E-013

クローズドキャプションを利用した被写体特定手法の検討

A Study on Video Object Detection Using Closed-Captions

三浦 菊佳† 山田 一郎† 住吉 英樹† 八木 伸行†
Kikuka Miura Ichiro Yamada Hideki Sumiyoshi Nobuyuki Yagi

1．はじめに

放送局では、効率的な番組制作のために、過去に放送された番組や映像素材を効果的に二次利用する環境の整備が求められている。番組制作者が過去の番組から必要とするシーンだけを検索する際、時間軸上のどの区間にどのような内容が描かれているかを記述したメタデータが重要な役割を果たす。そこで我々は、字幕放送の字幕データ（以後、クローズドキャプションと呼ぶ）を利用して、メタデータを自動生成する研究を進めている[1]。

これまでに、クローズドキャプションを応用した映像中の被写体を特定する手法として、Satoh らは Name-It[2]を提案している。この手法では、顔画像解析、オープンキャプション抽出処理を行い、クローズドキャプションの情報と組み合わせて高精度に映像中の人物を特定している。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するシステム Google Video[3] を公開している。このシステムは、検索語に関連する番組の提示が目的のため、映像の被写体特定まで行っていない。

我々は、クローズドキャプション中の映像内容を説明する文体に着目し、これを抽出することで映像中の被写体が何かを知ることができると考えた。本稿では、クローズドキャプションに現れる映像中の被写体を説明する典型的な表現と、決定木を用いた被写体抽出実験について述べる。

2．クローズドキャプションの特徴

総務省では、2007 年までに付与可能なすべての放送番組にクローズドキャプションを付与することを目標に掲げており、近年クローズドキャプションが付与された番組が急激に増加している[4]。

クローズドキャプションは、番組中の出演者の発話内容やナレーションをもとに作成されており、番組を説明する有効なテキスト情報と考えられる。表 1 にクローズドキャプションの例を示す。クローズドキャプションには、テキスト以外に、画面に提示された時刻情報が含まれており、映像とリンクしている。

表 1．クローズドキャプションの例

提示時刻 ：	テキスト
00:25:58	：ほかにもアリ塚を利用する鳥がいます。
00:26:02	：これはアナホリフクロウ。
00:26:05	：小鳥や昆虫を狙う草原のハンターです。
00:26:12	：アリ塚のそばに穴を掘って暮らしています。
00:26:19	：アナホリフクロウは昼も夜もこうしてアリ塚の上に止まっています。

クローズドキャプションは、映像とともに情報を伝えるため、映像内容を具体的に説明する記述が多く存在する。この部分を利用すれば、映像中の被写体が何であるか取り出すことができる。しかし、ただ単語を抜き出すだけでは目的の被写体を精度よく取り出すことはできない。例えば、ライオンの映像が目的でも、ライオンが脱走しましたと伝えるアナウンサーの顔が抽出される可能性がある。

映像内容を説明する文体には一定の特徴があると考えられる。例えば、被写体の存在を説明するような部分では、体言止めなどの表現が多用されると考えられる。そこで、実際に放送されたクローズドキャプションを対象として、文末が以下の特徴を持つ文を抽出し、各文の最後部の文節に含まれる名詞（以下、最終名詞と呼ぶ）が、対応する映像の被写体となっているか調査した。

①具象物名詞で体言止め
②具象物名詞＋断定の助動詞（「です」）

このとき対象とする最終名詞は、目で見て手に触れられる具象物名詞のみに限定した。NHK のテレビ番組 2 種類 4 番組分について調査した結果を表 2 に示す。ここでは、被写体が具象物であるカットのみを対象とし、適合率は、〔最終名詞が被写体であるカット数〕／〔①または②に該当する文の最終名詞に対するカット数〕、再現率は、〔最終名詞が被写体であるカット数〕／〔被写体が具象物であるカット数〕を表す。

表 2．①または②に該当する文の最終名詞が対応する映像の被写体である割合

番組名	適合率	再現率
地球・ふしぎ大自然 I	30／31 (96.8%)	30／269 (11.2%)
地球・ふしぎ大自然 II	20／20 (100%)	20／195 (10.3%)
ためしてガッテン I	18／20 (90.0%)	18／129 (14.0%)
ためしてガッテン II	22／24 (91.7%)	22／154 (14.3%)

適合率の結果より、体言止めの名詞や断定の助動詞「です」が後続する名詞は、高い確率で映像被写体となることがわかる。しかし、再現率は低く、この特徴だけでは被写体特定が不十分だと言える。

3．決定木を用いた被写体特定

前章で、映像内容を説明する記述には一定の特徴があることを示した。本章では、前述の①②以外の文体特徴も利用して、映像の被写体となる名詞を機械学習により抽出する実験について述べる。機械学習には、Quinlan の C4.5 決定木学習アルゴリズム[5] を用いる。決定木は、事例の属性とその属性値、判定値から分岐規則を生成して表現される。このアルゴリズムは、入力データに含まれるノイズに強く、学習データが多いほど正解判定精度が高くなるとい

