2005－DBS－137（Ⅱ）（55）
2005／7／14

# 異メディアコンテンツの差異情報に基づく対話文自動生成

灘本明代[†]　田中克己[†,††]

TV 番組と類似する Web ページよりそれらの差異情報を抽出し，そしてその差異情報に基づいた対話文を自動生成し CG キャラクターと音声合成を用いて比較テレビ番組のようなコンテンツを生成する機構の提案を行う．本論文では，TV ニュースと Web のニュースページを対象とし，これら構造の異なるコンテンツから主題語と内容語からなる話題構造を抽出し，その話題構造からトピックグラフを生成する．そしてそのトピックグラフをメイントピック，サブトピックに分け，これらの類似・相違関係より差異情報を取得する．本論文では視点差異情報と話題の詳細・広がり差異情報の 2 つの差異情報を提案する．また，抽出された差異情報に基づいた対話文の生成手法の提案も行う．

# Automatic Creation for Dialog based on Difference Information of Difference Media

Akiyo Nadamoto[†] and Katsumi Tanaka[†,††]

We have proposed a new way of automatic creation for dialog based on difference Information between TV-program and Web content. In this paper, our target content is TV-news and Web-news. We extract topic structures which consist of subject terms and content terms, and generate topic graph based on our topic structures from each media. Our topic graph consists of main-topic and sub-topics. We extract difference information between TV-news and Web-news based on main-topic and subtopics. We propose two kinds of difference information in this paper. One is viewpoint-difference-information; the other is topic-detail-spread-difference. We also propose automatic creation for dialog based on our proposed two kinds of difference information.

## 1. はじめに

インターネットの普及に伴い，Web はテレビや新聞等と同様に一般的なメディアとなっている．これに伴い，我々は様々な情報をテレビや新聞のみならず Web からも取得している．例えば日々報道されるニュースは Web が普及する以前はテレビや新聞から取得していたが，現在はこれらメディアのみならず Web 上のニュースサイトからも容易に取得することが可能になっている．しかしながら，ニュースは報道の視点や編集により種々の捉え方ができ，ひとつの事件や事故も報道機関が異なれば視点が異なり，視聴者や読者は同じニュースでも異なった印象を持つ．ましてや，テレビのニュース番組では音声だけでなく映像やインタビューを踏まえて伝えるのに対し，Web のニュースサイトは文章と画像を用いて伝えたりと，報道するメディアが異なると伝える手法も異なり，利用者にとって同じニュースソースでも異なった印象を持つ場合がある．インターネットにおけるブロードバンドの普及に伴い，PC 上でテレビや映像を見ることができるようになり，我々は同一 PC 上でテレビを視聴しながら Web を閲覧することが可能となってきている．このように，物理的にはニュース番組とニュースサイトを比較することが以前より容易になってきている．しかしながら，テレビと Web のように，異なるメディアの異なる報道機関から発信される同一のニュースソースを比較しようとした場合，テレビのニュースを視聴しそして Web のニュースを読み，どこが異なっているのか頭で考え把握しなければならない．そこで我々は，テレビと Web の類似しているコンテンツから差異情報を抽出し，自動提示してくれるシステムがあると便利であると考え，CWTB（Comparative Web and TV Browser）を提案する．

† 独立行政法人情報通信研究機構
メディアインタラクショングループ
Interactive Communication Media Contents Group, National Institute of Information and Communications Technology
†† 京都大学大学院情報学研究科社会情報学専攻
Department of Social Informatics, Graduate School of Informatics, Kyoto University

テレビ番組は動画像と音声からなる時系列データであり,「見る」「聞く」といった受動的な操作によりコンテンツを取得するインタフェースである．一方，Web コンテンツは，画像と文字列からなる 2 次元のウィンドウベースのコンテンツであり，利用者に「読む」「スクロールする」「クリックする」等の能動的な操作を要求するインタフェースである．このように，形式の異なるテレビ番組と Web コンテンツを比較し，その差異情報を提示する場合，以下に示すように 3 つのアプローチが考えられる．

- テレビ型提示方法
  利用者がテレビ番組を視聴ている時，システムはそのテレビ番組と類似する Web ページを取得後それらの差異情報を抽出し，利用者にテレビ番組のように抽出された差異情報を提示する方法.
- Web ブラウザ型提示方法
  利用者が Web コンテンツを閲覧している時，システムは閲覧している Web コンテンツと類似するテレビ番組を蓄積されたテレビ番組群から取得後それらの差異情報を抽出し，利用者にその抽出された差異情報を Web コンテンツのように変換して提示する方法.
- パラレル提示方法
  利用者がテレビ番組を視聴している時，システムはそのテレビ番組と類似する Web コンテンツを取得し，利用者にテレビ番組と同時に取得した Web コンテンツを提示するとともに，その差異情報も提示する方法.

