知能と複雑系 120－6
（2000．6．2）

アノテーションに基づく知的マルチメディア処理

長尾 確
日本アイ・ビー・エム (株)
東京基礎研究所

白井 良成[1]
慶應義塾大学
政策・メディア研究科

橋田 浩一
電子技術総合研究所

概要

ビデオなどのマルチメディアデータは、テキストデータに比べて、内容に基づく処理が極めて困難である。そこで、われわれは、マルチメディアデータにアノテーションと呼ばれるメタ情報を付与して、それを利用した検索や要約について研究を行なっている。このアノテーションは、主に、文書データの言語構造をXML形式のタグを埋め込むことによって明示化するもので、マルチメディアデータの場合は、その内容を記述する文書に対して意味構造を与えることになる。ここでのマルチメディアデータの内容記述には、シナリオにおけるト書きのような情景描写と登場人物の台詞に相当するトランスクリプト、および、フレーム内に登場するオブジェクトの記述からなる。われわれは、アノテーションをコンテンツの意味内容を機械的に理解するための手段と位置付け、アノテーションに基づくさまざまなコンテンツ加工を行なっている。それをセマンティック・トランスコーディングと呼ぶ。セマンティック・トランスコーディングは、今のところ、マルチメディアデータを含むオンラインコンテンツの要約、翻訳、音声化などを扱っている。マルチメディアデータの要約は、内容記述に文書要約の手法を適用して要約文書を生成し、要約文書に関連する音声や動画像を組み合わせることによって行なう。本稿では、マルチメディアデータに対するアノテーションの枠組みについて議論し、それを実現するツールや、アノテーションを利用したビデオ要約システムを紹介する。また、ビデオ要約以外のマルチメディア処理に関しても簡単に触れる。

Intelligent Multimedia Processing Based on External Annotations

Katashi Nagao
IBM Research, Tokyo Research Lab.
1623-14 Shimotsuruma, Yamato,
Kanagawa 242-8502, Japan
nagao@trl.ibm.co.jp

Yoshinari Shirai
Keio University
5322 Endou, Fujisawa,
Kanagawa 252-8520, Japan
way@sfc.keio.ac.jp

Kôiti Hasida
Electrotechnical Laboratory
1-1-4 Umezono, Tukuba,
Ibaraki 305, Japan
hasida@etl.go.jp

Abstract

This paper proposes a method for processing online multimedia data. The method employs metadata or annotations on text and video. We have developed techniques for semi-automatic video annotation using a text describing the content of the video. Our video annotation includes segmentation of video, generation of a video transcript using speech recognition, linking of video segments with corresponding text segments, and interactive naming of people and objects in video frames. In this paper, we classify annotations into three categories. One is linguistic annotation which helps the transcoder understand the semantic structure of textual elements. The second is commentary annotation which helps the transcoder manipulate non-textual elements such as images and sounds. The third is multimedia annotation, which is a combination of the above two types. All types of annotation are described using XML (Extensible Markup Language). We call the entire process "semantic transcoding" because we deal with the deep semantic content of data with annotations. The current semantic transcoding process mainly handles text and video summarization, language translation, and speech synthesis. This paper concentrates on text and video summarization. We also briefly describe other multimedia processing implemented in our semantic transcoding system.

[1] 現在、NTTコミュニケーション科学基礎研究所

1 はじめに

インターネットが発展し、動画像や音声をデジタル化するツールが一般に利用可能になるにつれて、マルチメディアデータは、最も重要なオンライン情報ソースになりつつある。しかし、ビデオなどのマルチメディアデータは、テキストデータと異なり、内容に基づく処理が困難である。そこで、われわれは、マルチメディアデータを柔軟に活用するための手段を提供する。それは、アノテーションと呼ばれる手法に基づいている。

アノテーションは、コンテンツの表現力を向上すると同時に、その利用法において重要な役割を果たす。その一つが、コンテンツ適応 (コンテンツをユーザーの都合に合わせてカスタマイズすること) である。われわれは、アノテーションに基づくコンテンツ適応の仕組みを実現した。それをセマンティック・トランスコーディングと呼んでいる [5]。

セマンティック・トランスコーディングは、基本的にテキストコンテンツの処理を中心としたものであるが、その手法はビデオやイメージなどの非テキストコンテンツの加工にも応用され、マルチメディアデータを含む一般的なドキュメントに適用できる。

