社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2006－ＮＬ－174（５）
2006／7／27

AdaBoostを利用した字幕テキストからの定型表現文章区間抽出

山田一郎* 三浦菊佳* 住吉英樹* 八木伸行* 奥村学† 徳永健伸‡

*ＮＨＫ放送技術研究所 〒157-8510 東京都世田谷区砧 1-10-11
† 東京工業大学精密工学研究所 〒226-8503 横浜市緑区長津田 4259
‡ 東京工業大学大学院情報理工科学研究科〒152-8552 東京都目黒区大岡山 2-12-1
E-mail: yamada.i-hy@nhk.or.jp

あらまし テレビ番組のナレーションでは、「場所紹介」や「人物紹介」など特定の事柄を表現するために同じような言い回しが多用される。このような言い回しを含む文章区間が抽出できれば、対応する番組映像区間の場所紹介や人物紹介といったメタデータを付与することができる。そこで本稿では、番組のクローズドキャプションを対象として定型表現を含む文章区間を抽出する手法を提案する。提案手法では、複数文のテキストデータから木構造を生成して、木構造間の類似性を木構造に含まれる部分木の類似度により評価する。この結果を弱学習器としたAdaBoostアルゴリズムにより学習を行い定型表現か否かの判定を行う。紀行番組のクローズドキャプションを対象として、場所を映像とともに説明する定型表現文章区間を抽出する実験を行い、提案手法の有効性を確認した。

キーワード メタデータ，特定表現抽出，クローズドキャプション，木構造，アダブースト

Detection of Text Sections which contain typical forms from Closed Captions using AdaBoost Algorithm

Ichiro YAMADA* Kikuka MIURA* Hideki SUMIYOSHI* Nobuyuki YAGI*
Manabu OKUMURA† and Takenobu TOKUNAGA‡

*NHK Science & Technical Research Laboratories 1-10-11 Kinuta, Setagaya-ku, Tokyo, 157-8510 Japan
† Precision & Intelligence Laboratory, Tokyo Institute of Technology 4259 Nagatsuda, Midori-ku, Yokohama, 226-8503
‡ Department of Computer Science, Tokyo Institute of Technology 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552
E-mail: yamada.i-hy@nhk.or.jp

Abstract In the closed captions, there are a lot of typical expressions to express specific things, for example, first introduction of a guest in a talk show or explanation of a place in travel program. Such information helps us to put metadata to the corresponding scenes. This paper proposes a method to extract a section including typical expressions. The first step generates tree structures from inputted section of sentences and evaluates the similarities between those tree structures. We use these similarities as weak learners of adaboost algorism to judge whether the section of sentences includes typical expressions or not. In the experiment of detecting sections including typical expressions which explain a place with video targeting closed captions of TV programs concerned with travel, we show the effectiveness of our method.

Keyword Metadata, Typical expression extraction, Closed-caption, Tree Structure, AdaBoost Algorithm

1. はじめに

近年、放送局では番組を蓄積・管理するシステムが普及し、NHKにおいてもNHKアーカイブス[1]として約59万本もの番組が蓄積されるようになった。このうち、約5千本は公開ライブラリとして利用されているが、その他は番組制作のために参照している程度で、十分に活用されているとは言えない。そこで、放送された番組を映像百科事典などの新たなコンテンツとし

て有効利用するため、我々は、番組のどの区間に何が映っているかというセグメントメタデータ情報を自動付与する研究に取り組んでいる。これまでに、映像に映っている被写体をクローズドキャプションから抽出する手法を提案してきた[2]。この手法では、クローズドキャプション中に出現する具象物名詞が被写体であるか否かを、統語構造を手がかりとした統計手法により判定している。しかし、被写体が映っている文章区間を特定する処理までは行っていない。

