論　文

テレビニュース字幕の語義属性解析のための辞書作成

井手　一郎[†]　浜田　玲子[††]　坂井　修一[†††]　田中　英彦[†††]

Compiling dictionaries for semantic attribute analysis of television news caption

Ichiro IDE[†], Reiko HAMADA[††], Shuichi SAKAI[†††], and Hidehiko TANAKA[†††]

あらまし　日々放送される映像量の増加につれ，将来の再利用や検索を前提として，それらを整理して蓄積する必要性が高まっている．とりわけ，重要性と利用価値の点から，ニュース映像への索引付けに対する期待は大きい．映像中のテキスト情報を利用して適切な自動索引付けを実現するためには，従来手法で一般に採用されてきた単純な索引抽出・付与手法では不十分であり，索引候補の語義属性を考慮して取捨選択することが重要である．そこで本稿では，テレビニュース映像中の字幕（名詞句）の語義属性を解析するのに必要な辞書を作成するために，テキストコーパス及び類義語辞書から一定の条件に基づいて語を抽出して収集した過程を紹介する．また，実際のニュース映像中に出現する語の語義属性解析を通して，辞書の性能評価を行なった結果もあわせて紹介する．既存の固有名詞辞書及び時相名詞辞書を併せて用いた評価実験の結果，79〜93%の再現率，41〜71%の適合率が得られた．この結果において適合率は低いものの，索引候補を取捨選択する際には，再現率の方が重要であるため，作成した辞書を実際の索引付けの際の解析に利用するのは現実的であると考えられる．

キーワード　索引付け，字幕，語義属性，接尾名詞，辞書

1. まえがき

放送される映像量の増加につれ，それらを整理して蓄積し，再利用や検索に供する必要性が高まっている．なかでも，内容の重要性や利用価値の点から，ニュース映像への索引付けの需要は高い．現在，このような作業は主に人手で大雑把に行なわれているが，増加する量に対応し，きめ細かな検索要求に応え得る索引付けの粒度を確保するために，自動化への期待は高い．

筆者らは，画像情報とテキスト情報を統合的に利用することにより，ニュース映像への自動索引付けの実現を目指している．きめ細かな検索要求に応え得る索引付けを行なうためには，既存手法の多くに見られるような単純で大雑把な索引付けのみでは不十分であり，索引候補の取捨選択のための語義属性解析が重要となる．

たとえば，Carnegie Mellon大学におけるInformediaプロジェクト[18]のNews-on-Demandシステム[17]に代表されるように，このような統合メディア処理によるニュース映像への自動索引付けの様々な試みが行なわれている．しかし，それらの多くは，自動索引付けへの要請をある程度までは満たすものの，出現頻度などの統計情報に基づく索引抽出や，語句の出現タイミングに基づく単純な索引付けであり，索引と画像内容との対応を必ずしも保証しない．このような手法でも，話題単位の大雑把な索引付けには適用できるが，ショットなどのより粒度の細かい単位への索引付けにおいては，対応する範囲内の画像と索引が適当に対応しないことが考えられる．

筆者らは，このような問題点を背景として，図1に示すような，索引と画像内容の属性別対応を考慮した自動索引付けシステムの構築を目指している．図1において，索引付けは画像の属性及びテキストの属性が一致するもの同士の対応を考慮して行なわれる．このシステムの中の自然言語処理部において，索引候補となるテキストの語義属性の解析が必要となる．本稿では，このような目的のために必要となるニュース映像中の字幕（ニュース字幕）の語義属性解析のための辞

[†] 国立情報学研究所，東京都
National Institute of Informatics, 2–1–2 Hitotsubashi, Chiyoda-ku, Tokyo 101–8430, Japan
[††] 東京大学大学院工学系研究科，東京都
Graduate School of Engineering, The University of Tokyo, 7–3–1 Hongo, Bunkyo-ku, Tokyo 113–8656, Japan
[†††] 東京大学大学院情報理工学系研究科，東京都
Graduate School of Information Science and Technology, The University of Tokyo, 7–3–1 Hongo, Bunkyo-ku, Tokyo 113–0033, Japan

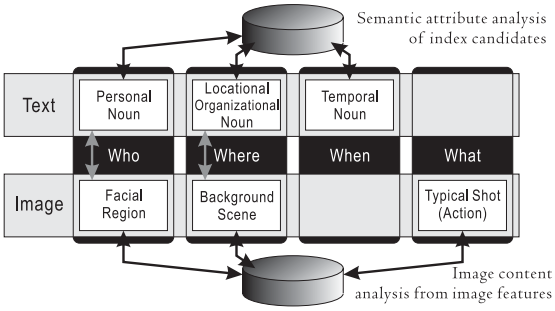


図1 索引と画像内容の属性別対応を考慮した索引付け
Fig. 1 Indexing considering correspondences between indices and image contents.