コンテンツ適応以外のアノテーションの利用法として知識発見がある。これは、アノテーションを含む大量のコンテンツから、機械的に何らかの発見をさせようとするものである。従来の検索エンジンのように、キーワードから複数の Web ページを検索するのではなく、ユーザーのある要求を満たすような情報を複数のコンテンツを合成して作り出すのである。たとえば、IBM の製品に関する一年分の Web 情報から、IBM のその一年の製品戦略に関するサマリーを生成する、ということが実現できる。

現在のところ、アノテーションに基づいて、内容的に類似するコンテンツを収集し、それぞれのサマリーを含む一つのドキュメントを生成する程度のことが可能になっている。知識発見に関しては、まだまだ多くの研究が必要であるが、アノテーションによって大きく促進されることは間違いがない。

2 アノテーション

従来の Web は一枚の平面上に存在するグラフとして捉えることができる。われわれは、Web を平面から立体に拡張する手法を提案する。それは、外部アノテーション (各々のコンテンツの外部に存在するアノテーション) によって実現できる。

図 1 はわれわれのイメージする Web の上位構造を表している。

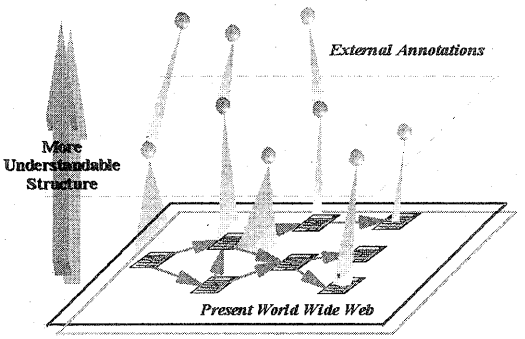

図 1: WWW 上に構築される上位構造

Web の上位構造とは、コンテンツに対するメタコンテンツ、さらにそのメタコンテンツに対するメタコンテンツという具合に、Web をより立体的に捉える構造である。ここでは、そのようなメタコンテンツを外部アノテーション (以下では、単にアノテーションと表記する) として一般化する。アノテーションは、一般に、XML (Extensible Markup Language)[9] 形式のデータとして表現される。

2.1 言語的アノテーション

言語的アノテーションは、テキストコンテンツの意味構造に関するアノテーションである。それは、文節間の係り受け、代名詞の指示対象、多義語の意味など、かなり細かい情報を含む。このタイプのアノテーションは、ドキュメントの内容理解に大きく貢献し、テキストの加工以外にも、たとえば、内容検索や知識発見などに利用される。

言語的アノテーションは、GDA (Global Document Annotation)[3] の規定するタグセットに基づいている。GDA は多言語間に共通な意味的・語用論的タグをドキュメントに付与することにより、その機械的な内容理解を可能にし、ドキュメントの検索・要約・翻訳を実用的なレベルで実現するとともに、ドキュメントの作成・公開 (共有化)・再利用を考慮した統合的なプラットフォームを構築して、世界的に普及させようという、壮大なプロジェクトである。われわれのセマンティック・トランスコーディング・プロジェクトは GDA を現在の

Webのアーキテクチャ上で利用可能にし、さまざまなサービスと連動させることによって、GDAの思想をより具体的な形で浸透させようとする試みの一つと位置付けられる。

GDAタグ付きドキュメントは、たとえば以下のようなものである。

```
<su><adp rel="loc"><adp rel="pos">人間の
</adp><np sense="0f2e4c">細胞</np>には、
</adp><np syn="p"><np><vp><adp><adp><np
sense="0f74e9">自動車</np>でいえば</adp>
<adp rel="iob">アクセルに</adp>当たり、
</adp><adp rel="obj"><np id="a1" sense=
"3be2c7">がん</np>を</adp><adp rel="gra">
どんどん</adp>増殖する</vp><n>「<namep
id="a2"><np eq="a1" sense="3be2c7">がん
</np><n id="a3" sense="3bf4d0">遺伝子</n>
</namep>」</n></np>と、<np><adp><np rel=
"pos" sense="107ab3">ブレーキ</np>役の
</adp><n>「<namep id="a4"><np eq="a1"
rel="obj" sense="3be2c7">がん</np><n
sense="10d244 3cf57c">抑制</n><n eq="a3"
sense="3bf4d0">遺伝子</n></namep>」</n>
</np></np>がある。</su>
<su><adp rel="cnd"><adp rel="sbj"><np>
<adp rel="pos"><np eq="a2 a4" sense=
"0face2">双方</np>の</adp>バランス</np>
が</adp>取れていれば</adp>問題はない。</su>
```

