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

2006－NL－171 (1)
2006／1／12

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

三浦菊佳　山田一郎　住吉英樹　八木伸行

NHK 放送技術研究所
〒157-8510 東京都世田谷区砧1－10－11
Tel. 03－5494－3145

本論文では、テレビ番組のクローズドキャプションに含まれる映像内容を説明した文の特徴を利用して、映像中に現れる主要な被写体を推定する手法を提案する。主被写体を表現するときの文の特徴抽出に Quinlan の C4.5 決定木学習アルゴリズムによる機械学習を用い、得られたプロダクションルールの予測精度を指標として、映像カットごとに1つの被写体名詞を抽出する。動物や自然を題材とする20番組を対象とした被写体名詞の抽出実験では、適合率57.6%、再現率40.7%の精度が得られ、手法の有効性を確認した。

A Method of Detecting Principal Video Objects Using Closed Captions

Kikuka MIURA, Ichiro YAMADA, Hideki SUMIYOSHI, and Nobuyuki YAGI

Science and Technical Research Laboratories
Japan Broadcasting Corporation
1-10-11 Kinuta, Setagaya-ku, Tokyo, 157-8510, JAPAN
Tel: 03-5494-3145

This paper proposes a method of detecting principal objects in video material using features of sentences in the associated TV closed captions that describe the video content. We use Quinlan's C4.5 decision tree for machine learning to extract the features of sentences that describe principal objects, and extract a noun for every shot by using the values of predicted accuracy of production rules. In an experiment using 20 TV programs dealing with animals or nature, the method attained 57.6% precision and 40.7% recall. Experimental analysis demonstrates the effectiveness of the method.

1. はじめに

放送局では、多種多様な番組が大量に制作されている。NHKは、NHKアーカイブス[1]として約50万本もの番組を蓄積している。そのうち約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二次利用しているのみで、十分活用されているとは言いがたい。そこで、我々は映像の百科事典など新たなコンテンツサービスに利用できないかと考えている。このような映像の活用を進めるためには、番組のどの区間に何（被写体）が映っているのかという情報（メタデータ）が必要であるが、保存コンテンツのほとんどにメタデータが付与されていないのが現状である。また、すべての番組に自動的にメタデータを付与することは容易ではない。被写体情報を抽出する手法についてはさまざまな研究が行われているが、映像認識技術を用いて被写体情報を抽出することは非常に困難であり、言語処理からのアプローチが現実的である。

総務省では、聴覚障害者のために2007年までに付与可能なすべてのテレビ番組で字幕放送を行うことを目標に掲げており、字幕放送番組が急激に増加している[2]。クローズドキャプションは、字幕放送に対応した機器を通して見ることのできる字幕のデータで、表示/非表示を受信機側で切り替えることができる。また、番組中の出演者の発話内容やナレーションから、字数制限して読みやすく作成されており、映像内容を説明した文章を多く含んでいる。

我々は、これまでにクローズドキャプションを利用した被写体特定手法[3]を提案している。この手法では、クローズドキャプションに出現するすべての名詞について被写体であるか否かを判定しているため、同一映像カット（映像の切り替わり点により区切られる映像区間）に複数の被写体が抽出されることもあった。そのため、目的の被写体が背景の一部として映っているといった不要な映像も抽出されていた。そこで本稿では、番組の映像カットごとに登場する主要な被写体（以下、主被写体）を推定することを目的とし、C4.5決定木学習アルゴリズムにおけるプロダクションルールを用いてクローズドキャプションから主被写体を表す名詞候補を抽出する手法を提案する。映像の主被写体を表現する名詞がその区間に提示されるクローズドキャプション中に現れる場合を対象とし、動物や自然を題材とした番組に適用した実験とその結果について報告する。