テレビ型提示方法とパラレル提示方法はリアルタイム放送及び蓄積されたテレビ番組に対応できるのに対し，Web ブラウザ型提示方法は類似テレビ番組の検索を行うため，蓄積されたテレビ番組にのみ対応可能である．また，パラレル提示方法はテレビと Web とを同時に視聴・閲覧を行わなければならないため，利用者にとって負担が大きいことが予想される．そこで本論文では，テレビ型提示方法を用いた CWTB の提案を行う．本論文では，これまで我々が提案してきた Web2Talkshow[1] の技術を利用し，キャラクターアニメーションと音声合成によりテレビ番組との Web コンテンツの差異情報を対話文により利用者に伝えることを行う．また，日々のニュースを対象とし，ニュース番組とニュースサイトの比較提示を行うことを提案する．一般にニュース番組は複数のニュースからなり，ニュースサイトは 1 つのニュースは 1 つの Web ページで提示されている．そこで，本論文ではテレビのニュース番組の複数ニュース各々を TV ニュースと呼び，Web 上のニュースページを Web ニュースと呼ぶ．時系列データの分割方式は種々提案されており [2)3)]，本論文では TV ニュースはあらかじめ分割されているものとする．また，TV ニュースの文字放送データをその TV ニュースのメタデータとし，このメタデータから話題構造を抽出し，類似 Web ニュースの取得及び比較を行う．図 1 にシステムの流れと画面イメージ図を示す.

CWTB の特徴を以下に示す.

- 構造の異なる TV ニュースと Web ニュースからのキーワードに基づく話題構造の抽出
- 抽出された各々の話題構造に基づく TV ニュースと Web ニュースの比較及び差異情報の抽出.
- 差異情報に基づく対話文の自動生成

CWTB を使用することにより，ユーザはテレビ番組を見るように容易に TV ニュースと Web ニュースの相違を知ることができるようになる．

以下，2 章では類似 Web ニュースの取得と比較を，3 章では対話文生成について述べ，4 章でまとめと今後の課題について述べる．

## 2. 類似 Web ニュースの取得と比較

### 2.1 TV ニュースからの話題構造の抽出

ニュースは時系列データであり，時々刻々と変化している．これら時系列データであるニュースの話題を抽出する方法は種々提案されている [4)5)]．本論文では，TV ニュースと Web ニュースの差異情報を抽出する最初の一歩とし，これらニュースページの時間変化による話題構造は考慮しないものとする．

これまで我々は複数の Web ページの比較を行う CWB[6] の提案において，小山らの提案する [7] 主題語と内容語からなる話題構造を利用し類似 Web ページの比較を行ってきた．ここで，主題語は単語の出現頻度が高く且つページのタイトル，サブタイトルに含まれる名詞句であり，内容語は単語の出現頻度が高く且つ Web ページ内の内容を示す文章に含まれる名詞句と定義している．主題語はそのページの特徴を示す単語群であり，ユーザの興味を引く単語群である．また，内容語はそのページの内容を示す単語群である．本論文では，この主題語，内容語からなる話題構造を用いて TV ニュースと Web ニュースの比較を行うことを提案する．しかしながら，TV ニュースは Web ニュースと異なり，時系列データであるとともにメタデータに構造がなく，テキストデータの集まりである．また，TV ニュースにはタイトルが画面上に表示されているが，我々がメタデータとして使用する文字放送