テレビ番組のナレーションでは、「場所紹介」や「人物紹介」など特定の事柄を表現するために同じような言い回しが多用される。例えば、表1に示すクローズドキャプション中では、矩形で囲まれた部分が「場所」を映像とともに説明している。最初に体言止めにより「オンフルール」という町の位置情報を説明し、次に町の詳細を断定の助動詞「です」を使って説明している定型的な表現である。このような文章区間を抽出することができれば、対応する番組映像区間に「場所：オンフルール」というメタデータを付与することができる。そこで本稿では、番組のクローズドキャプションを対象として定型表現を含む文章区間を抽出する手法を提案する。提案手法では、複数文のテキストデータから木構造を生成して、木構造間の類似性を評価する。この結果を弱学習器としたAdaBoostアルゴリズム[3]により学習を行い定型表現か否かの判定を行う。

以下、2章で関連研究についてまとめ、3章では定型表現を含む文章区間の抽出処理の詳細を説明する。4章では、NHKで放送された「わが心の旅」という紀行番組のクローズドキャプションから、場所を映像とともに説明する定型表現を含む文書区間を抽出する実験と評価を行い、最後にまとめと今後の課題について述べる。

2. 関連研究

クローズドキャプションから特定の事項を表現する文章区間を抽出する手法として、まず文章内容の区切れ目を特定してから、各区間で特定の事項を表現しているかを判定するアプローチが考えられる。Hearstは、テキストに含まれる単語の出現頻度から隣接ブロック間の類似度を計算し、この値の変化から内容の区切れ目を推定する手法を提案した[4]。また、望月らは、単語の語彙的結束性や接続詞、修飾語などの表層的な手がかりに基づき内容の区切れ目を推定する手法を提案した[5]。しかし、本稿で対象とする一つの番組に付与されたクローズドキャプションでは、番組開始から終了まで同じテーマについて論じることが多いため、重要な単語は番組全体に均等に出現する傾向がみられ、単語の集合のみを特徴としたこれらの手法では、内容の区切れ目を推定することは難しい。

単語集合の特徴だけでなく、構文構造を考慮したテキスト解析の手法としてCollinsらによりTree Kernelが提案されている[6]。この手法では、テキストに含まれる共通部分木の数により類似性を評価しているが、部分木は膨大な数となるため処理速度の問題があげられている。そこで、市川らはTree Kernelを近似する高速処理可能な手法を提案した[7]。また、工藤らは部分木を素性とするdecision stumps[8]とそれを弱学習器としたboostingアルゴリズムを提案し、製品レビュー文や新聞記事のテキスト分類の実験を行っている[9]。これらの部分木を特徴として利用する手法では、ノードの飛び越えを許さない部分木の完全一致を類似度判定の基準としているため、結果として局所的な部分木しか特徴として利用されないことが多い。また、複数文にまたがる類似性評価は行われていない。

本稿では、ノードの飛び越えを許した部分木を利用して木構造間の類似度を弱学習器として利用し、boostingによる学習を行う。ノードの飛び越えを許すことにより、構文木で遠く離れて位置する文節間の特徴なども考慮した類似性が評価でき、さらには、複数文を対象とした文集合の類似性評価も可能となる。

3. 定型表現抽出手順

本手法では、メタデータとして利用できる被写体を表す単語をキーとしたとき、このキーとなる単語が一つ以上存在する一文以上のテキストを処理対象とする。表1の例では、場所を表す「オンフルール」がキーと

表1　クローズドキャプション例（矩形で囲まれた部分は「場所」を説明する定型的な表現区間）

提示時間	クローズドキャプション
08:29:03	絵は　全然描きませんからって→
08:29:09	まっ　こんなとこですかね。
08:29:12	やっぱり　絵を描かなくてよかったかもしれませんね。
08:29:46	セーヌ川を挟み　ル・アーブルの対岸に位置する港町　オンフルール。
08:29:53	今なお中世の古い家並みが残る　町です。
08:29:59	18歳の時　モネは　パリに出て画家を　目指しますが　美術学校の　入学試験に合格しませんでした。
08:30:11	実家に戻る事を　強要した父親の意向に反して　なおも　パリにとどまって絵の勉強を　し続けた　モネ。

