

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

Automatic Creation for Dialog based on Difference Information of Difference Media

灘本 明代[†] 田中 克己^{†,‡}

Akiyo NADAMOTO Katsumi TANAKA

TV番組と類似するWebページから差異情報を抽出し、その差異情報に基づいた対話文を自動生成し、比較TV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。

We propose a new way of automatic creation for dialog based on difference information between TV-program and Web content. We extract topic structures which consist of subject terms and content terms, and generate topic graph based on our topic structures from each media. We extract difference information between TV-news and Web-news based on main-topic and subtopics. We propose two kinds of difference information. 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

現在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。ブロードバンドの普及に伴い、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、テレビのニュース番組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。このように、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。CWTB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を用いてTV番組と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。本論文では、ニュース番組とニュースサイトを対象とし、これら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構成され、ニュースサイトは1つのニュースは1ページの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ニュースソースごとに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。

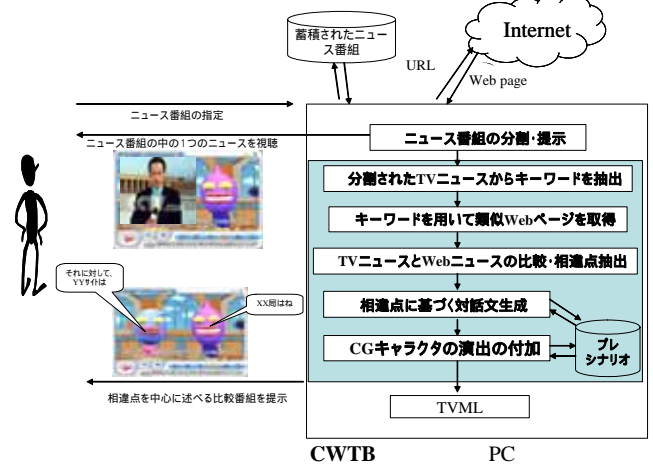


図1 CWTBのシステムの流れ

Fig.1 System Flow of CWTB

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

これまで我々は複数のWebページの比較を行うCWB[4]の提案において、小山ら[5]の提案する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。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ニュースの原稿の性質に注目し、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[6]。これによると、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

[†] 正会員 独立行政法人情報通信研究機構
nadamoto@nict.go.jp

[‡] 正会員 京都大学大学院情報学研究所社会情報学専攻
tanaka@dl.kuis.kyoto-u.ac.jp

準備の体勢を取らせることにある。Lead-in では、原則としてニュースの要約ということは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、その後間をとり、再びそのニュースの詳細な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と内容からなっている。このタイトルは Web ニュースの内容を顕著に示すものであり、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。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 $TS(i) \ i \in \{1, \dots, n\}$ 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 TV における単語 t の出現頻度を示し、 $weight(t)$ は品詞による単語の重みを示し、 α は閾値を示す。

● 内容語

内容語 $TC(j) \ j \in \{1, \dots, m\}$ 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 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。

主題語抽出領域

高橋尚子選手が、復活を目指して、16日、アメリカに向けて出発しました。

高橋選手はこれまで10年間指導を受けてきた小出義雄監督の下から離れ、今月からは新たに千葉市に練習拠点を移してトレーニングを始めています。そして秋のマラソンでの復活を目指し、本格的に標高の高い場所でのトレーニングを行うため、16日成田空港から、アメリカのコロラド州に向けて出発しました。髪を短く切って空港に現れた高橋選手は、「3、4か月後の秋のマラソンに走る時にちょうどいい長さになるように髪を切りました。心機一転のつもりです。秋には皆さんに元気な走る姿が見せられるよう頑張ってきます」と話していました。高橋選手は今年9月から11月までの間に国内か海外で行われるマラソンに出場する予定で、それまではアメリカでのトレーニングを続けることにしています。

内容語抽出領域

図2 TVニュースのメタデータ

Fig.2 Metadata of TV News

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 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

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最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。

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以下のように生成する。

(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 $TW(s(i), c(j))$ を決定する。

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

ここで n は $s(i)$ の出現頻度を m 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 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。

(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 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,メインピックとする.その他の連結成分をサブピックとする.図2のトピックグラフを図3(b)に示す.