これらは統語構造を表わしており、各エレメント(タグで囲まれた部分)は統語的構成素である。ここで、<su>は一文の範囲を表し、<n>, <np>, <vp>, <adp>, <namep>は、それぞれ名詞、名詞句、動詞句、形容詞句/形容動詞句(前置詞句、後置詞句を含む)、固有名詞句を表す。syn="p"は等位構造(たとえば上の「〜がん遺伝子と〜がん抑制遺伝子」)を表わす。等位構造の定義は、係り受け関係を共有するということである。特に何も指定がない場合は、たとえば、<np><adp rel="x">A</adp><n>B</n> </np>はAがBに依存関係があることを表す。また、rel="x"は<adp>エレメントの関係属性を表している。また、sense="*"は語義属性を表している(属性値としては、たとえばEDR単語辞書[2]の概念識別子が利用できる。また、一語が複数の語義を持つ場合は、属性値が複数になる)。

一般にGDAドキュメントの構造は図2のようになる。

つまり、GDAドキュメントはネットワーク構造を成しており、そのリンクには、タグの入れ子構造よって

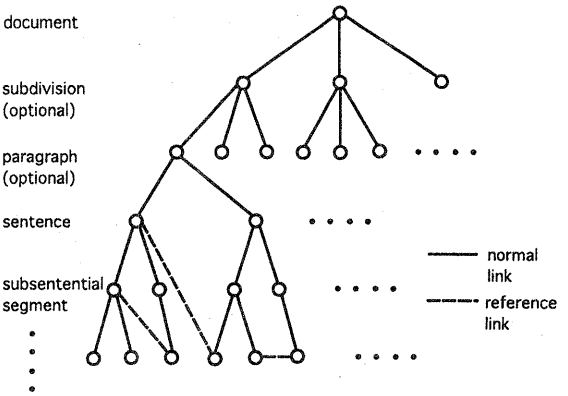

図2: GDAドキュメントの構造

定義される関係と参照関係の2種類がある。

また、GDAのタグ集合は10項目以上からなるが、さしあたり、そのうちで自動タグ付け作業が比較的大変だと思われる、統語構造、文法・意味関係、語義、照応、修辞関係という5項目だけを扱っている。GDAタグセットの詳細については、http://www.etl.go.jp/etl/nl/gda/を参照のこと。

文法機能(主語、目的語、間接目的語)、主題役割(動作主、被動作者、受益者など)、および修辞関係(理由、結果など)は関係属性によって表示する。関係属性はrel="*"という形で表される。主語、目的語、および間接目的語の主題役割の判断は難しいことが多いので、文法機能(sbj、obj、iob)を用いる。

属性の内、id="*"はID属性を示す。ID属性はそのエレメントのユニークな識別子である。また、先行詞のID属性の値を照応詞のeq属性の値にすることにより、照応(anaphora)や共参照(coreference)を表示する。たとえば、上の例で、「双方のバランスが取れていれば問題はない。」の「双方」は「がん遺伝子」と「がん抑制遺伝子」を指し示すことが、属性eq="a2 a4"によって表示されている。

このようなタグ付けは多くの労力を要すると思われるが、アノテーションエディターと呼ばれるツールにいくつかの自然言語処理モジュール(構文・意味解析、照応解析など)を統合することによって、人間の負担を極力減らせるように工夫している。

このエディターを用いて、ユーザーは言語構造(構文や意味に関する構造)をテキストに関連付けたり、ドキュメント内の任意のエレメントにコメントを付けたりすることができる。言語構造は、まず自動的に生成されるが、その構造に曖昧さが含まれる場合は、それ