なる単語に該当する。まず、対象テキストに対して人手により定型表現が含まれるか否か判定して、学習データを生成する。学習データから部分木を抽出し、木構造間の類似度を基準とした弱学習器を生成する。次に、AdaBoostアルゴリズムにより、どの弱学習器が正例と負例の分別力があるかを判定しながら学習する。テストデータ中のキーとなる単語の周辺の複数文に対して学習結果を適用することにより、定型表現を含む文章区間か否かを判定する。以下に、部分木抽出、類似度評価、AdaBoostアルゴリズム、そして定型表現部分の抽出手法について記す。

3.1. 部分木抽出

入力テキストを一文ごとに構文解析して、各ノードを文節により構成する構文木を生成する。クローズドキャプション中の文の区切れ目は句点、疑問符、感嘆符などにより判断できる。各文の根ノードの親ノードに最上位ノードを生成し、最上位ノードから各文の構文木へは順序付きのアークで結んだ木構造を生成する。順序付きアークは文の出現順序を考慮した木構造間の類似度評価で利用する。表1の矩形で囲まれた区間の入力テキストを木構造に変換した例を図1に示す。次

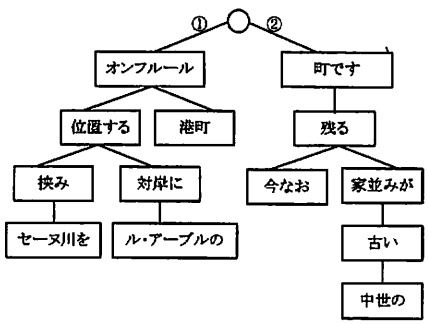

図1. 木構造生成例

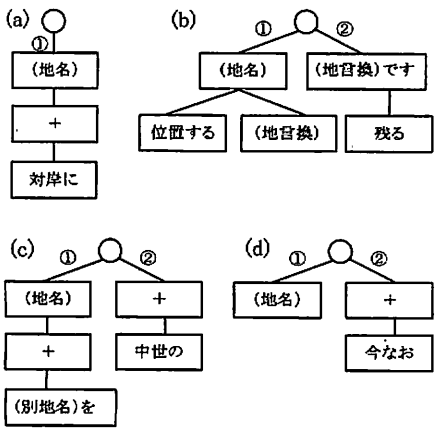

図2. 木構造から抽出された部分木（一部）

に、学習データ中の正例として与えられた木構造からキーとなる単語を含む部分木を生成する。この処理で、キーとなる地名、キーとなる地名以外の地名、地名の言い換え表現は単語表記そのものを利用しないで、"(地名)"、"(他地名)"、"(地言換)"という表記で抽象化して部分木を生成した。また、部分木の作成の際にノードの飛び越えを許し、飛び越えたノードは"+"の記号で置き換え1つ以上のノードとのマッチングを許した。図1に示した木構造から生成される部分木の一部を図2に示す。

部分木生成時にキーとなる地名以外にいくつのノード（文節）を利用するかにより、弱学習器の特徴が決定するが、利用するノード数が多い場合は計算量が膨大になる。もし、100個のノードから50個抽出する場合、その組み合わせ数は 1.0×10^{29} 個を超え計算が困難となる。しかし、本実験で対象とするような定型表現は数個のノードにより表現できると考え、今回はノード数を制限する（本実験では4個）ことにより、生成する部分木の数を計算可能な値とした。

3.2. 類似性評価

抽出した部分木と、学習データに含まれるテキストから生成される木構造との類似度は、部分木に含まれる葉ノードから根ノードまでの全リスト構造を抽出し、その各リスト構造が対象とする木構造に含まれる割合を基準として定義する。部分木 t と木構造 x の類似度 $sim(t,x)$ は以下の式とする。

$$sim(t,x)=\frac{1}{N(t)}\sum_{t_i\in t}\frac{1}{L(t_i)}\sum_{st\in t_i}\max_{sx\in x}(C^d\times sim'(st,sx)) \quad (1)$$

t_i：部分構造 t に含まれる i 番目の文
st：t_i に含まれる葉ノードから根ノードまでのリスト
sx：x に含まれる葉ノードから根ノードまでのリスト
$sim'(st,sx)$：st が sx に含まれる割合。リストに含まれる主辞と付属語を分割して計算。
$N(t)$：t に含まれる文数
$L(t_i)$：t_i に含まれるリスト数
C：キーとなる単語を基準とした文位置の差に与えるペナルティ値（本実験では0.5）
d：キーとなる単語を地名のある文を基準とした文位置の差

