6J-04

末尾の名詞に着目した TV 放送中の字幕の意味属性解析手法

井手 一郎, 田中 英彦
{ide,tanaka}@mtl.t.u-tokyo.ac.jp
東京大学大学院 工学系研究科 *

1 はじめに

昨今の放送媒体の多様化などに伴い、日々提供される映像情報は増加している。しかし、これらの情報の検索や再利用に必要となる自動的な索引付けは、いまだ実用化の域に達していない。本稿では、このような自動的索引付けの際に必要となる字幕の意味属性の解析手法について論じる。具体的には、末尾に存在する名詞に着目して、字幕が人, 場所, 日時のいずれを指すかを解析する手法を紹介する。既に末尾に存在する名詞を利用して未知語 (人名) を抽出する試み [2] は行われているが、本手法ではその手掛かりとして用いられている末尾に存在する名詞そのものを抽出する。なお、索引付けを行う対象としては、TV 番組の中でも特に検索や再利用の需要が高いと思われるニュース番組を扱う。

2 ニュース番組中の自然言語情報

TV 番組中の自然言語情報には字幕, 文字放送, 主音声, 副音声のようなものがある。映像の索引付けにあたって、将来的にはこれらの情報を総合的に利用することを考えているが、現時点ではキーワードになり得る重要な語が最も密に含まれる字幕を用いることにする。

2.1 字幕の文法的特徴

一般に字幕は、単なる名詞の連なりや、省略が多い文であることが多く [3]、通常の自然言語処理が対象とする文と較べて文法的に特殊である。このため、字幕や新聞の見出しのような特殊な自然言語情報を処理する手法の確立が望まれる。本稿でも、このような特徴を考慮した解析を行う。

2.2 字幕の意味的特徴

字幕には装飾や変形の激しいものも存在するので、本稿ではそのような極端なものを除いたものを扱う。なお、以下の議論では、ある 30 分のニュース番組中に登場する全 134 件の字幕について解析する。

表 1に字幕を人手で分類したものを示す。このうち、1, 2, 5 はそのまま、3, 4, 7 は名詞を切り出すことによって、即キーワード候補となり得る。前者だけでも、全字幕のうち 62%がキーワード候補として直接利用可能である。しかし、意味属性に関する情報をもたないキーワード候補は索引付けに利用しにくいため、以下で議論するように、字幕が (1) 人, (2) 場所, (3) 日時, (4) その他のいずれを表すのかを解析することが必要である。

1	撮影場所	27%
2	映像中の人物	23%
3	その他 (タイトルなど抽象的なものを含む)	13%
4	発言 (要旨や翻訳を含む)	13%
5	撮影日時	12%
6	技術的情報 (*ex.* 「中継」)	9%
7	映像内容の描写	3%

表 1: TV ニュース中の字幕の分類 (一部 [4])

3 字幕の意味属性の解析

ここでは、自然言語的な解析により字幕に対して (1) 人, (2) 場所, (3) 日時, (4) その他 の意味属性の付与を行う。従来も、一般の文章に対して文脈や格の解析を行ったり [1]、字幕の出現位置の情報を利用したり [3] して同様の解析を行う試みがなされているが、後者に関しては番組依存なので汎用性の問題があり、前者に関しては **2.1** で記したような字幕の性質上、そのような解析は行えないので、各々の字幕単独、特に末尾に存在する名詞に注目することにした。

たとえば、「橋本」だけではこの名詞が人か場所かは人間でも分らないが、「橋本**首相**」ならば人であり、「橋本**市**」ならば場所であることが分る。このように、多くの場合末尾に存在する名詞 (接尾語を含む) により字幕全体の意味属性が確定する。

そこで、RWC テキストデータベース中の、人手による修正済みの形態素解析された文からなるコーパスで

---

* "A Method to Analyze Semantic Attributes of TV Captions by Suffixes"
Ichiro Ide, Hidehiko Tanaka
University of Tokyo, Graduate School of Engineering
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113 Japan

ある、RWC-DB-TEXT-95-2 を利用して、末尾に存在して (1) 人, (2) 場所 の意味属性を決定する名詞を収集した。同コーパスは 1994 年 1 年間の毎日新聞の記事から適当な 2,000 文を選択して形態素解析したものであり、ニュース番組の字幕に登場するものと近い語彙が多く含まれていると考えられる。

しかし、これだけの語彙では字幕を一般的に解析するには不十分なので、分類語彙表を利用することとし、収集した単語が項目中に含まれるものに含まれる全単語を末尾に存在して人を示す名詞の辞書とすることにした。