をインタラクティブに解消することができる。言語構造を修正するために、自動的に解析された構造をわかりやすく表示するための工夫を行なっている。

言語的アノテーションは図3の右下に表示されている画面上の操作によって修正できる。

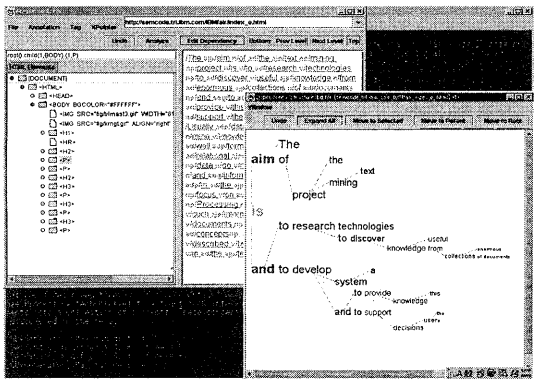

図 3: 言語的アノテーションの修正画面

このエディターを使って人間がインタラクティブに解析した部分は、次の機会に再利用されるので、それによって解析の精度が少しづつ上がっていくことになる。解析の精度が上がれば、それだけ人間の負担が減ると思われるので、将来的にはタグ付けのコストは十分に小さくなるだろう。

2.2 マルチメディアデータへの応用

われわれのアノテーション手法はビデオなどのマルチメディアデータにも適用できる。ビデオは今後インターネットの主要な情報リソースになっていくと思われる。それは、テレビが新聞よりも多くのアテンションを集められるように、動画像の持つ魅力はテキストやイメージの持つそれよりも一般に大きいからである。さらに、最近はテレビをハードディスクに録画したり、ビデオカメラがテープではなくディスクに映像を記録できるようになってきたため、デジタル化された映像を容易に作成・入手できるようになってきたためである。このようにオンライン情報におけるビデオの割合が増えるにしたがって、それを検索したり、要約したりする技術の必要性が高まってくるのは明らかである。

われわれのビデオアノテーションは、自動的にシーンの検出を行ない、それらのシーンとやはり自動的に生成されたテキストの関連付けを行なって、さらに人や物などのフレーム内のオブジェクトと言語表現を関連付けていく、という形で行なわれる。それぞれのプロセスでは、ユーザー (アノテーター) が自由に介入して、インタラクティブに変更・修正が行なわれる。

われわれはビデオのトランスクリプトを自動的に生成して、半自動的にビデオアノテーションを作成するシステムを開発した。図4はビデオアノテーションエディターの画面例である。

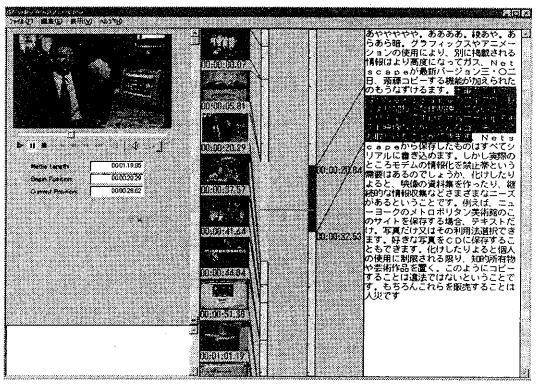

図 4: ビデオアノテーションエディターの画面例

このシステムは、ビデオの各フレームのカラーヒストグラムの差分検出に基づいてシーンの変わり目を認識し、シーンに関する記述の作成を支援する。シーン記述は開始時間、終了時間、シーンタイトルから成る。さらに、画像内のオブジェクトを指定すると、その前後のフレーム列をスキャンして、そのオブジェクトのトラッキングを行ない、オブジェクトがビデオに現れる開始時間と終了時間、およびフレーム内の移動軌跡を調べる。これらは、ビデオオブジェクト記述として、やはり XML エレメントとして表現される。音楽などの、言語表現に置き換えられないオーディオデータに関しても、同様に開始時間と終了時間を調べ、オーディオオブジェクト記述として表現できる。シーン記述、ビデオオブジェクト記述、オーディオオブジェクト記述をデータエレメントと呼ぶ。

一方、システムはビデオの音声部を音声認識し、トランスクリプト (書き起こし文書) の作成を支援する。これに、シナリオにおけるト書きのような情景描写を追加して、ビデオの内容に関する文書を作成する。その後、先ほどのアノテーションエディターを用いて文書解析を行ない、ビデオに対する言語的アノテーションを作成する。