図2(b)に示す部分木 t との類似度を求める例を図3に示す。この例では、葉ノードから根ノードまでのリストが、部分木 t から3つ、木構造 x から4つ取り出されている。最も類似しているリスト構造間の類似度 $sim'(st,sx)$ をそれぞれ求めることにより、部分木 t

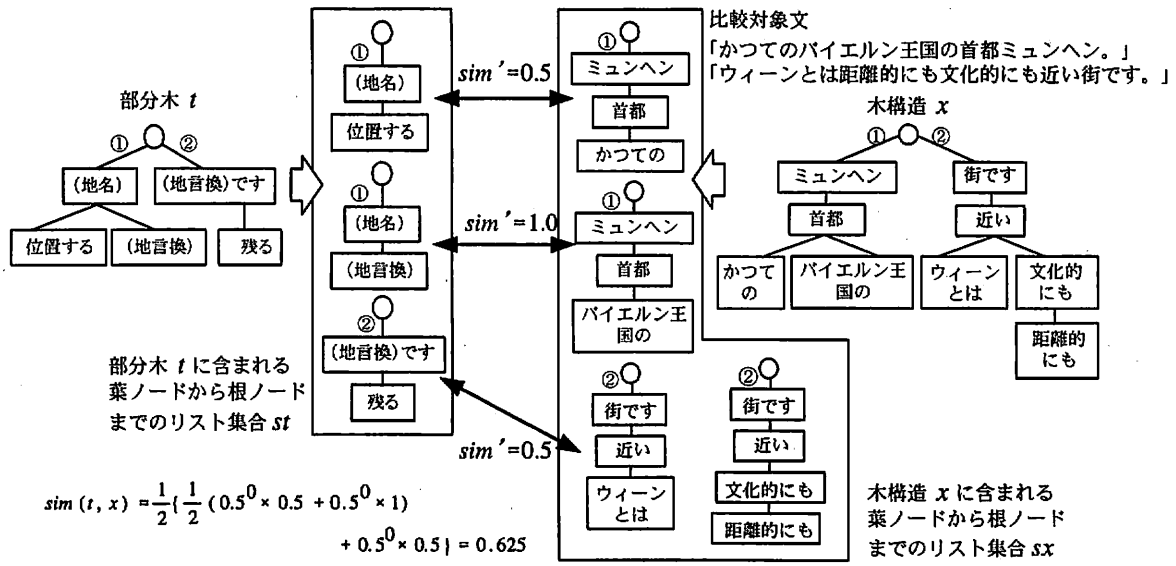

図 3. 部分木と比較対象文との類似度計算例

と木構造 x の類似度 $sim(t,x)=0.625$ と算出することができる。

部分木 t と学習データに含まれるテキストとの類似度の分布を 0～1 の間にマッピングする。ある閾値より大きいものを正例、小さいものを負例、もしくはその逆とした関数を作成し、エラーが最小となる点を θ_t とする。出力のクラスラベルを $y \in \{\pm 1\}$ としたとき、部分木 t と閾値 θ_t に対する弱学習器 $h_t(x)$ は以下のように定義できる。

$$h_t(x) = \begin{cases} y & sim(t,x) \geq \theta_t \\ -y & sim(t,x) < \theta_t \end{cases} \quad (2)$$

3. 3. AdaBoost による学習

学習データに含まれるテキストから抽出した部分木によって大量の弱学習器が生成される。この弱学習器を AdaBoost の機械学習に利用する。本手法では図 4 に示すアルゴリズムによる学習を行う。

