論　文

大域的な文章構造の類似性を利用したクローズドキャプション中の定型的な文章区間の抽出

山田　一郎[†]　三浦　菊佳[†]　河合　吉彦[†]　住吉　英樹[†]
八木　伸行[†]　奥村　学[††]　徳永　健伸[†††]

Extraction of Text Sections which Contain Typical Expressions in Closed Captions Using Similarity of Global Sentences Structures

Ichiro YAMADA[†], Kikuka MIURA[†], Yoshihiko KAWAI[†], Hideki SUMIYOSHI[†], Nobuyuki YAGI[†], Manabu OKUMURA[††], and Takenobu TOKUNAGA[†††]

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に対して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．そこで本論文では，番組のナレーションとほぼ同じ内容のテキストデータであるクローズドキャプションから定型的な文章区間を統計的に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，木構造から飛び越えを許した大域的な範囲をカバーする部分木を抽出する．この部分木を比較対象として，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 アルゴリズムにより学習を行い定型的な文章区間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的な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，定型文章区間抽出，クローズドキャプション，木構造，アダブースト

1. まえがき

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いえない．そこで，放送された番組を映像百科事典 [2] 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 [3] 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，番組のナレーションとほぼ同じ内容のテキストデータである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 [4], [5]．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，統語構造を手掛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区間は，映像解析処理による映像ショットの切換点を利用しているため，複数の映像ショットにまたがるような区間は抽出できない．

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では，方形で囲まれた部分が「場所」を映像とともに説明している．第 1 文では，体言止めされた単語「オンフルール」に対し，その位置情報を連体修飾節により説明して，次の文で「オンフルール」の詳細を説明し，「オンフルール」の言い換え表現で

[†] NHK 放送技術研究所，東京都
Science & Technical Research Laboratories, NHK, 1–10–11 Kinuta, Setagaya-ku, Tokyo, 157–8510 Japan
[††] 東京工業大学精密工学研究所，横浜市
Precision & Intelligence Laboratory, Tokyo Institute of Technology, 4259 Nagatsuda, Midori-ku, Yokohama-shi, 226–8503 Japan
[†††] 東京工業大学大学院情報理工科学研究科，東京都
Department of Computer Science, Tokyo Institute of Technology, 2–12–1 Ookayama, Meguro-ku, Tokyo, 152–8552 Japan

表 1　クローズドキャプション例（方形で囲まれた部分は「場所」を説明する定型的な文章区間）
Table 1　Examples of closed captions in which a section including typical expressions is boxed off.

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。

ある「町」と断定の助動詞「です」を利用した定型的な場所紹介の表現である．このように場所を説明している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．このような特定の事項を表現する文章区間を抽出する一手法として，特定表現のテンプレートを作成し，その一致度を指標とした解析的なアプローチが考えられる．しかし，あらゆる表現に対するテンプレートの生成には限界があるため網羅性に欠け，更に抽出対象ごとに手作業によりテンプレートを作成しなければならず多大な労力を要する．

本論文では，番組のクローズドキャプションを対象として，特定の事柄を表現する際の定型的な文章区間を統計的に抽出する手法を提案する．クローズドキャプションから区間の抽出を行うため，複数の映像ショットにまたがるような区間も抽出可能となる．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，木構造から飛び越えを許した大域的な範囲をカバーした部分木を抽出する．この部分木を比較対象として，木構造間の類似性を評価する．類似性評価結果をベースとした判定関数を弱学習器とした AdaBoost アルゴリズム [6] により学習を行い定型的な文章区間か否かの判定を行う．実験では，特定の事柄として場所を映像とともに説明する定型的な文章区間を対象とした．場所を映像とともに説明するシーンは，放送される番組には頻出しており，また，抽出される区間は映像百科事典としての有用性も高いと考えられる．