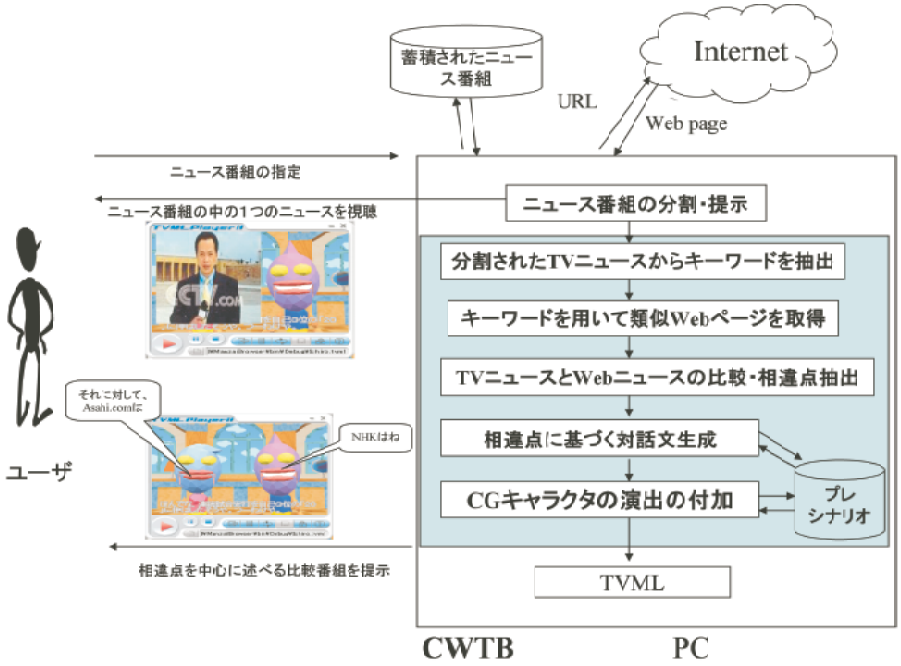


**図 1** CWTB システムの流れ
Fig. 1 System flow of the CWTB

にはこのタイトルが含まれていることがほとんどない．そこで，我々は，TV ニュースの原稿の構造に注目し，TV ニュースから話題構造を抽出することを提案する．

「パッケージニュース」と呼ばれるニュースは Lead-in, Body, Interview, Cut, Stand upper, Sign-off の構成になっている [8]．8) によると，Lead-in は普通は大体２０秒以下といわれ，短いものだと５秒程度であり，最も重要な役目は，視聴者の関心を引きつけ，続くリポートへの準備の体勢を取らせることにある．テレビのニュース番組の視聴者は新聞を読むように，分からなかったところを後戻りして理解することはできないため，Lead-in では，原則としてニュースの要約ということはまずありえないと述べられている．実際の TV ニュースでは，一般的に最初の 1 文もしくは 2 文をアナウンサーが読み上げ，その後アナウンサーが間をあけその後新たにニュースの内容をはじめから述べるか，もしくは現場からの報告やインタビュー映像に切り替わる．これに対し，Web ニュースはタイトル（サブタイトル）と内容からなっている．このタイトルはその Web ニュースのリンク元のアンカー文字列となってりるのが一般的である．アンカー文字列は一般的にリンク先のページを顕著に示す文または単語の集合となっている場合がおおく，つまりはタイトルは Web ニュースの内容を顕著に示すものであったり，ユーザの関心をひきつけるものであると考えられる．これらのことより，TV ニュースの最初の 1 文もしくは 2 文と Web ニュースのタイトルはユーザ（もしくは視聴者）の関心をひきつけるという性質が似ていると考え，我々は TV ニュースの主題語をメタデータの最初の 1 文もしくは 2 文から抽出することを提案する．TV ニュースの主題語，内容語の抽出方法を以下に示す．

- 主題語
  メタデータの最初の 1 文を主題語を抽出する領域である主題語抽出領域とする．また，次の文が最初の文との接続詞,「したがって」「そして」などで構成されている場合，前の最初の 1 文とつながりのある文であると考え，次の文も主題語抽出領域とする．一つの TV ニュース全体で単語の出現頻度を求め，単語の出現頻度と品詞による重み付けによる特徴ベクトルがある閾値以上の単語で且つ決定した主題語抽出領域の中にある単語を主題語 $TS(i)$ $i \in \{1, \ldots, n\}$ とする．
  つまりは，主題語は
  $$tf(t) \times weight(t) > \alpha$$
  で決定された単語群の中から主題語抽出領域に含まれる単語となる．ここで，$tf(t)$ は TV ニュース $TN$ における単語 $t$ の出現頻度を示し，$weight(t)$

