言語的アノテーションに基づくマルチメディア要約

長尾 確
日本アイ・ビー・エム (株)
東京基礎研究所
nagao@trl.ibm.co.jp

白井 良成
慶應義塾大学
政策・メディア研究科
way@sfc.keio.ac.jp

橋田 浩一
電子技術総合研究所
hasida@etl.go.jp

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 [3]。

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。

2 アノテーション

アノテーションは、コンテンツに対するメタコンテンツであり、XML (eXtensible Markup Language) 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。

言語的アノテーションは、GDA (Global Document Annotation)[2] 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。

一般に GDA ドキュメントの構造は図 1 のようになる。

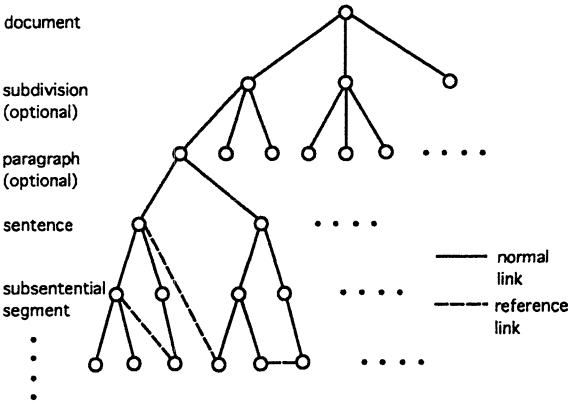

図 1: GDA ドキュメントの構造

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。

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。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2はビデオアノテーションエディターの画面例である。

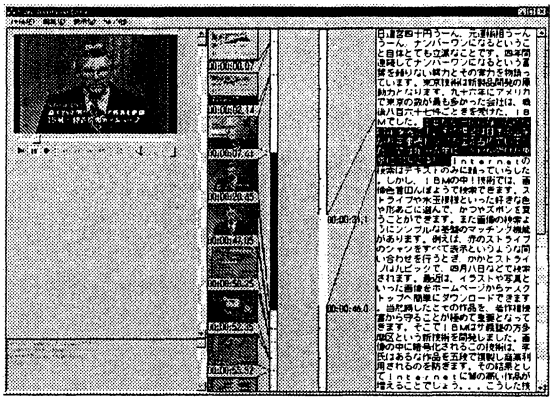

図2: ビデオアノテーションエディターの画面例

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、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。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。

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

ここでのマルチメディア要約では、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、というやり方を用いている。

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[4]の手法を用いている。

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。

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[5]。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。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。

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の部分を要約して、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。

図3は要約機能付きのビデオプレイヤーの画面例である。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。

4 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ドキュメントから重要な部分を抽出するのに大いに役に立つ。

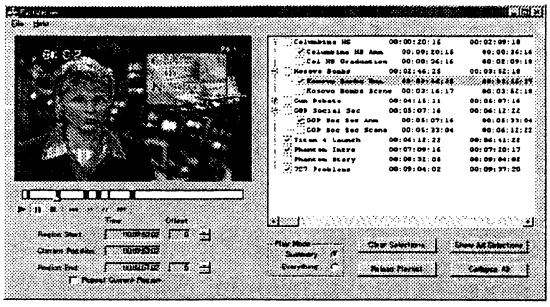

図3: 要約機能付きビデオプレイヤーの画面例

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て、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だろう。

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。参加者の川喜田佑介氏、細谷真吾氏、有賀征爾氏に感謝します。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] Koiti Hasida et al. Global Document Annotation. http://www.etl.go.jp/etl/nl/gda/.

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

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

[5] 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.