以下，**2.** で関連研究についてまとめ，**3.** では定型的な文章区間抽出処理の詳細を説明する．**4.** 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的な文章区間を抽出する実験と評価を行う．最後に **5.** で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従来研究として，ベクトルスペースモデル [7] を利用する手法が挙げられる．この手法では，特定の事項を表現する文章区間に出現する単語から特徴ベクトルを生成し，このベクトルとの類似度が高い文章区間を，特定の事項を表現する文章区間として抽出する．また，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も考えられる．Hearst は，テキストに含まれる単語の出現頻度をベースとして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 [8]．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掛りに基づき内容の区切れ目を推定する手法を提案した [9]．しかし，本論文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見られ，単語の集合のみを特徴としたこれらの手法では，類似する文章区間の抽出や，内容の区切れ目推定は難しい．

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 [10]．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挙げられている．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 [11]．また，工藤らは部分木を素性とする decision stumps [12] 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 [13]．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．

本論文では，ノードの飛び越えを許した部分木を生成し，木構造間の類似度を弱学習器として利用した boosting による学習を行う手法を提案する．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，更には，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. 定型的な文章区間抽出手順

本手法では，キーとなる単語を一つ選択し（例えば映像の被写体を表す単語など），この単語が一つ以上存在する 1 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弱学習器の信頼度を示す重みを学習する．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，定型的な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 1 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 の方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と任意の数のノードを含む部分木を生成する．この処理で，キーとなる地名，キーとなる単語以外の地名，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)”，“(別地名)”，“(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 “+” の記号で置き換え一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．図 2 (a) は「対岸に〜オンフルール。」，(b) は「〜位置する港町オンフルール。〜残る町です。」の部分を図 1 の木構造から抽出した部分木であり，「オンフルール」は “(地名)” に，「港町」「町」は “(地言換)” に抽象化されている．

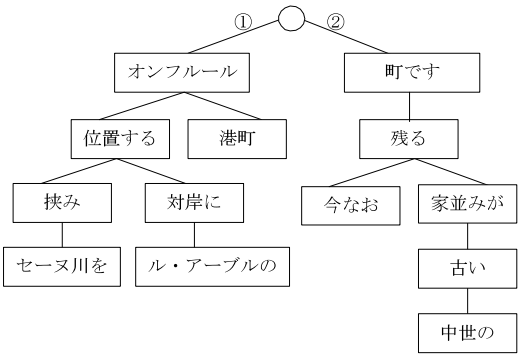

図 1 木構造生成例
Fig. 1 Example of tree structure generation.

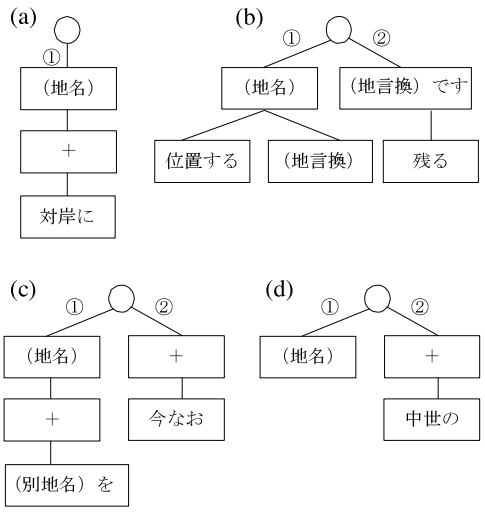

図 2 木構造から抽出された部分木（一部）
Fig. 2 Part of sub-tree structure extracted the tree structure of Fig. 1.

クローズドキャプション中では，キーとなる単語が出現した以降の文で「この町は」など場所を映像とともに説明する特徴的な表現が現れることがある．キーとなる単語から離れて位置するような単語も定型的な

文章区間抽出には重要な役割を果たすと考えられる．部分木生成時に飛び越えを許さないような従来手法では，キーとなる単語と，キーとなる単語から離れて位置するような単語との関係のみを利用することが難しい．例えば図 1 のキーとなる単語「オンフルール」と「中世の」という単語の関係を利用する場合，「オンフルール」，「町です」，「残る」，「家並みが」，「古い」，「中世の」というノードからなる部分木を生成しなければならなかった．この場合，次節で定義する類似度計算において「中世の」という単語以外の単語の「町です」，「残る」，「家並みが」，「古い」にも影響を受けるため，キーとなる単語「オンフルール」と「中世の」という単語の関係のみを考慮できない．また，工藤らの手法 [13] では，類似度を利用しないで部分木の完全一致による decision stumps を利用するため，多くのノード数をもつ部分木は，学習データに大量に出現しない限り boosting アルゴリズムで重みは小さな値が与えられ，結果的に考慮されなくなる．部分木生成時に飛び越えを許すことにより，図 2 (d) のように「オンフルール」，「＋」，「中世の」というノードのみからなる部分木を生成でき，キーとなる単語「オンフルール」と「中世の」という単語の関係のみを考慮することが可能となる．提案手法では，対象文章に含まれるこのような関係をすべて考慮できるため，文章の大域的な範囲を考慮した類似性評価手法となる．

