自然言語処理による放送コンテンツ解析手法に関する研究

山 田 一 郎

目次

1 章　序論

本章では，自然言語処理による放送コンテンツ解析手法に関する研究の背景，研究の位置付け，関連研究，特徴について言及し，本論文の大要を説明する．

1.1　本研究の背景

近年，放送局では番組を蓄積・管理するシステムが普及し，放送した番組映像などの放送コンテンツを大量に蓄積できる環境が整備されてきた．NHK においても 2007 年 6 月現在，過去に放送された約 61 万番組が NHK アーカイブスに蓄積されている．このような放送コンテンツには有益な情報が多く含まれており，放送コンテンツを効率的かつ効果的に二次利用することが課題となっている．放送コンテンツを二次利用するために，放送コンテンツの内容をシーンごとに詳細に説明したメタデータと呼ばれる情報が重要な役割を果たす．

メタデータとは「データのためのデータ」と定義されており，データそのものではなく，データに関連する情報を指す．例えば放送番組に対するメタデータとして，EPG(Electronic Program Guide)と呼ばれる電子番組ガイドが挙げられる．EPG には，番組のタイトル，サブタイトル，放送日，放送時間，ジャンルなど，番組に関連する情報が含まれており，ハードディスクレコーダーなどの予約録画に利用されている．EPG は，番組ごとにまとめて付与されているため，番組のどの時間帯で何が起きているかという情報までは分からない．

メタデータに関する国際規格 MPEG7 では，映像や音声のコンテンツに対して，その映像区間，音声区間の内容を詳細に説明するメタデータを付与するタグセットが規定されている．放送番組における区間ごとの詳細説明を，我々はコンテンツベースドメタデータと呼んでいる．本論文では，以後，コンテンツベースドメタデータを，単にメタデータと呼ぶ．番組コンテンツに対して，このようなメタデータが付与されれば，映像検索や映像フィルタリングなどを容易に実行でき，将来の様々なサービスを実現する可能性が生まれる．しかし，大量の番組に対してメタデータを人手で生成する作業には，大変な労力を要する．実際に，どのようなサービスが可能となり，人手を介して作成する意義を明確に示さなければ，放送局におけるメタデータ付与は難しく，現状では，放送番組に対して区間ごとのメタデータは付与されていない．また，スポーツ中継などの生放送番組に対しては，人手を介してもリアルタイムにメタデータを付与することが困難となる．

放送では，番組中でアナウンサーが話した言葉を文字化した「クローズドキャプション」と呼ばれるテキストデータが，番組とともに各家庭まで送られている．総務省が発表した平成 18 年度における字幕放送の実績では，総放送時間に占める字幕放送時間の割合は NHK で 43.1%，在京キー5 局の平均は 32.9%，クローズドキャプション付与可能な番組に占めるクローズドキャプションが付与された番組放送時間の割合は NHK で 100%，在京キー5 局の平均は 77.8%と報告されており，クローズドキャプションが付与される番組の割

合は年々増加している．この字幕放送は，アナログ放送では専用の字幕放送受信機が必要であったが，デジタル放送ではリモコンの字幕ボタンを押すことにより，容易に視聴することができる．

クローズドキャプションは対応する映像中の内容を説明することが多く，メタデータ生成のための重要な情報源となり得る．そこで本論文では，自然言語処理技術により，実際に放送された番組に付与されたクローズドキャプションなどを解析して，映像中の内容を特定し，メタデータを自動付与する手法について論じる．メタデータにより，放送コンテンツの要約視聴や，放送コンテンツに含まれる映像を利用した映像百科事典など，様々なアプリケーションを実現することができる.また放送番組には,社会情勢や因果関係知識,人間が持っている常識のような知識や，単語の上位概念や関連単語などの辞書的な語彙的知識など有益な情報も大量に含まれ，このような知識を自動獲得できれば，メタデータ生成時の精度向上が期待できる．さらには，放送番組を二次利用する効果的なアプリケーションにも繋がる．そこで，テキストコンテンツを解析して，番組で扱われる有益な情報や知識を自動獲得する手法についても論じる．

1.2　本研究の位置付けと関連研究

番組に対するメタデータ生成のための研究は，これまでに数多く行われてきた．図 1.1 に，メタデータ生成アプローチの分類と研究課題を示す．放送コンテンツには，映像，音声，そしてクローズドキャプションなどのテキストが含まれるため，これらを自動解析する手法として，映像・画像処理，音声処理，自然言語処理に分類できる．映像・画像処理では,映像の切り替わりを検出するカット点検出[1][2][3],ニュース番組でアナウンサーがニュース原稿を読むスタジオのショット（アンカーショット）検出[4],顔画像検出[5],顔

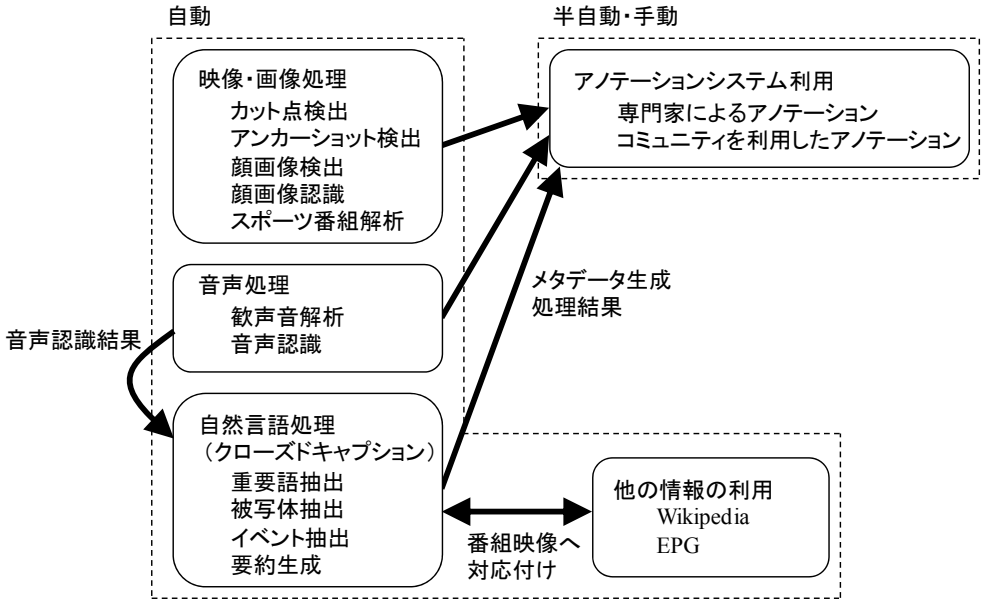


図1.1　メタデータ生成アプローチの分類

画像認識[6]，そして，スポーツ番組を選手やボールの位置情報や，ショットの切り替えなどの情報により解析するスポーツ番組解析[7][8][9][10]など，数多くの研究が行われている．また，近年，米国の NIST(National Institute of Standards and Technology)と DTO(Disruptive Technology Office)の主催で，映像コーパスを用いた，情報検索のための競争型のワークショップの TRECVID(TREC Video Retrieval Workshop)[11]が行われており，ニュース映像とクローズドキャプションを解析することにより「船が映っているシーン」，「2 人以上の人間が歩いているシーン」などの高次特徴を抽出するタスクが行われている．カット点検出やアンカーショット検出，顔画像検出では，ある程度の精度が得られているが，他の処理では，映像解析のための複雑なモデルを用いるため，リアルタイム処理が難しく，精度にも課題が残る．

音声処理では，スポーツ番組などにおける歓声音を解析することによる重要シーン抽出[12]，ニュース番組やスポーツ番組に対する音声認識処理[13][14][15]などが行われている．重要シーン抽出では，重要か否かを判定する目的のため，番組の内容まで特定することはできない．また，音声認識処理で得られるテキストデータは，番組内容を直接説明している訳ではないため，そのままではメタデータとして利用は難しい．そこで，自然言語処理による詳細なテキスト解析が必要となる．

自然言語処理では，音声認識結果や番組のクローズドキャプションなどのテキストコンテンツを処理対象として，メタデータを生成する．Chang らは，放送された番組中のコメントから音声認識処理によりキーワードとなる単語を抽出し，歓声音解析処理，画像解析処理を利用してアメリカンフットボールのタッチダウンシーンを抽出する手法を提案した[16]．Lazarescu らは，キーワードと画像解析処理を利用して，アメリカンフットボールの選手の動きを抽出する手法を提案した[17]．新田らは，クローズドキャプションを利用して，キーワード列による検索を行うことにより，アメリカンフットボールの試合進行部分を抽出し，画像解析処理結果を利用することにより，各プレイを抽出している[18]．また，佐藤らは，ニュースを解析して映像中の人物が誰かを推定する Name-It を提案した[19]．この手法では，顔画像解析処理により，映像中の顔を検索し，クローズドキャプション中の人物を表す単語との整合を取ることにより，高精度にニュース映像中の人物を特定している．これらの処理では，処理対象のテキストから重要語を抽出する処理に留まり，言語情報を十分に利用できていない．

本研究では，テキスト情報のみを解析して，より高度なメタデータである映像の被写体や，発生したイベントを抽出することを対象とする．また，ニュース記事を要約してニュース番組に対するメタデータとして利用する手法も提案する．より高度なメタデータを生成する処理として，柴田らは，映像情報と言語情報を利用して，料理番組中のトピックを推定する手法を提案している[20]．この手法では，番組のクローズドキャプションを解析して談話構造を抽出し，隠れマルコフモデルに基づいた状態遷移により，料理番組で，今，何が起こっているかというトピックを推定している．このようなモデルは番組の型がある程度決まっており，状態遷移が記述できるような対象であれば有効と考えられる．

映像や音声，クローズドキャプション以外の情報を利用して，メタデータを生成する手法も提案されている．奥岡らは，Wikipedia エントリに出現するニュースと，放送されたニュースとを関連付けることにより，ニュース番組に変遷などの構造を与える手法を提案している[21]．河合らは EPG に含まれる要約文を利用して，番組の要約を生成する手法を提案している[22]．これらは，番組に関連する Wikipedia や EPG の要約文が存在する場合，とても有効な手法と考えられる．

また，番組に対するメタデータを半自動，もしくは全て手作業で付与するためのアノテーションシステムに関する研究も行われている．佐野らは，映像・画像処理，音声処理，自然言語処理で解析した結果を統合して，最終的には人が判断して確度の高いメタデータを付与するメタデータエディタを提案している[23]．山本らは，インターネット上に映像共有サービスを提供し，映像シーンに関連付けられたコメントやブログエントリーなどの情報を獲得する Synvie というシステムを公開している[24]．例えば Synvie 上で公開されている映像をユーザのブログに引用して，その映像に対してコメントを記述することにより，Synvie は引用された映像区間とコメントを収集でき，該当区間のメタデータとして利用できる．自動処理では 100%正確なメタデータを生成することは難しいため，リアルタイムにメタデータを必要としないような番組には，人手を介してメタデータを生成するアプローチは現実的な手法となり得る． この場合でも，自動処理により生成するメタデータも有効に活用できると考えられる．

長尾らは，番組のクローズドキャプションや関連するテキストに対して，自動処理の結果を人手で修正することにより言語構造や語彙情報をアノテーションとして関連付ける仕組みを提案し，ダイジェストや翻訳などのアプリケーションを精度良く生成できることを示している[25]．このような付加的な情報は，番組の作り手から誤解なく情報を伝達するための手段として重要と考えられる．このアプローチでは，テキストコンテンツに対して，複数の候補の中から正解を選択してアノテーションを作成する手間がかかるが，語彙情報などは大量のテキストデータを用いることで処理の精度を向上させることも可能と考えられる．本論文では，語彙情報などの知識の獲得についても提案する．