書の作成過程について述べ，実際のニュース字幕の解析を通じて，その性能を評価する．

なお，ここでいう語義属性解析とは，具体的には名詞句を4つの語義属性，すなわち,（1）人物,（2）場所・組織,（3）時相,（4）その他，に分類することと等価である．また，本稿ではニュース字幕の解析を主眼とした評価を行なうが，作成された語義属性別接尾名詞辞書は，固有表現の抽出など情報検索分野への応用や，他の類似した索引付け手法 [6], [12], [15] への適用，更にはより一般的な自然言語理解のための利用も考えられる．

以下，第2章では語義属性解析手法について，第3章では辞書の作成過程について，第4章では実際のニュース字幕への適用による評価について述べ，最後に第5章でまとめる．

2. 末尾の名詞に着目したニュース字幕解析

ここでは，まずニュース字幕の性質を示してから，辞書を用いた語義属性解析の手法を紹介する．

字幕（オープンキャプション）以外にも，音声，文字放送字幕（クローズドキャプション）など，映像には様々なテキスト情報が含まれる．一般に，これらの情報源には多くの情報が含まれるものの，内容的に冗長であり，重要な語句を抽出しなければならない．更には，音声認識技術の性能は向上しつつあるものの，音声を正確なテキスト情報源として利用し得る水準には達していない．文字放送字幕はこのような音声認識技術の欠点を補い得る可能性があり，実際に欧米では広く普及している．しかし，日本ではニュース映像に対しては，ようやく特定話者（アナウンサ）に限定した音声認識による部分的な提供が始まったばかりであり [1]，いまだ本格的に利用できる段階にはない．

一方，字幕は重要な情報を簡潔に表現しており，他の情報源に対しては必要となる重要語句抽出の処理が省ける．しかし，背景の画像の上に重なって表示されること，走査線の本数の少なさ（NTSC放送方式の場合で525本）に比例して文字あたりの解像度が低いことにより，既存のOCR（光学的文字認識）手法を単純には適用できない．そのため，字幕に対する文字認識に特化したいくつかの研究が行なわれており [3], [9]，一定の成果が得られている．

以上のような理由により，ここでは字幕を索引付けのためのテキスト情報源として直接利用することを考え，その解析のために必要な辞書の作成について述べる．なお，以下の実験では人手で書き下した字幕を利用しているが，将来は上記のような様々な研究成果の利用や，データ放送などの手段により入手することを考えている．

2.1 字幕の言語的性質

字幕は他の一般のテキストと比較して，文法的にも内容的にも特殊な性質を有する．

まず文法的には，体言止めや名詞列などの形をとる名詞句がほぼ半数を占める．ここで属性分類を行ないたい，人物，場所・組織，時相を表す字幕はほとんどがこのような名詞句であるため，名詞句を解析対象とする．

また内容的には，人物，場所・組織，時相を表す字幕は全体の48.3%，1ショットあたり0.39件存在し(注1)，索引付けするのに比較的十分な情報量が得られる．

2.2 名詞句の語義属性解析

前節で述べたような言語的性質を考慮し，名詞句からなる（1）人物,（2）場所・組織,（3）時相に関する字幕を解析する．ここで，通常この種の解析では独立したものとして扱われる場所と組織とを分けなかったのは，映像に対する索引付けにおいて，画像内容との対応を考慮する際，組織名が画像中の具体的な場所を示すこと（例：「〜銀行」が組織を示すのと同時に銀行の建物そのものの画像と対応する）が多く，実質的に場所と明確に区別する必要性が少ないためである．

ここではまず，既存の名詞（句）解析手法として，本研究の目的にある程度適用可能な関連研究を紹介し，次にここで採用する字幕の特徴を考慮した語義属性解析手法について述べる．

(注1)：370分間のニュース映像に出現した2,842件の字幕を人手で分析した結果

表1　接尾名詞による語幹の語義決定の例
Table 1　Example of determination of stem's meaning by the suffix.

森			…	森	［名詞-？］			
森	＋	首相	…	森	［名詞-固有名詞-人名］	＋	首相	［名詞-普通名詞］
森	＋	町	…	森	［名詞-固有名詞-地名］	＋	町	［名詞-普通名詞］

2.2.1　関 連 研 究

類似した語義属性解析問題として，英語を対象としたMessage Understanding Conference（MUC）や日本語を対象としたInformation Retrieval and Extraction Exercise (IREX) の課題として定義された，固有表現抽出（Named entity task [16], [20]）がある．この課題では，参加者に提示されたテキストに対して，人物，場所，組織，時相，数値など（このうち前三者は固有名詞に限定）に関する表現を抽出してタグ付けすることを課している．

このような課題に対して，たとえば一般的なテキスト中において，文脈を解析することにより固有名詞の語義を解析する手法 [14] などがある．これらの課題や手法は，人物，場所，組織に関する表現の抽出及び分類を行なう点においては本研究で必要とする技術に近いが，いずれも固有表現，なかでも固有名詞に解析対象を限定している点で異なる．このような固有名詞のみの解析と，普通名詞のみからなる一般の名詞句を含めた解析とでは，本質的に異なる困難さをはらむ．補助的な情報なくして,「森」などの固有名詞単体でその語義を決定することは困難であり，那須川の手法 [14] や固有表現抽出の様々な手法のように，文内の格構造や周囲の文の文脈から推測することになる．

更に，渡辺らはニュース映像における字幕の出現位置と文法的特徴を考慮した字幕解析手法を提案しているが [19]，出現位置を含めたデザインは番組により異なり，汎用性に欠けるという問題がある．

2.2.2　接尾名詞に基づく解析

以上の字幕の性質と既存手法の問題点を考慮し，ここでは，周囲の文脈や格構造を参照できない字幕に対して，単独で語義属性を解析する手法を採用する．

一般に特定の名詞が普通名詞であるか固有名詞であるか，固有名詞の場合でも人名，地名，組織名のいずれであるかを，文脈を考慮せずにその名詞単独から判断することは人間でも著しく困難である．実際には，表1に例示するように，末尾の名詞（接尾名詞）により語幹及び名詞句全体の語義が決定されると考えられる．この例では,「森」単独では語義を決定できないが,