まず、Step1 において学習データ全てに対する重み D を均等に与える。最初のループでは Step2 において、最も誤り率 ε が少ない弱学習器が選択される。Step3 では、選択された弱学習器で誤って判定されたデータに対する重み $D(i)$ に大きな値が与えられ、次の繰り返し処理では $D(i)$ を考慮した誤り率 ε を考慮することにより、誤ったデータを正確に分類するような弱学習器が選ばれる。この際、$D_{t+1}(i)$ は $\varepsilon_t=0.5$ となるような値に更新されている。Step2 と Step3 を繰り返すことにより全ての弱学習器に対しての重み α が計算され、Step5 では、それらの和により判定を行うことにより、精度の高い分類器を構築することができる。

3. 4. 定型表現部分の抽出

学習の結果得られる最終仮説を利用して、学習データとは異なるテストデータから、定型表現を含む文章区間の抽出を行う。まず、テストデータからキーとなる単語を抽出し、その単語が出現する前後数文を処理対象として、最終仮説 $H(x)$ を計算する。$H(x)=1$ の時、対象区間は定型表現部分であると判断できる。しかし、負例には特徴が少ないため、定型表現を含まない文章区間は、定型表現を含むと誤判定される可能性がある。そこで、最終仮説 $H(x)=1$ と判定された事例に対して、

学習データ:

$(x_1, y_1), \cdots, (x_N, y_N)$, $y_i = \{1, -1\}$

Step1: $D_1(i) = 1/N$ に初期化

Step2: $\varepsilon_t = \sum_i D_t(i) \mid h_t(x_i) - y_i \mid$ が最小となる部分木（弱学習器）を選択

Step3: 抽出された弱学習器により学習データに対する重み D_t を更新

$$\alpha_t = \frac{1}{2}\log\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$$

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & if\ y_i = h_t(x_i, \theta) \\ e^{\alpha_t} & otherwise \end{cases}$$

Z_t は正規化定数

Step4: Step2 と Step3 の処理を部分木が無くなるまで繰り返す

Step5: 最終仮説 $H(x) = sign(\sum_t \alpha_t h_t(x))$

図 4. AdaBoost による学習アルゴリズム

再度、AdaBoost による学習を行い判定する。この際、学習で利用しなかった負例に対して誤って定型表現を含むと判定されたものから負例データを選択し、正例はそのままとした学習による最終仮説を利用する。この処理を数段繰り返すことにより、精度向上が期待できる。

また、ある文章区間で $H(x)=1$ となる場合は、その前後の文を含めた区間でも同様に $H(x)=1$ と判定される。この場合は、$H(x)$に含まれる関数の値 $\sum_t \alpha_t h_t(x)$により定型表現部分の区間を判定し、文を追加した時にこの値が増加するときのみ、その文を定型表現部分に追加する。この処理により、キーとなる単語と定型表現を含む文章区間が抽出される。

4. 場所を映像とともに説明する定型表現文章区間抽出実験

提案手法を検証するため、NHK で放送された紀行番組「わが心の旅」のクローズドキャプションを対象として、「場所」に関する情報を映像とともに説明している定型的な表現部分を抽出する実験を行った。形態素解析辞書に「地名」として登録されている単語を「キーとなる単語」、その単語が場所を映像とともに説明している場合を正例、場所を映像とともに説明していない場合を負例として 15 番組に対して人手により正解データを付与した。このうち 10 番組を学習データ、5 番組をテストデータとした。学習データに含まれる負例の数は正例に比べて多いため、正例と同数の 29 個を無作為に選択した。負例における区間は、正例と同じ平均文数となるように調整し、利用するノード数を 4 個として学習を行った。この結果、14929 個の弱学習木が生成された。

形態素解析辞書に「地名」として登録されている単語を「キーとなる単語」としてテストデータから抽出し、その前 2 文、後 7 文から、単語のある文を含む任意の連続文を処理対象文章とした。この処理対象文章が定型表現を含むか否かを最終仮説により判定した。抽出結果の一部を図 5 に示す。図中の矩形で囲まれた部分が提案手法により抽出された定型表現を含む区間、下線部の単語が「キーとなる単語」である。

[抽出例 1]