1.3 本論文の特徴と概要

本論文では，言語情報から，放送された番組に対してメタデータを自動付与する手法について主に論じる．対象とする番組は，情報系番組，生放送スポーツ番組，ニュース番組など，番組の型が明確に決まっていないものとする．本論文では，メタデータを自動付与するために，番組に付随して放送されるクローズドキャプションなどの言語情報を統計的アプローチにより解析することを特徴とする．また，各番組に対して自動付与されたメタデータを利用することにより，テレビ番組のコンテンツを効果的に二次利用するためのアプリケーションの提案も行う．

本章に続く，本論文の概要は以下の通りである．

2 章では，本論文で用いる AdaBoost アルゴリズム，GibbsBoost アルゴリズム，Support Vector Machine, 最大エントロピー法，EM アルゴリズムなどの機械学習アルゴリズムについて，その概要を説明する．

3 章では，情報系番組を対象としたメタデータ自動生成手法について述べる．情報系の番組のクローズドキャプションでは，「場所紹介」や「人物紹介」など特定の事柄を表現するために同じような言い回しが多用される．このような言い回しを含む文章区間が抽出できれば，対応する番組映像区間の場所紹介や人物紹介といったメタデータを付与することができる．局所的な特徴しか利用されない従来法の問題点を改善し，大域的な文章構造の類似性を利用する手法と，さらに，2 種類のサンプリング処理を行うことにより処理時間の問題点を改善した手法を提案する．

4 章では，生放送スポーツ中継番組のアナウンスコメントを解析することにより，メタデータを自動生成する手法について述べる．生放送スポーツ中継番組のアナウンスコメントには，実際に発生したイベントに対する説明と，発生したイベントとは直接関係しない補足的な説明が存在する．実際に発生したイベントに対する説明は，対応する映像に対するメタデータとして有益な情報となる．サッカー中継番組を対象として，この 2 種類の説明の分類手法と分類結果を利用したメタデータ生成手法を言及する．

5 章では，ニュース番組を対象としたメタデータ自動生成手法について述べる．ニュースは社会の情勢や最新の流行など，豊富な情報が含まれているため，二次利用の有用性が高いと考えられる．放送局では大量のニュース記事データを電子化して蓄積するようになり，これらの効率的な管理，活用が急務となっている．そこで本章では，ニューステキストを解析して，管理するための話題推定，話題要約手法について説明し，この結果を利用したアプリケーションについて述べる．

6 章では，放送コンテンツ中のテキストデータからの知識獲得手法について述べる．最初に大量のテキストデータを解析する際に問題となる未知語処理について述べる．次に，従来研究ではほとんど行われていない語の典型的な機能・目的や起源などの語彙知識を自動獲得する手法について述べる．これらの処理で得られる未知語の上位語推定結果や単語間の関係は，メタデータを自動生成する処理の精度向上のための知識として利用できる．また，番組で難しい用語が使われる場合，語彙を説明するフレーズが出現する．そこで，用語とその説明を抽出し，用語とその説明間の意味関係を分類する手法について言及する．語彙知識の一つとして，物事の「原因－結果」の関係を示す因果関係知識がある．この因果関係知識は，人間の思考において重要な役割を果たし，大量に因果関係知識を蓄積できれば，「何故」といった質問に対する答えを推論により導きだすことが可能となる．健康に関する番組を対象として，専門的な知識となる因果関係知識の獲得手法についても言及する．用語の説明や因果関係知識は，放送された番組を二次利用する効果的なアプリケーションに有用と考えられる．

7 章では，本研究の成果をまとめ，これにより結言とする．

2 章　本論文で使用する機械学習アルゴリズムの概要

放送コンテンツ中のテキストデータを統計的に解析するために使用する AdaBoost アルゴリズム[26]，GibbsBoost アルゴリズム[27]，Support Vector Machine[28]，最大エントロピー法[29]，そして，EM アルゴリズム[30]について，その概要を説明する．

2.1　はじめに

本論文では，番組に付随して放送されるクローズドキャプションを統計的に解析することにより，番組に対して，その映像区間，音声区間の内容を詳細に説明するメタデータを自動付与する手法について論じる．クローズドキャプションは，ナレーションとほぼ同等のテキストであり，このようなテキストには一定のパターンが存在する．機械学習を使って，クローズドキャプションを統計的に解析することにより，このようなパターンを抽出することが可能となり，各パターンに対応する映像にメタデータを付与することができる．

本論文で使用する機械学習アルゴリズムの長所と短所，さらには本論文においての適用対象を表 2.1 に示す．

表 2.1　本論文で使用する機械学習アルゴリズム

機械学習アルゴリズム	長所	短所	本論文における適用対象(章，節番号)
AdaBoost アルゴリズム	汎化能力高	ノイズの多いデータには過学習を起こす可能性有 計算量多（学習時）	定型表現文章区間の抽出(3.2)
GibbsBoost アルゴリズム	計算量少（学習時）	計算量多（実行時）	定型表現文章区間の抽出(3.3)
Support Vector Machine	汎化能力高 少量の学習データでも効果的	計算量多（学習時）	サッカー中継番組におけるコメントの分類処理(4)
最大エントロピー法	確率値を直接計算可能	素性間の相互作用を考慮する必要がある	語彙知識獲得処理(6.3) 用語と説明文の関係判定処理(6.4)
EM アルゴリズム	尤度が単調に増加することが保証され，振る舞いが安定	初期値により局所解に収束する可能性有	因果関係抽出処理(6.5)

2.2　AdaBoost アルゴリズム

AdaBoost アルゴリズムは Boosting アルゴリズム[31]の一つであり，必ずしも高精度ではない識別関数である弱学習器を大量に組み合わせることにより，データを高精度に 2 つの

クラスへ分類する組み合わせ学習アルゴリズムである．まず，大量の弱学習器を用意し，この弱学習器をデータに対する重みを考慮した誤り率により選択する．選択された弱学習器の信頼度を訓練集合に対する誤り率から求め，さらに，訓練集合に含まれる各データに対する重みを，正確に分類されたデータに対しては減少，誤分類されたデータに対して増加させることにより更新しながら逐次的に学習を行う．この処理では，誤分類されたデータに対して重みを増加させているため，次の処理では，これまでに選択された弱学習器で誤分類されたデータを間違えないような弱学習器が選択されやすくなる．最終的に，弱学習器の信頼度を加味した多数決で，どのクラスに分類するか決定する．

訓練集合である学習データとして，入力 x_i に対して出力 y_i となる弱学習器を N 個与え，以下のアルゴリズムにより学習を行い，最終仮説により判定する．

学習データ: $(x_1, y_1), \cdots, (x_N, y_N)$

入力 $X=\{x_1, x_2, \ldots, x_N\}$

出力 $Y=\{y_1, y_2, \ldots, y_N\}$, $y_i=\{+1, -1\}$

Step1: 繰り返し回数 t=1 における学習データに対する重み D_t を初期化

$$D_t(i) = 1/N \quad , t=1$$

Step2: 学習データに対する誤り率εを式(2.1)により計算し，εが最小となる弱学習器を選択

$$\varepsilon = \sum_{i=1}^{N} D_t(i) \, | h_t(x_i) - y_i | \tag{2.1}$$

$h_t(x_i)$: t 番目に選択された弱学習器の入力 x_i に対する判定結果

Step3: 選択された弱学習器により弱学習器に対する信頼度αを式(2.2)により計算し，学習データ i に対する重み $D_t(i)$を式(2.3)により更新．新たな重み $D_{t+1}(i)$ は，誤り率ε=0.5 となる値としている．

$$\alpha_t = \frac{1}{2} \log\left(\frac{1-\varepsilon}{\varepsilon}\right) \tag{2.2}$$

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & if \ y_i = h_t(x_i) \\ e^{\alpha_t} & otherwise \end{cases} \tag{2.3}$$

Z_t は正規化定数

Step4: Step2 と Step3 の処理を弱学習器が無くなるまで繰り返す

Step5: 最終仮説 $H(x)$は t 番目に選択された弱学習器の信頼度 α_t と判定結果 $h_t(x)$との積を t=1~T について足し合わせ，その符号により判定．

$$H(x) = \mathrm{sgn}\left(\sum_t \alpha_t h_t(x)\right) \tag{2.4}$$

図 2.1 に，AdaBoost アルゴリズムの概念図を記す．

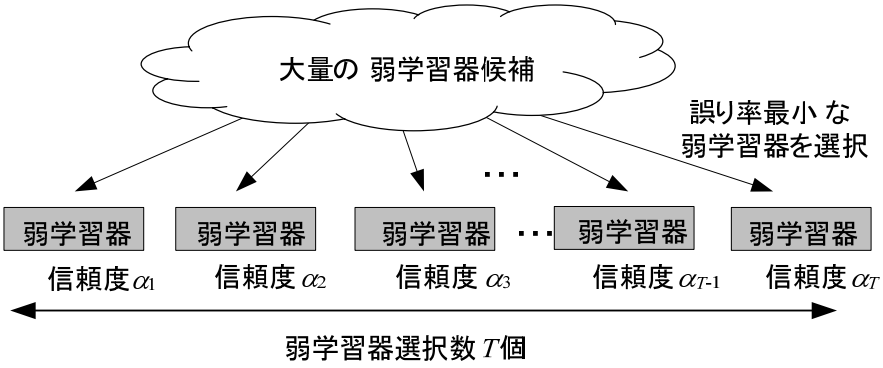


図 2.1　AdaBoost アルゴリズムの概念図

自然言語処理において，複数文のテキストを構文解析し，その依存関係を取りだして特徴として利用する場合，その量は膨大なものになる．しかし，その特徴のほとんどが無用なものと考えられる．AdaBoost アルゴリズムは，大量の弱学習器から解析に効果のある弱学習器を選択して利用するため，このような役に立たない特徴が大量に存在するデータに対しても頑健に働く．そこで，本論文では AdaBoost アルゴリズムを，複数文のテキストを構文解析した結果得られる依存構造の特徴を統計的に解析する処理で利用する．

2.3　GibbsBoost アルゴリズム

GibbsBoost アルゴリズムは Boosting アルゴリズムの一つであり，弱学習器を効果的かつ効率的にサンプリングすることで弱学習器集合を複数生成する．AdaBoost アルゴリズムでは，全ての弱学習器の信頼度を評価していたが，弱学習器の数が膨大な量となる場合は相当な計算量となってしまう．さらに AdaBoost アルゴリズムでは，逐次的に処理を行い，各処理で最も誤り率の低い弱学習器を選択するため，局所解に陥る可能性もある．そこで GibbsBoost アルゴリズムでは，弱学習器を効果的かつ効率的にサンプリングすることで計算量を飛躍的に減少させ，さらには弱学習器集合による判定器を複数生成することにより局所解に陥ることを防いでいる．

Boosting アルゴリズムの判別関数は式(2.5)で表される．

$$F(x;\Theta_t)=\sum_{t'=1}^{t}\alpha_{t'}h(x;\theta_{t'}) \tag{2.5}$$

ここで，x は観測データ，h は弱学習器，α_t は t 番目の弱学習器の信頼度，θ_t は t 番目にどの弱学習器を選択するかを決めるパラメータである．また，$\Theta_t=(\alpha_1,\ldots,\alpha_t,\theta_1,\ldots,\theta_t)$ とする．学習処理では，入力 x_i と出力 y_i からなる学習データ $\{y_i,x_i\}_{i=1}^N$ が与えられた時，損失関数 $\sum_{i=1}^N L(y_iF(x;\Theta_t))$ を最小化する Θ_t を使用する弱学習器の数まで逐次的に求める．Θ_t の決定により，入力 x に対する判別関数 F の値が 0 より大きい場合+1 を返し，0 以下の場合 -1 を返す 2 値判別器を実現できる．

GibbsBoost アルゴリズムでは，Boosting アルゴリズムの損失関数に対応するエネルギー

関数 $L(z)$を用いて，パラメータ Θ_t に対する確率分布 $P_t(\Theta_t)$ を式(2.6)により定義する．

$$P_t(\Theta_t) \propto \pi(\Theta_t)\prod_{i=1}^{N}\exp\left(-\beta_t L(y_i \frac{F(x_i;\Theta_t)}{\sqrt{t}})\right) \tag{2.6}$$