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．

$$sim(t,x) = \frac{1}{NS(t)} \sum_{t_i \in t} \frac{1}{L(t_i)} \sum_{st \in t_i} \max_{sx \in x} C_1^d \times sim'(st,sx) \quad (1)$$

$$sim'(st,sx) = \frac{com_{main}(st,sx) + C_2 \times com_{att}(st,sx)}{NW_{main}(st) + C_2 \times NW_{att}(st)}$$

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．主辞と付属語は分割．

sx：x に含まれる葉ノードから根ノードまでのリスト．主辞と付属語は分割．

$NS(t)$：t に含まれる文数

$L(t_i)$：t_i に含まれるリスト数

C_1：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

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

$com_{main}(st,sx)$：st と sx の共通する主辞のノード数（出現順序考慮）

$com_{att}(st,sx)$：st と sx の共通する付属語ノード数（出現順序考慮）

$NW_{main}(st)$：st に含まれる主辞のノード数

$NW_{att}(st)$：st に含まれる付属語ノード数

C_2：付属語に与える重み

部分木 t と木構造 x では，一つの文節が一つのノードとして扱われているが，ここから取り出すリスト st と sx では，文節中の主辞は同文節の付属語を修飾するとして主辞と付属語を別々の項として扱う．付属語は，文節の最終形態素を含み連続する助詞，助動詞を抜き出し利用する．例えば，「ウィーンとは」という文節では，付属語として格助詞「と」と係助詞「は」が抜き出され，「ウィーン」と「とは」の二つがリストの項となる．$sim'(st,sx)$ はリスト st とリスト sx の類似度を示し，リスト中での出現順序を考慮した共通単語数を基準としている．

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 から三つ，木構造 x から四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 と木構造 x の類似度 $sim(t,x) = 0.675$ と算出することができる．このとき，文位置の差に与えるペナルティ値 $C_1 = 0.5$，付属語に与える重み $C_2 = 0.5$ としている．

部分木 t と学習データ（正例，負例）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．あるしきい値 θ に対して θ より大きい値を正例，小さい値を負例とする 2 値判別関数を考え，マッピングされた学習データ（正例，負例）との整合性を判定する．θ より小さい値で正例がマッピングされている場合，または θ より大きい値で負例がマッピングされている場合がエラーとなる．また，θ より大きい値を負例，小さい値を正例とした 2 値判別関数も同時に考え，この場合のエラーもカウントする．この二つの