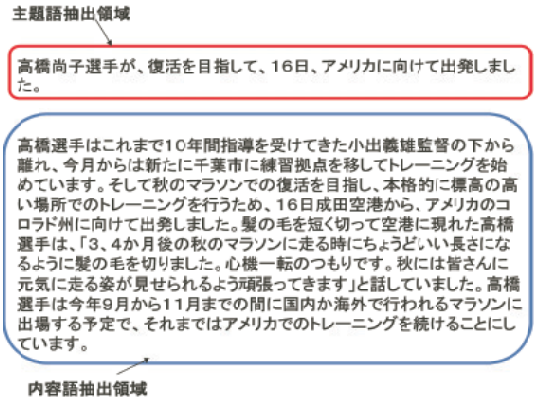


**図 2** TV ニュースのメタデータ
Fig. 2 Metadata of TV News

は品詞による単語の重みを示し，$\alpha$ は閾値を示す．

- 内容語
  内容語 $TC(j)$ $j \in \{1, \ldots, m\}$ は主題語抽出領域以外の TV ニュースを対象とする領域である内容語抽出領域より抽出される．つまりは，単語の特徴ベクトルの値が閾値 $\alpha$ 以上である単語の内，主題語以外の単語群を内容語とする．

図 2 の場合，この TV ニュース全体における単語の特徴ベクトルの値が $\alpha$ 以上の単語が値の大きい順に
{ 高橋，トレーニング，アメリカ，選手，マラソン，秋，髪，毛，尚子，小出，千葉，コロラド }
であった場合，主題語は
{ 高橋，アメリカ，選手，尚子 }
となり，内容語は
{ トレーニング，マラソン，秋，髪，毛，小出，千葉，コロラド } となる．

### 2.2 類似 Web ニュースの取得

先に求めた話題構造の主題語はその TV ニュースの特徴を示す単語群であると考え，類似 Web ニュースの取得には，TV ニュースの主題語を検索キーワードとする．そして検索結果の中で，TV ニュースが発信された時間ともっとも近い時間に発信されている Web ニュースを類似 Web ニュースとする．

このとき，TV ニュースでは「アメリカ」という単語が Web ニュースでは「米国」となっているなど，TV ニュースと Web ニュースでは同じ意味を指し，単語の異なる言葉がある．このような言葉は，あらかじめ辞書を作成し，単語を置き換えて対応することを行う．

図 2 の場合，{ 高橋，アメリカ，選手，尚子 } が検索キーワードとなる．検索には GoogleAPI を使用する．

### 2.3 トピックグラフの生成

TV ニュースと類似 Web ニュースの比較を行うために，各々のコンテンツの話題構造からトピックグラフを生成し，そのトピックグラフに基づいた比較を行う．主題語はそのコンテンツの特徴を示す単語群であることより，その主題語と内容語との関係からそのコンテンツの話題の構造を示すトピックグラフを生成することを行う．主題語と内容語が単語同士が隣接している場合はそれらの単語の関係が強いと考える．また，同一文内にあり，且つ係り受け関係がある場合も，さらにその単語同士の関係は強いと考えられる．これにより本論文では，各々のコンテンツ内の文章における主題語と内容語の位置関係と係り受け関係から重みを求め，主題語，内容語を節点とする，重みつき無向グラフを生成する．実際には，トピックグラフの生成は以下の手順にて行う．

( 1 ) 主題語と内容語との関係を示すグラフの生成

( 2 ) 主題語同士の位置関係より (1) で求めたグラフの連結成分の結合

#### (1) 主題語と内容語との関係を示すグラフの生成

トピックグラフの重み $TW(s(i), c(j))$ を以下のように決定する．

$$TW(s(i), c(j)) = \sum_{k=1}^{nn \times mm} \frac{1}{wd} \times pw$$

ここで nn は $s(i)$ の出現頻度を mm は $c(j)$ の出現頻度を示し，$wd$ は $s(i)$ と $c(j)$ との単語間の距離を示す．また，$pw$ は $s(i)$ と $c(j)$ との係り受け関係を示す．

単語間の距離は，隣接する単語同士の距離を 1 とし，単語間に n 個の単語がある場合は 1+n とする．$pw$ は $s(i)$ と $c(j)$ が同一文にあった場合の係り受け構造により決定され，$s(i)$ が $c(j)$ に係るもしくは係られる場合は 2 の値とし，その他の場合は 1 とする．