ガウディは　どのようにして建築と出会い　造形の世界を　究めていったのでしょうか？

バルセロナの南西　地中海に面して広がるタラゴナ平原。　← 正解区間
オリーブや　ブドウの畑が続くのどかな田園地帯です。　← 正解区間

１８５２年　ガウディはリウドムスという村で　鍋や釜を　作る　職人の子として生まれました。

[抽出例 2]

日本の場合ですと　川のように歴史が流れていく訳です。
こちらは　古いところから新しいところへ　過去が忘れられるという事がないのです。
２，０００年の歴史が積み重なった城塞都市。

カルカソンヌは　都市全体が石造りの歴史書そのものです。　← 正解区間
城壁の中では　中世からの町が今も　そのままに生きています。
町の中には　美しいステンドグラスで有名な教会があります。　← 正解区間

これまで　中世は暗黒時代と思われていました。

図 5. 場所を説明する定型表現区間抽出例

4.1. 実験結果の評価

キーとなる単語が、判定結果と正解データでともに「場所を説明する文章区間」、または「場所を説明しない文章区間」に出現しているときを正解として結果の評価を行った。テストデータとした 5 番組には形態素解析辞書に「地名」として登録されている名詞が 230 個含まれ、そのうちの 16 個が実際に映像とともに場所を説明していた。評価結果を表 2 に示す。

表 2　提案手法による判定評価結果

キー単語 (学習の繰り返し数)	適合率	再現率	F 値
場所説明 (1 回)	16/98 (16.3%)	16/16 (100%)	0.281
場所説明でない (1 回)	132/132 (100%)	132/214 (61.7%)	0.763
場所説明 (2 回)	15/54 (27.8%)	15/16 (93.8%)	0.429
場所説明でない (2 回)	175/176 (99.4%)	175/214 (81.8%)	0.895
場所説明 (3 回)	13/32 (40.6%)	13/16 (81.3%)	0.542
場所説明でない (3 回)	195/198 (98.4%)	195/214 (91.1%)	0.947

「キー単語が場所を説明・繰り返し数 1 回」における結果では、適合率が 16.3%と低い。しかし学習を繰り返すことにより適合率が向上し、適合率と再現率の調和平均である F 値も向上している。場所を説明しないキーとなる単語はテストデータ中に 214 個出現しており、この判定結果の精度は繰り返し 3 回で F 値 0.947 と良好な結果が得られている。

次に、繰り返し処理を 3 回行った後に場所を説明する文章区間と判定された 13 箇所に対して、どの程度人手により付与した正解区間と一致しているか評価を行った。この 13 箇所は、人手により合計 27 文が正解区間として抽出されていた。この 27 文の正解区間のうち、提案手法により抽出された文の割合を示す再現率は 81.5%(22 文/27 文)、提案手法により抽出した区間に含まれる文中で正解データの区間に含まれている割合を示す適合率は 68.8%(22 文/32 文)であった。提案手法では、対象とする文数が増えても部分木からの類似度は減少しない。そのため、余分に文を抽出する傾向が見られた。対象文数の増加に対するペナルティを類似度に与えることにより改善可能と考えられる。

4.2. ノードの飛び越えと木構造間の類似度を利用する効果

ノードの飛び越えを許した木構造を生成し、かつ、木構造間の類似度を利用した弱学習器を利用する効果を検証するために、工藤らの手法[8]を利用して定型表現区間抽出実験を行った。工藤らの手法では木構造の分類問題を扱うため、定型表現を含む文章区間の木構造と含まない木構造の分類に適用できる。この手法では、ノードの飛び越えを許さず、木構造間の類似度を用いず、木構造が一致するか否かのみを弱学習器として利用している。木構造 x, t、出力クラスラベルを $y \in \{\pm 1\}$ としたとき、分類を行うための decision stumps は以下のように定義される。

$$h_t(x) = \begin{cases} y & t \subseteq x \\ -y & otherwise. \end{cases} \quad (3)$$