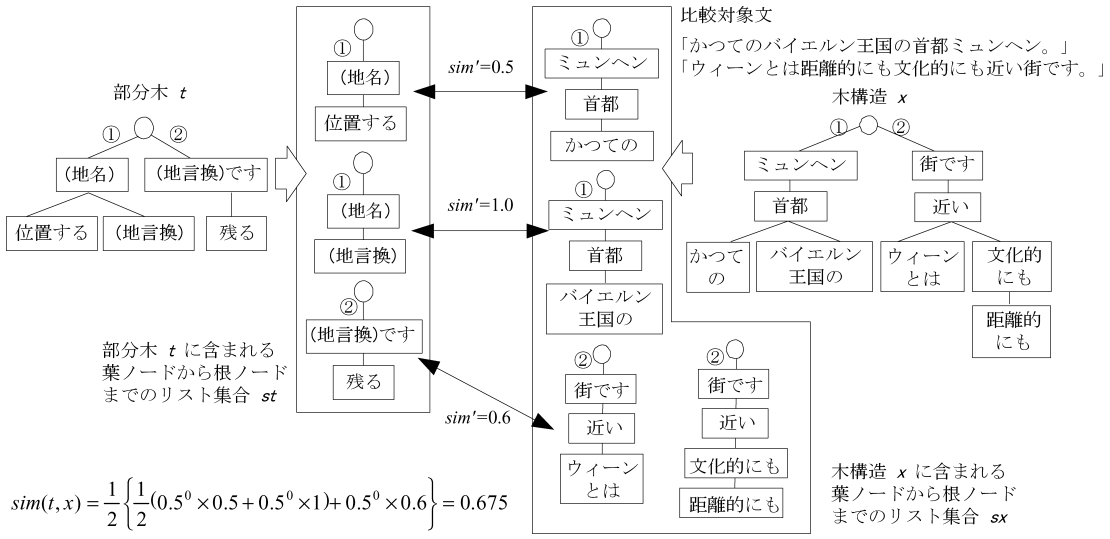

図 3　部分木と比較対象文との類似度計算例
Fig. 3　Examples of similarity calcuration between a sub-tree and targeted sentences.

2 値判別関数のしきい値を 0〜1 間で動かし，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 としきい値 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．

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

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に対する重み $D_1(i)$ を均等に与える．Step2 では，最初のループで最も誤りが少ない弱学習器が選択され，以降，学習データに対する重みを考慮した誤り率 ϵ が最も少ない弱学習器が選択される．Step3 では，t 回目のループにおいて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 $D_t(i)$ に大きな値が与えられ，次の繰返し処理では $D_t(i)$ 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 $\epsilon_t = 0.5$ となるような値に更新される．Step2 と Step3 を繰り返すことによりすべての弱学習器に対しての重み α が計算され，Step5 では，弱学習器に対する重みを考慮した多数決による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．

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:　$\varepsilon_t=\sum_i D_t(i)\,|\,h_t(x_i)-y_i\,|$ 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　$D_t(i)$，　$i=1\sim N$　を更新

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

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

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
Fig. 4　AdaBoost learning algorithm.

3.4　定型的な文章区間の抽出

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，定型的な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．$H(x)=1$ のとき，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，定型的でない文章区間は，

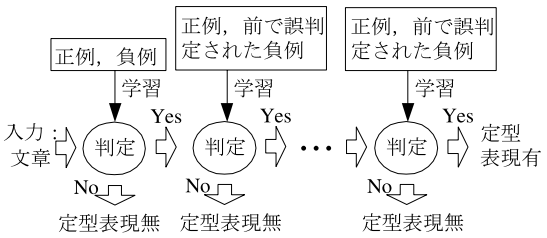

図 5　カスケードによる判定処理の繰返し
Fig. 5　Detection cascade.

定型的であると誤判定される可能性がある．Viola らは顔画像検出処理において，判定処理を何段もカスケードすることにより適合率を向上させる手法を提案している [14]．そこで本手法でも，図 5 に示すように最終仮説 $H(x) = 1$ と判定された事例に対して，再度，AdaBoost による学習を行い判定する．この際，前の学習で利用しなかった負例に対して誤って定型的な文章区間と判定されたものから，次の学習で利用する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．判定処理をカスケードして複数回行うことにより，適合率向上が期待できる．

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により定型表現部分の区間を判定し，文を追加したときにこの値 $\sum_t \alpha_t h_t(x)$ 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的な文章区間が抽出される．判定処理を何段もカスケードすることにより各処理で文章区間が抽出され，最終的にすべての処理において正と判定される区間を定型的な文章区間とする．

4. 場所を映像とともに説明する定型的な文章区間抽出実験

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 60 番組に対して人手により正解データを付与した．このデータを無作為に二つに分割し，片方を学習データ，残りをテストデータとし

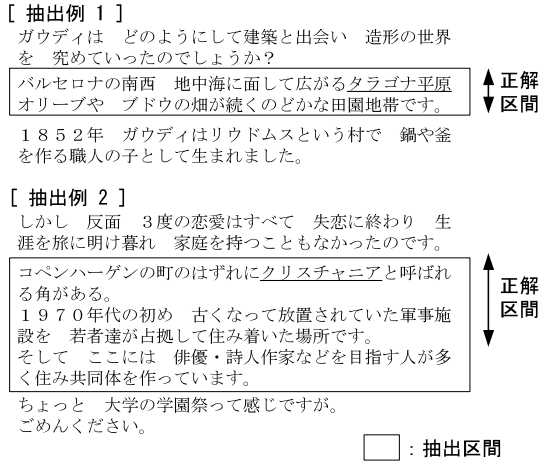

図 6　場所を説明する定型表現区間抽出例
Fig. 6　Example of experimental result of detecting sections including typical expressions which explain a place.