$TW(s(i), c(j))$ がある閾値 $\beta$ 以上の時，その主題語と内容語を連結し，グラフを作成する．この時，TV ニュースの主題語抽出領域の文は Web ニュースでのタイトルと同じ意味合いを示すと考え，TV ニュースのトピックグラフの重み付け計算は内容語抽出領域のみで行う．また，同様に Web ニュースの場合も，タイトル，サブタイトルを除いた部分から重み付け計算を行う．このようにして，TV ニュースと Web ニュース各々の主題語と内容語の関係を示すグラフを生成する．図 2 の TV ニュースから生成した主題語と内容語の関係を示すグラフを図 3(a) に示す．

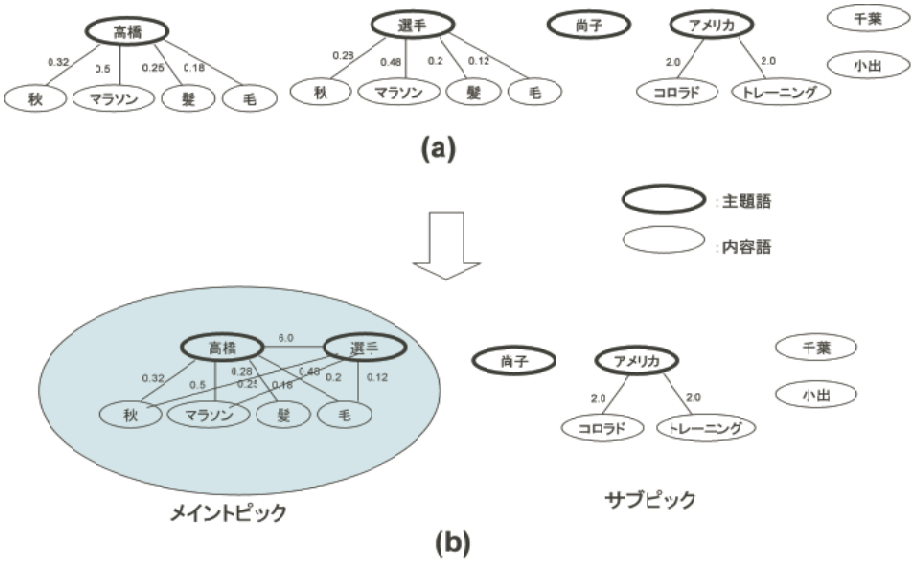


**図 3** トピックグラフの例
Fig. 3 Example of Topic Graph

**(2) 連結成分の結合**

(1) と同様に主題語同士の重み付けを求め，その重みがある閾値 $\beta$ 以上の時，作成した主題語と内容語の関係を示すグラフの連結成分を結合しトピックグラフを生成する．このとき，もっとも多い節点数を持つ連結成分がそのコンテンツの中心となる話題の集合と考え，メイントピックとする．その他の連結成分をサブトピックとする．図 2 のトピックグラフを図 3(b) に示す．

このようにして，TV ニュースと類似 Web ニュース各々のトピックグラフを生成する．

### 2.4 TV ニュースと類似 Web ニュースの比較

生成したトピックグラフに基づいて，TV ニュースと類似 Web ニュースを比較し，その差異情報を抽出する．複数のコンテンツの差異情報の抽出では，視点の相違やコンテンツの詳細度の相違，感情の相違など種々の差異情報が考えられる．本論文では，差異情報の抽出のはじめの一歩として，TV ニュースと類似 Web ニュースの視点の差異情報と話題の広がり及び詳細の差異情報をトピックグラフのメイントピックとサブトピックの関係から取得することを提案する．

**視点差異情報**

視点差異情報とは，TV ニュースと類似 Web ニュースの書かれている視点が異なると思われる情報である．つまりは，視点差異情報とはコンテンツ全体における差異情報であるといえる．メイントピックはそのコンテンツの中心となる話題を示す単語の集合であり，つまりはそのコンテンツがどのように書かれているか，コンテンツ作成者の視点であると考え，このメイントピックに注目して視点差異情報を取得することを行う．