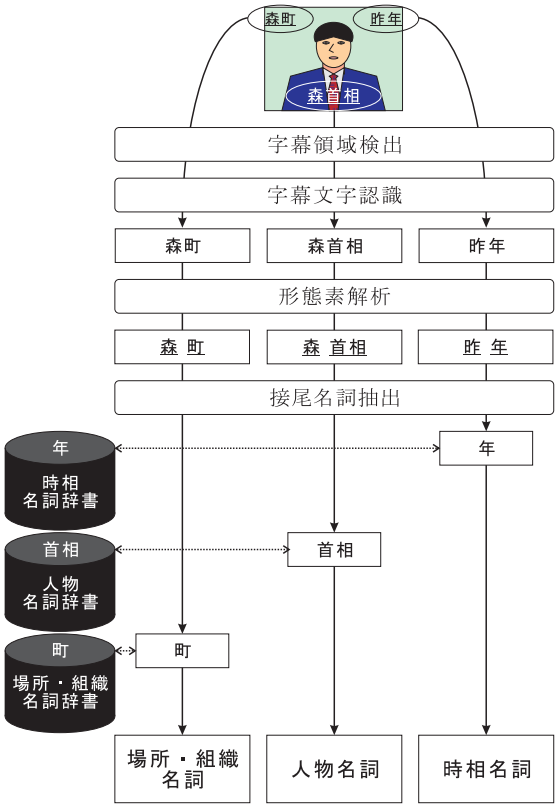


図2　接尾名詞に注目した字幕の語義属性解析の例
Fig. 2　Example of caption analysis referring to suffixes.

「首相」と結合することにより人名となることを,「町」と結合することで地名となることを示している．このように，一般に日本語において，接尾名詞を解析することにより名詞句全体の語義を決定できる．このような仮定に基づく研究は，固有名詞を含む名詞句の認識においても行なわれている [7], [13]．本研究では，このような性質を仮定したうえで，接尾名詞に注目することにより，名詞句からなるニュース字幕の語義属性を解析する．

図2にこのような方針による字幕の語義属性解析の例を示す．まず，字幕に形態素解析を施すことにより接尾名詞を切り出し，切り出された接尾名詞を語義属

性別接尾名詞辞書中の語と比較することにより解析する．

なお，この例や以下の議論では接尾名詞に重点をおくが，本手法では固有名詞を含む名詞句のみならず，「俳優」や「台所」のように，単独で人物や場所を示す普通名詞をも収集，解析対象とする．

3. 語義属性別接尾名詞辞書の作成

ここでは，前章で述べた接尾名詞に注目したニュース字幕の語義属性解析手法を実現するために必要となる語義属性別接尾名詞辞書の作成について述べる．具体的には,（1）人物,（2）場所・組織,（3）時相を示す接尾名詞を集めた辞書を各々作成する．

前述のように，純粋な接尾名詞以外に,「俳優」,「台所」,「今日」のように単独で人物，場所・組織，時相を示す名詞も収集対象とする．なお，固有名詞は，単独では語義決定が困難であることから，収集対象から除外した．

辞書の作成にあたり，以下の2つのテキストコーパスを利用した．これらのテキストコーパスは人手で形態素解析された結果が本文にタグ付けされており，形態素の区切りや種別を正しいものとみなして利用した．

- RWC テキストデータベース（第2版）[2] 中の RWC-DB-TEXT-95-2

1994年版の毎日新聞記事の27,418文からなる．

- 京都大学テキストコーパス（第2版）[11]

1995年版の毎日新聞記事の19,956文からなる．

これらを合わせた47,374文から一定の条件のもとに接尾名詞を抽出した．

以下の各項では，各々の語義属性別接尾名詞辞書の作成手順と収集結果について述べる．

3.1 人物名詞の収集

本節では,「人物名詞」，すなわち,「〜博士」のように人物を示す接尾名詞や「俳優」のように単独で人物を示す普通名詞の収集手順と結果について述べる．

3.1.1 コーパスからの抽出

人物名詞を表2に示す基準に従ってコーパスから抽出した．ここでは，接尾辞「ら」及び「たち」が通常複数の人間（のみ）を表すのに用いられることを利用した．

また，表2の最下段にこの基準による抽出例を，表3に抽出された語の一部を頻度の高い順に列挙する．

3.1.2 類義語辞書による語彙の拡張

表2に示した基準によるコーパスからの抽出により136件の接尾名詞が抽出されたものの，汎用的なニュース字幕解析に適用するには語彙が不十分である．そこで，類義語辞書を用いて語彙を拡張した．拡張にあたっては，類義語辞書のなかで，抽出された語が含まれるものと同一分類に含まれる語は全て該当する語義分類に属すると判断した．

類義語辞書としては,「『分類語彙表』形式による語彙分類表（増補版）」[8]（以下，新分類語彙表）を用いた．新分類語彙表は，87,743語（異なり数70,858語）を4大分類項目の下の842小分類項目，10,334分類段落に分類した概念分類辞書である．最小分類単位である分類段落には，平均6.8個の語が含まれる．

表4に辞書の作成過程で得られた語彙数の推移を示す．最終的に，1,795件の人物名詞が収集された．

表3 抽出された人物名詞（上位30件）
Table 3 List of extracted personal nouns (Top 30).