(3)式は、木構造 t が木構造 x の部分構造となっている場合($t \subseteq x$)に出力 y を返す関数であり、提案手法における(2)式に対応する。この式を弱学習器とした boosting による学習を行うことにより、入力となる木構造 x が定型表現区間か否かを判定できる。提案手法と同様に、負例に対して誤って定型表現を含むと判定されたものから再度負例データを選択し、正例はそのままとした学習を繰り返した。キーとなる単語が場所を説明する場合の判定評価結果を表3に示す。

表3 既存手法による抽出評価結果

繰り返し回数	適合率	再現率	F値
1回	16/135 (11.9%)	16/16 (100%)	0.212
2回	13/52 (25.0%)	13/16 (81.3%)	0.382
3回	13/37 (35.1%)	13/16 (81.3%)	0.491

表3の結果は、いずれの繰り返し回数でも表2のキー単語が場所説明である場合の結果を下回っており、ノードの飛び越えを許し、木構造間の類似度を考慮した提案手法の有効性が確認できた。

4.3. 考察

提案手法による実験結果では、AdaBoost による学習を3回繰り返しても適合率は40.6%であり、依然、多くの誤抽出が残されている。誤抽出の例を図6に示す。例1では、「地中海」というキーとなる単語に対する説明区間として2文が誤抽出されている。実際には、この区間は「タラゴナ平原」に対する説明区間である。提案手法では、キーとなる単語による定型表現区間から生成する木構造と、同じ区間にある別のキーとなる単語に対する木構造が類似するため、弱学習器の類似度も同様の傾向が見られて誤判定されてしまう。同一区間において複数のキーとなる単語が出現する場合は、最終仮説の値による比較などにより絞込み処理を行う必要性が考えられる。また誤抽出例2の区間では、人によっても「カタルーニャ音楽堂」が映像とともに説明されているか否かの判断は難しい。このような部分は、機械による解析も困難と考えられる。

[誤抽出例 1]
バルセロナの南西　地中海に面して広がるタラゴナ平原
オリーブや　ブドウの畑が続くのどかな田園地帯です
[誤抽出例 2]
カタルーニャ音楽堂は ガウディのライバルといわれたモンタネールの 代表作です
華やかな装飾によって カタルーニャの繁栄を、表現しました

図6. 提案手法による誤抽出例

5. まとめ

本稿では、クローズドキャプションから定型表現を含む文章区間を抽出する手法を提案した。ノードの飛び越えを許した木構造間の類似度を取り入れることにより、遠く離れた位置にある単語間の関係も考慮した処理を実現した。場所を映像とともに説明する定型表現を含む文章区間を抽出する実験により、既存手法より良好な結果が得られ一定の分別能力があることを示した。

今回の実験では「場所」に関する情報を映像とともに説明している定型的な表現部分を抽出対象としたが、今後、他の定型表現に対しての実験を行う予定である。

文　献

[1] NHK アーカイブス (http://www.nhk.or.jp/nhk-archives/)

[2] 三浦, 山田, 住吉, 八木: クローズドキャプションを利用した映像主被写体の推定手法, 情処学会研究報告 NL171-1, Vol.2006, No.1, pp1-6(2006)

[3] Freund, Y. and Schapire, R.E.: A decision theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences, Vol.55, No.1, pp.119-139(1996)

[4] Hearst, M.A.: Multi-paragraph segmentation of expository text. In ACL'94 Proceedings, pp8-16(1994)

[5] 望月, 本田, 奥村: 複数の知識の組み合わせを用いたテキストセグメンテーション, 情処学会研究報,NL109-7, pp47-54(1994)

[6] Collins, M. and Duffy, N. Convolution Kernels for Natural Language. In Proceedings of NIPS2001(2001)

[7] 市川, 橋本, 徳永, 田中: テキスト構文構造類似度を用いた類似文検索手法, 情処学会研究報 FI-079, Vol.2005, No.42

[8] Schapire, R.E. and Singer, Y.:BoosTexter: A boosting-based system for text categorization. Machine Learning, 39(2/3), pp135-168(2000)

[9] 工藤, 松本: 半構造化テキストの分類のためのブースティングアルゴリズム, 情処論文誌, Vol.45, No.9, pp2146-2156(2004)