実際には，メイントピックとサブトピックの特徴ベクトルからユークリッド距離を求めて差異情報を取得する．2 つのコンテンツ間のメイントピックの類似度 $Sim_m$ は以下のように決定される．

$$Sim_m = \sqrt{F_a(1)^2 + \ldots + F_a(i)^2 + \ldots + F_a(n)^2}$$

$$F_a(i) = f_t(t_i) - f_w(w_i)$$

$$f_t(t_i) = tf(t_i) \times weight(t_i)$$

ここで，$f_t(t_i)$ は TV ニュースのメイントピックに含まれる単語の特徴ベクトルの要素であり，$f_w(w_i)$ は類似する Web ニュースのメイントピックに含まれる単語の特徴ベクトルの要素である．同様に，TV ニュースと類似 Web ニュースのサブトピック間の類似度及び TV ニュースのメイントピックと類似 Web ニュースのサブトピックの類似度，TV ニュースのサブトピックと類似 Web ニュースのメイントピックの類似度を求める．これらの類似度がある閾値以上のものを類似しているとし，閾値以下のものを相違しているとする．この類似，相違関係より，視点差異情報を取得する．

メイントピックとサブトピックの関係から以下の 3 種類の差異情報を定義する．

- 視点類似
  メイントピック同士が類似している場合，その TV ニュースと類似 Web ニュースは視点類似である

と定義する．

- 注視点相違
  メイントピック同士は類似していないが，メイントピックとサブトピックが類似している場合，その TV ニュースと類似 Web ニュースは注目すべき視点が異なると考え，注視点相違であると定義する
- 視点相違
  メイントピック同士が類似しおらず，且つメイントピックとサブトピックも類似していない場合，その TV ニュースと類似 Web ニュースは視点相違であると定義する．

**話題の詳細・広がり差異情報**

視点差異情報では TV ニュースと類似 Web ニュース各々のコンテンツ全体がどのようにかかれているかの差異情報を取得することを提案した．それに対し，TV ニュースと類似 Web ニュースのどの部分がより詳細にかかれているか，またどの部分の話題が広がってるかの差異もある．そこで，本論文では TV ニュースと類似 Web ニュースの話題の詳細，広がりの差異を提案する．

- 話題の詳細差異情報
  メイントピックはコンテンツ $D$ の中心となる話題の構造，つまりはコンテンツ $D$ のテーマについて述べている単語群であると考え，その節点数はコンテンツ $D$ における話題の詳細度を示すと考える．TV ニュースのメイントピックと最も類似度の高い類似 Web ニュースのトピックグラフの連結成分との節点の差分が話題の詳細差異情報であるとする．
- 話題の広がり差異情報
  サブトピックはテーマとなる話題の周辺の話題であると考え，その節点は話題の広がりであるとする．類似 Web ニュースのサブトピックの連結成分が TV ニュースのいずれのサブトピックの連結成分と類似していない場合，その類似 Web ニュースのサブトピックの連結成分の節点が話題の広がり差異情報であるとする．また，同様に TV ニュースのサブトピックの連結成分が類似 Web ニュースのいずれのサブトピックの連結成分と類似していない場合，その TV ニュースのサブトピックの連結成分の節点も話題の広がり差異情報であるとする．

## 3. 対話文の生成

本論文では，TV ニュースと類似する Web ニュースの違いをユーザに示すために，キャラクターアニメーションと音声合成による提示方法を提案する．この時，2 人のキャラクターが対話を用いてその差異情報を提示することにより，よりわかりやすく 2 つのコンテンツの差異を示すことができると考え，差異情報に基づいた対話文の生成を行う．ここで，完全な自動で対話文を生成することは困難であるため，プレシナリオと呼ぶ対話のフレームワークを XML で記述した台本を用いる．差異情報の取得で提案した視点差異情報は各々のコンテンツ全体の視点に対する差異情報である．つまりは，視点差異情報の種類により対話全体の雰囲気を変えることを考え，プレシナリオを視点類似，注視点類似，視点相違の 3 種類作成し，そのプレシナリオに基づいた対話を生成することを行う．これまで我々が提案してきたプレシナリオの構成[1)] に基づき，CWTB のプレシナリオもイントロ，メイン，まとめの 3 つのパートから構成する．また，話題の詳細・広がり差異情報はどの単語が違うのかを抽出しているため，プレシナリオに書かれた対話のタイプを決定し，抽出した異なる単語及び文に基づいた比較対話文を生成する．表 1 にプレシナリオのタグを示す．