なお、(3) 日時に関しては、収集するまでもなく、年, 月, 日, 時, 分, 秒が数字か一部の接頭辞 (昨, 翌 など) の後に続く場合に日時を示すとみなす。

3.1 末尾に存在して人を示す名詞の収集

以下の手順で末尾に存在して人を示す名詞を RWC-DB-TEXT-95-2 から収集した。

1. 人のみに付き得る接尾辞**ら**か**たち**を探す
2. 発見されたら、その前の単語が**普通名詞**か**接尾辞**であれば、その名詞を収集する

たとえば、「橋本 (固有名詞) 首相 (名詞) ら (接尾)」からは「首相 (名詞)」が収集される。

このようにした結果、260 種の名詞が収集された。次に、260 種の名詞のうち明らかに不適なもの 11 種を除去し、さらに分類語彙表に含まれるもの 178 種が含まれる項目中の全単語を収集して、1,999 種の名詞の辞書を得た。

この辞書を、**2.2** で挙げた字幕を JUMAN により形態素解析したものの集合に対して適用したところ、88%の適合率、81%の再現率が得られた。誤検出の原因は、**形態素分割の失敗**と**語の多義性**であり、検出洩れの原因は、**固有名詞単独での解析不能**と**辞書の語彙不足**であった。

3.2 末尾に存在して場所を示す名詞の収集

以下の手順で末尾に存在して場所を示す名詞を RWC-DB-TEXT-95-2 から収集した。

1. **地域を表す固有名詞**を探す
2. それに続く名詞を辿る
3. 名詞が途切れて、次に場所を表す格助詞**から**, **で**, **に**, **へ**, **より**, **にて**があれば、最後の名詞を収集する

たとえば、「横田 (地域) 町 (接尾) で (格助詞)」からは「町 (接尾)」が収集される。

このようにした結果、389 種の名詞が収集された。次に、人を示す名詞と重複するもの (*ex.* 「アメリカ (地域) 大統領 (名詞) より (格助詞)」で抽出されてしまう「大統領 (名詞)」など)、および明らかに不適なもの 186 種を除去し、さらに分類語彙表に含まれるもの 141 種が含まれる項目中の全単語を収集して、2,572 件の名詞の集合を得た。

この辞書を、**2.2** で挙げた字幕を JUMAN で形態素解析したものの集合に対して適用したところ、83%の適合率、58%の再現率が得られた。誤検出の原因は、**分類語彙表の分類との不整合**であり、検出洩れの原因は、**固有名詞単独での解析不能**と**形態素解析の失敗**であった。

4 おわりに

本稿では、TV ニュースデータベースの索引付けに必要となる、字幕の意味属性解析の手法を提案し、実際の字幕に適用して有効性を評価し、誤検出や検出洩れの原因について考察した。その結果、字幕が人を示す場合は、ほぼ実用的な精度が得られた。一方、場所を示す場合は、実用的な精度は得られておらず、末尾に存在し得る名詞の収集法に改良の余地がある。

また、本稿では触れていないが、今後はタイトルや発言要旨などの字幕からのキーワード抽出の研究も行う。なお、本研究の全体像については、[5] を参照されたい。

謝辞

JUMAN は京都大学長尾研究室と奈良先端科学技術大学院大学松本研究室にて開発されたフリーソフトウェアであり、分類語彙表は国立国語研究所の成果物である。RWC テキストデータベースは技術研究組合新情報処理開発機構の成果物であり、同機構の許可の下に利用した。

本研究の遂行にあたって、角田達彦博士には全般にわたる適切な助言を、永松健司氏には自然言語処理に関する思慮深い助言をいただいたことを深く感謝する。

参考文献

[1] 那須川 哲哉; **文脈情報を利用したキーワード語義決定**; 人工知能学会第 11 回全国大会,17-01; 予稿集 pp.348-349; *Jun., 1997.*

[2] 久光 徹, 丹羽 芳樹; **辞書と共起情報を用いた新聞記事からの人名獲得**; 情報処理学会技術研究報告,97-NL-118-1; Vol.97, No.29, pp.1-6; *Mar., 1997.*

[3] 渡辺 靖彦, 岡田 至弘, 長尾 眞; **TV ニュースで用いられるテロップの意味解析**; 情報処理学会技術研究報告,97-NL-116-16; Vol.96, No.89, pp.107-114; *Nov., 1996.*

[4] 角田 達彦, 大石 巧, 渡辺 靖彦, 長尾 眞; **キャプションと記事テキストの文字列照合による報道番組と新聞記事との対応づけの研究**; 情報処理学会論文誌; Vol.38, No.6, pp.1149-1162; *Jun., 1997.*

[5] 井手 一郎, 田中 英彦; **画像・言語情報の統合的利用による映像データの自動的インデクシングの試み**; 言語処理学会第 3 回年次大会,A5-2; 予稿集 pp.485-488; *Mar., 1997.*