データエレメントと言語的アノテーションに現れる言語エレメントはデータ (言語エレメントの場合には発

話) そのものを表わすと同時にそのデータが表わす事物 (意味内容) も表わす。

データエレメントと言語エレメントの間には2項関係を表わすリンクが定義できる。これらの2項関係には、第1項と第2項の各々について、データに言及するか、それが表わす事物に言及するかに応じた4種類 (データ-データ、データ-意味内容、意味内容-データ、意味内容-意味内容) がある。

これにより、以下のような異種エレメント間の関係が自然に記述できる。

1. シーン記述と言語エレメントがデータとして同じものを指している場合は、両者の間にデータ-データリンクを定義する。(例. アナウンサーが話しているシーンと、「アナウンサーが「...」と言った。」という記述)
2. オブジェクト記述と言語エレメントが同じ対象 (内容) を指している場合は、両者の間に意味内容-意味内容リンクを定義する。(例. 画面上の人物と、「この人物は...」という記述の「この人物」)
3. オブジェクト記述によって表現されるデータを言語エレメントが参照している場合は、データ-意味内容リンクを定義する。(例. ある曲のオブジェクト記述と「この曲は...」という記述の「この曲」)

このうち、シーン記述と言語エレメントのデータ-データリンクは、音声認識部が、認識した単語の開始・終了時間を出力するため、ほぼ自動的に生成することができる。

もちろん、ビデオに関してはこれ以外にもさまざまな試みがなされている。その一つが現在規格の策定が進められている MPEG-7 である [4]。MPEG-7 は ISO/IEC に属する Moving Picture Experts Group (MPEG) によって標準化活動が行なわれている新しい規格で、マルチメディアコンテンツ記述という新しい仕様を含んでいる。このコンテンツ記述はわれわれのアノテーションと同様に、ビデオデータに直接含まれないデータ (いわゆるメタデータ) によって検索や要約を容易にする仕組みを提供する。さらに、ビデオを再生するデバイスのスペックに応じて、画像の解像度を変えたり、色情報を減らしたり、音声の帯域を制限したりすることも考慮されている。さらに、オブジェクトレベルの記述というのがあり、シーンに登場する人物や物や場所などの情報も付け加えることが可能になるそうである。

MPEG-7 の仕様が確定するのを待ってから作業を始めるのでは遅いので、われわれはまずさまざまな試みを行なって、タイミングを見て MPEG-7 の規格と統合するつもりである。

ビデオへのアノテーションは、いわゆるビデオの編集に比べて複雑な情報処理を含むため、人間の行なう部分も多少複雑になるが、自動処理の精度も徐々に上がっていくと思われるので、将来は編集ではなくアノテーションによってビデオを再利用する形式が一般的になると思われる。

3 マルチメディア要約

セマンティック・トランスコーディングはコンテンツに付加されたアノテーションを用いたトランスコーディングであり、ユーザー情報を用いたコンテンツ適応の機能を有する。これは、コンテンツサーバーからユーザーのブラウザーへ至る通信経路上で行なわれ、トランスコーディングを実現するモジュールはプロキシー上に実装されている。トランスコーディングには、テキスト文書の要約・翻訳・音声化、マルチメディアデータの要約・文書化などが含まれている。

ここでは、テキストとビデオの要約について詳しく述べる。マルチメディア要約は、マルチメディアデータに対して作成された言語的アノテーションにテキスト要約の手法を適用して得られた結果から、マルチメディアデータそのものの要約を生成する、というやり方によって実現されている。

3.1 テキスト要約

テキスト要約に関しては、筆者らが以前に発表した GDA に基づく要約 [6] の手法を用いている。

一般に、要約には深い意味処理と多くの背景知識が必要である。しかし、これまでの研究の多くは表層的な手がかりやドキュメントの持つ何らかのスタイルに関するヒューリスティックスを用いるものであった。

たとえば、文の重要性を判断するのに使われる特徴素としては、文の長さ、キーワードの出現回数、時制、文のタイプ (たとえば、事実 (〜である)、推測 (〜だろう)、主張 (〜べきだ) など)、修辞関係 (たとえば、理由、例示など)、文頭からの位置、文末からの位置などがある。これらの大部分は、特に深い処理を行なわなくても、ある程度抽出できるものであり、それゆえ、これに基づく処理は非常にロバストである。