2. 関連研究

これまでに、クローズドキャプションを利用した映像被写体を特定する手法として、Satohらによる Name-It[4]がある。この手法では、顔画像解析とオープンキャプション（専用の機器を通さずに見ることのできる通常の字幕）の認識処理を組み合わせて高精度に映像中の人物を特定している。しかし、ニュース映像の人物のみを対象としているため、あらゆる被写体に適用するのは難しい。

また、Google社は、クローズドキャプションを利用して番組を検索し、代表画像とともに提示するするシステム Google Video[5]を公開している。このシステムは、検索語をクローズドキャプションに含む番組の提示は行っているが、必ずしも検索語が被写体として抽出されているとは言えない。

一方、望月らはクローズドキャプションのデータフォーマットを利用したシーン検索[6]を提案している。これは、文字の色や記号などクローズドキャプションに含まれるテキスト以外の情報から、発話者や音の種類などを抽出する。テキスト自体の解析処理を行っていないため、被写体の抽出までには至っていない。

3. クローズドキャプションの予備調査

表1にクローズドキャプションの例を示す。本稿では、クローズドキャプションデータのうち、番組開始時刻からの「時:分:秒.フレーム」で表される画面提示開始時刻と、そこで表示されるテキストのデータを対象とする。この時刻により映像と対応付けることができる。

クローズドキャプションには映像内容を具体的に説明する文章が存在する。このような文章を見つけることができれば、映像に現れる被写体を推定できると考えられる。しかし、被写体を特定するために目的の言葉を抜き出すキーワードマッチングだけでは精度良く取り出すことができない。例えば、表1

表1. クローズドキャプションの例

画面提示時刻	テキスト
00:22:53.22	スイギュウの母親が気づきました。
00:23:04.25	ライオンの位置を確認するといちもくさんに走り出します。
00:23:18.26	スイギュウは深みに向かって逃げ込みます。
00:23:23.11	泳いで後を追うライオン。
00:23:28.28	スイギュウの巨体は水をはね飛ばし群れで駆け抜けます。
00:23:37.02	ライオンはあっという間に離されてしまいました。
00:23:41.22	水の深いところに逃げ込まれてはライオンも手が出ません。
00:23:47.05	湿地の深みを上手に利用してスイギュウが逃げきりました。

のクローズドキャプションの例において、実際の映像で主被写体がライオンであるのは4行目の文に対応するカットのみである。他の文に対応するカットでは、ライオンという単語が出現しているにもかかわらずスイギュウが主被写体であり、ライオンは主被写体ではない。つまり、ライオンという単語がクローズドキャプションに存在する映像を単純に抜き出すだけでは目的外の映像被写体を抽出してしまう可能性がある。そこで、映像中の被写体を説明している文を抽出し、解析して主被写体名詞を抽出しなくてはならない。

例えば、「眠ってばかりのコアラ。」といった体言止めの文や、「これはライオンです。」といった文末が断定の助動詞の文は、映像内容を説明していると見受けられる。NHK で放送された「地球・ふしぎ大自然」20 番組のクローズドキャプションを対象として、この特徴を持つ具象物名詞が、対応する映像カットの主被写体となっていたか調査した結果を表2 に示す。ここで具象物名詞とは、目で見て手で触れられるものを表現した名詞を指し、国立国語研究所の分類語彙表[7]の上位4桁を判断基準とした。

表 2. “体言止め”、“名詞+「です」”が
映像の主被写体を表現する割合

適合率	再現率	F 値
479 / 748 (64.0%)	479 / 2776 (17.3%)	0.272

適合率の結果から、体言止めの具象物名詞や断定の助動詞「です」が後続する具象物名詞は、映像カットの主被写体となる傾向があることがわかる。しかし再現率が低いことより、ほかの文体表現による主被写体の説明が行われていると考えられる。精度を改善するためには、このほかの特徴も抽出する必要がある。