順位	抽出された語	頻度	順位	抽出された語	頻度
1	者	75	14	書記長	8
2	氏	57	14	客	8
3	さん	51	14	員	8
4	会長	32	19	代表	7
5	長	26	20	容疑者	6
6	議員	21	20	職員	6
7	人	16	20	助教授	6
7	教授	16	20	局長	6
9	家	15	20	関係者	6
10	幹部	14	20	官	6
11	首相	13	20	委員長	6
12	被告	10	27	大使	5
13	業者	9	27	社長	5
14	長官	8	27	幹事	5
14	相	8	27	課長	5

表4 収集された人物名詞の語彙数の推移：分類項目及び分類段落は新分類語彙表中語が属するものの数を示す
Table 4 Number of collected personal nouns: Numbers of classes and sub-classes indicate those that extracted words listed in the thesaurus belong to.

	コーパスからの抽出	新分類語彙表による拡張
新分類語彙表中語	117	1,776
（分類段落数）	125	125
（小分類項目数）	59	59
新分類語彙表外語	19	19
総語彙数	136	1,795

3.2 場所・組織名詞の収集

本節では,「場所・組織名詞」，すなわち,「〜駅」のように場所や組織を示す接尾名詞や「台所」のように単独で場所や組織を示す普通名詞の収集手順と結果について述べる．

表2　コーパスからの人物名詞抽出条件：以下の語と形態素に関するパターンを同時に満たす単語列から，下線部の語を抽出する.「＊」は，該当箇所の単語及び形態素が何でも良いことを示す．RWC コーパスと京大コーパスでは形態素体系が異なるため，各々における条件を示す．抽出例は日本語形態素解析システム JUMAN により形態素解析した結果を示す.

Table 2　Personal noun extraction rules: Extract the underlined word from a sequence that matches the following patterns of both words and morphemes. “＊” denotes that any word or morpehme is applicable. Patterns for both RWC corpus and Kyoto University corpus are shown, since they have different morphological systems. The extraction example shows the result from automatic morphological analysis performed by the Japanese morphological analysis system JUMAN.

語	＊ …	＊	「ら」 「たち」
RWC形態素	［名詞-＊］ …	［名詞］ ［名詞-非自立-＊］ ［名詞-接尾］	［名詞-接尾］
京大形態素	［名詞-＊］ ［接尾辞-名詞性名詞接尾辞］ …	［名詞-普通名詞］ ［接尾辞-名詞性名詞接尾辞］	［接尾辞-名詞性名詞接尾辞］
抽出例	田中 一郎 ［名詞-人名］ ［名詞-人名］	教授 ［名詞-普通名詞］	ら ［接尾辞-名詞性名詞接尾辞］

表5　コーパスからの場所・組織名詞抽出条件
Table 5　Locational/organizational noun extraction rules.

語	＊	＊ …	＊	「から」 「で」 「に」 「へ」 「より」 「にて」
RWC形態素	［名詞-固有名詞］ ［名詞-固有名詞-組織］ ［名詞-固有名詞-地域-＊］	［名詞-＊］ …	［名詞］ ［名詞-非自立-＊］ ［名詞-接尾］	［助詞-格助詞］
京大形態素	［名詞-地名］ ［名詞-組織名］	［名詞-＊］ ［接尾辞-名詞性名詞接尾辞］ …	［名詞-普通名詞］ ［接尾辞-名詞性名詞接尾辞］	［助詞-格助詞］
抽出例	東京 ［名詞-地名］	大学 工学部 ［名詞-普通名詞］ ［名詞-普通名詞］	構内 ［名詞-普通名詞］	へ ［助詞-格助詞］
人物名詞混入例	東京 ［名詞-地名］	大学 工学部 ［名詞-普通名詞］ ［名詞-普通名詞］	教授 ［名詞-普通名詞］	へ ［助詞-格助詞］

3.2.1　コーパスからの抽出

場所・組織名詞は表5に示す基準に従ってコーパスから抽出した．ここでは，格助詞「から」,「で」,「に」,「へ」,「より」,「にて」が場所や方向を示すのに用いられることを利用した.

また，表5の下から2段目に，この基準による抽出例を示した.

3.2.2　人物名詞の削除

表5に示した基準は緩やかなため，これに従った抽出では，最下段に例示するように，本来は人物名詞である語が混入してしまう．これは，たとえば，格助詞「へ」が抽出例では方向を示すのに対し，人物名詞混入例では目的格の人物を指すことに示されるように，助詞の用法の多様性によるものである．このようにして混入した人物名詞を除去するために，ここで抽出された語から，3.1節で抽出された人物名詞を削除する.

表6に人物名詞を削除した後の，抽出された語の一部を頻度の高い順に列挙する.

3.2.3　類義語辞書による語彙の拡張

人物名詞の削除後に696件の接尾名詞が残ったものの，汎用的な字幕解析に適用するには語彙が不十分である．そこで，3.1節で述べたように，類義語辞書を用いて語彙を拡張した.

表7に辞書の作成過程で得られた語彙数の推移を示す．最終的に，7,908件の場所・組織名詞が収集された.

3.3　時相名詞収集

本節では,「時相名詞」，すなわち,「〜後」のように時相を示す接尾名詞や「今日」のように単独で時相を

表6 抽出された場所・組織名詞（上位30件）
Table 6 List of extracted locational/organizational nouns (Top 30).