**表 1** プレシナリオの XML タグ
Table 1 Pre-scenario Content Tags

| 構造タグ | |
|---|---|
| Initialize | プレ台本はこのタグにより囲まれる．type 属性により視点差異情報のタイプを指定する |
| Intro | イントロを示すタグである．Intro タグは始めの挨拶と，TV ニュースのテーマを述べる対話で構成される． |
| Dialogue | メインを示すタグである．Dialogue タグは対話文を自動生成する部分であり，各対話の骨組みからなる． |
| Conclusion | まとめを示すタグである．Conclusion タグは終わりの挨拶を示す対話で構成されている． |
| Content Tags | |
| line | キャラクターの台詞を示すタグである．chara 属性でどのキャラクターが話す台詞なのかを指定する． |
| question | 質問・応答のフレームワークを指定するタグである．type 属性により話題の広がり，詳細の差分のタイプを決定する． |

### 3.1 視点差異情報によるプレシナリオの決定

以下に視点差異情報によるプレシナリオの内容を示す．

**視点類似**

TV ニュースと類似する Web ニュースの視点が類似しているため，おだやかな対話文を生成するプレシナリオを記述する．2 つのメインテーマが類似している

ので，対話は主にメインテーマの話題の詳細度の差異情報を中心とする．また，サブテーマが相違している場合，この相違しているサブテーマつまりは話題の広がり差異情報についても述べる．

**注視点相違**

注視点が異なる場合，コンテンツ作成者が注目する点が違うため，2 人のキャラクターが対決するようなプレシナリオを記述する．例えば，一方がある X-TV 局ファンに対し，もう一方がアンチ X-TV 局であるといったような対戦型対話を生成する．また，この場合 2 つのコンテンツのメインテーマ同士が相違し且つメインテーマとサブテーマが類似しているので，各々のメインテーマの概要を述べ，また類似しているメインテーマとサブテーマの差分情報である話題の詳細差異情報について述べる．

**視点相違**

この場合，注視点相違よりさらにコンテンツ間の情報が相違しているため，注視点相違よりより過激な対決型プレシナリオを記述する．すべての差異情報を対話にしたのでは，あまりにも差異情報が多くなる場合があるため，メイントピック同士の差異情報つまりは話題の詳細差異情報について述べる．

### 3.2 話題の詳細・広がり差異情報による対話文生成

視点差異情報によりプレシナリオのタイプを決定した後，話題の詳細・広がり差異情報に基づきプレシナリオに記述した対話のタイプを決定し，差異情報となる単語を含む文から対話文を生成する．この時，対話の順番は差異情報となる単語の TV ニュースにおける出現順とし，TV ニュースにない単語は Web ニュースにおける出現順により決定される．

**話題の詳細差異情報による対話文生成**

話題の詳細差異情報は，TV ニュースのメイントピックと類似する Web ニュースの連結成分の差分情報である．つまりは，TV ニュースでは語られていない情報で且つ TV ニュースでは中心となる話題に関連する単語群であると考えられる．そこで，対話のきっかけとして，TV ニュースのメイントピックの主題語を含む文について述べた後，話題の詳細差異情報を強調するような対話を生成する．例えば，図 2 において，TV ニュースのメイントピックと類似している類似 Web ニュースの連結成分が { 高橋，選手，ファンテン，マラソン，髪，毛 } の場合，TV ニュースの詳細差異情報は { 秋 } であり，類似 Web ニュースの詳細差異情報は { ファンテン } となり { 高橋，選手，マラソン，髪，毛 } は共通の単語である．この時,「高橋」と「選手」が主題語であるため，この 2 つの単語を含む文を TV ニュースから抽出し対話のきっかけとして用いる．以下に視点類似の場合の話題の詳細差異情報による対話例を示す．

ボブ：髪の毛を短く切って空港に現れた高橋選手の話だけど…

マリ：ふーん．シドニー五輪女子マラソン金メダルの高橋尚子（ファイテン）でしょ．

ボブ：ほー．よく知ってるね．じゃ秋にどうしたのか知ってる？

マリ：知らないよ．秋にどうしたの？

ボブ：もっと詳しいところまで知らないとね．秋のマラソンでの復活を目指すんだって．

マリ：へー．

ここでプレシナリオでは

ボブ：$TVnews1 の話だけど

マリ：ふーん．$Webnews_d1 でしょ．

ボブ：ほー．よく知ってるね．じゃ$TVdiff1 にどうしたのか知ってる？

マリ：知らないよ．$TVdiff1 にどうしたの？

ボブ:もっと詳しいところまで知らないとね．$TVnews_d1 だって．

マリ：へー．

となる．ここで，$TVnews1 は TV ニュースと Web ニュースの共通の単語のうち主題語をいくつか含む TV ニュースの文であり，$Webnews_d1 は Web ニュースの詳細差異情報を含む Web ニュース上の文である．また，$TVdiff1 は TV ニュースの詳細差異情報であり，$TVnews_d1 は$TVdiff1 を含む TV ニュースの文である．