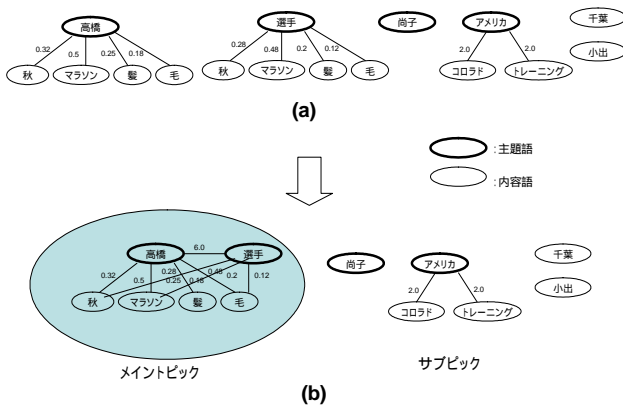


図3 トピックグラフの例
Fig.3 Example of Topic Graph

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ピックとサブ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,メインピックとサブ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.

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

$$F_a(i) = F_t(t_i) - F_w(w_i)$$

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

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ニュースのサブピック間の類似度及びTVニュースのメインピックと類似Webニュースのサブピックの類似度,TVニュースのサブピックと類似Webニュースのメインピックの類似度を求め

る.これらの類似度がある閾値以下のものを類似しているとし,閾値以上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.メインピックとサブ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:メインピック同士は類似していないが,メインピックとサブ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:メインピック同士が類似しおらず,且つメインピックとサブ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っ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っているか,またどの部分の話題が広がっているかの差異もある.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

- 話題の詳細差異情報

メイン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ニュースのサブピックの連結成分がTVニュースのいずれのサブ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ニュースのサブピックの連結成分が類似Webニュースのいずれのサブ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. 対話分の生成

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.ここでは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.プレシナリオを視点差異情報の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.

3.1 視点差異情報によるプレシナリオの決定

- 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。
- 注視点相違: 注視点異なる場合、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。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
マの差分である話題の詳細差異情報について述べる。
- 視点相違: 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、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、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

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。

<対話例>
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: ほう。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かね。秋のマラソンでの復活を目指すんだって。
マリ: へー。
<プレシナリオ>
ボブ: \$TVnews1の話だけだ
マリ: ふーん。\$Webnews¥_d1でしょ。
ボブ: ほう。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かね。\$TVnews¥_d1だって。
マリ: へー。

(a)

<対話例>
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じゃない!!
<プレシナリオ>
マリ: ¥\$Webnews¥_s1何を¥\$Webnews¥_v1か知ってる?
ボブ: 知らないよ。
マリ: ¥\$Webnews¥_w1なんだよ。
ボブ: そんなこと知ってどうなるの? 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の話題なんじゃない!!

(b)

図4 プレ台本例

Fig.1 Example of Pre-Scenario

- 話題の詳細差異情報による対話文生成

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。図4(a)に視
点類似の場合の話題の詳細差異情報による対話例とそのプ
レシナリオを示す。

- 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られる。したがって、これ
も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}の場合の例を図4(b)に示す。

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 TV 番組のようなコン
텐츠を生成する機構である CWTB の提案を行った。本論文
では CWTB の最初の一步であり、例えば視点差異情報はコン
텐츠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。

【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-Independent 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] 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.
- [5] 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
- [6] <http://akasaka.cool.ne.jp/kakeru3/bs3.html>

灘本 明代 Akiyo NADAMOTO

独立行政法人情報通信研究機構勤務。2002年神戸大学大学院
自然科学研究科博士後期課程修了, 博士(工学)。マルチ
メディアコンテンツの情報配信, 閲覧に関する研究に従事。
情報処理学会, 日本データベース学会会員。

田中 克己 Katsumi TANAKA

京都大学大学院情報学研究所社会情報学専攻教授。1976年
京都大学大学院前期博士課程修了, 工学博士。主にデータ
ベース, マルチメディアコンテンツ処理の研究に従事。IEEE
Computer Society, ACM, 人工知能学会, 日本ソフトウェア
科学会, 情報処理学会, 日本データベース学会会員。