たクロスバリデーションによる 2 回の実験を行った．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た．部分木生成時に選択するノード数が多い場合は計算量が膨大になるため，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，今回は利用するノード数を 4 個として学習を行った．この結果，2 回の実験ではそれぞれ 56899 個，56092 個の弱学習木が生成された．また，カスケードによる判定処理の繰返し処理では，処理対象データに対して，学習データとして利用できる負例データが確保できた 4 回行った．

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」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的な文章区間か否かを最終仮説により判定した．抽出結果の一部を図 6 に示す．図中の方形で囲まれた部分が提案手法により抽出された定型的な文章区間，下線部の単語が「キーとなる単語」である．

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合計

1972 個含まれ，そのうちの 196 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．

「キー単語が場所を説明・カスケード数 1 回」における結果では，適合率が 21.8%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 値もカスケード数 4 回で 0.557 まで向上している．場所を説明しないキーとなる単語はテストデータ中に 1776 個出現しており，この判定結果の精度はカスケード数 4 回で F 値 0.916 と良好な結果が得られた．

次に，カスケードによる判定を 4 回行った後に場所を説明する文章区間と判定された 175 箇所に対して，人手により付与した正解区間とどの程度一致しているか評価を行った．結果を表 3 に示す．

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 82.1%，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 50.2%であった．提案手法では，4 回の判定を行いすべての処理で正と判定された区間を抽出している．各判定において学習データが異なるため抽出される区間にも差が生じ，すべての判定で正と判定される区間は短くなる傾向が見られた．この影響で再現率が多少低い値となった．

表 2　提案手法による判定評価結果
Table 2　Evaluation results of judgements using proposed method.

キー単語（カスケード数）	適合率	再現率	F 値
場所を説明する区間（1 回）	192/882 (21.8%)	192/196 (98.0%)	0.356
場所を説明しない区間（1 回）	1086/1090 (99.1%)	1086/1776 (61.1%)	0.758
場所を説明する区間（2 回）	183/581 (31.5%)	183/196 (93.3%)	0.471
場所を説明しない区間（2 回）	1378/1391 (99.1%)	1378/1776 (77.6%)	0.870
場所を説明する区間（3 回）	181/516 (35.1%)	181/196 (92.3%)	0.508
場所を説明しない区間（3 回）	1441/1456 (99.0%)	1441/1776 (81.1%)	0.892
場所を説明する区間（4 回）	175/432 (40.5%)	175/196 (89.2%)	0.557
場所を説明しない区間（4 回）	1519/1540 (98.6%)	1519/1776 (85.5%)	0.916

表 3　文章区間の抽出精度
Table 3　Extraction accuracy of text sections.

適合率	再現率
230 文/280 文 (82.1%)	230 文/458 文 (50.2%)

4.2　既存手法との比較

提案手法の有効性を検証するため，既存手法を応用して定型表現区間を抽出する二つの実験を行った．以下に抽出手法と実験結果の詳細を記す．

4.2.1　ベクトルスペースモデルを利用した定型表現区間抽出

学習データの正例区間に含まれる単語と負例区間に含まれる単語をベクトルの要素とした正例ベクトルと負例ベクトルを生成する．ベクトルの要素となる単語は，自立語（名詞，動詞，形容詞，副詞など）に限定し，ベクトルの要素の値は，単語の出現頻度 TF と単語の逆文書頻度 IDF の積である TFIDF 値 [15] とする．例えば正例ベクトルを生成する場合，TF は正例区間に出現する自立語の出現頻度，IDF は対象自立語のすべてのクローズドキャプション中における逆文書頻度である．