順位	抽出された語	頻度	順位	抽出された語	頻度
1	市	163	16	会談	28
2	内	108	16	選	28
3	側	78	18	地裁	27
4	大会	60	19	町	25
5	県	46	20	市場	23
6	市内	43	20	間	23
7	駅	39	22	地区	22
7	政府	39	22	会	22
7	戦	39	24	館	20
10	問題	38	25	沖	19
11	署	36	25	会議	19
12	地方	31	25	地域	19
13	区	30	28	村	18
14	大震災	29	28	場	18
14	国内	29	28	軍	18

表7 収集された場所・組織名詞の語彙の推移
Table 7 Number of collected locational/ organizational nouns.

	コーパスからの抽出	人物名詞の削除	新分類語彙表による拡張
新分類語彙表中語	674	607	7,819
（分類段落数）	764	697	697
（小分類項目数）	318	307	307
新分類語彙表外語	91	89	89
総語彙数	765	696	7,908

表8 コーパスからの時相名詞抽出条件
Table 8 Temporal noun extraction rules.

語	＊	「から」 「に」 「より」
RWC **形態素**	［名詞-副詞-＊］ ［名詞-非自立-副詞可能-＊］	［助詞-格助詞］
京大形態素	［名詞-副詞的名詞］	［助詞-格助詞］
抽出例	ころ ［名詞-副詞的名詞］	から ［助詞-格助詞］

示す普通名詞の収集手順と結果について述べる．

3.3.1 コーパスからの抽出

時相名詞は表8に示す基準に従ってコーパスから抽出した．ここでは，格助詞「から」,「に」,「より」が時相を示すのに用いられることを利用した．また，表8には記していないが，京都大学テキストコーパスに限って,［名詞-時相名詞］の形態素をあらかじめ付与された全単語も併せて抽出した．

また，表8の最下段にこの基準による抽出例を，表9に抽出された語の一部を頻度の高い順に列挙する．

3.3.2 類義語辞書による語彙の拡張

表8に示した基準によるコーパスからの抽出により

表9 抽出された時相名詞（上位30件）
Table 9 List of extracted temporal nouns (Top 30).

順位	抽出された語	頻度	順位	抽出された語	頻度
1	ため	465	16	ほか	30
2	時代	145	17	ころ	27
3	中	128	18	戦後	25
4	間	123	19	当時	24
5	前	117	19	時期	24
6	時	116	19	まま	24
7	ところ	79	22	とき	23
8	うち	65	23	結果	22
9	際	54	24	今年	20
10	場合	47	24	現在	20
11	上	46	26	度	16
12	直後	45	26	直前	16
13	以来	43	26	長期	16
14	日	40	29	以降	15
15	後	31	29	なか	15

表10 収集された時相名詞の語彙の推移
Table 10 Number of collected temporal nouns.

	コーパスからの抽出	新分類語彙表による拡張
新分類語彙表中語	136	2,124
（分類段落数）	213	213
（小分類項目数）	104	104
新分類語彙表外語	20	20
総語彙数	156	2,144

156件の接尾名詞が抽出されたものの，汎用的な字幕解析に適用するには語彙が不十分である．そこで，3.1節で述べたように，類義語辞書を用いて語彙を拡張した．

表10に辞書の作成過程で得られた語彙数の推移を示す．最終的に，2,144件の時相名詞が収集された．

4. ニュース字幕の解析による辞書の性能評価

前章で作成した語義属性別接尾名詞辞書を用いて，実際のニュース字幕を解析し，辞書の性能評価を行なう．

4.1 実験条件

字幕解析は，図2に示した手順に従って行なった．本稿では，字幕領域検出と文字認識は目視と人手により書き下し，形態素解析には日本語形態素解析システムJUMAN [10] を用いた．JUMANは形態素の境界とその属性または品詞を出力するので，ここでは解析対象の字幕の末尾が名詞であれば，各語義属性別接尾名詞辞書と比較して解析した．

一方，JUMAN自身も固有名詞辞書及び時相名詞辞書を内蔵しており，末尾に人名，地名，組織名の固有

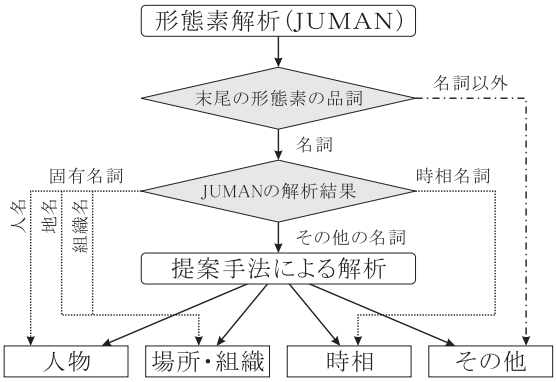


図3　JUMANと提案手法を組み合わせた解析処理
Fig. 3　Analysis process by combination of JUMAN and the proposed method.

