1P-01

ニュース映像中の複数テキスト情報源からの重要語抽出*

井手一郎† 坂井修一‡ 田中英彦‡
ide@nii.ac.jp, {sakai,tanaka}@mtl.t.u-tokyo.ac.jp
† 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

ニュース映像への自動索引付けへの需要が高まるなか，筆者らは画像内容と索引の対応を重視し，画像とテキストの属性毎の対応による索引付けを行っている．そのためには，索引候補のテキストから，属性（一般にニュース映像の検索に必要と考えられる，いわゆる 4W：人物（Who），場面（Where），時相（When），行為（What）付きの重要語抽出が必要だが，従来 [1] 情報源として利用してきたオープンキャプション（テロップ）のみでは情報量が不足していた．一方，主音声やそれをディジタルテキスト化したクローズドキャプション（字幕放送）は，そのまま索引候補を抽出するには冗長である．そこで本稿では，索引候補の不足を補うために，オープンキャプションを手がかりにして，クローズドキャプションから属性付きの重要語抽出を行う手法を提案する．

2 ニュース映像からの重要語抽出

ここでは，ニュース映像中のテキスト情報源の種類と性質についてまとめた後に，それらを複合的に利用した属性付きの重要語抽出手法を紹介する．

2.1 ニュース映像中のテキスト情報源

ニュース映像に含まれるテキスト情報源としては，キャプション（字幕）と，画像中の看板や書類に記されているテキストがある．後者は認識が困難なほか，必ずしも画像内容を代表するものではないため，本研究では利用しない．キャプションは，画像上に挿入されるタイミングの違いにより，以下の 2 種類に分類され，異なる特徴をもつ．

オープンキャプション（OC）

送信側で画面上に画像として重ね合わせて表示されるもので，重要な情報を簡潔に示す．内容に基づき，表 1のように分類される．

- 画像内容と関連が強い重要な情報が多い．
- 体言止めや助詞の省略などによる簡潔な表記．

表 1: オープンキャプションの種別と出現割合（ニュース映像 370 分中の 2,842 件を人手で分類したもの）

種別		割合	特徴
(1)	場所・組織	28.11%	名詞列
(2)	人物	13.19%	名詞列
(3)	タイトル	9.43%	話題の冒頭に出現，半文章的
(4)	時相	7.00%	名詞列
(5)	発言	5.52%	要約・翻訳，文章的
(6)	放送技術	3.52%	名詞列，内容と無関係
(7)	描写	1.55%	名詞句
(8)	その他	31.68%	

- 単位時間あたりの情報量が少ない[1] ．

クローズドキャプション（CC）

走査線の隙間に挿入されるデータが受信側でテキスト化されて表示されるもので，主に主音声を書き下したもの．

- 画像内容と具体的に関係ない情報も多く，冗長．
- 原則として音声を書き下した文章．
- 単位時間あたりの情報量が多い．

また，主音声や副音声を書き下したものもテキスト情報源としての利用が考えられるが，ここでは主音声の一部（アナウンサの原稿朗読箇所）を書き下した CC のみを利用し，副音声はニュースにおいては一般に英語翻訳音声であるために利用しない．

2.2 複数テキスト情報源からの重要語抽出

OC の半数程度は 4W に関する内容を端的に表しているが（表 1の（1），(2），(4），(7）が相当），ショット単位への索引付けを行うには量的に不十分である．そこで，OC を手がかりに，不足する 4W に関する属性付き重要語を CC 中から抽出することを考える．

2.2.1 関連研究

OC と CC を組み合わせた手法として，佐藤らによる英語ニュース映像に対する OC の誤認識訂正及び重要語抽出手法 [3] がある．この手法では，OC と CC に共通して出現する語（類義語を含む）を重要語として抽出しているが，不足する情報を相補的に抽出していないほか，文単位の重要度評価による冗長な情報の排除は行っていない．

また，タイトル等を利用した本文からの重要文抽出手法 [4, 5, 6] も存在するが，属性レベルでの他メディアの相補的利用や，時間的な同期を考慮したものは見られない．

[1] 著者らの統計で 6〜8 件/分，国立国語研究所の調査 [2] に基づく推計で 9〜10 件/分程度．OC が豊富な聴覚障害者対象のニュースにおいては，主音声との文字比で 4〜7 割だが [7] ，通常はこれよりも大幅に少ない．

* "Keyword extraction from various text sources in news video"
† Ichiro IDE:
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
‡ Shuichi SAKAI, Hidehiko TANAKA:
Graduate School of Engineering, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

