8M-03

キャプションを複合的に利用したニュース文からの重要文抽出*

小野 晋太郎 †, 井手 一郎 ††, 坂井 修一 ‡, 田中 英彦 ‡
onoshin@mtl.t.u-tokyo.ac.jp, ide@nii.ac.jp, {sakai,tanaka}@mtl.t.u-tokyo.ac.jp
† 東京大学工学部 †† 国立情報学研究所 ‡ 東京大学大学院工学系研究科

1 はじめに

近年の映像データの増加に伴い、有用な情報を計算機を用いて効率的に取捨選択する必要が高まっている。特にニュース映像については、いわゆる 4W(時相、場所、人物、事象・行為) をキーとする検索需要が見込まれ、これらを重要情報として抽出しておく必要性は高い。ニュース映像には字幕 (オープンキャプション, 以下 OC) による話題のタイトルや、様々な付加的情報が付随しているが、これらは断片的で雑多であり、4W 情報を適切に表現しているとは言い難い。

新聞記事やレポートなどの紙面上の文書に対し、それに付随するタイトルや見出しを用いて内容の要約・重要文抽出を行った研究例としては [1] や [2] がある。ニュース映像において、新聞などにおける見出しに相当する情報は画面下部に表示されるタイトルであるが、これは見出しよりも更に短く簡素であるなど、性質が異なる。また、ニュース映像においては、紙面上の文書と異なり、本文と OC の表示される時間関係を考慮できる。

本稿では、OC とニュース本文全体の関連性を利用し、効率的に 4W 情報を示すための重要文抽出を行うことを目的とする。

2 ニュース映像中のテキスト情報源

ニュース映像中の主なテキスト情報源にはキャプション (字幕) があり、画像に挿入される方法及び性質の違いから、以下の 2 種類に分類される。

- オープンキャプション (OC):
  送信側が予め画面上に画像として重ね合わせて表示するもの。半数程度が人物 (Who)、場面 (Where)、時相 (When) のいずれかに関する重要情報を簡潔に表す [6] が、表示と同時に画像や音声を含む映像全体を見ることを前提としており、単独ではニュース内容を十分に表現し得ない。
- クローズドキャプション (CC):
  アナウンサの原稿朗読箇所を書き下したもの[1]。日本では文字放送の形で提供され、視聴者は画面に重ねて見ることができるほか、容易にディジタルテキストとして利用できる。単位時間当たりの情報量が多く、冗長なのが特徴である。

OC は背景の画像に埋め込まれて与えられるため、そのままテキストデータとして利用することはできないが、自動認識に関する研究 [4] が行われているほか、放送のディジタル化に伴い、将来的にディジタルテキストとして送信される可能性もあることから、本研究では人手により書き下したものを用いる。

3 キャプションを複合的に利用した重要文抽出

前述の OC, CC の特性を考慮し、以下では OC を手がかりに CC から重要文を抽出することを考える。手法の全体像を図 1に示す。以下に各々の処理について述べる。

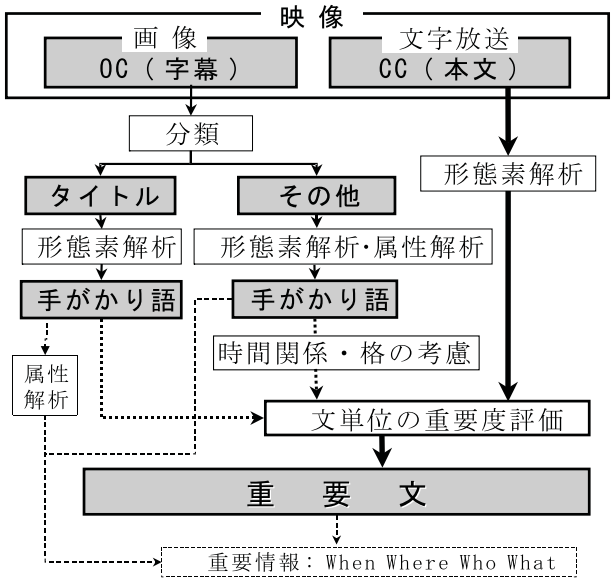

図 1: 重要文抽出機構の全体像

*"Extraction of important sentences from closed caption using open caption in news video"
Shintaro Ono†, Ichiro Ide††, Shuichi Sakai‡, Hidehiko Tanaka‡
† Faculty of Engineering, ‡ Graduate School of Engineering, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
†† National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

[1]現在、NHK「ニュース 7」(225 分/週) では、スタジオの特定のアナウンサによる原稿朗読箇所に限り、高精度 (公称 95% [3]) の音声認識システムにより、リアルタイムでの提供が行われている。