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する．処理対象文章に対しても，出現単語をベクトルの要素，単語の TFIDF 値をベクトルの要素の値とした対象文章ベクトルを生成する．この対象文章ベクトルと正例ベクトル，負例ベクトルとのコサイン距離を求める．負例ベクトルより正例ベクトルとのコサイン距離が小さい場合に，対象文章を正と判定する．一つの「キーとなる単語」に対して抽出した複数の処理対象文章のうち一つでも正と判定された場合，「キーとなる単語」の周辺に場所を説明する文章区間があると判定する．提案手法で利用した正例と，4 回の学習で利用したすべての負例により正例ベクトル，負例ベクトルを生成して行ったクロスバリデーションによる判定結果を表 4 に示す．

4.2.2　ノードの飛び越えと木構造間の類似度を利用しない手法による定型表現区間抽出

提案手法と同様に，正例の文章区間に含まれる文章

表 4　ベクトルスペースモデルを利用した定型表現区間抽出実験の評価結果
Table 4　Evaluation results of judgements using a vector space method.

キー単語	適合率	再現率	F 値
場所を説明する区間	183/816 (22.4%)	183/196 (93.4%)	0.362
場所を説明しない区間	1143/1156 (98.9%)	1143/1776 (64.4%)	0.780

から，各文に対する構文木を統合した木構造を生成し，この木構造から弱学習器として利用するための部分木を抽出する．このとき，部分木におけるノードの飛び越えは許さない．また，弱学習器生成では，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する．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．

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

式 (3) 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，提案手法における式 (2) 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．この手法は，工藤らの手法 [13] を複数文の文章区間に適用したものと等価である．提案手法と同様に，負例に対して誤って定型的な文章区間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．前節と同じデータを対象としたクロスバリデーションによる判定結果を表 5 に示す．

この実験では，負例データが提案手法で利用したものと異なる．偶然，提案手法で使用した負例データが分別に適している可能性もあるため，単純に比較するには適切でない．しかし，提案手法と同じ負例データを使用すると，この手法で既に負と判定されているデータも選択されることがあり，不利な条件となる．実際に，提案手法と同じ負例データを使用して 4 回のカスケードの実験をした結果，場所を説明する区間では適合率 24.2%，再現率 93.4%，F 値 0.384 と，表 5 の結果を下回る値であった．そのため，ここでは同じ学習データによる実験でなく，データ作成手法を同じとした実験を比較対象としている．

表 5　ノードの飛び越えと木構造間の類似度を利用しない手法による抽出実験の評価結果
Table 5　Evaluation results of judgements using a modified conventional method which dose not use inconsecutive nodes in a tree and similality between trees.

キー単語（カスケード数）	適合率	再現率	F 値
場所を説明する区間（1 回）	193/1240 (15.6%)	193/196 (98.5%)	0.269
場所を説明しない区間（1 回）	729/732 (99.6%)	729/1776 (41.0%)	0.581
場所を説明する区間（2 回）	186/629 (29.6%)	186/948 (94.8%)	0.451
場所を説明しない区間（2 回）	1333/1343 (99.3%)	1333/1776 (75.1%)	0.855
場所を説明する区間（3 回）	183/589 (31.1%)	183/196 (93.4%)	0.466
場所を説明しない区間（3 回）	1370/1383 (99.1%)	1370/1776 (77.1%)	0.867
場所を説明する区間（4 回）	183/569 (32.2%)	183/196 (93.4%)	0.478
場所を説明しない区間（4 回）	1390/1403 (99.1%)	1390/1776 (78.3%)	0.874

4.3 考　　察

提案手法による結果（表 2）とベクトルスペースモデルを利用した手法による結果（表 4）を比較すると，カスケード数が 2 回以上において提案手法の方が良好な F 値が得られている．ベクトルスペースモデルは提案手法の 4 回の学習で利用したすべての負例データと正例データを利用しているため，提案手法の方が有効であると判断できる．

また提案手法による結果は，いずれのカスケード数でも表 5 に示すノードの飛び越えと木構造間の類似度を利用しない手法による結果の F 値を上回っている．ノードの飛び越えを許した部分木生成と，木構造間の類似度を考慮した弱学習器を利用する提案手法の有効性が確認できた．