4. 主被写体抽出処理

本章では、前章で調査した特徴も含め、主被写体を説明する文の特徴を抽出して主被写体の推定を行うために、機械学習を用いる。手順を図 1 に示す。まず、映像カットごとに主被写体を表す名詞の正解値が与えられた学習データから特徴を抽出し、Quinlan の C4.5 決定木学習アルゴリズム[8]に入力する。この結果、クローズドキャプションに含まれる各名詞が主被写体であるか否かを判定する決定木が生成される。次に、生成された決定木から、優先順位付けされたプロダクションルールを生成する。このプロダクションルールを利用してテストデータの名詞について主被写体であるか否かを判定し、主被写体候補を抽出する。その際、1映像カットにつき複数主被写体候補がある場合、予測精度を指標に主被写体名詞を1つに絞る。

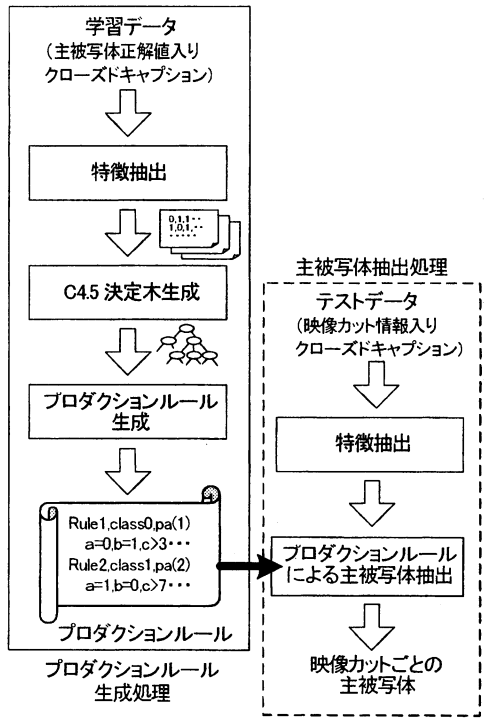

図 1. 主被写体抽出処理手順

以下に、特徴抽出、決定木生成、プロダクションルール生成、主被写体抽出について説明する。

4.1 特徴抽出

映像内容を説明する文の特徴を抽出するために、クローズドキャプションの各文に含まれるすべての自立語名詞を対象として表 3 に示す属性に対する属性値を付与する。表中の①により文の格構造の特徴を抽出し、②により文末の文体を判定する。また、主被写体となる重要な名詞には補足説明が多用されると考えられるため⑤を属性とする。「これ」などの指示詞は映像中の被写体を参照することが多いと報告されている[9]ため⑦の特徴を考慮し、さらに「ある」「いる」などの動詞に係る名詞は被写体の存在を表すと考えられるため⑧を設ける。

4.2 決定木生成

抽出した特徴を用いて「主被写体である」「主被写体でない」の2クラスに分類する決定木を生成する。映像カットごとに主被写体が特定された正解値を持つ学習データを対象として、クローズドキャプション中の全自立語名詞に 4.1 の特徴抽出処理を行い、C4.5 決定木学習の入力とする。決定木学習ではまず全ての属性を分岐条件の候補とし、分岐前後の情報利得比が最大になる属性を分岐条件とする。この処理を繰り返すことにより、中間ノードに分岐条件が配置され、葉ノードに学習データの事例が配置され

表 3. クローズドキャプションから抽出した特徴

属性	属性値
①対象名詞が含まれる名詞句の付属語の種類	は格 or NOT, が格 or NOT, の格 or NOT, に格 or NOT, を格 or NOT, と格 or NOT, へ格 or NOT, で格 or NOT, 副助詞 or NOT
②文末の文体, 句の係り先	体言止め or NOT, 断定の助動詞「です」が後続 or NOT, 対象名詞句の係り先が最終文節 or NOT
③対象名詞の種類	固有名詞 or NOT, サ変名詞 or NOT, 形容動詞語幹 or NOT, 数量名詞 or NOT, 番組中の新語 or NOT, 具象物名詞 or NOT
④対象名詞の重要度	正の数値(TFIDF 値)
⑤対象名詞句を修飾する文節数	自然数(0, 1, 2, …)
⑥主語の有無	主語有 or 主語無
⑦指示語の有無	指示語有 or 指示語無
⑧存在を表す動詞の有無	存在を表す動詞有 or 存在を表す動詞無