式(2.6)では，損失関数 $L(z)$からの影響が t に比例して増えないよう $\sqrt{t}$ により抑制している．β_t は，統計力学における温度の逆数を示す係数であり，確率分布 $P_t(\Theta_t)$の分散を抑制する．β_t が大きい場合，Θ_t は損失関数の和が小さくなるような値に集中し，β_tが小さい場合，$\pi(\Theta_t)$と似た分布となる．β_t には, Boltzmann annealing[32]をベースとした値と, Cauchy annealing[32]をベースとした値などを利用できる．

[Boltzmann annealing]

$$\beta_t = \beta_0 \log(t+e) \tag{2.7}$$

[Cauchy annealing]

$$\beta_t = \beta_0 (t+1) \tag{2.8}$$

$\pi(\Theta_t)$はパラメータ Θ_t に対する事前確率分布であり，式(2.9)により定義される．

$$\pi(\Theta_t) = \prod_{t'=1}^{t}\pi_\theta(\theta_{t'})\pi_\alpha(\alpha_{t'}) \tag{2.9}$$

$\pi_\theta(\theta_{t'})$は，t' 番目の弱学習器の候補を選択する事前分布であり，先見的情報に基づいて決定する．$\pi_\alpha(\alpha_{t'})$は t' 番目の弱学習器の信頼度に対する事前分布であり，ここでは正規分布とする．

GibbsBoost アルゴリズムにおける判別関数は，使用する弱学習器数を T 個としたとき，$F(x;\Theta_T)$ に対してパラメータ Θ_T に対する確率分布 $P_T(\Theta_T)$による期待値により，式(2.10)で定義される．

$$F_{ave,T}(x) = \int F(x;\Theta_T)P_T(\Theta_T)d\Theta_T \tag{2.10}$$

この値は，解析的に解を求めることが困難である．そこで逐次モンテカルロ法[33]を利用し，Θ_T を有限個数だけサンプリングする．サンプリングの処理手順を図 2.2 に示す．

逐次モンテカルロ法では，パラメータ Θ_t のサンプリングと，その重み(importance weight)の計算のために提案分布Qを使用する．Qの値が大きい部分から多くサンプリングし，Qの値が小さい部分からはあまりサンプリングを行わない．そして，Qをもとに取り出した M 個のサンプル $\Theta_t^{(j)}$ (j=1～M)に対する重み $w^{(j)}$ (importance weight)を計算する．この重み $w^{(j)}$

弱学習器($t = 1$ to T)において，Step1～Step3 を繰り返す

Step1:　t-1 個の弱学習器が線形結合された弱学習器集合 j (j = 1～M)に対して，t 番目の弱学習器を提案分布 Q に基づきサンプリング

$$(\alpha_t^{(j)}, \theta_t^{(j)}) \sim Q(\alpha_t, \theta_t; \Theta_{t-1}^{(j)})$$

Step2: Step1 において選択された $\Theta_t^{(j)}$ の Importance weight　$w^{(j)}$ を計算

$$w^{(j)} \propto \frac{P_t(\Theta_t^{(j)}; \beta_t)}{P_{t-1}(\Theta_{t-1}^{(j)}; \beta_{t-1}) Q(\alpha_t^{(j)}, \theta_t^{(j)}; \Theta_{t-1}^{(j)})}$$

ここで，$\sum_{j=1}^{M} w^{(j)} = 1$

Step3: 確率 $w^{(j)}$ によって $\Theta_t^{(j)}$ をリサンプリング

図 2.2　GibbsBoost アルゴリズムにおけるサンプリング処理手順

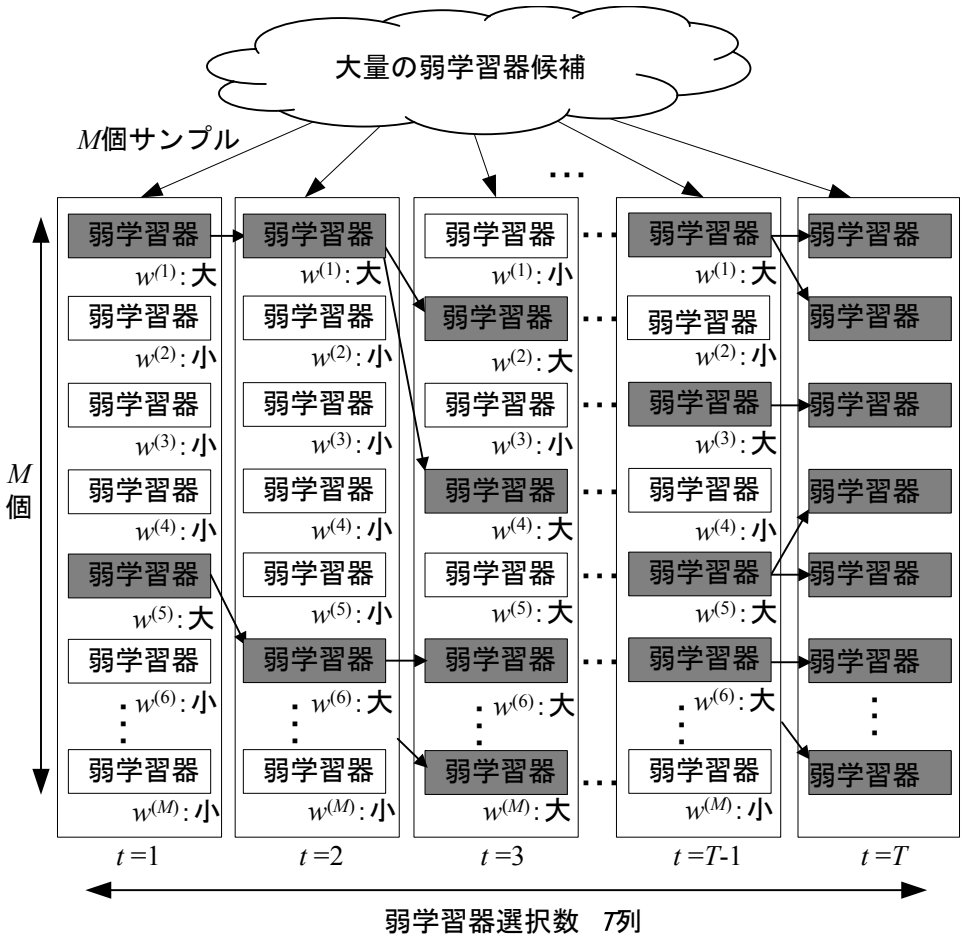


図 2.3　GibbsBoost アルゴリズムの概念図
（網目の掛かったものが重み $w^{(j)}$ の値により選択された弱学習器）

は，選ばれた弱学習器に対する $P_t(\Theta_t)$ の値により算出される．この重み $w^{(j)}$ の値を利用して，選ばれた $\Theta_t^{(j)}$ からリサンプリングを行う．$\Theta_t^{(j)}$ が選択される確率は，重み $w^{(j)}$ の値に比例した値とする．大きな重みを持つ $\Theta_t^{(j)}$ は何度も選択され，小さな重みを持つ $\Theta_t^{(j)}$ は選択される可能性は少ない. 確率的にサンプリングを行っているため, 大きな重みを持つ $\Theta_t^{(j)}$ でも選択されない場合があり，逆に小さな重みを持つ $\Theta_t^{(j)}$ が選択される場合もある．この処理を, 線形結合する弱学習器数 t=1～T のそれぞれの場合において逐次的に行うことにより，T 個からなる弱学習器が M 個生成される（合計 T×M 個の弱学習器）．これらの弱学習器集合がサンプリングされた結果となる．

提案分布 $Q(\alpha_t,\theta_t;\Theta_{t-1})$ は，弱学習器の選択を決定する分布 $Q(\theta_t)$ と，選ばれた弱学習器の信頼度を決定する分布 $Q(\alpha_t;\theta_t,\Theta_{t-1})$ の積 $Q(\alpha_t,\theta_t;\Theta_{t-1})=Q(\alpha_t;\theta_t,\Theta_{t-1})Q(\theta_t)$ で表現できる．$Q(\theta_t)=\pi_\theta(\theta_t)$ とし，$Q(\alpha_t;\theta_t,\Theta_{t-1})$ は学習データに対するエラーレートから算出する．

M 個のサンプリング結果を利用して，$F_{avs,T}(x;\Theta_T)$ の和を計算する．M 個のサンプルを $\{\Theta_T^{(j)}\}_{j=1}^{M}$ とすると，式(2.10)は式(2.11)で近似される．

$$F_{ave,T}(x) \approx \frac{1}{M}\sum_{j=1}^{M} F(x;\Theta_T^{(j)}) \tag{2.11}$$

この値の正負を判別基準とすることにより，2 値判別が可能となる．図 2.3 に GibbsBoost アルゴリズムの概念図を示す.

本論文では, 複数文のテキストを構文解析した結果得られる依存構造の特徴を, AdaBoost アルゴリズムを利用して統計的に解析する手法を提案している． AdaBoost アルゴリズムを利用すると計算量が膨大となるため，処理の効率化をはかる用途で GibbsBoost アルゴリズムを利用する.

2.4 Support Vector Machine

Support Vector Machine(以後 SVM)は，データ空間を高次元に写像し，正例データと負例データを分割する超平面を見つけることにより 2 つのクラスに分類する機械学習アルゴリズムである．SVM では，全学習データを超平面の決定に使用せずに，識別に効果的と判定されたサポートベクターと呼ばれる学習データのみを利用する．そのため，他の学習モデルと比較して汎化能力が高く過学習しにくい.

n 次元の素性ベクトルからなるデータ x_i (i=1~N)と，その出力 y_i を訓練集合となる学習データとして与える.

学習データ: $(x_1, y_1),\cdots,(x_N,y_N)$

入力 $\mathbf{x}=\{x_1,x_2,\ldots,x_N\}$, $x_i \in \mathrm{R}^n$, N:データ数, n:素性数

出力 $\mathbf{Y}=\{y_1,y_2,\ldots,y_N\}$, $y_i=\{+1,-1\}$

データを分割する超平面を N 個の要素をもつ行列 $\mathbf{w}$ とデータ $\mathbf{x}$ との内積を利用して, 式

(2.12)で表現する．

$$\mathbf{w}\cdot\mathbf{x}+b=0 \qquad \mathbf{w}\in R^n, b\in R \tag{2.12}$$

$f(\mathrm{x})=\mathbf{w}\cdot\mathbf{x}+b$ とおくと，データ x_i が正例（y=+1）の場合は $f(x_i)\geq 1$，負例（y=-1）の場合は $f(x_i)\leq -1$ となる．$-1<f(x_i)<1$ の区間は関数マージンと呼ばれ，正例，負例ともに属さない領域となる．

超平面と学習データ中のベクトルとのマージンが最大となる超平面を選択する．ある 1 点 x_i から超平面までの距離は式(2.13)で与えられる．

$$r=\frac{|\mathbf{w}\cdot x_i+b|}{\|\mathbf{w}\|}=\frac{|f(x_i)|}{\|\mathbf{w}\|} \tag{2.13}$$

$|f(\mathrm{x}_i)|\geq 1$ であるので，超平面に最も近い正例と最も近い負例の間のマージンは以下の式(2.14)で与えられる．

$$\min_{x_i;y_i=+1}\frac{|f(x_i)|}{\|\mathbf{w}\|}+\min_{x_i;y_i=-1}\frac{|f(x_i)|}{\|\mathbf{w}\|}=\frac{1}{\|\mathbf{w}\|}+\frac{1}{\|\mathbf{w}\|}=\frac{2}{\|\mathbf{w}\|} \tag{2.14}$$

学習データに対して，この値を最大化する超平面を見つける．ここで $2/\|\mathbf{w}\|$の最小化の代わりに，$\|\mathbf{w}\|^2/2$ の最大化を考える．学習データ x_i に対してラグランジュ未定係数α_i としたラグランジュ関数は式(2.15)定義できる．

$$L(\mathbf{w},b,\alpha_i)=\frac{1}{2}\|\mathbf{w}\|^2-\sum_{i=1}^{N}\alpha_i[y_i(\mathbf{w}\cdot\mathbf{x}+b)-1] \tag{2.15}$$