3.1 OC に対する処理

OC からは、CC 中の文単位の重要度を評価する際の手がかり語を選定する。OC は以下のように分類される。

1. タイトル
2. タイトル以外の名詞句 (列)、動詞: 場所・組織、人物、時相

これらのうち 1. 中の名詞と動詞、及び 2. を手がかり語として利用する。品詞の解析には日本語形態素解析システム JUMAN [5] を用いる。2. には「中継」などなどの放送技術に関連した用語なども含まれるが、これらは語尾に注目した名詞句の属性解析手法 [6] により取り除く。

3.2 CC に対する処理

CC(本文) に対しては、OC から得た手がかり語をもとに個々の文の重要度を評価し、重要文を抽出する。

- タイトルとの関連性の定量化
  ニュースのタイトルと内容が近い文は重要と考えられることから、前節で得たタイトル中の手がかり語と同じ語を含む文に対し、得点を与える。
- 同期して表示される OC との関連性の定量化
  ニュース本文に対して同期して表示される OC と内容が近い文は重要と考えられることから、同期する OC 中の手がかり語と同じ語を含む文に対し、得点を与える。また、ある文に対し、同期する OC 中の固有名詞がその文の主語となっていれば (地名の場合は「○○では」の形で文中に含まれていれば) 更に得点を加える。

更に、TF-IDF 法により、珍しいと判断された語を含む文にも得点を与える。同じ語を含むか否かを判定する際は、同義語・略語辞書を参照し、例えば「大蔵大臣」と「蔵相」が異なる語と判断されるのを防ぐ。

4 実験

以上の手法を計算機上に実装し、2000 年の NHK「ニュース７」163 記事に対して実験を行った。そのうち 1 記事の結果を表 1に示す。

得られた結果を主観的に評価したところ、1 文どうしの細かな重要度の順位については改善すべき点が見受けられるものの、全体的に重要と思われる文は上位に、一般的でニュース性のない文や雑多な情報の多い文は下位に判定される傾向が見られた。また、表 1で最下位と判定された文は第 1 文である。一般的に新聞やニュースでは第 1 文が特に重要とされているが、本実験では形式的な第 1 文に得点を与えないことで第 1 文以外の文が意味的により重要な場合も正しい評価を行うことができる。

表 1: 重要文抽出実験結果の例

タイトル	式根島で震度 5 弱
順位	CC の文
1 位	気象庁の観測によりますと、今日午後 5 時 44 分頃新島神津島近海の深さ 10 キロの所を震源とする、マグニチュード 4.9 の地震があり、式根島で震度 5 弱、新島と利島で震度４、神津島と伊豆大島で震度３を観測しました。
2 位	伊豆諸島では地震活動が活発な状態と落ち着いた状態を繰り返していて、けさ 6 時 52 分にも式根島で震度 5 強を観測したほか、夕方の強い地震の後も、式根島や神津島などで震度４や３の地震が続いています。
3 位	伊豆諸島ではきょう地震活動が活発になっていまして、けさ式根島で震度 5 強を観測したのに続いて、午後 6 時前にも式根島で震度 5 弱を観測する強い地震がありました。
...	
最下位	伊豆諸島の地震からお伝えします。

5 おわりに

本稿では、OC と CC の時間的・意味的な関連度を定量化することにより重要文を抽出する手法を提案し、実験によりその有効性を確認した。今後は重要度評価の精度向上、および語単位での 4W 情報の抽出について検討する。

参考文献

[1] 瀬戸喜巳, 井手一郎, 坂井修一, 田中英彦: “見出しからの制約による新聞記事からの重要文抽出”, 第 58 回情処全国大会, Vol.2, pp.73-74 (Mar 1999)

[2] 仲尾由雄: “見出しを利用した新聞・レポートからのダイジェスト情報の抽出”, 情処研報, NL-117-17 (Jan 1997)

[3] 日本放送協会: “NHK INFORMATION「最新! 技術情報」”, http://www.nhk.or.jp/pr/marukaji/m-giju028.html

[4] 茂木祐治, 有木康雄: “ニュース映像中の文字認識に基づく記事の索引付け”, 信学技報, PRU95-240 (Mar 1996)

[5] 京都大学大学院情報学研究科知能情報学専攻言語メディア研究室: “日本語形態素解析システム JUMAN 第 3.6 版”, http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/juman.html

[6] Ichiro Ide, Reiko Hamada, Shuichi Sakai, Hidehiko Tanaka: “Semantic analysis of television news captions reffering to suffixes” Proc. IRAL'99, pp.37-42 (Nov 1999)