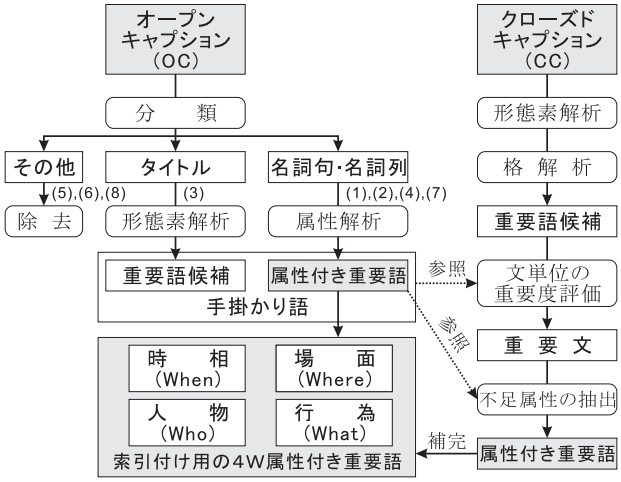

図 1: OC と CC の相補的利用による属性付き重要語抽出

2.2.2　OC と CC の相補的利用による重要語抽出

図 1 に示す機構により，OC と CC を相補的に利用した属性付き重要語抽出を行う．

まず，OC について以下の処理を行う：

- **分類**

  表 1 の分類に基づき，以下の 3 通りに分類する：

  - 名詞句・名詞列 … (1),(2),(4),(7) に相当
  - タイトル … (3) に相当
  - その他 … (5),(6),(8) に相当 (利用せず)

  分類の際には，表1 に示した特徴を利用するほか，文字飾りの存在などの出現形態 [8] も参考にする．

- **タイトルからの重要語抽出**

  タイトルは話題 (シーン) の内容を最も端的に表した文章である．そこで，話題中の CC から重要文・重要語を抽出するための重要語を抽出しておく．

- **名詞句・名詞列の属性解析**

  (6) を除くほとんどの名詞句・名詞列の OC は (1),(2),(4),(7) のいずれかに属する．OC は画像内容と関連が強い重要な情報を示すため，属性を解析することにより，4W に関する属性付き重要語が抽出できる．名詞句・名詞列の属性解析には，末尾の名詞に基づく解析手法 [1] を利用する．

以上の処理から得られる属性付き重要語のみではショット単位での索引付けを行うには不十分である．そこで，OC から得られる手がかり語を用いて，不足している属性の重要語を CC から抽出して補完する．そのために，CC について以下の処理を行う：

- **重要語候補の抽出**

  予め重要語候補を抽出し，格解析を行っておく．

- **OC からの手がかり語による文単位の重要度評価**

  OC からの手がかり語は 1 シーン (ニュースの場合，話題に相当) 内のみ有効とし，以下の 2 通りを用いる：

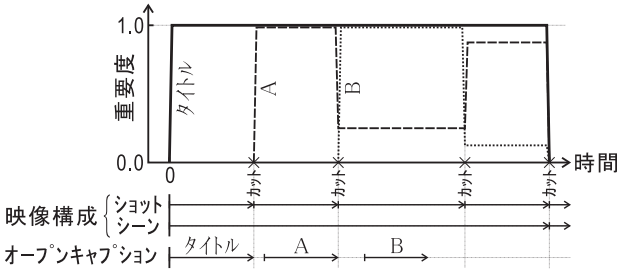

図 2: 画像的類似度に基づく OC の重要度の時間的変化

  - タイトルから抽出される重要語候補
    シーン内の重要度を一定 (1.0) に設定．CC 文の重要度評価の際に格を考慮せず．
  - 4W の属性付き OC (属性付き重要語)
    重要度はその OC の出現ショット内では 1.0，以降のショットでは出現ショットとの画像特徴の類似度に応じた値に設定．CC 文の重要度評価の際に，属性に応じて格を考慮．

  図 2 に OC からの手がかり語の重要度に関する時間的変化の例を示す．このように時間的に変化する重要度付きの手がかり語と，CC から得られる重要語との比較により，CC 文の重要度を決定する．

- **属性付き重要語の抽出**

  以上のように得られた重要文中から，格などを考慮して，不足していた属性の重要語を抽出する．

このようにして，不足する属性付き重要語を補完する．

3　おわりに

本稿では，OC から得られる情報のみでは不足する属性の重要語を，OC の出現形態や属性を考慮して CC から抽出する手法を提案した．今後は，未実装部分の実現手法の検討及び実装・評価を行う．

参考文献

[1] Ide, I., Hamada, R., Sakai, S., and Tanaka, H.: "Semantic analysis of television news captions referring to suffixes", *Proc 4th intl workshop on information retrieval with Asian languages*, pp.37-42 (Nov 1999).

[2] 国立国語研究所："テレビ放送の語彙調査 I –方法・標本一覧・分析–"，国立国語研究所報告，Vol.112 (1995)．

[3] 佐藤，金出："文字認識と異種情報の対応関係に基づいたニュース放送からの情報抽出"，情処論文誌，Vol.40，No.12，pp.4266-4276 (Dec 1999)．

[4] 瀬戸，井手，坂井，田中："見出しの制約による新聞記事からの重要文抽出"，第 58 回情処全大，Vol.2，pp.73-74 (Mar 1999)．

[5] 仲尾："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)．

[6] 吉見，奥西，山路，福持："表題へのつながりに基づく文の重要性"，自然言語処理，Vol.6，No.1，pp.43-57 (Jan 1999)．

[7] 若尾，江原，白井："テレビニュース番組の字幕に見られる要約の手法"，情処研報，NL-122-13 (Nov 1997)．

[8] 渡辺，岡田，長尾："TV ニュースで用いられるテロップの意味解析"，情処研報，NL-116-16 (Nov 1996)．