$L(\mathrm{w},b,\alpha_i)$ を w で偏微分した結果，以下の式(2.16)が導き出される．

$$\mathbf{w}=\sum_{i=1}^{N}y_i\alpha_i\mathrm{x}_i \tag{2.16}$$

α_i を求めることができれば，この式から超平面を決めるベクトル w が求まることが分かる．α_i=0 となるデータ x_i はベクトル $\mathbf{w}$ には全く影響しない．つまり，α_i>0 のデータ x_i のみから超平面が決定される．超平面の決定に使われるα_i>0 の学習データをサポートベクターと呼ぶ．

α_i は，以下の目的関数を最大化するものとして求める．

$$L(\mathbf{w},b,\alpha_i)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j y_i y_j(\mathrm{x}_i\cdot\mathrm{x}_j) \tag{2.17}$$

求められたα_iから，識別関数は式(2.18)で与えられる．

$$H(x) = \mathrm{sgn}(\sum_{i=1}^{N} \alpha_i y_i (\mathrm{x}_i \cdot \mathbf{x}) + b) \tag{2.18}$$

SVM は，超平面を決めるためにサポートベクターと呼ばれる特徴のみを利用する．そのため，汎化能力が高く，少量の学習データに対しても頑健な解析処理が可能である．本論文では, 学習データを大量に作成することが困難であった生放送スポーツ番組のコメント解析の際に，SVM を利用する．

2.5 最大エントロピー法

最大エントロピー法とは，与えられた制約の下で，情報量の尺度となるエントロピーを最大化するモデルを推定する手法である．エントロピーとは，ある確率変数において，その値を特定するために必要な情報量のことを言い，例えばサイコロでは，1 の目が異常に多く出るサイコロより，全ての目が 1/6 の確率で出るサイコロのほうが，エントロピーは大きい値を取る．モデルのエントロピーを最大化することにより，与えられた制約以外の未知変数に対して尤もらしい確率値を割り振ることができる．

学習データを与えたときの素性関数を以下のように定義する．

学習データ： $(x_1, y_1), \cdots, (x_N, y_N)$

入力 $X=\{x_1,x_2,\ldots,x_N\}$

出力 $Y=\{y_1,y_2,\ldots,y_N\}$

素性関数： $F = \{f_i : (x, y) \mapsto \{0,1\}, i \in \{1,2,...,n\}\}$

例えば，素性関数$f_i(x_j,y_j)$は，ある特徴を(x_j,y_j)が持つときに 1 の値，持たないときに 0 の値を返す．

最大エントロピー原理を満たす確率モデルは式(2.19)で表すことができる．

$$P_\Lambda(x, y) = \frac{1}{z_\Lambda} \exp(\sum_i \lambda_i f_i(x, y)) \tag{2.19}$$

$$z_\Lambda = \sum_{x,y} \exp(\sum_i \lambda_i f_i(x, y))$$

$P_\Lambda(x, y)$ を求める式(2.19)では，右辺の $\sum_i \lambda_i f_i(x, y)$ の値を基準として(x, y)の出現確率を求めている．$\sum_i \lambda_i f_i(x, y)$ の値を非負とするため，exp 関数を用い，さらには確率値とするため z_Λ により正規化を行っている． λ_i は，素性関数 $f_i(x, y)$ に対する重みを表すパ

Step1: $\Lambda=\{\lambda_1,...,\lambda_n\}$に適当な初期値を与える

Step2: $\Delta=\{\delta_1,\ldots,\delta_n\}$を次式より求める

$$\sum_{x,y} P(x,y) f_i(x,y) \exp(\delta_i \sum_{i=1}^{n} f_i(x,y)) = \sum_{x,y} \frac{C(x,y)}{N} f_i(x,y)$$

$C(x,y)$: (x,y)の出現回数

Step3: Λ を更新

$$\lambda_i = \lambda_i + \delta_i$$

Step4: Λ の値が収束するまで Step2, Step3 を繰り返す

図 2.4　最大エントロピー法のパラメータ λ を求めるための
反復スケーリング法アルゴリズム

ラメータであり，最大エントロピー法では，この重みのパラメータをモデルのエントロピーを最大化するように学習して $P_\Lambda(x,y)$ を求める．この処理では図 2.4 に示す反復スケーリング法[34]などが用いられる．

自然言語を対象とする場合，同時確率分布 $P_\Lambda(x,y)$ より，条件付きモデル $P_\Lambda(x|y)$ を利用することが多い. 例えば, 出現した単語の履歴から後続する単語を予測する場合には，条件付モデルが必要となる．条件付きモデルは，式(2.20)で表すことができる．この式中のパラメータ λ も，同時確率分布と同様に反復スケーリング法により求めることができる．

$$P_\Lambda(x|y) = \frac{1}{z_\Lambda(x)} \exp(\sum_i \lambda_i f_i(x,y)) \tag{2.20}$$

$$z_\Lambda(x) = \sum_y \exp(\sum_i \lambda_i f_i(x,y))$$

最大エントロピー法は，確率値で結果が出力されるため，複数の出力がある場合，その出力間の比較を容易に比較することができ，さらには，他の統計量と組み合わせてランク付けなどを行う処理にも向いている．本論文では，複数の出力を持つ関係抽出と，相互情報量と組み合わせてランク付けを行う語彙知識獲得で利用する．

2.6　EM アルゴリズム

EM アルゴリズムは，観測データから内部のパラメータが直接決定できないような問題に対して，逐次，モデルの尤度を増加させるようなパラメータを推定する手法である．

尤度とは，観測データから内部パラメータを推定するときに，どの程度尤もらしいかを表し，式(2.21)で定義される．

$$L(\theta) = \prod_{i=1}^{N} P_\theta(x_i) \tag{2.21}$$

ここで，θ は内部パラメータを指す．最尤推定法とは，この尤度 $L(\theta)$を最大にするパラメータ $\hat{\theta}$ を推定する方法で，多くの確率モデルでは上記の式を直接用いず，対数をとり，対数尤度を最大化する．

$$\hat{\theta} = \arg\max_{\theta}(\log L(\theta)) \tag{2.22}$$

構造が単純なモデルは対数尤度を最大化する最尤推定法で内部パラメータを求めることができるが，複雑なモデルとなると，直接，最尤推定法を用いることができず，このような場合に EM アルゴリズムが有効となる．EM アルゴリズムでは，繰り返しにより，逐次，モデルの対数尤度を増加させるように内部パラメータを推定する．EM アルゴリズムの手順を図 2.5 に示す．

Step1: 内部パラメータ θ に適当な初期値を与える

Step2: モデルの対数尤度を増加させるように，内部パラメータ θ を新たな内部パラメータ $\bar{\theta}$ に更新．θ が収束するまで，以下の Estep と Mstep を繰り返す．

Estep : $Q(\theta,\bar{\theta})$ を計算

$$Q(\theta,\bar{\theta}) = \sum_{y} P_\theta(y \mid x_i) \log P_{\bar{\theta}}(x_i, y)$$

Mstep : θ を新しいパラメータ値 $\bar{\theta}$ に更新

$$\theta = \arg\max_{\bar{\theta}}(\log Q(\theta,\bar{\theta}))$$

図 2.5　EM アルゴリズム

EM アルゴリズムは，少量のラベル付きの訓練データを利用して，大量のテストデータのラベルを推定する処理で使われる．本論文では，少量の因果関係が付けられたデータを基にして，大量のテキストデータから，他の因果関係を抽出する処理で利用する．

3 章　情報系番組を対象としたメタデータ自動生成

本章では，情報系の番組のクローズドキャプションを入力とし，特定の事柄を表現するための定型表現区間を抽出することにより「場所紹介」のシーンに対するメタデータを自動生成する手法について述べる．

3.1　はじめに

情報系の番組のクローズドキャプションでは，「場所紹介」や「人物紹介」など特定の事柄を表現するために同じような言い回しが多用される．このような言い回しは，映像内容を説明する目的があるため，放送局や番組に依存することは少ない．例えば，表 3.1 に示すクローズドキャプション中では，矩形で囲まれた部分が「場所」を映像とともに説明している．最初に体言止めにより「オンフルール」という町の位置情報を説明し，次に町の詳細を断定の助動詞「です」を使って説明している定型的な表現である．

表 3.1　クローズドキャプション例
（矩形で囲まれた部分は「場所」を説明する定型的な表現区間）

提示時間	クローズドキャプション
08:29:03	絵は　全然描きませんからって→
08:29:09	まっ　こんなとこですかね．
08:29:12	やっぱり　絵を描かなくてよかったかもしれませんね．
08:29:46	セーヌ川を挟み　ル・アーブルの対岸に位置する港町　オンフルール．
08:29:53	今なお中世の古い家並みが残る　町です．
08:29:59	18歳の時　モネは　パリに出て画家を　目指しますが　美術学校の　入学試験に合格しませんでした．
08:30:11	実家に戻る事を　強要した父親の意向に反して　なおも　パリにとどまって絵の勉強を　し続けた　モネ．

そこで，情報系の番組のクローズドキャプションから表 3.1 に示すような特定の事柄を表現するための定型表現区間を抽出する手法を提案する．抽出された定型表現区間が，情報系の番組に対する「場所紹介区間」や「人物紹介区間」などのメタデータとなり，番組映像検索に有用な情報となる．

このような区間を抽出するために，従来から提案されているベクトルスペースモデル[35]を利用する手法が考えられるが，単語の出現の有無を特徴として利用するため，定型表現区間の構文的な特徴をとらえることができないため，高精度な解析ができない．単語の特徴だけでなく，構文構造を考慮したテキスト解析の手法として Collins らにより Tree Kernel が提案されている[36]．また，工藤らは部分木を素性とする decision stumps[37]とそれを弱学習器とした boosting アルゴリズムを提案し，製品レビュー文や新聞記事のテキスト分類の実験を行っている[38]．これらの手法では，テキストに含まれる共通部分木の数により類似性を評価しているが，ノードの飛び越えを許さない部分木の完全一致を類似度判定の基準としているため，結果として局所的な部分木しか特徴として利用されないこと

が多い．また，複数文にまたがる類似性評価は行われていない．

3.2 では，文章に含まれる大域的な情報まで加味した特徴量を利用して AdaBoost アルゴリズム[26]による学習を行うことにより，文章間の類似性を評価し，番組のクローズドキャプションから定型表現区間を抽出する手法を提案する．AdaBoost アルゴリズムによる学習を行う手法では，大域的な情報まで加味した特徴量を利用するため，文章の特徴量が爆発的に増え，計算時間に問題が生じてしまう問題点が残されていた．そこで，3.3 では，爆発的に増えた弱学習器を，その事前情報によりサンプリングすることで処理時間の効率化を図り，さらには弱学習器系列を複数生成して判定を行うことにより効果的な処理結果を導く GibbsBoost アルゴリズム[27]を用いた手法を説明する．

放送された番組を提案手法により解析することにより，場所紹介をしている映像区間のクリップ集を収集できる．3.4 では，この映像クリップを利用して，ユーザが興味のある場所について映像とともに調べることができるマルチメディア百科事典を紹介し，最後に今後の展望について言及する．

3.2 大域的な情報を利用した定型表現文章区間抽出

提案する手法では，キーとなる単語を一つ選択し（例えば映像の被写体を表す単語など），この単語が一つ以上存在する一文以上のテキストを定型表現文章区間の候補とする．例えば，「場所」を映像とともに説明する定型表現文章区間では，場所名を表す単語が文章区間に出現すると考えられる．そこで，表3.1では，場所を表す「オンフルール」をキーとなる単語とし，この単語を含む一文以上のテキストが定型表現文章区間の候補となる．この候補が，定型表現文章区間であるか否かを，AdaBoostアルゴリズムによって判定する．