確かに、これらは現在の技術をもって実用的なシステムを作ることに成功している。しかしながら、どれ

ほどのヒューリスティックスをもってしても、要約の品質の向上はすぐに限界にきてしまうだろう。いずれにしても、内容に基づく処理は必須であろう。

ここでは、GDA タグから得られる、文の構成要素の重要度 (活性値) を用いた要約を提案する。この手法は、ドキュメントのドメインやスタイルに関するヒューリスティックスを用いていないため、GDA タグの付いたものならどのようなドキュメントにも適用可能であり、また、文より細かい単位で重要度を計算しているので、一文をさらに短くすることも可能である。

要約のアルゴリズムは以下のようになっている。

1. 照応 (共参照) 表現とその先行詞の間で活性値が等しくなり、それ以外では活性値が減衰するように活性拡散を実行する。
2. 活性拡散が終了した時点で、平均活性値の大きい順に文を選択する。
3. 選択された文の必須要素を抽出する。必須要素になりうるのは、以下のエレメントとする。
   - エレメントの主辞 (head)
   - sbj、obj、iob、pos (所有)、cnt (内容)、cau (原因)、cnd (条件)、sbm (主題) の関係属性を持つエレメント
   - 等位構造 (syn="p"を持つエレメント) が必須要素の場合は、それに直接含まれるエレメント
4. 文の必須要素をつなげて文の骨格を生成し、要約に加える。照応表現の先行詞が要約に含まれない場合は照応表現を先行詞で置き換える。
5. 要約が指定された分量に達したときは終了する。まだ余裕がある場合は、次に活性値の高い文と省略したエレメントの活性値を比較して、高い方を要約に加える。

要約はそれを行なう人間の知識によっても変わってくるが、それを読む人間の興味などによっても変わってくるべきであろう。システムの情報処理を特定の個人に適応させることをパーソナライゼーション (個人化) という。GDA タグを用いた要約は、内容に基づく一般的な手法によるものなので、個人の興味や嗜好のようなパーソナライゼーションのための情報を取り入れれば、その個人に特化した要約を行なうことができる。つまり、読み手が知りたい内容を含むように要約を生成することができるのである。

そのための手法として、以下のことを実現している。

- 要約を開始する時点で、任意のキーワードを入力できる。

  要約システムは、キーワードと関連するドキュメント中の単語を重要語として処理する。重要語を含むエレメントは初期活性値をかなり大きくした上で活性拡散を行なう。

- 要約システムがユーザーの興味を学習できる。

  これは情報検索やフィルタリングなどを含む一般のシステムのパーソナライゼーションにも関連することである。これを実現するために、ユーザーの行動履歴を保持してユーザーモデリングを行ない、ユーザーが好んで読むドキュメントの傾向を把握するメカニズムを開発している。これによって、要約システムは特にユーザーからの入力がなくても、そのユーザーに特化した要約を生成することができる。

3.2 ビデオ要約

ビデオの要約はテキストの要約と同様に盛んに研究されている。古くは CMU の Infomedia で、ビデオに含まれるさまざまな属性を自動抽出して、より重要な部分を選択している [8]。たとえば、ビデオに現れる文字情報、人の顔、シーンの変わり目、クローズドキャプションと呼ばれる字幕情報などを利用して、あらかじめリストアップされた重要な固有名詞の出現頻度や、TF*IDF 法と呼ばれる情報検索の手法を用いてキーワードの重要度を計算し、そのキーワードの現れるシーンをつなぎ合わせて要約とする。

また、IBM アルマデン研究所の CueVideo はビデオのキーフレームを並べて表示して、人間がどれかを選択すると、その部分のビデオを再生することによって、人間がビデオ全体を見る手間を減らしている [1]。また、音声のみを再生して、画像は静止画をシーンが変わるごとに変化させることによって、ダウンロードする情報の容量を少なくする工夫もなされている。このとき音声の再生スピードを変化させることによって、早口にしたり、ゆっくり聞き取りやすくすることもできる。CueVideo は遠隔教育におけるビデオの利用に焦点を当てて研究が進められており、教育に使われるビデオを効果的に見せるためのさまざまな手段が開発されている。たとえば、ある講義のビデオとその講義の資料 (PowerPoint などのスライドファイル) を自動的にリンクして、再生時に連動させることもできる。また、ビ