†NHK 放送技術研究所（知能情報処理）

う利点を持つ。本手法では、分岐接点における枝分岐数が 3 以上の場合にも適応できる多分木を使用した。

3.1 属性の抽出

クローズドキャプションを、1 文ごとに分割し、各文に含まれるすべての自立語名詞を対象として、4 種類の属性（要素数 29 ）を付与した。その内訳を表 3 に示す。係り受け解析には南瓜[6] を使用し、分類番号は、国立国語研究所の分類語彙表[7] の上位 2 桁 を用いた。表中の「文節の機能語情報」とは、格助詞、係助詞、助動詞などの有無のほかに、係り先の情報が含まれる。「文中での位置」とは、文の最終文節からの文節番号であり、機能語が存在せず、最終文節の場合に体言止めと判断できる。「分類番号の種類」は、具象物名詞か否かの判断指標のために設けた。

表 3．使用属性と要素数

属性	要素数
文節の機能語情報	9
文中での位置	11
固有名詞の有無	2
分類番号の種類	7
計	29

3.2 実験

NHK のテレビ番組「地球・ふしぎ大自然」6 番組分（1 番組 42 分 30 秒）のクローズドキャプション（オープニング、インターミッション、エンディングを除く 7294 名詞 ）を実験対象とした。このうち、学習データを 5 番組分、テストデータを残りの 1 番組とし、クロスバリデーションにより計 6 回の試行を行った。クローズドキャプションに出現するすべての名詞に対して、人手により被写体であるか否かを判定し、正解データとした。このとき、名詞が含まれるクローズドキャプションに対応する映像の同一カット内で、その名詞が主被写体として映っていた場合を「被写体である」と判断した。

3.3 結果

生成された決定木により、テストデータを評価した結果を表 4 に示す。ここでは、6 回分の試行により得られた出現数をすべて足し合わせて、適合率と再現率を算出した。被写体であると判定された名詞の適合率は 55.8% 、再現率は 17.2% であり、改善の余地が残される。

表 4．実験結果

判定	適合率	再現率
被写体である	217／389 (55.8%)	217／1259 (17.2%)
被写体でない	5863／6905 (84.9%)	5863／6035 (97.1%)

3.4 考察

決定木は、分岐後の集合のクロスエントロピーが小さくなる分岐規則が上位に位置する特徴を持っているため、深さの浅い分岐規則が有力な条件と考えられる。学習により生成された決定木を詳しく見てみると、6 回の試行により生成された 6 種類の決定木はすべて、深さ 1 に「分類番号の種類」、深さ 2 に「文節の機能語情報」が分岐規則として選ばれていた。

最上位の分類規則が「分類番号の種類」である原因の一つは、番組の分野が偏っていたことにあった。自然現象の用語や、未知語に分類されるオオアリクイやコメツキムシといった動物の名称が多数を占めたことで、データの偏りによる過学習になったと考えられる。加えて、学習データが少なかったこともあげられる。決定木の性質上、より多くの学習データを用意することで精度の向上が見込まれる。

また、番組中には具象物が被写体でも「平和」や「仲良し」など、番組制作者の意図では抽象概念を表す映像もある。実験対象の全カットを調査した結果、具象物そのものを意図して撮影されたカットは 76.8% を占めた。被写体を具象物として説明している割合が高く、判定対象とした名詞が、具象物名詞か否かを正しく判断することが有効と考えられる。今回は分類語彙表の上位 2 桁 のみで判定したため、抽象名詞と具象物名詞が同一カテゴリーに混在しており、正確に具象物名詞を分類できていない。今後は、具象物名詞か否かという属性を設ける予定である。

深さ 2 の分岐規則において「文節の機能語情報」として設定した要素の中では、【体言止め】、【後続が断定の助動詞であるもの】という 2 章で調査した文体に加えて【が格かつ最終文節に係るもの】が「被写体である」と判定されており、これらの要素が被写体を説明する典型的な表現となると考えられる。

4．まとめ

本稿では、決定木学習を用いて、クローズドキャプションから映像中の被写体を説明する典型的な表現を抽出する検討を行った。抽出実験の結果、被写体であると判定された名詞は、適合率 55.8% 、再現率 17.2% と良好とは言えず、これは偏りのある番組内容や分類語彙表のカテゴリーを属性とした学習データにより最上位の分岐規則が分類番号となったためと考えられる。今後、偏りのない学習データを増やし、属性を見直すことで、改善をはかる予定である。

【参考文献】

[1] 山田、小早川、三浦、住吉、八木、崔：クローズドキャプションを対象とした因果関係知識抽出の検討, FIT2005

[2] 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)

[3] Google Video（http://video.google.com/）

[4] 総務省：平成 1 5 年度の字幕放送の実績 http://www.soumu.go.jp/s-news/2004/040806_3.html#02

[5] Quinlan, J.R : C4.5 Programs for Machine Learning, Morgan Kaufmann (1993)

[6] 工藤、松本：チャンキングの段階適用による係り受け解析, 情報処理学会論文誌, Vol.43, No.6, pp.1834-1842 (2002)

[7] 国立国語研究所：分類語彙表 増補改定版 (2004)