学習処理では，まず，テキストの一部に対して人手により定型表現が含まれるか否か判定して，学習データを生成する．学習データから部分木を抽出し，依存構造木間の類似度を基準とした弱学習器を生成する．次に，AdaBoostアルゴリズムにより，どの弱学習器が正例と負例の弁別力があるかを判定しながら弱学習器の信頼度を示す重みを学習する．テストデータ中のキーとなる単語の周辺の複数文に対して学習結果を適用することにより，定型的な文章区間か否かを判定する．以下に，複数文にまたがる特徴の抽出，抽出した文章の特徴となる部分木間の類似性評価，そして，「場所」を映像とともに説明する定型表現文章区間の抽出とその実験について記す．

3.2.1 複数文にまたがる特徴の抽出

入力テキストを一文ごとに構文解析して，各ノードを文節により構成する依存構造木を生成する．クローズドキャプション中の文の区切れ目は句点，疑問符，感嘆符などにより判断できる．各文の最終文節となる根ノードの親ノードに最上位ノードを生成し，最上位ノードから各文の依存構造木へは順序付きのアークで結んだ依存構造木を生成する．順序付きアークは文の出現順序を考慮した依存構造木間の類似度評価で利用する．表 3.1 の矩

形で囲まれた区間の入力テキストを依存構造木に変換した例を図 3.1 に示す．図では，1 文目の「セーヌ川を挟み　ル・アーブルの対岸に位置する港町　オンフルール．」を構文解析し，その根ノード「オンフルール」と，2 文目の「今なお中世の古い家並みが残る　町です．」の根ノード「町です」を結ぶ最上位ノード(図の○部分)を生成している．この最上位ノードから伸びるアークには，1 文目の根ノード「オンフルール」へは①，2 文目の根ノード「町です」には②の番号が振られている．

次に，人が定型文章区間であるか否か判定することにより生成した学習データを用い，この学習データ中の正例として与えられた依存構造木からキーとなる単語と任意の数のノ

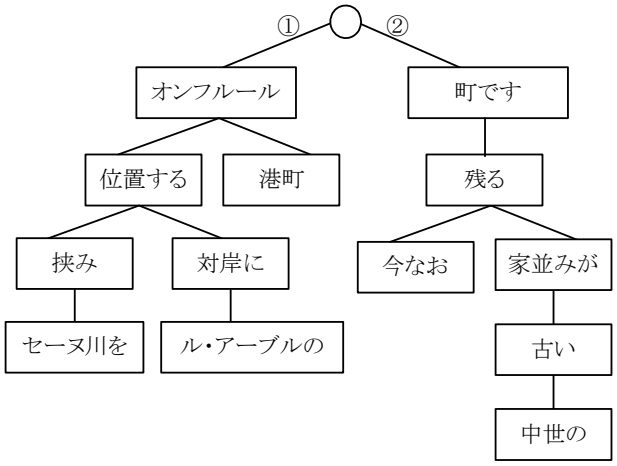


図 3.1　依存構造木生成例

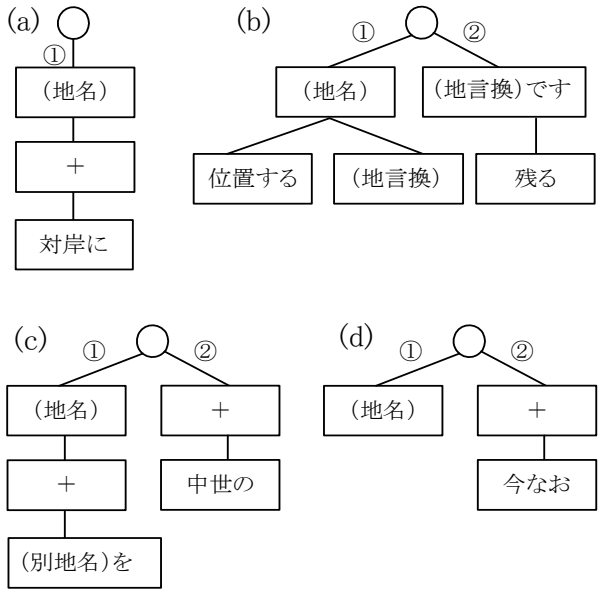


図 3.2　依存構造木から抽出された部分木（一部）

ードを含む部分木を生成する．この処理で，キーとなる地名，キーとなる単語以外の地名，地名の言い換え表現は単語表記そのものを利用しないで，“(地名)”，“(別地名)”，“(地言換)”という表記で抽象化して部分木を生成する．また，部分木の作成の際にノードの飛び越えを許し，飛び越えたノードは“+”の記号で置き換え 1 つ以上のノードとのマッチングを許す．図 3.1 に示した依存構造木から生成される部分木の一部を図 3.2 に示す．図 3.2(a)は「対岸に～オンフルール．」，図 3.2(b)は「～位置する港町オンフルール．～残る町です．」の部分を図 3.1 の依存構造木から抽出した部分木であり，「オンフルール」は“(地名)”に，「港町」「町」は“(地言換)”に抽象化されている．図 3.2 の(b)，(c)，そして(d)は，2 文にまたがる部分木となっている．この処理で抽出した部分木を，複数文にまたがる特徴として利用する．

クローズドキャプション中では，キーとなる単語が出現した以降の文で「この町は」など場所を映像とともに説明する特徴的な表現が現れることがある．キーとなる単語から離れて位置するような単語も定型的な文章区間抽出には重要な役割を果たすと考えられる．部分木生成時に飛び越しを許さないような従来手法では，キーとなる単語と，キーとなる単語から離れて位置するような単語との関係のみを利用することが難しい．例えば図 3.1 のキーとなる単語「オンフルール」と「中世の」という単語の関係を利用する場合，「オンフルール」，「町です」，「残る」，「家並みが」，「古い」，「中世の」というノードからなる部分木を生成しなければならなかった．この場合，次節で定義する類似度計算において「中世の」という単語以外の単語の「町です」，「残る」，「家並みが」，「古い」にも影響を受けるため，キーとなる単語「オンフルール」と「中世の」という単語の関係のみを考慮出来ない．提案手法では，単語から離れて位置するような単語間の特徴を考慮するため，飛び越しを許す部分木を利用する．

また，工藤らの手法[38]では，類似度を利用しないで部分木の完全一致による decision stumps を利用するため，多くのノード数を持つ部分木は，学習データに大量に出現しない限り boosting アルゴリズムで重みは小さな値が与えられ，結果的に考慮されなくなる．

部分木生成時に飛び越えを許すことにより，図 3.2(d)のように「オンフルール」，「＋」，「今なお」というノードのみからなる部分木を生成でき，キーとなる単語「オンフルール」と「今なお」という単語の関係のみを考慮することが可能となる．

提案手法では，対象文章に含まれるこのような関係を全て考慮できるため，文章の大域的な範囲を考慮した類似性評価手法となる．

3.2.2 抽出した文章の特徴となる部分木と依存構造木間の類似性評価

抽出した部分木と，学習データに含まれるテキストから生成される依存構造木との類似度は，部分木に含まれる葉ノードから根ノードまでの全リスト構造を抽出し，その各リスト構造が対象とする依存構造木に含まれる割合を基準として定義する．部分木 t と依存構造木 x の類似度 $sim(t,x)$ は以下の式(3.1)とする．

$$sim(t,x)=\frac{1}{N(t)}\sum_{t_i\in t}\frac{1}{L(t_i)}\sum_{st\in t_i}\max_{sx\in x}(C^d\times sim'(st,sx)) \qquad (3.1)$$

t_i ：部分木tに含まれるi番目の文

st ：t_iに含まれる葉ノードから根ノードまでのリスト

sx ：xに含まれる葉ノードから根ノードまでのリスト

$sim'(st,sx)$ ： stがsxに含まれる割合．リストに含まれる主辞と付属語を分割して計算．

$N(t)$ ：tに含まれる文数

$L(t_i)$ ：t_iに含まれるリスト数

C ：キーとなる単語を基準とした文位置の差に与えるペナルティ値（本実験では0.5）

d ：キーとなる単語を基準とした文位置の差

図3.2(b)に示す部分木tとの類似度を求める例を図3.3に示す．この例では，葉ノードから根ノードまでのリストが，部分木tから3つ，依存構造木xから4つ取り出されている．最も類似しているリスト構造間の類似度$sim'(st,sx)$をそれぞれ求めることにより，部分木tと依存構造木xの類似度$sim(t,x)=0.625$と算出することができる．

部分木tと学習データに含まれるテキストとの類似度の分布を0〜1の間にマッピングする．マッピングした正例と負例を弁別する閾値を考え，弁別した際の誤りが最小となる点をθ_tとする．出力のクラスラベルを$y\in\{\pm1\}$としたとき，部分木tと閾値θ_tに対する弱学習

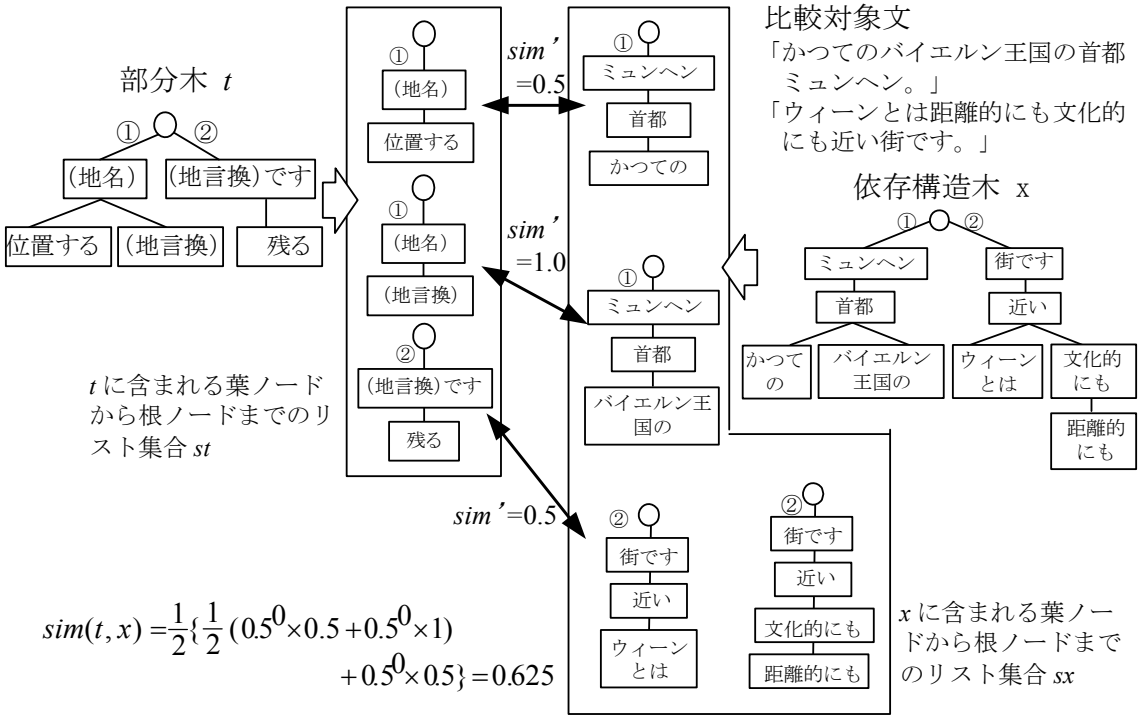


図3.3　部分木と比較対象文との類似度計算例

器 $h_t(x)$ は式(3.2)により定義できる．

$$h_t(x) = \begin{cases} y & sim(t,x) \geq \theta_t \\ -y & sim(t,x) < \theta_t \end{cases} \tag{3.2}$$

ここで，出力のクラスラベル $y \in \{\pm 1\}$ は，部分木 t と学習データに含まれるテキストとの類似度がある閾値 θ_t より大きいものを正例，小さいものを負例とした場合に $y=1$，逆に，ある閾値 θ_t より大きいものを負例，小さいものを正例とした場合に $y=-1$ とする．