名詞が存在する字幕，及び一部の時相名詞に関してはJUMAN の出力結果を解析結果として優先的に採用し，JUMANによって固有名詞とも時相名詞とも判定されなかった名詞のみを対象として提案手法により解析した．このような各辞書の利用手順について，図3に示す．

実験に用いた字幕は約370分間のニュース映像に出現した2,546件である．なお，話題の冒頭のタイトルなど名詞句ではない字幕と，CGやフリップによる模式図などに現れる字幕については，実験対象から除外した．これは，これらの字幕は直接画像内容を描写することが少ないためであり，これらは画像認識や自然言語処理により容易に弁別できる．

なお，評価の際の正解は第三者によって与えられたものを用いた．複数の属性の可能性があると解析された場合は，正解が含まれていれば正答と判断した．

4.2　実験結果と考察

以上の条件のもとに実験を行なった結果を示し，考察を述べる．

4.2.1　実 験 結 果

解析結果を表11，12，13に示す．表中,「提案手法による解析」に示す数値は，図3中の提案手法による解析処理の結果を,「JUMANによる解析」に示す数値はJUMANによる解析結果を示す．前述のように，末尾に人名，地名，組織名の固有名詞が存在する字幕，及び一部の時相名詞に関してはJUMANにより解析される.「総合」は両者により解析された総合的結果を示す．

以下のように定義される，適合率と再現率を評価尺度として用いた．各々の解析結果に対する適合率は，

表11　作成した辞書による字幕解析結果（人物）：N_c，N_m，N_oは各々正答，誤答，洩れの数を表す
Table 11　Result of personal caption analysis by the created dictionary: N_c, N_m, and N_o stand for numbers of correct, mistaken, and oversight answers, respectively.

	N_c	N_m	N_o	適合率	再現率
提案手法による解析	210	54	—	79.6%	68.9%
JUMANによる解析	33	45	—	42.3%	10.8%
総合	243	99	62	**71.1%**	**79.7%**

表12　作成した辞書による字幕解析結果（場所・組織）
Table 12　Result of locational/organizational caption analysis by the created dictionary.

	N_c	N_m	N_o	適合率	再現率
提案手法による解析	307	491	—	38.5%	46.9%
JUMANによる解析	220	12	—	94.8%	33.6%
総合	527	503	127	**51.3%**	**80.5%**

表13　作成した辞書による字幕解析結果（時相）
Table 13　Result of temporal caption analysis by the created dictionary.

	N_c	N_m	N_o	適合率	再現率
提案手法による解析	142	221	—	39.1%	83.5%
JUMANによる解析	17	0	—	100.0%	10.0%
総合	159	221	11	**41.8%**	**93.5%**

各解析処理単独での性能を示し，再現率は，総合的結果の再現率に対する各解析処理の貢献度を示している．

$$\text{適合率} = \frac{\text{正答数}\,(N_c)}{\text{正答数}\,(N_c) + \text{誤答数}\,(N_m)}$$

$$\text{再現率} = \frac{\text{正答数}\,(N_c)}{\text{正答数}\,(N_c) + \text{洩れ数}\,(N_o)}$$

4.2.2　考　　察

実験の結果，いずれの語義属性においても高い再現率が得られた．このことは，画像処理など他の処理からの制約により，特定の語義の索引を求める際に重要な要素である．人物を示す字幕以外における適合率は，本手法単体としても総合的にも全体に低かった．これは，辞書作成時に不適格語が混入したためと，語義の多様性により提案手法では正しく解析できなかったためと考えられる．

しかしながら，人物を示す字幕の解析については，再現率，適合率のいずれにおいてもほぼ実用的な解析性能が得られた．また，適合率については，同じ日本語を扱ったIREXの結果[21]よりも値が低いが，再現率については匹敵しうる値を示している．このことは，より多くの情報を拾い上げ，過剰に存在するものは後の画像処理や統合処理の際に取捨選択することを想定

表14　誤答と洩れの原因（異なり数の比率を示す）
Table 14　Reasons of mistakes and oversights.

	原因	比率
（a）	作成した辞書中の普通名詞の語彙不足	44%
（b）	JUMAN固有名詞辞書・時相辞書の語彙不足	39%
（c）	語義の多様性	9%
	例）「小錦-関」（人物）と「箱根-関」（場所・組織）	
（d）	JUMANによる形態素解析の誤り	6%
	例）○「6年-生」（人物）→×「6-年生」（その他）	
（e）	作成した辞書中の不適格語の存在	2%

している，映像への索引付けという応用例においては，十分実用的な精度である．

また，JUMANによる解析では固有名詞と一部の時相名詞の属性解析のみしかできないが，提案手法と組み合わせることにより，固有名詞を含む含まないにかかわらず，末尾の名詞が普通名詞である名詞句についても解析できるため，再現率で46%から83%の範囲で解析性能が向上している．

誤答と洩れの理由として，主に表14に示すような5通りのものが考えられる．

原因（a）と（e）は作成した辞書の性能を直接反映したものである．しかし，原因（a）のうちの4割程度はより細かい形態素に分解されていれば正しく解析できたものである．このような場合への対処法として，現在の完全一致による照合ではなく，最長後方一致による照合を採用すれば良いと思われるが，逆に誤答を増やすおそれもある．原因（e）は，表2，5，8に示した抽出条件が緩いために生じた不適格語の存在による問題と，類義語辞書を用いた語彙の拡張の際に混入した不適格語の存在による複合的な問題であるが，辞書を人手で修正することにより解消される．前者の問題による不適格語の多くは，抽出条件中で想定されたものとは異なる助詞の用法に起因して抽出されたため，この問題を解消するためには，コーパス中の形態素情報に助詞の用法に関して記述されることを期待したい．また，後者の問題による不適格語の多くは，用いた類義語辞書の分類方針と提案手法で必要とする分類が一致しないことによる問題であり，より分類方針の近い辞書を採用することにより，改善が期待される．