た決定木が生成される。

4.3 プロダクションルール生成

4.2 で得られた決定木を属性の分岐条件と分類クラスからなるルール記述の集合であるプロダクションルールに変換する。初期値として全ての葉に対応するルールを生成する。この結果、葉の数だけルールが生成される。次に、各ルールから予測精度が低下しない範囲で分岐条件を除外する。ルール i に対する予測精度 pa は以下の式で算出される。

$$pa(i) = 100\,(1 - pe(i))$$

$$pe(i) = \left\{ p \mid \sum_{j=0}^{E} \binom{N}{j} p^{j} (1-p)^{N-j} = CF,\ 0 \le p \le 1 \right\}$$

E ：ルール における誤り事例数
N ：ルール に適合する総事例数
CF ：枝刈り度 (0.25)

pe は悲観的誤り率を表し、CF とは枝刈りの度合いを示す定数で標準値を用いている。

分岐条件の除去により生じた重複ルールを除去し、残ったルールに対し予測精度の降順に優先順位付けを行う。この優先順位付けされたルール集合をプロダクションルールとする。プロダクションルールは決定木の枝刈りと違い、中間ノードに存在する分岐条件も除去することができ、また人間の目で見て理解しやすいという利点を持つ。さらに、予測精度の値により各ルールに優先順位付けできるため、本手法では決定木により直接判定を行わずプロダクションルールを採用した。

4.4 主被写体抽出

学習データから生成されたプロダクションルールをテストデータに適用し、主被写体を抽出する。ここでは、クローズドキャプションに対応する映像カット情報が付与されたテストデータを用い、映像カットごとに主被写体を抽出する。まず、テストデータの各名詞に対してプロダクションルールに含まれるルールを優先順位順に適用し、最初に合致したルールが持つ分類クラスを判定値とする。主被写体と判定された名詞が 1 つの映像カットに 1 つであった場合は、その名詞を主被写体として出力し、複数主被写体があった場合は、最も予測精度の高いルールが適用された名詞を出力する。予測精度が同値であった場合は、複数の選択を許し同値のものをすべて出力する。反対に、映像カット内に主被写体と判定された名詞が 1 つも存在しない場合は「主被写体なし」と出力する。

5. 主被写体抽出実験

提案手法の有効性を検証するために、NHK で放送された番組「地球・ふしぎ大自然」を対象とした主被写体抽出実験を行った。この番組は、世界の動物や植物などの紹介を趣旨に制作されており、本稿の目的である主被写体を説明する記述が多く、実験に適していると考えたため使用した。以下、詳細を述べる。

5.1 プロダクションルールによる主被写体抽出実験

実験用データ作成のため、「地球・ふしぎ大自然」20 番組分のクローズドキャプションの各文にカット番号を人手で付与した。なお、オープニング（番組の導入部）、エンディング（結び）、インターミッション（転換部で、次のコーナーの紹介などが含まれる）は番組ごとに類似した表現が多用されるため除外した。

このクローズドキャプションに含まれる全ての自立語名詞に対して、人手により映像カットごとに主被写体であるか否かを与え学習データとした。このとき、主被写体を映像カットごとに 1 つだけ絞りこんで正解値を付与した。この結果、全自立語名詞 19520 個の正解値の内訳は、2776 個 (14.2%) が「主被写体である」、16744 個が「主被写体でない」であった。映像カットに対応するクローズドキャプションに主被写体となる名詞が存在しない場合も多くみられ、4351 カット中 1575 カット (36.2%) に主被写体となる名詞が存在しなかった。