学習データに含まれるテキストを構文解析して生成した依存構造木から抽出できる部分木の数は，膨大な数となり，全ての部分木を学習で利用することは不可能と考えられる．本手法では，ノード数により部分木の数を制限し，後述する実験ではノード数が 4 個以下の全ての部分木を抽出し，式(3.2)に示す弱学習器を生成している．

3.2.3 「場所」を映像とともに説明する定型表現文章区間の抽出

学習データに含まれるテキストから抽出した部分木によって大量の弱学習器が生成される．この弱学習器を 2.2 で概説した AdaBoost アルゴリズムの入力とし，機械学習を行い，どの弱学習器が効果的であるかを判定する処理を行う．

学習の結果得られる最終仮説を利用して，学習データとは異なるテストデータから，定型的な文章区間の抽出を行う．まず，テストデータからキーとなる単語を抽出する．実験では，「場所」を映像とともに説明する定型表現文章区間の抽出を行うため，キーとなる単語は場所名を表す単語とする．次に，単語が出現する前後数文からなる区間を入力とする．実験では，キーとなる単語が出現する文とその前 2 文，後ろ 7 文の合計 10 文から，キーとなる単語が出現する文を含む任意の連続する文を処理対象としている．この場合， 1 文から 10 文で構成される区間が AdaBoost アルゴリズムの入力 x となる．この 1 文以上からなる区間の入力 x に対して，AdaBoost アルゴリズムの最終仮説 $H(x)$を計算する．$H(x)$=1 の時, 対象区間は定型表現部分であると判断できる．しかし, 負例には特徴が少ないため, 定型的でない文章区間は，定型的であると誤判定される可能性がある．

Viola らは顔画像検出処理において, 判定処理を何段もカスケードすることにより適合率を向上させる手法を提案している[5]．そこで本手法でも，図 3.4 に示すように最終仮説

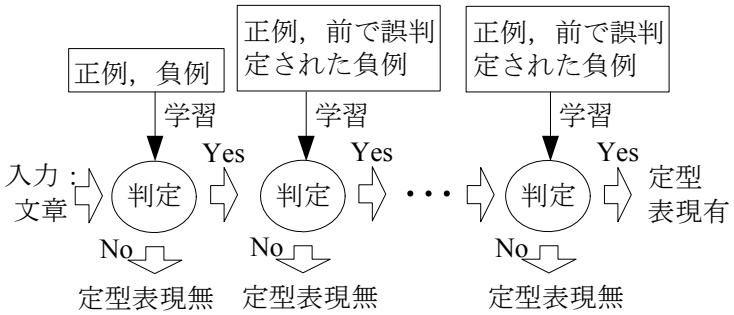


図 3.4　カスケードによる判定処理の繰り返し

$H(x)=1$ と判定された事例に対して，再度，AdaBoost による学習を行い判定する．この際，前の学習で利用しなかった負例に対して誤って定型的な文章区間と判定されたものから，次の学習で利用する負例データを選択し，正例はそのままとした学習による最終仮説を利用する．判定処理をカスケードして複数回行うことにより，適合率向上が期待できる．

また，ある文章区間で $H(x)=1$ となる場合は，その前後の文を含めた区間でも同様に $H(x)=1$ と判定される．この場合は，$H(x)$に含まれる関数の値 $\sum_t \alpha_t h_t(x)$ により定型表現部分の区間を判定し，文を追加した時にこの値が増加するときのみ，その文を定型表現部分に追加する．この処理により，キーとなる単語と定型表現を含む文章区間が抽出される．

3.2.4 「場所」を映像とともに説明する定型表現文章区間の抽出実験

提案手法を検証するため，NHK で放送された紀行番組「わが心の旅」のクローズドキャプションを対象として,「場所」に関する情報を映像とともに説明している定型的な表現部分を抽出する実験を行った．形態素解析辞書に「地名」として登録されている単語を「キーとなる単語」とし，その単語を含む区間が，場所を映像とともに説明している場合を正例, 場所を映像とともに説明していない場合を負例として 60 番組(1 番組は 45 分)に対して人手により正解データを付与した．このデータを無作為に 2 つに分割し，片方を学習データ，残りをテストデータとした交差検定を行った．学習データに含まれる負例の数は正例に比べて多いため，正例と同数を無作為に選択した．負例における区間は，正例と同じ平均文数となるように調整した．部分木生成時に選択するノード数が多い場合は計算量が膨大になるため，対象とするような定型表現は数個のノードにより表現できると考え，今回は利用するノード数を 4 個として学習を行った．この結果，2 回の実験ではそれぞれ 56899 個，56092 個の弱学習器が生成された．

また，カスケードによる判定処理の繰り返し処理では，処理対象データに対して，学習データとして利用する負例データが確保可能な 4 回行った．

形態素解析辞書に「地名」として登録されている単語を「キーとなる単語」としてテストデータから抽出し，その「キーとなる単語」を含む区間が定型的な文章区間か否かを最終仮説により判定した．抽出結果の一部を図 3.5 に示す．図中の矩形で囲まれた部分が提案手法により抽出された定型的な文章区間，下線部の単語が「キーとなる単語」である．

キーとなる単語が，判定結果と正解データでともに「場所を説明する文章区間」，または「場所を説明しない文章区間」に出現しているときを正解として結果の評価を行った．テストデータとした番組には形態素解析辞書に「地名」として登録されている名詞が合計 1972 個含まれ，そのうちの 196 個が実際に映像とともに場所を説明していた．評価結果を表 3.2 に示す．表中の F 値とは，適合率と再現率の調和平均を示し，式(3.3)で定義される．

$$F\text{値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}} \tag{3.3}$$

［ 抽出例 1 ］

ガウディは　どのようにして建築と出会い　造形の世界を　究めていったのでしょうか？

バルセロナの南西　地中海に面して広がるタラゴナ平原
オリーブや　ブドウの畑が続くのどかな田園地帯です。

↕ 正解区間

１８５２年　ガウディはリウドムスという村で　鍋や釜を作る職人の子として生まれました。

［ 抽出例 2 ］

しかし　反面　３度の恋愛はすべて　失恋に終わり　生涯を旅に明け暮れ　家庭を持つこともなかったのです。

コペンハーゲンの町のはずれにクリスチャニアと呼ばれる角がある。
１９７０年代の初め　古くなって放置されていた軍事施設を　若者達が占拠して住み着いた場所です。
そして　ここには　俳優・詩人作家などを目指す人が多く住み共同体を作っています。

↕ 正解区間

ちょっと　大学の学園祭って感じですが。
ごめんください。

□ : 抽出区間

図3.5　場所を説明する定型表現区間抽出例(下線がキーとなる単語)

表3.2　提案手法による判定評価結果

キー単語 (カスケード数)	適合率	再現率	F 値
場所を説明する区間 (1 回)	192/882 (21.8%)	192/196 (98.0%)	0.356
場所を説明しない区間 (1 回)	1086/1090 (99.1%)	1086/1776 (61.1%)	0.758
場所を説明する区間 (2 回)	183/581 (31.5%)	183/196 (93.3%)	0.471
場所を説明しない区間 (2 回)	1378/1391 (99.1%)	1376/1776 (77.6%)	0.870
場所を説明する区間 (3 回)	181/516 (35.1%)	181/196 (92.3%)	0.508
場所を説明しない区間 (3 回)	1441/1456 (99.0%)	1441/1776 (81.1%)	0.892
場所を説明する区間 (4 回)	175/432 (40.5%)	175/196 (89.2%)	0.557
場所を説明しない区間 (4 回)	1519/1540 (98.6%)	1519/1776 (85.5%)	0.916

「キー単語が場所を説明・カスケード数 1 回」における結果では，適合率が 21.8%と低い．しかし学習を繰り返すことにより適合率が向上し，適合率と再現率の調和平均である F 値もカスケード数 4 回で 0.557 まで向上している．場所を説明しないキーとなる単語はテストデータ中に 1776 個出現しており，この判定結果の精度はカスケード数 4 回で F 値 0.916 と良好な結果が得られた．

提案手法による実験結果では，カスケード数を 4 回としても適合率は 40.5%であり，依然，多くの誤抽出が残されている．場所を映像とともに説明していると誤抽出されてしまった例を図 3.6 に示す．誤抽出例 1 では，「地中海」というキーとなる単語に対する説明区間として 2 文が誤抽出されている．実際には，この区間は「タラゴナ平原」に対する説明区間である．提案手法では，キーとなる単語による定型表現区間から生成する依存構造木と，同じ区間にある別のキーとなる単語に対する依存構造木が類似するため，弱学習器の類似度も同様の傾向が見られて誤判定されてしまう．そこで，同一区間において複数のキーとなる単語が出現する場合は，最終仮説の値が大きいものに絞込みを行う処理を行った．カスケード数4回における評価結果を表3.3に示す．場所を説明する区間の適合率が6.2%，F 値でも 0.044 向上している．

誤抽出例 2 の区間では，過去についての言及区間である．提案手法では，この文中の過去を表す文節「覆われていたのです」に対して，付属語として最終形態素を含み連続する助詞と助動詞しか取り出されないため「です」しか考慮されず，過去を示す助動詞「た」が考慮されない．文章区間から依存構造木を生成する際に，過去形なども考慮するよう改善が必要と考えられる．

[誤抽出例 1]
バルセロナの南西　地中海に面して広がる　タラゴナ平原.
オリーブや　ブドウの畑が続くのどかな田園地帯です.

[誤抽出例 2]
民話を育んだのは　豊かな森.
中世まで　ヨーロッパは　豊かな森に覆われていたのです.

[誤抽出例 3]
１４世紀　コルドは　実際に　革製品の取り引きで　賑わったと言います.
今　コルドは工芸の伝統を受け継いで　芸術の町として生きようとしています.

図3.6　提案手法による誤抽出例

表3.3　最終仮説の評価式の値を利用した評価結果

キー単語	適合率	再現率	F 値
場所を説明する区間	165/353 (46.7%)	165/196 (98.5)	0.601
場所を説明しない区間	1588/1619 (98.1%)	1588/1776 (89.4%)	0.935

また誤抽出例 3 では，人によっても「コルド」が映像とともに説明されているか否かの判断は難しい．正解データを付与した者とは異なる被験者により，クローズドキャプションのみから映像とともに場所を説明している区間か否かを判定する実験を行った．その結果，提案手法により場所を説明する区間と誤抽出された 188 区間のうち，44 区間(23.4%)で人間でも同様に，場所を映像とともに説明していない区間と判定できなかった．このような部分は，機械による解析も困難と考えられる．

次に，カスケードによる判定を 4 回行った後に，場所を説明する文章区間と判定された 175 箇所に対して，人手により付与した正解区間とどの程度一致しているか評価を行った．結果を表 3.4 に示す．

表3.4　文章区間の抽出精度

適合率	再現率
230 文/280 文 (82.1%)	230 文/458 文 (50.2%)

提案手法により抽出した区間に含まれる文中で正解データの区間に含まれている割合を示す適合率は 82.1%，正解区間のうち提案手法により抽出された文の割合を示す再現率は 50.2%であった．提案手法では，4 回の判定を行い全ての処理で正と判定された区間を抽出している．各判定において学習データが異なるため抽出される区間にも差が生じ，すべての判定で正と判定される区間は短くなる傾向が見られた．この影響で再現率が多少低い値となった．

3.2.5　既存手法との比較

提案手法の有効性を検証するため，ベクトルスペースモデル[35]を利用する手法と，ノードの飛び越えと部分木ー依存構造木間の類似度を利用しない手法による 2 つの実験を行った．ベクトルスペースモデルを利用した定型表現区間抽出手法では，まず，学習データの正例区間に含まれる単語と負例区間に含まれる単語をベクトルの要素とした正例ベクトルと負例ベクトルを生成する．ベクトルの要素となる単語は，自立語（名詞，動詞，形容詞，副詞など）に限定し，ベクトルの要素の値は，単語の出現頻度 TF と単語の逆文書頻度 IDF の積である TFIDF 値[39]とする．正例ベクトルを生成する場合，TF は正例区間に出現する自立語の出現頻度，IDF は対象自立語の全てのクローズドキャプション中における逆文書頻度である．例えば，表 3.1 の矩形で囲まれた区間のみを正例としてベクトルを生成する場合，以下の要素を持つベクトル $\vec{p}$ が生成され，各要素の値は，その TFIDF 値となる．