一方，原因（c）は語義を扱う際に直面する本質的な問題であり，本手法のような表層的な解析手法では容易に解決できず，原因（b）と（d）については，形態素解析ツールの語彙及び解析性能の向上をまつしかない．

ここで，コーパスからの抽出を行なわずに，類義語辞書から適切な分類項目を人手で選択した辞書（理想的な辞書）を用いた結果を，表11，12，13の結果と比較したものを表15に示す．これは，提案手法を用いて語義属性解析を行なった場合の性能の上限との比較を示すものである．理想的な辞書を用いても正答数はほぼ向上しないことから，原因（a）は頻度としては大きな問題になっていないことが分かる．一方で，誤答数は大幅に減少していることから，原因（e）が頻度としては大きな問題となっていることが分かる．また，理想的な辞書を用いても洩れ（N_o）がほとんど改善されないことから，不足語彙を補うためには，用いた類義語辞書に未収録の語彙を補完せねばならず，提案手法を用いてより大規模なコーパスからさらに語を抽出することが必要であることが示唆される．

表15　人手で類義語辞書から分類項目を選択して作成した辞書による字幕解析結果（括弧内は表11，12，13との比較）
Table 15　Result of caption analysis by manually created dictionary.

	N_c	N_m	N_o	適合率	再現率
人物	243	93	62	72.3%	79.7%
	(±0)	(-6)	(±0)	(+1.2%)	(±0.0%)
場所・組織	534	172	120	75.6%	81.7%
	(+7)	(-331)	(-7)	(+24.3%)	(+1.2%)
時相	159	29	11	84.6%	93.5%
	(±0)	(-192)	(±0)	(+42.8%)	(±0.0%)

5. むすび

本稿では，ニュース映像の自動索引付けに必要となる，字幕の語義属性解析のための辞書作成の過程およびその性能評価について述べた．具体的には，字幕（名詞句）の末尾の名詞（接尾名詞）に着目して，人物，場所・組織，時相のいずれを示すかを解析するために，まず各語義属性を示す接尾名詞をコーパスから抽出した辞書を作成した．その結果，人物名詞辞書には1,795語，場所・組織名詞辞書には7,908語，時相名詞辞書には2,144語，合計11,847語の語彙からなる辞書が得られた．同様の辞書を新分類語彙表の分類段落を人手で選択することにより作成することも考えられるが，自動で作成することにより，人手で作成する労力が省けるうえ，事例に基づく裏付けが得られ，また新分類語彙表非収録語を収集できる利点がある．ここでは，人物名詞として19語，場所・組織名詞として89語，時相名詞として20 語の合計128語が非収録語として収集できた．

次に，実際のニュース字幕の解析に適用して提案手

法を評価した．その結果，解析結果の適合率は低かったものの，索引付けの際に重要となる再現率は高い値を示し，実用性が示された．なお，理想的な辞書を用いた比較実験の結果でも，辞書の語彙不足が改善されず，索引付けの際に重要視される再現率を向上させるために必要な類義語辞書非収録語の補完のためにも，提案手法を用いてより大規模なコーパスからさらなる語の収集を行なうことの必要性が示唆された．

なお，実際に作成された辞書の内容（該当分類項目及び分類語彙表外語の一覧）については[5]を，本稿で作成した辞書を用い，図1に示すような機構に基づき，実際のニュース映像へ索引付けした結果については[4]を参照されたい．

謝辞　「『分類語彙表』形式による語彙分類表（増補版）」[8]は，国立国語研究所との間のモニタ契約のもとに，「RWCテキストデータベース」[2]は技術研究組合新情報処理開発機構（RWCP）との間のライセンス契約のもとに使用した．

実験に用いた字幕の映像からの書き下しは，平博司氏及び織田友恵氏の多大な労力の賜でり，ここに謝意を表す．

文　献

[1] 安藤彰男，今井　亨，小林彰夫，本間真一，後藤　淳，清山信正，三島　剛，小早川健，佐藤庄衛，尾上和穂，世木寛之，今井　篤，松井　淳，中村　章，田中英輝，都木　徹，宮坂栄一，磯野春雄：“音声認識を利用した放送用ニュース字幕製作システム”，信学論（D-II），vol.J84-D-II, no.6, pp.877–887，June 2001.

[2] 技術研究組合新情報処理開発機構（RWCP），“RWCテキストデータベース第2版，” July 1998.

[3] 堀　修，三田雄志，“テロップ認識のための映像からのロバストな文字部抽出法，” 信学論（D-II），vol.J84-D-II, no.8，pp.1800–1808，Aug. 2001.

[4] I. Ide, R. Hamada, S. Sakai, and H. Tanaka, “An attribute based news video indexing,” Proc. ACM Multimedia 2001 Workshops –Multimedia Information Retrieval–, pp.70–73, Oct. 2001.

[5] 井手一郎，“映像への自動索引付けに関する研究 –統合メディア処理による索引付けとそのニュース映像への適用–，” 博士請求論文，東京大学大学院工学系研究科電気工学専攻，Dec. 1999.

[6] 井手一郎，山本晃司，浜田玲子，田中英彦，“ショット分類に基づく映像への自動的索引付け手法，” 信学論（D-II），vol.J82-D-II，no.10，pp.1543–1551，Oct. 1999.