デオのシーンを検索するのに、任意の単語やフレーズを入力すると、音声認識を利用してその言葉を含む部分を抽出してリストアップし、そのうちのどれかを選択するとその部分を再生する、という通常のテキスト検索と同様のことがビデオに対して行なえる。

同じくIBMワトソン研究所の開発したVideoZoomは、ビデオの画像の解像度を動的に変化させ、荒い画像のビデオから徐々に鮮明にしていったり、荒い画像のビデオをまずダウンロードして、細かく見たいところのみについて差分の情報を追加していくことができる[7]。これも、ネットワークやデバイスの制約に依存して、ビデオコンテンツを加工するトランスコーディングの一種と言える。

これらのビデオ処理はアノテーションを用いないので、一度実装すれば利用するのは簡単であるが、ビデオをさまざまな形で再利用するのには問題がある。われわれはビデオが今後重要な情報ソースになることを確信しているので、検索や要約に限定されない、さまざまな再利用を可能にする枠組みをできるだけ早めに用意しておきたいと考えている。

われわれのビデオ要約は、まずアノテーションに含まれるトランスクリプトなどのテキストを要約して、その要約に対応するビデオシーンを抽出することによって行なわれる。これは、トランスクリプトに対する言語的アノテーションが対応するシーンへのリンクを含んでいるため、トランスクリプトの要約に含まれるシーンを選択して、時間順に並べることでビデオの要約が生成できる。要約は、トランスコーディングプロキシー上で行なわれる。その提示法には、要約部分のみを抽出してストリーミングで配信するやり方と、SMIL (Synchronized Multimedia Integration Language) という形式で表現された、ビデオ内のすべてのシーンのタイムコードおよび要約に含まれるシーンを記述した内容リストをビデオデータと共に配信するやり方の2通りが存在する。

図5は要約機能付きのビデオプレイヤーの画面例である。これは、SMIL形式の内容リストを参照し、MPEG形式のビデオデータを再生するソフトウェアである。ビデオ画面の下にあるスライダーバーの濃い色の部分が要約に相当する。また、右のウィンドウには、シーンのタグ構造と、要約に含まれるシーン (チェックされているもの) が表示されている。

テキスト要約におけるパーソナライゼーションは、そのままビデオ要約にも適用される。つまり、キーワードなどを入力すると、トランスクリプトにおいて該当する部分の活性値が高くなり、関連するシーンが要約に含まれる。

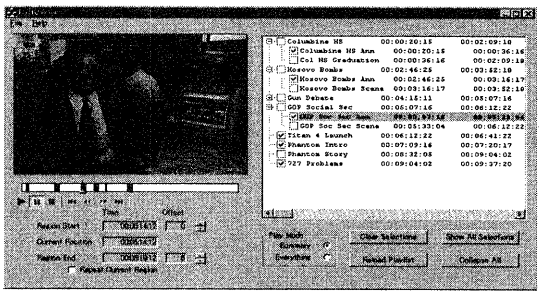

図5: 要約機能付きビデオプレイヤーの画面例

現在では、ビデオのシーンをオリジナルの時間順に並べているが、活性値に応じて順番を並び替えることも可能である。これによって、より重要なシーンを先に見ることができるだろう。ただし、シーン間に参照などの依存関係があるときは、それを考慮して順番を決定しなければならない。このようなシーン間の依存関係もアノテーションによって定義される。

4 その他のマルチメディア処理

ビデオ要約以外のマルチメディア処理には、たとえば、ビデオの再生機能を持たないクライアントデバイスがビデオをアクセスした場合に行なう、ビデオデータからHTML文書への変換や、複数ビデオ間で関連のあるシーンを抽出してまとめる自動編集、さらに、トランスクリプトを翻訳して字幕化し、ビデオと連動して表示するビデオ翻訳がある。