提案手法による実験結果では，カスケード数を 4 回としても適合率は 40.5%であり，依然，多くの誤抽出が残されている．場所を映像とともに説明していると誤抽出されてしまった例を図 7 に示す．誤抽出例 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

[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]
民話を育んだのは　豊かな森.
中世まで　ヨーロッパは　豊かな森に覆われていたのです.
[誤抽出例 3]
１４世紀　コルドは　実際に　革製品の取り引きで　賑わったと言います.
今　コルドは工芸の伝統を受け継いで　芸術の町として生きようとしています.

図 7　提案手法による誤抽出例
Fig. 7　Example of extraction error.

表 6　最終仮説の評価式の値を利用した評価結果
Table 6　Evaluation results of judgements using a value of final hypothesis.

キー単語	適合率	再現率	F 値
場所を説明する区間	165/353 (46.7%)	165/196 (84.1%)	0.601
場所を説明しない区間	1588/1619 (98.1%)	1588/1776 (89.4%)	0.935

同様の傾向が見られて誤判定されてしまう．そこで，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，最終仮説の値が大きいものに絞込みを行う処理を行った．カスケード数 4 回における評価結果を表 6 に示す．場所を説明する区間の適合率が 6.2%，F 値でも 0.044 向上している．

誤抽出例 2 の区間では，過去についての言及区間である．提案手法では，この文中の過去を表す文節「覆われていたのです」に対して，付属語として最終形態素を含み連続する助詞と助動詞しか取り出されないため「です」しか考慮されず，過去を示す助動詞「た」が考慮されない．文章区間から木構造を生成する際に，過去形なども考慮するよう改善が必要と考えられる．

また誤抽出例 3 では，人によっても「コルド」が映像とともに説明されているか否かの判断は難しい．正解データを付与した者とは異なる被験者により，クローズドキャプションのみから映像とともに場所を説明している区間か否かを判定する実験を行った．その結果，提案手法により場所を説明する区間と誤抽出された 188 区間のうち，44 区間（23.4%）で人間でも同様に場所を映像とともに説明していない区間と判定できなかった．このような部分は，機械による解析も困難と考えられる．

5. むすび

本論文では，クローズドキャプションから定型的な文章区間を抽出する手法を提案した．複数の文からなる文章の特徴を的確にとらえるため，文章から，構文解析結果により単語をノードにもつ木構造を生成した後，ノードの飛び越えを許した木構造間をその特徴候補として抽出し，木構造間の類似度を評価することにより，木構造で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的な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，弱学習器を生成する基となる部分木のノード数を処理時間の問題から 4 個以下に制限した．逐次モンテカルロ法を用いた GibbsBoost [16] などのアルゴリズムを応用することにより，複数文から生成した木構造から部分木を効果的に選択でき，ノード数を増加させることが可能と考えられる．今後，ノード数を増やした実験にも取り組む予定である．また，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，他の定型表現や他の番組に対しての実験と検証を進めていく．

文　献

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

[2] “マルチメディア百科事典—膨大な映像資産の有効活用に向けて，” NHK 技研 R&D，no.99, p.58, Jan. 2006.

[3] 八木伸行，佐野雅規，“メタデータの規格，” NHK 技研 R&D，no.95, pp.12–21, Jan. 2006.

[4] 三浦菊佳，山田一郎，住吉英樹，八木伸行，“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 NL 研報，NL171-1, pp.1–6, Jan. 2006.

[5] K. Miura, I. Yamada, H. Sumiyoshi, and N. Yagi, “Automatic generation of multimedia encyclopedia from TV programs by detecting principal video objects using closed captions,” IEEE International Symposium on Multimedia Information Processing and Retrieval (MIPR2006), pp.873–880, Dec. 2006.

[6] Y. Freund and R.E. Schapire, “A decision theoretic generalization of on-line learning and an application to boosting,” J. Comput. Syst. Sci., vol.55, no.1, pp.119–139, 1996.

[7] G. Salton, The Vector Space Model, Automatic Text Processing, pp.312–325, Addison-Wesley Publishing, 1989.

