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の検討

山田一郎[*1]　三浦菊佳[*1]　住吉英樹[*1]　八木伸行[*1]　奥村学[*2]　徳永健伸[*3]
[*1] ＮＨＫ放送技術研究所, [*2] 東京工業大学精密工学研究所
[*3]東京工業大学大学院情報理工科学研究科
E-mail: yamada.i-hy@nhk.or.jp

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 45 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、我々は放送された番組を、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の研究に取り組んでいる。放送された番組を効果的に二次利用するためには、番組のどの区間に何が映っているかというメタデータ情報が必要となる。これまでに我々は、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[5]。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[6]。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。

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）

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

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[7]。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[8]。また、工藤らは部分木を素性とする decision stumps 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。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、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

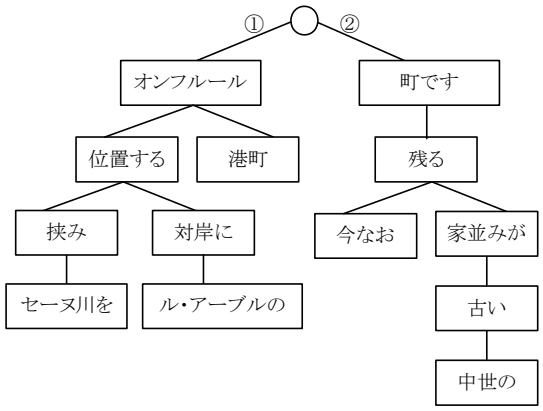

図 1. 木構造生成例

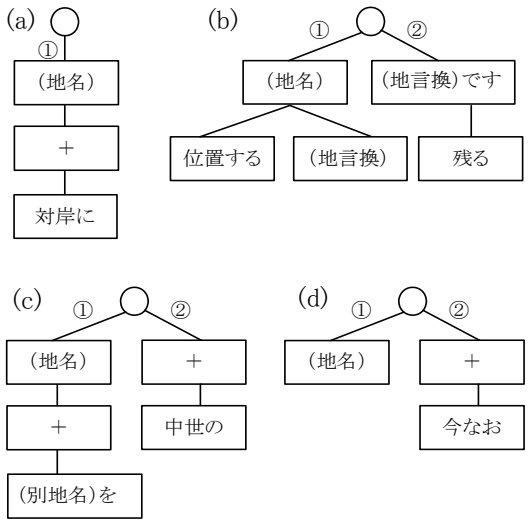

図 2. 木構造から抽出された部分木（一部）

と考えられる。本実験で対象とするテキストはクローズドキャプションのため長文は少なく定型表現を含む文章区間は数文以内と考えられるため、選択対象のノード数は、それほど多くはならない。さらに、弱学習器として利用する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。

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

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とした）
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

類似度が一定値以上か否かを判断基準とすることにより、部分木tと閾値θ_tを変数に持つ弱学習器$h(t,\theta_t)$を生成する。

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 3 に示すアルゴリズムによる学習を行う。

最初のループでは Step2 において、最もエラーが少ない弱学習器が選ばれる。Step3 では、選択された弱学習器で誤っ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。この処理を繰り返すことにより全ての弱学習器に対してαが計算され、それらの和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: $\varepsilon_t=\sum D_t(i)\,|\,h_t(t_i,\theta)-y_i\,|$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を更新

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

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

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 3. AdaBoost による学習アルゴリズム

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、再度、AdaBoost による学習による判定処理を行う。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$\sum\alpha_t h_t(x)$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」を説明する定型表現区間抽出実験

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。14 番組に対して人手により正解データを付与し、10 番組を学習データ、4 番組をテストデータとした。学習データに含まれる正例は 29 個あったため、負例も学習データから無作為に 29 個選択し、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。評価では、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た。最初の判定評価結果を表 2 に示す。「場所を説明する」と判定された文章区間に対して、新たな最終仮説により判定した 2 回目の判定評価結果を表 3 に示す。

表 2　場所説明単語判定評価結果

キー単語の位置	適合率	再現率
場所を説明する文章区間	13/76 (17.1%)	13/14 (92.9%)
場所を説明しない文章区間	121/122 (99.2%)	121/184 (65.8%)

表 3　2 つの最終仮説を利用した場所説明単語判定評価結果

キー単語の位置	適合率	再現率
場所を説明する文章区間	12/46 (26.1%)	12/14 (85.7%)
場所を説明しない文章区間	150/152 (98.7%)	150/184 (81.5%)

表 2 では、場所を説明する文章区間にあるキーとなる単語の適合率が悪い。表 3 の結果では、場所を説明する文章区間において調和平均を表す F 値で 11.2 の向上が見られた。この処理を繰り返すことにより、精度の向上が見込まれる。

次に、正解であった場所を説明する文章区間に対して、抽出された区間がどの程度正例データにおける区間と一致しているか調査を行った。正解データの区間に含まれる文中のうち、提案手法により抽出された文の割合を示すカバー率は 93.8%(30 文/32 文)、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抽出精度は 61.2%(30 文/49 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。

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」に関する情報を映像とともに説明している定型表現を含む文章区間抽出する実験により、一定の分別能力があることを示したが、適合率が低いという問題が残された。

今後、判定処理を複数段繰り返す実験を進めるとともに、類似度の改善を検討していく予定である。

【参考文献】

[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 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] 栗岡, 柳川, 福田, 長尾: サーバ型放送, 映情学誌, Vo.58, No.5, pp.647-650,

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

[6] 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)

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

[8] 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

[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)