**話題の広がり差異情報による対話文生成**

話題の広がり差異情報はサブトピック同士の比較から求められているため，一方のコンテンツでは述べられていないが，他方のコンテンツでは述べられている新しく且つ重要度がそんなに高くない話題であると考える．したがって，こんなことまで知っているという自慢に似た対話を生成することを行う．Web ニュースの話題の広がり差異情報が { スタッフ，チーム Q} であった場合，以下のような対話を生成する．

マリ：高橋が何を結成したか知ってる？

ボブ：知らないよ．

マリ：高橋は専属スタッフ３人と「チームＱ」を結成したんだよ．

ボブ：そんなこと知っててどうなるの？このニュースには関係ないんじゃない？

マリ：関係あるよ，だってこのニュースは高橋の話題なんでしょう!!

ここでプレシナリオでは
マリ：$Webnews_s1 何を$Webnews_v1 か知ってる？
ボブ：知らないよ．
マリ：$Webnews_w1 んだよ．
ボブ：そんなこと知っててどうなるの？このニュースには関係ないんじゃない？
マリ:関係あるよ，だってこのニュースは$Webnews_s1 の話題なんでしょう!!
ここで$Webnews_s1 は Web ニュースの話題の広がり差異情報を含む文の主語であり．$Webnews_v1 は述語である．$Webnews_w1 は話題の広がり差異情報である．

## 4. ま と め

本論文では，TV ニュースと類似する Web ニュースの差異情報を抽出し，その差異情報に基づいた対話文を自動生成し CG キャラクターと音声合成を用いて比較テレビ番組のようなコンテンツを生成する機構である CWTB の提案を行った．実際には，構造の異なる TV ニュースと Web ニュースから主題語と内容語からなる話題構造を抽出し，その話題構造からトピックグラフを生成した．そしてそのトピックグラフより視点差異情報と話題の詳細・広がり差異情報の 2 つの差異情報を求める手法の提案を行った．抽出された差異情報に基づき対話文の生成手法の提案も行った．本論文では CWTB の最初の一歩であり，例えば視点差異情報はコンテンツ全体における視点差異情報を取得しているが，実際にはコンテンツ内においても様々な視点から述べられているニュースがある．今後は，コンテンツ内における視点の変化を抽出することを行う予定である．また，対話文生成においても，表層的な対話を生成しているが，概念辞書を用いるなどしてさらに内容に踏み込んだ対話文生成をしてゆきたい．

## 参 考 文 献


1) 灘本明代, 田中克己,「対話文自動生成による Web コンテンツの受動的視聴」, 情報処理学会研究報告, Vol.2004, No.72 2004-DBS-134(I)，pp.183-190 2004 年 7 月．
2) Utiyama M., Isahara H,, "A Statistical Model for Domain-Ondependent Text Segmentation.", ACL/EACL, 2001, pp.491-498
3) 馬強，田中克己, "話題構造に基づく放送と Web コンテンツの統合のための検索機構", 情報処理学会論文誌：データベース Vol.45 No.SIG 10 (TOD23)，pp.18-36, 2004
4) J.Kleinberg, "Bursty and Hierarchical Structure in Streams", Proc. of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp91-101, Edmonton, Alberta, Canada, July 2002.
5) J.Allan, R.Papka, V.Lavrenko,"On-line New Event Detection and Tracking", Proc. of 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.37-45, Melbourne, Australia, August 1998.
6) Akiyo Nadamoto and Katsumi Tanaka,“ A Comparative Web Browser (CWB) for Browsing and Comparing Web Pages ”, Proceedings of the 12th International World Wide Web Conference (WWW2003) , pp.727-735, Budapest, Hungary, May 2003.
7) 小山 聡，田中 克己,“ 話題の階層構造を反映した Web 検索手法の提案 ”，情報処理学会研究報告，Vol.2002, No.67 2002-DBS-128，pp.465-472 2002 年 7 月
8) http://akasaka.cool.ne.jp/kakeru3/bs3.html