[8] M.A. Hearst, “Multi-paragraph segmentation of expository text,” 32nd. Annual Meeting of the Association for Computational Linguistics, pp.8–16, 1994.

[9] 望月　源，本田岳夫，奥村　学，“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 NL 研報，NL109-7, pp.47–54, Sept. 1994.

[10] M. Collins and N. Duffy, “Convolution kernels for natural language,” Proc. NIPS2001, pp.625–632, 2001.

[11] 市川　宙，橋本泰一，徳永健伸，田中穂積，“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 FI 研報，FI079, pp.39–46, May 2005.

[12] R.E. Schapire and Y. Singer, “BoosTexter: A boosting -based system for text categorization,” Mach. Learn., vol.39, no.2/3, pp.135–168, 2000.

[13] 工藤　拓，松本裕治，“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，” 情処学論，vol.45, no.9, pp.2146–2156, Sept. 2004.

[14] P. Viola and M. Jones, “Fast and robust classification

using asymmetric AdaBoost and a detector cascade," Advances in Neural Information Processing System 14, pp.1311–1318, MIT Press, Cambridge, MA, 2002.
[15] G. Salton and M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill Book Company, 1983.
[16] Y. Nakada, Y. Mouri, Y. Hongo, and T. Matsumoto, "GibbsBoost: A boosting algorithm using a sequential Monte Carlo approach," IEEE Machine Learning for Signal Processing Workshop 2006, pp.259–264, 2006.

（平成 18 年 11 月 8 日受付，19 年 2 月 6 日再受付）

山田　一郎

1991 名大・工・情報工学卒．1993 同大大学院修士課程了．同年 NHK 入局．1996 より放送技術研究所にて自然言語処理を利用した情報抽出，メタデータ生成，知識獲得の研究に従事．2003〜2004 スタンフォード大客員研究員．現在，放送技術研究所専任研究員．映像情報メディア学会，情報処理学会，言語処理学会各会員．

三浦　菊佳

2002 慶大・理工・物理情報工学卒．同年 NHK 入局．2004 より放送技術研究所にて，自然言語処理，情報抽出の研究に従事．映像情報メディア学会会員．

河合　吉彦　（正員）

2001 大阪大学大学院情報数理系専攻修士課程了．同年 NHK 入局．2005 から放送技術研究所にてメディア情報処理などの研究に従事．

住吉　英樹　（正員）

1980 広島県立広島工業高校電気科卒．同年，同年 NHK 入局．広島放送局を経て，1984 より放送技術研究所にて，コンピュータを応用した番組制作システムの研究に従事．現在，放送技術研究所専任研究員．工博．映像情報メディア学会会員．

八木　伸行　（正員）

1980 京都大学大学院電気工学専攻修士課程了．同年 NHK 入局．甲府放送局，放送技術研究所，技術局，編成局を経て，現在，放送技術研究所知能処理グループリーダー．2005 から東京工業大学特任教授（兼任）．画像・映像・メディア情報処理，コンピュータアーキテクチャ，コンテンツ制作技術，ディジタル放送などの研究開発に従事．工博．映像情報メディア学会，情報処理学会各会員．

奥村　　学　（正員）

1984 東工大・工・情報工学卒．1989 同大大学院博士課程了．同年，東工大工学部情報工学科助手．1992 北陸先端科学技術大学院大学情報科学研究科助教授，2000 東工大精密工学研究所助教授，現在，同大学精密工学研究所准教授．工博．自然言語処理，知的情報提示技術，語学学習支援，テキストマイニングに関する研究に従事．情報処理学会，人工知能学会，AAAI，言語処理学会，ACL，認知科学会，計量国語学会各会員．

徳永　健伸

1983 東工大・工・情報工学卒．1985 同大大学院理工学研究科修士課程了．同年（株）三菱総合研究所入社．1986 東工大大学院博士課程入学．現在，同大大学院情報理工学研究科准教授．博士（工学）．自然言語処理，計算言語学，情報検索などの研究に従事．情報処理学会，人工知能学会，言語処理学会，計量国語学会，Association for Computational Linguistics，ACM SIGIR 各会員．