情報と共に，スピーチエンジンを用いることで任意の音声を再生することが可能となる．動画 5 は，映像情報と音声情報を独立に再生することができるという本手法のメリットを用いた場合の最も良い要約結果として，人間の要約者が作成したものである．すなわちこれは，本手法を採用した場合の最良解と考えることができ，本手法を用いて自動的に生成した動画 3，動画 4 を比較すると，結果から，全ての場合において動画 5 は動画 3，動画 4 の結果以上となっているのがわかる．また，動画 3，動画 4 はトピックによって結果の違いが激しいが，動画 5 はどのトピックでも安定した好結果を残しており，10 秒に要約した場合と 30 秒に要約した場合の結果の差が小さい．

最後に，従来手法における正解である動画 1，動画 2 と，本手法を採用した場合の最良解である動画 5 を比較すると，動画 5 の結果の方が良好であることがわかる．これはすなわち，結果の提示方法にという点については，従来のように単に動画区間を再生する方法よりも，本手法の方が優れているということであり，視聴者が内容をこれまでよりも深く理解できる可能性を広げたことを示している．

6. まとめと今後の予定

本論文ではニュース動画を対象にして，動画内容を予め記述したメタデータを用い，自然言語処理の手法を適用した新しい要約手法を提案した．これにより，動画中に含まれる映像と音声を分離してそれぞれを要約し，再合成することが可能となり，両方の一貫性を考慮した自然な要約動画を生成することが可能となった．今回提案した手法では，予め与えられたアルゴリズムでのみ重要部分を選択したが，今後の方向性としては，ユーザの嗜好や要求に応じたパーソナライズ要約や，スポーツやドラマといったニュース以外の動画も扱えるようにしていくことで，より汎用的かつ実用的なシステムになっていくと考えられる．

[文献]

[1] 伊藤一成，斎藤博昭: 汎用アノテーション記述言語 MAML，情報処理学会研究報告 FI74−9，pp.63-69，(2004)．

[2] 橋田浩一: GDA 意味的修飾に基づく多用途の知的コンテンツ，人工知能学会論文誌，Vol.13，No.4，pp.528-535，(1999)．

[3] 日本語係り受け解析器 Cabocha ホームページ http://cl.aist-nara.ac.jp/~katu-ku/software/cabocha/

[4] 吉見毅彦，奥西稔幸，山路孝浩，福持陽士: 表題へのつながりに基づく文の重要度評価，自然言語処理，Vol.6，No.1，pp.43-57，(1999)．

[5] 野村雄司，伊藤一成，斎藤博昭: GDA タグを用いたテキスト自動要約，言語処理学会第 9 回年次大会，pp.206-209 (2003)．

[6] 馬場口登，栄藤稔，佐藤真一，安達淳，阿久津明人，有木康雄，越後富夫，柴田正啓，全炳東，中村裕一，美濃導彦，松山隆司: 映像処理評価用映像データベースについて，電子情報通信学会技術研究報告，PRMU2002-30，pp. 69-74，(2002)．

伊藤 一成　Kazunari ITO

現在，慶應義塾大学大学院理工学研究科後期博士課程在学中．自然言語処理及びマルチメディア情報処理に関する研究に従事．日本データベース学会，情報処理学会，電子情報通信学会，電気学会，人工知能学会，各学生会員．

酒井 康旭　Yasuaki SAKAI

慶應義塾大学大学院理工学研究科前期博士課程修了．現在 NTT コミュニケーションズ（株）勤務．在学中はマルチメディア情報処理に関する研究に従事．

斎藤 博昭　Hiroaki SAITO

慶應義塾大学工学部数理工学科卒業．現在同大理工学部情報工学科専任講師．工学博士．自然言語処理，音声言語理解に興味を持つ．日本データベース学会，情報処理学会，言語処理学会，日本音響学会，電子情報通信学会，ACL 各会員．