$\vec{p}$ = {セーヌ川, 挟む, ル・アーブル, 対岸, 位置する, 港町, オンフルール,
今, 中世, 古い, 家並み, 残る, 町 }

形態素解析辞書に「地名」として登録されている単語を「キーとなる単語」としてテストデータから抽出し，その前 2 文，後 7 文から，単語のある文を含む任意の連続文を処理対象文章とする．処理対象文章に対しても，出現単語をベクトルの要素，単語の TFIDF 値をベクトルの要素の値とした対象文章ベクトル$\vec{t}$を生成する．この対象文章ベクトル$\vec{t}$と正例ベクトル$\vec{p}$，負例ベクトル$\vec{n}$とのコサイン距離を求める．コサイン距離は式(3.4)により定義される．

$$\cos(\vec{p},\vec{t}) = \frac{\vec{p}\cdot\vec{t}}{|\vec{p}||\vec{t}|} \tag{3.4}$$

負例ベクトルより正例ベクトルとのコサイン距離が小さい場合に，対象文章を正と判定する．一つの「キーとなる単語」に対して抽出した複数の処理対象文章のうち一つでも正と判定された場合,「キーとなる単語」の周辺に場所を説明する文章区間があると判定する．提案手法で利用した正例と，4 回の学習で利用した全ての負例により正例ベクトル，負例ベクトルを生成して行った交差検定による判定結果を表 3.5 に示す．

表3.5　ベクトルスペースモデルを利用した定型表現区間抽

キー単語	適合率	再現率	F 値
場所を説明する区間	183/816 (22.4%)	183/196 (93.4%)	0.362
場所を説明しない区間	1143/1156 (98.9%)	1143/1776 (64.4%)	0.780

また，ノードの飛び越えと部分木－依存構造木間の類似度を利用しない手法では，提案手法と同様に，正例の文章区間に含まれる文章から，各文に対する構文解析結果を統合した依存構造木を生成し，この統合した依存構造木から弱学習器として利用するための部分木を抽出する．この時，部分木に含まれるノードの数は，提案手法と同様に 4 つ以下に制限するが，部分木におけるノードの飛び越えは許さない点で提案手法と異なる．また，弱学習器生成では，部分木－依存構造木間の類似度を用いず，部分木が依存構造木の部分構造か否かのみを判定する decision stumps を弱学習器として利用する．依存構造木 *x*，部分木 *t*, 出力クラスラベルを$y \in \{\pm 1\}$としたとき，分類を行うための decision stumps は式(3.5)で定義される

$$h_t(x) = \begin{cases} y & if\ t \subseteq x \\ -y & otherwise \end{cases} \tag{3.5}$$

ここで，出力のクラスラベル$y \in \{\pm 1\}$は，部分木tが依存構造木の部分構造であるものを正例，部分構造でないものを負例とした場合に$y = 1$，逆に，部分構造であるものを負例，部分構造でないものを正例とした場合に$y = -1$とする．

表3.6　ノードの飛び越えと部分木－依存構造木間の類似度を利用しない手法による抽出実験の評価結果

キー単語 (学習の繰り返し数)	適合率	再現率	F 値
場所を説明する区間 (1 回)	193/1240 (15.6%)	193/196 (98.5)	0.269
場所を説明しない区間 (1 回)	729/732 (99.6%)	729/1776 (41.0%)	0.581
場所を説明する区間 (2 回)	186/629 (29.6%)	186/196 (94.9%)	0.451
場所を説明しない区間 (2 回)	1333/1343 (99.3%)	1333/1776 (75.1%)	0.855
場所を説明する区間 (3 回)	183/589 (31.1%)	183/196 (93.4%)	0.466
場所を説明しない区間 (3 回)	1370/1383 (99.1%)	1370/1776 (77.1%)	0.867
場所を説明する区間 (4 回)	183/569 (32.2%)	183/196 (93.4%)	0.478
場所を説明しない区間 (4 回)	1390/1403 (99.1%)	1390/1776 (78.3%)	0.874

この式は，部分木 t が依存構造木 x の部分構造となっている場合，すなわち $t \subseteq x$ である時に出力 y を返す関数である．この式を弱学習器とした boosting による学習を行うことにより，入力となる依存構造木 x が定型表現区間か否かを判定できる．この手法は，工藤らの手法[38]を複数文の文章区間に適用したものと等価である．

提案手法と同様に，負例に対して誤って定型的な文章区間と判定されたものから再度負例データを選択し，正例はそのままとした学習を繰り返した．前節と同じデータを対象とした交差検定による判定結果を表 3.6 に示す．

この実験では，負例データが提案手法で利用したものと異なる．偶然，提案手法で使用した負例データが弁別に適している可能性もあるため, 単純な比較は適切でない. しかし，提案手法と同じ負例データを使用すると，この手法で既に負と判定されているデータも選択されることがあり，不利な条件となる．実際に，提案手法と同じ負例データを使用して 4 回のカスケードの実験をした結果, 場所を説明する区間では適合率 24.2%, 再現率 93.4%, F 値 0.384 と，表 3.6 の結果を下回る値であった．そのため，ここでは同じ学習データによる実験でなく，データ作成手法を同じとした実験を比較対象としている．

提案手法による結果(表 3.2)とベクトルスペースモデルを利用した手法による結果(表 3.5)を比較すると，カスケード数が 2 回以上において提案手法の方が良好な F 値が得られている．ベクトルスペースモデルは提案手法の 4 回の学習で利用した全ての負例データと

正例データを利用しているため，提案手法のほうが有効であると判断できる．

また提案手法による結果は，いずれのカスケード数でも表 3.6 に示すノードの飛び越えと部分木－依存構造木間の類似度を利用しない手法による結果の F 値を上回っている．ノードの飛び越えを許した部分木生成と，部分木－依存構造木間の類似度を考慮した弱学習器を利用する提案手法の有効性が確認できた．

3.3 GibbsBoost アルゴリズムによる文章類似性評価

3.2 で述べた手法により，依存構造木で遠く離れて位置する文節間の特徴なども考慮した類似性が評価でき，さらには，複数文を対象とした文集合の類似性評価も可能となる．しかし，依存構造木から抽出した大量の部分木を基として弱学習器を生成するため，比較対象となる依存構造木から抽出される部分木の数は膨大になるうえ，最終的には類似性評価に使われない無駄な部分木も弱学習器として生成されてしまい効率が悪い．そこで，2.3 で概説した GibbsBoost アルゴリズム[27]を利用して，爆発的に増えた弱学習器をサンプリングすることで処理時間の効率化を図り，さらには弱学習器系列を複数生成して頑健に判定を行う手法を提案する．図 2.3 に示した概念図のように，大量の弱学習器から判定で利用する弱学習器候補のサンプリングと，弱学習器候補から最終的に利用する弱学習器をサンプリングする，2 種類のサンプリング処理を行うことにより効率的な処理が可能となるとともに，弱学習器系列を複数生成した処理により頑健性が向上する．

3.3.1 GibbsBoost アルゴリズムによる文章類似性評価

GibbsBoost アルゴリズムによる文章類似性評価処理では，比較対象となる複数文に対して，複数文にまたがる特徴を 3.2 で述べた手法により抽出し，文章が定型的な文章区間に含まれるか否かを，抽出した特徴を利用した類似性評価により判定する．判定のための弱学習器は，式(3.2)で与えられる．この弱学習器が膨大な量となるため，その弱学習器の有効性を示す事前確率によりサンプリングを行い，処理の効率化を図る．

図 3.1 などに示されるような依存構造木では，各文の根ノードに主節の述部があり，その下のノードには，主節の述部に直接係る連体節または主節の格要素が位置する．これらは文を比較する上で重要な要素と考えられる．そこで，根ノードに近いノードほど選択される確率が高くなり，さらに，選択されたノード間の距離が近いほど選択される確率が高くなるような事前分布を考える．依存構造木 θ において部分木 θ_t が選択される確率 $\pi_\theta(\theta_t)$ を，式(3.6)の通り定義する．

$$\pi_\theta(\theta_t) \propto \prod_{n\in\theta_t} C_1^{depth(n)} \times \prod_{n_1,n_2\in\theta_t,n_1\neq n_2} C_2^{length(n_1,n_2)} \tag{3.6}$$

$depth(n)$：ノード n の根ノードからの深さ

$length(n_1,n_2)$：ノード n_1 とノード n_2 間のノード数

C_1：ノードの深さに対するペナルティ(本実験では 0.9)

C_2：2 つのノード間の距離に対するペナルティ(本実験では 0.95)

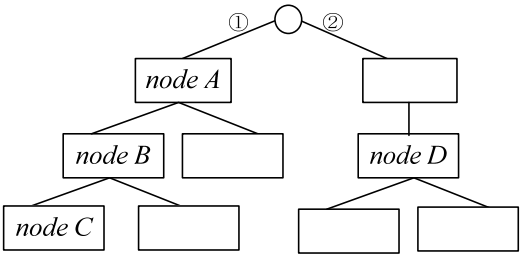


図 3.7　依存構造木例

図 3.7 に示す依存構造木例では，1 つのノードからなる部分木の{*nodeA*}と{*nodeC*}，２つのノードからなる部分木の{*nodeB*, *nodeC*}と{*nodeC*, *nodeD*}に対する事前分布 $\pi_\theta(\theta_t)$ は以下の値となる．

$$\pi_\theta(nodeA) \propto 0.9^1 = 0.9$$
$$\pi_\theta(nodeC) \propto 0.9^3 = 0.729$$
$$\pi_\theta(nodeB, nodeC) \propto 0.9^2 \times 0.9^3 \times 0.95^0 = 0.590$$
$$\pi_\theta(nodeC, nodeD) \propto 0.9^2 \times 0.9^3 \times 0.95^4 = 0.481$$

根ノードに近い{*nodeA*}は葉ノードの{*nodeC*}より $\pi_\theta(\theta)$ の値が大きく，連続するノードの{*nodeB*, *nodeC*}は，遠くに位置するノード{*nodeC*, *nodeD*}より $\pi_\theta(\theta)$ の値が大きくなる．

学習データの正例の依存構造木から生成された大量の部分木（弱学習器）に対して，式(3.6)の値により一定数 *M* 個をサンプリングし(図 2.2 Step1)，次に選択された *M* 個に対して Importance weight を計算する(図 2.2 Step2)．この Importance weight の値によって，利用する弱学習器を決定する．Importance weight の値が高いほど，次の処理でも利用される確率が高くなる．さらに，改めて *M* 個の弱学習器をサンプリングし，同様の処理を *T* 回繰り返すことにより，*T* 個の弱学習器がカスケードしされた列が *M* 個出来る．これを利用して，最終的に式(2.11)により 2 値判別を行う．

3.3.2　「場所」を映像とともに説明する文章の判別実験

NHK で放送された紀行番組「わが心の旅」のクローズドキャプションを対象として，入力文章が，「場所」を映像とともに説明している定型的な文章と，場所を表す単語があるが場所を映像とともに説明していない文章のどちらに類似しているか判別する実験を行った．3.2.4 の実験で用いたデータの一部から，「場所」を映像とともに説明している定型的な文章 154 区間を抜き出し，学習データの正例とした．負例も，正例と同数だけ人手により無作為に抽出した．この学習データを 2 つに分け，一方を学習データ，他方をテストデータとした交差検定による実験を行った．使用するノード数を 3 個とした時，1 回の試行では，平均 10655 個の弱学習器が生成された．

弱学習器の変数 Θ_t に対する確率分布 $P_t(\Theta_t)$ として利用した式(2.6)の Gibbs 方程式において係数 β_t が存在する．この係数 β_t により，どの弱学習器が選択されるかが変動する．