主被写体であるか否かの正解値が付けられた 19 番組を学習データ、残り 1 番組分のテストデータと

し、クロスバリデーションにより合計20回の主被写体抽出の評価実験を行った。この際、特徴抽出処理では、係り受け解析には南瓜[10]を使用した。評価結果の平均値を表4に示す。

表4. プロダクションルールによる主被写体抽出実験結果

	適合率	再現率	F値
主被写体有	1131 / 1963 (57.6%)	1131 / 2776 (40.7%)	0.477
主被写体無	1265 / 2440 (51.8%)	1265 / 1575 (80.3%)	0.630

表4の「主被写体有」に対するF値は0.477であり、表2に示す体言止めと断定の助動詞を手掛かりとした手法のF値0.272と比べて有効性が確認できた。特に、再現率では23.4%向上しており、プロダクションルールを用いた手法により表2の2つの文末表現以外の特徴も抽出できたといえる。

5.2 考察

本研究では、クローズドキャプションというテキスト情報から映像カットの主被写体を推定することを目的としている。しかし番組では、クローズドキャプションからは推測できない意外な映像を扱う場合や、映像内容について言及しない場合もあり、クローズドキャプションのみから全ての主被写体を推定することは難しいと考えられる。そこで、人手によりどの程度まで主被写体が推定可能か検証実験を行った。正解データの作成者とは異なる被験者が、クローズドキャプションのみから主被写体を推定した結果と、映像を見て作成した正解データと比較した。5番組を対象とした結果を表5に示す。この値が、自然言語処理によるアプローチの限界と考えられる。表5と表4を比較すると、主被写体有では適合率の上限が64.4%のところ57.6%、再現率では73.7%のところ40.7%であり、提案手法により良好な結果が得られているといえる。

表5. 人による主被写体抽出結果

	適合率	再現率	F値
主被写体有	474 / 736 (64.4%)	474 / 643 (73.7%)	0.684
主被写体無	316 / 485 (65.2%)	316 / 578 (54.7%)	0.595

次に、実験で得られたルールの検証を行った。ここでは、学習データごとに生成されるプロダクションルールが異なるため、一つの番組に対して生成されたものを対象とした。表6に主被写体抽出に成功した事例数の上位4ルールと判定例を示す。表中のルール番号は、ルールが持つ優先順位を示し、小さい番号ほど優先度が高い。Rule27とRule9は、具象物名詞に断定の助動詞「です」を伴うものと体言止めとなるルールである。Rule28とRule31は、「が格」、「は格」により文の主語になる具象物である。これらが主被写体の特徴的な表現であることがわかる。

表7に、主被写体を表現していない名詞が主被写体と誤判定された事例数上位4ルールと判定例を示す。Rule27は、具象物名詞に断定の助動詞「です」を伴うルールだが、誤判定例の名詞「姿」は、この前に出現した単語の言い換え表現となっており、誤判定してしまった。言い換え表現の抽出処理、元表現の推定処理が必要と考えられる。また、人が読んでも映像を見ないと判断できない文章表現も多く含まれていた。提案手法で誤判定された80名詞のうち、表5に示した被験者による実験では31名詞に誤りが見られた。これらは、言語処理のみからでは解決不可能だと考えられる。

さらに誤判定の原因の一つに、「AのB」という表現の解析の難しさが挙げられる。例えば、「この植物の正体は何でしょうか？」では、主被写体である「植物」を抽出することができなかった。もし「この植物は何でしょうか？」であれば「植物」を主被写体として抽出できる。提案手法の属性だけでは、この問題解決は困難と考えられ、新たな属性を与える必