[7] 加藤直人，浦谷則好，相沢輝昭，中瀬純夫，“英日機械翻訳における固有名詞処理，” 第40回情処学全大 2F-2, pp.421–422，March 1990.

[8] 国立国語研究所，“「分類語彙表」形式による語彙分類表（増補版）［増補モニタ用電子化データ］，” March 1996.

[9] S. Kurakake, H. Kuwano, and K. Odaka, “Recognition and visual feature matching of text region in video for conceptual indexing,” Proc. SPIE Conf. 3022: Storage and Retrieval for Image and Video Databases V, pp.368–379, 1997.

[10] 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 より入手，Dec. 1998.

[11] 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，“京都大学コーパス第2.0版，” http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/corpus.html より入手，June 1998.

[12] Y. Nakamura, and T. Kanade, “Semantic analysis for video contents extraction –spotting by association in news video–,” Proc. Fifth ACM Intl. Multimedia Conf. (ACM Multimedia ’97), pp.393-402, Nov. 1997.

[13] 長瀬友樹，“形態素タイプを用いた日本語構文解析前処理，” 第41回情処学全大 1S-6，vol.3，pp.109–110，Sep. 1990.

[14] 那須川哲哉，“文脈情報を利用したキーワード語義決定，” 第11回人工知能学全大 17-1，pp.348–349，July 1997.

[15] S. Satoh, Y. Nakamura, and T. Kanade, “Name-It: Naming and detecting faces in news videos,” IEEE MultiMedia, vol.6, no.1, pp.22-35, March 1999.

[16] B. M. Sundheim, “Named entity task definition, version 2.1,” Proc. Sixth Message Understanding Conf. (MUC-6), pp.317–332, Nov. 1995.

[17] H. D. Wactler, A. G. Hauptmann, and M. J. Witbrock, “Informedia News-on-Demand: Using speech recognition to create a digital video library,” CMU Tech. Rep. CMU-CS-98-109, Carnegie Mellon University, 1998.

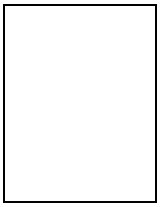
[18] H. D. Wactler, M. G. Christel, Y. Gong, and A. G. Hauptmann, “Lessons learned from building a Terabyte digital video library,” IEEE Computer, vol.32, no. 2, pp.66–73, Feb. 1999.

[19] 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，” 情処学自然言語処理研報 96-NL-116，pp.107-114，Nov. 1996.

[20] “付録H:NE定義，” IREXワークショップ予稿集, pp.264–273，Sep. 1999.

[21] “付録J：NE評価結果，” IREXワークショップ予稿集，pp.286–294，Sep. 1999.

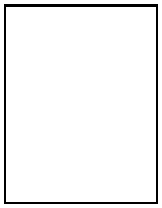
（平成13年8月24日受付，12月13日再受付）



井手　一郎　（正員）

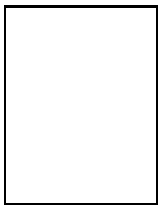
平6東大・工・電子卒．平8同大学院工学系研究科情報工学専攻修士課程了．平12同研究科電気工学専攻博士課程了．博士（工学）．同年より国立情報学研究所助手．自然言語処理，映像理解，統合メディア処理

に興味をもっている．平7第51回情報処理学会全国大会奨励賞受賞．人工知能学会，情報処理学会，IEEE Computer Society，ACM各会員．



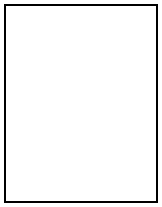
浜田　玲子　（学生員）

平10東大・工・電子情報卒．平12同大学院工学系研究科電気工学専攻修士課程了．修士（工学）．現在同専攻博士課程在学中，日本学術振興会特別研究員．平12第63回情報処理学会全国大会奨励賞受賞．情報処理学会会員．



坂井　修一　（正員）

昭56東大・理・情報科学卒．昭61同大学院工学系研究科情報工学専門課程了．工学博士．同年工業技術院電子技術総合研究所入所．この間平3〜4米国マサチューセッツ工科大学招聘研究員，平5〜6RWC超並列アーキテクチャ研究室室長．平8〜10筑波大学電子・情報工学系助教授．平10より東京大学大学院工学系研究科助教授．平13より同大学院情報理工学系研究科教授．計算機システム一般，特にアーキテクチャ，並列処理，スケジューリング問題，マルチメディアなどの研究に従事．平2情報処理学会論文賞，平3日本IBM科学賞，平7市村学術賞，平7ICCD Outstanding Paper Awardなど受賞．情報処理学会，IEEE，ACM各会員．



田中　英彦　（正員）

昭40東大・工・電子工学卒．昭45同大学院工学系研究科博士課程了．工学博士．同年東京大学工学部講師．昭46同助教授．昭62より同教授，現在同大学院情報理工学系研究科教授．この間昭53〜54ニューヨーク市立大学客員教授．計算機アーキテクチャ，並列処理，人工知能，メディア処理，自然言語処理，分散処理，CAD等の研究に興味をもっている．著書「非ノイマンコンピュータ」,「情報通信システム」，共著書「計算機アーキテクチャ」,「VLSIコンピュータI，II」,「ソフトウェア指向アーキテクチャ」．“New Generation Computing” 編集長．情報処理学会，人工知能学会，日本ソフトウェア科学会，IEEE，ACM各会員．