4.1 ビデオデータの文書化

ビデオからテキストとイメージを含む文書への変換は、もう一つの種類のビデオトランスコーディングである。もし、クライアントのデバイスがビデオを再生することができない場合、ユーザーはビデオのコンテンツにまったくアクセスできなくなってしまう。その場合、ビデオトランスコーダーはそれぞれのシーンを代表するイメージとそれぞれのシーンの内容を表すテキストを含めたドキュメントを作成してユーザーに提示することができる。また、生成されたドキュメントをテキストトランスコーダーを用いて要約あるいは翻訳することもできる。

4.2 シーンの関連性に基づくビデオの自動編集

複数のビデオデータから関連するシーンを抽出してつなぎ合わせ、一つのビデオクリップを生成することもできる。これは、言語的アノテーションを用いた内容の類似性検出に基づいている。言語的アノテーションによってシーンごとの重要なキーワードを選択し、固有名詞の場合はその語と同義のさまざまな表現、それ以外の場合は、語義を集めて各シーンのインデックスとする。内容の類似度は、インデックス間の重複の度合いによって定義される。類似度がある閾値を越えるようなシーンの集合を、もとのデータが作成された時間順に並び替え、連続したビデオデータとして生成する。もし、シーン間の参照関係がアノテーションによって記述されている場合は、参照元のシーンの直後に参照されているシーンを挿入する。

4.3 テキストの翻訳と字幕化によるビデオ翻訳

ビデオの翻訳は、トランスクリプトを翻訳し、もとの文が発話されるタイミングで、テロップとして表示することで行なわれる。あるいは、翻訳結果を音声合成し、ビデオの再生と音声再生を同期させることによって、他の言語のビデオを作成することもできる。この部分は、まだ実現されていないが、近い将来にわれわれのビデオプレイヤーに統合される予定である。

5 おわりに

われわれの次なるターゲットは大量なコンテンツからの知識発見である。アノテーションはそれぞれのコンテンツから重要な部分を抽出するのに大いに役に立つ。

マルチメディアデータを含むWebコンテンツの効率的な検索もわれわれのターゲットの一つである。この場合の検索の質問には単なるキーワードではなく、音声あるいはテキストの自然言語文を用いることができるだろう。

近い将来に、われわれはWebから情報を得るために、検索エンジンを用いるのではなく、知識発見エンジンを用いることになるだろう。その場合、ハイパーリンクを集めた大量のリストの代わりに、短時間で容易に理解できるように個人化されたサマリーを見ることができるようになると思われる。

謝辞

セマンティック・トランスコーディング・プロジェクトは筆者と慶応大SFCの学生との共同研究である。参加者の細谷真吾氏、川喜田佑介氏、有賀征爾氏、東中竜一郎氏に感謝します。また、ビデオアノテーションエディターの音声認識部については、IBM東京基礎研究所の西村雅史氏と伊東伸泰氏、言語解析については、同研究所の渡辺日出雄氏に協力していただきました。ここに記して感謝いたします。

参考文献

[1] A. Amir, S. Srinivasan, D. Ponceleon, and D. Petkovic. CueVideo: Automated indexing of video for searching and browsing. In *Proceedings of SIGIR'99*. 1999.

[2] Japan Electoronic Dictionary Research Institute. Electoronic Dictionary. http://www.iijnet.or.jp/edr/J_index.html.

[3] Koiti Hasida. Global Document Annotation. http://www.etl.go.jp/etl/nl/gda/.

[4] Moving Picture Experts Group (MPEG). MPEG-7 Context and Objectives. http://drogo.cselt.stet.it/mpeg/standards/mpeg-7/mpeg-7.htm.

[5] Katashi Nagao et al. Semantic Transcoding: Making the World Wide Web more understandable and usable with external annotations. *TRL Research Report*. IBM Tokyo Research Laboratory, 2000.

[6] Katashi Nagao and Koiti Hasida. Automatic text summarization based on the Global Document Annotation. In *Proceedings of COLING-ACL'98*. 1998.

[7] John R. Smith. VideoZoom: Spatio-temporal video browser. *IEEE Trans. Multimedia*. Vol. 1, No. 2, pp. 157–171, 1999.

[8] Michael A. Smith and Takeo Kanade. Video skimming for quick browsing based on audio and image characterization. *Technical Report CMU-CS-95-186*. School of Computer Science, Carnegie Mellon University, 1995.

[9] World Wide Web Consortium. Extensible Markup Language. http://www.w3.org/TR/PR-xml-971208.