表6. 主被写体抽出に成功した事例数上位4ルールと判定例

ルール番号 [該当数]	ルール	判定例 (下線部が主被写体と判定)
Rule 28 [14事例]	①が格/②係り先最終文節/③サ変名詞でない，具象物名詞/④TFIDF値>2.25	そして確かに山の頂には氷河が待っていました。
Rule 27 [12事例]	②断定助動詞/③形容動詞語幹でない，具象物名詞	大きなカバです。
Rule 9 [7事例]	②体言止め/③形容動詞語幹でない，具象物名詞/④TFIDF値>2.14/⑤修飾する文節数>0/⑥主語無/⑦指示語無	世界最長の大河、ナイル川。
Rule 31[7事例]	①は格/②係り先最終文節/③新語でない，具象物/④TFIDF値>20.2/⑦指示語無/⑧存在の動詞無	ロベリアは寒さから身を守るために閉じていた葉を開いていきます。

表 7. 主被写体を表現しない名詞が主被写体と誤判定された事例数上位4ルールと誤判定例

ルール番号 [該当数]	ルール	誤判定例 (下線部が主被写体と判定)
Rule 28 [12 事例]	①が格/②係り先最終文節/③サ変名詞でない, 具象物名詞/④TFIDF 値>2.25	山の寒さと赤道直下の太陽が源流の自然に魔法をかけます。
Rule 27 [9 事例]	②断定助動詞/③形容動詞語幹でない, 具象物名詞	てっぺんだけに葉がついた奇妙な姿です。
Rule 33 [5 事例]	①(の格・に格・を格・と格・で格)でない/②係り先最終文節/③固有名詞でない, サ変名詞でない, 数量名詞でない, 具象物名詞でない/④TFIDF 値>15.23	氷河期は生き物たちに試練を与えるとともに進化の原動力にもなったのです。
Rule 37 [5 事例]	①が格/④TFIDF 値>10.77/⑤修飾する文節数>0	ロベリアの葉が大好物です。

要がある。また、主被写体であると誤判定した結果と同じように、人が判断できない場合も、誤判定された 21 名詞のうち、16 名詞存在した。

6. まとめ

本稿では、クローズドキャプションから主被写体を表す名詞を抽出し、C4.5 決定木学習アルゴリズムにおけるプロダクションルールの予測精度を用いて映像カットごとの映像主被写体を推定する手法を提案した。動物や自然を題材とした 20 番組を対象とした実験を行った結果、適合率 57.6%、再現率 40.7%を得た。今後、言い換えの表現の抽出、「AのB」の表現の解析など、今回誤判定されたものについて改善をはかる予定である。また、ゼロ主語補完、ゼロ目的語補完、代名詞の照応解決などを考慮していくことにより、クローズドキャプション中に被写体名詞そのものが出現していなくても推定できるよう、研究を進めていく予定である。

なお、本稿を作成するにあたり、ご指導を賜りました東京工業大学徳永健伸助教授、奥村学助教授に深く感謝いたします。

【参考文献】

[1] NHK アーカイブス (http://www.nhk.or.jp/nhk-archives/)
[2] 総務省：平成 16 年度の字幕放送の実績 (http://www.soumu.go.jp/s-news/2005/050811_6.html)
[3] 三浦、山田、住吉、八木：クローズドキャプションを利用した被写体特定手法, 情報科学技術フォーラム一般講演論文集, E-013 ,pp145-146 (2005)
[4] Shin'ichi Satoh and Yuichi Nakamura and Takeo Kanade : Name-It ; Naming and Detecting Faces in Video by the Integration of Image and Natural Language Processing, IJCAI-97, pp.1488-1493 (1997)
[5] Google Video (http://video.google.com/)
[6] 望月、有安、佐野、住吉、井上：シーン検索時における字幕データ利用の一検討, 映情学年次大, 23-5, pp338-339 (2000)
[7] 国立国語研究所：分類語彙表 増補改定版 (2004)
[8] Quinlan, J.R : C4.5 Programs for Machine Learning, Morgan Kaufmann (1993)
[9] 徳永、西田、山田：情報番組におけるコソア系の指示詞の分布について, 計量国語学会第 49 回大会 (2005)
[10] 工藤、松本：チャンキングの段階適用による係り受け解析, 情処学論, Vol.43, No.6, pp.1834-1842 (2002)