また，サンプリングする弱学習器の数と弱学習器系列の長さも精度に影響する．実験で使用するこれらのパラメータの最適値を推定するため，学習データをテストデータとしたクローズドテストを行った．実験ではサンプリング時の乱数の影響を吸収するために 5 回の試行を行った．Boltzmann annealing (β_0={5.0, 10.0, 15.0, 20.0, 30.0})と Cauchy annealing (β_0={0.2, 0.5, 0.7, 1.0, 1.5})，サンプリングする弱学習器の数 M={500, 1000}，弱学習器系列の長さ T={500,1000}としたときの判別結果の正解率を表 3.7 と表 3.8 に示す. 実験の結果，サンプリングする弱学習器の数 M=1000，弱学習器系列の長さ T=1000 で β_0=0.7 の場合の Cauchy annealing の正解率が最良であった．

求められたパラメータの最適値を使用して，学習データとは異なるテストデータを利用した文章判別実験を行った．パラメータ値推定処理と同様に，サンプリング時の乱数の影響を吸収するために 5 回の試行を行った．

比較対象として，3.2 で説明した AdaBoost を利用した手法と，Naïve Bayes 分類器を利用

表 3.7　Boltzmann annealing による判別結果の正解率(平均)

	M=500 T=500	M=1000 T=1000
β_0=5.0	139.5/154 (90.6%)	142.1/154 (92.3%)
β_0=10.0	134.8/154 (87.5%)	144.0/154 (93.5%)
β_0=15.0	141.5/154 (91.9%)	142.6/154 (92.7%)
β_0=20.0	140.9/154 (91.5%)	140.8/154 (91.4%)
β_0=30.0	143.5/154 (93.2%)	147.9/154 (96.0%)

表 3.8　Cauchy annealing による判別結果の正解率(平均)
※網目部分が最良の正解率

	M=500 T=500	M=1000 T=1000
β_0=0.2	143.4/154 (93.1%)	145.5/154 (94.5%)
β_0=0.5	137.8/154 (89.5%)	145.4/154 (94.4%)
β_0=0.7	144.8/154 (94.0%)	148.2/154 (96.2%)
β_0=1.0	141.5/154 (91.9%)	145.2/154 (94.3%)
β_0=1.5	141.6/154 (91.9%)	147.4/154 (95.7%)

した手法[40]による実験を行った．AdaBoost を利用した手法では，弱学習器生成までは GibbsBoost を利用した提案手法と同じ処理を行い，式(3.2)による弱学習器を大量に生成する．生成した大量の弱学習器を対象として，重み付きの学習データに対する誤り率を最小とする弱学習器系列を選択し，学習データに対する重みを更新しながら弱学習器系列を決定する．得られた弱学習器と信頼度から，式(2.11)で表される判別関数により，入力文章が定型表現文章区間のクラスに属するか否かを判定した．弱学習器系列の長さは提案手法と同じ T=1000 として実験を行った．

Naïve Bayes 分類器を利用した手法では，文章 x に含まれる単語$\{w_1,w_2,\ldots,w_n\}$に対して，分類対象となるクラス C を決定する．ある文章 x がクラス C に属する確率は，文章 x 中の単語 w_i の生起を独立と仮定し，式(3.7)で定義される．

$$P(C \mid x) = \frac{P(C)P(x \mid C)}{P(x)} = \frac{P(C)\prod_{w_i \in x} P(w_i \mid C)}{P(x)} \propto P(C)\prod_{w_i \in x} P(w_i \mid C) \quad (3.7)$$

また，クラス C における単語 w_i の生起確率は，式(3.8)で定義できる．

$$P(w_i \mid C) = \frac{\sum_{x \in C} N(i,x) + \delta}{\sum_i \sum_{x \in C} N(i,x) + \delta|V|} \quad (3.8)$$

ここで，$N(i,x)$は，文章 x 中の単語 w_i の出現頻度，$|V|$はクラス C 中の単語の種類数を指す．δはラプラス法によるスムージング係数を示し，本実験ではクローズドデータを利用した予備実験で分類精度が最良となった 0.05 を使用した．最終的に，分類対象を定型表現文章区間である c_1，定型表現文章区間でない c_2 の 2 クラスとし，式(3.7)の値の大きいクラス $\hat{C}$ を入力文章 x の属するクラスと判定した．

場所説明をしている定型表現文章区間か否かを判別する実験の，3 つの手法による評価結果を表 3.9 に示す．

表 3.9　場所説明をしている定型区間判別評価結果

	適合率	再現率	F 値
場所を説明する区間 GibbsBoost(提案手法Ⅱ)	138.4/152.2 (90.9%)	138.4/154 (89.9%)	0.904
場所を説明しない区間 GibbsBoost(提案手法Ⅱ)	140.2/155.8 (90.0%)	140.2/154 (91.0%)	0.905
場所を説明する区間 AdaBoost(3.2 節の提案手法Ⅰ)	125/131 (95.4%)	125/154 (81.2%)	0.877
場所を説明しない区間 AdaBoost(3.2 節の提案手法Ⅰ)	148/177 (83.6%)	148/154 (96.1%)	0.894
場所を説明する区間 Naïve Bayes 分類器	133/184 (72.3%)	133/154 (86.4%)	0.787
場所を説明しない区間 Naïve Bayes 分類器	103/124 (83.1%))	103/154 (66.9%)	0.741

Xeon^{TM} 2.80GHz x 2 の PC を用いた各手法における平均学習時間(CPU Time)は, 以下の通りであった.

GibbsBoost (提案手法)	14 分 11 秒
AdaBoost	650 分 34 秒

GibbsBoostを利用した提案手法Ⅱは，Naïve Bayes分類器を利用した手法と比べて全体の精度で大幅に上回り，AdaBoost を利用した提案手法Ⅰとほぼ同程度の精度を保ちながら，学習時間は AdaBoost を利用した従来手法と比べて 45 倍以上の速さを実現していることがわかる.

3.4 生成したメタデータの応用

3.2，3.3 で説明した,「場所」を映像とともに説明する文章区間抽出処理を，放送局で蓄積された番組映像に対して適用することにより，番組映像に対して効率的にメタデータを付与することができる．このメタデータを利用して，自然を扱った番組の中からめったに行けない場所などのシーンを一覧表示し，検索が可能なマルチメディア百科事典[41]の試作を進めている. このシステムでは,「場所を映像とともに説明する映像区間」だけでなく,「動物が映っている区間」をクローズドキャプションから推定する処理結果[42][43]も利用している．この手法では，クローズドキャプションに言葉で現れた「もの」が被写体として出現する時の言葉の特徴を統計的に処理して判定している．図 3.8 にマルチメディア百科事典用のメタデータ生成例，図 3.9 に試作したマルチメディア百科事典を示す．図 3.8 では，映像とともに場所「セレンゲティ国立公園」を紹介する区間と，動物「チーター」と「ウサギ」が被写体として映っている区間が自動抽出されている．図 3.9 上部のマルチメディア百科事典のトップ画面では，自動抽出した映像クリップが，その項目ごとに表示されており，例えば「セレンゲティ国立公園」をクリックすると，選択した場所の映像区間を視聴することができる（図 3.9 下部）．NHK で放送された「地球ふしぎ大自然」20 番組を処理対象とした結果，地名は 38 種類（延べ 47 シーン），動物は 113 種類（延べ 419 シーン）を自動抽出できた．この番組は動物を紹介することを主とする番組のため，自動獲得できた場所の説明は 1 番組で 1~2 箇所程度であるが，他の紀行番組などにも適用することにより，多くの場所紹介シーンを自動抽出可能と考えられる.

3.5 おわりに

本章では，AdaBoost アルゴリズムを利用してクローズドキャプションから定型的な文章区間を抽出する手法を提案した．複数の文からなる文章の特徴を的確にとらえるため，文章から，構文解析結果により単語をノードに持つ依存構造木を生成した後，ノードの飛び越えを許した部分木をその特徴候補として抽出した．この部分木と学習データ中の依存構造木間の類似度を評価することにより，依存構造木で離れた位置にある単語間の関係も考

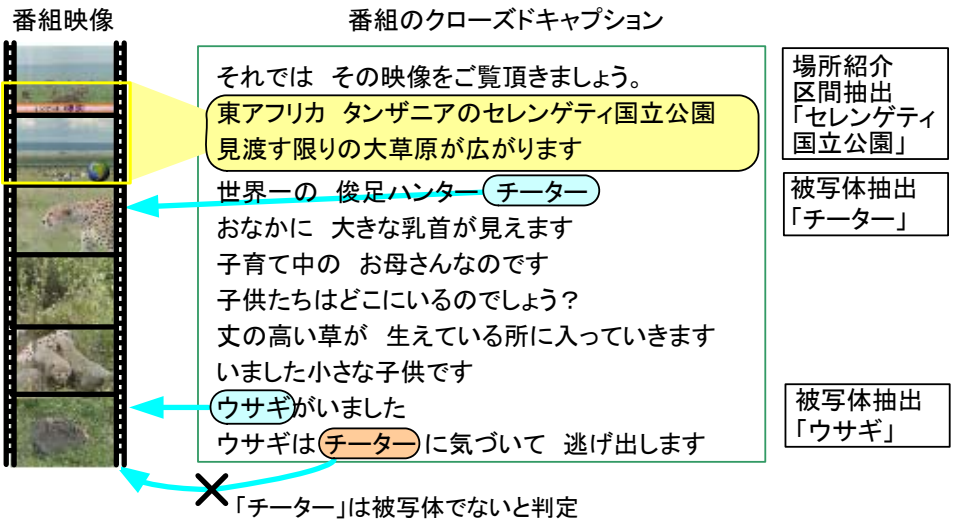


図 3.8 マルチメディア百科事典用のメタデータ生成例

図 3.9 マルチメディア百科事典

慮した処理を実現した．場所を映像とともに説明する定型的な文章区間を抽出する実験により，既存手法より良好な結果が得られ一定の弁別能力があることを示した．

AdaBoost アルゴリズムを利用した手法では，対象とする弱学習器が膨大な量となり計算時間に問題が生じる問題点が残されていた．爆発的に増えた弱学習器を，その事前情報によりサンプリングすることで処理時間の効率化を図り，さらには弱学習器系列を複数生成して判定を行うことにより効果的な処理結果を導く GibbsBoost アルゴリズムを用いた手法を提案し，定型区間の文章か否かを判別する精度を維持しながら，計算時間は 45 倍以上の速さを実現することを確認した．

提案手法により抽出した場所を映像とともに説明する定型的な文章区間により，放送番組に対して，場所の説明部分を示すメタデータを付与することができる．このようなメタデータを自動付与することにより，自然を扱った番組の中から珍しい動物やめったに行けない場所などの映像をまとめたマルチメディア百科事典を試作した．放送局で大量に蓄積している放送番組には貴重な映像が大量に含まれているために，他の番組での再利用や，教育用途などへ有効活用が望まれている．本手法により自動生成したマルチメディア百科事典は，学校教育だけでなく，趣味や生涯学習など，幅広い用途での利用が期待でき，一つの有効なアプリケーションとなり得ると考えられる．

4 章 生放送スポーツ番組を対象としたメタデータ自動生成

本章では，サッカー中継番組（以後，サッカー番組）におけるアナウンスコメントを入力として，サッカーの試合で発生するイベントを構成する区間を抽出し，さらに，各区間のイベント名とその主関与者となるイベント動作主を抽出することにより，サッカー中継番組のメタデータを自動生成する手法につい述べる．

4.1 はじめに

スポーツなどの生放送で中継される番組は，番組のハイライトシーンなどを選択して見るダイジェスト視聴などのアプリケーションが効果的と考えられる．そのため，放送局では，放送番組のどの部分にどのようなイベントが発生したかを示すメタデータの付与は重要な課題となっている．しかし，生放送スポーツ中継番組に対して，人手による均質なメタデータをリアルタイムに付与することは，相当なスキルと労力を必要とする．そこで，生放送スポーツ中継番組を対象としたメタデータ制作システムの開発を進めている[23]．このメタデータ制作システムでは，映像認識処理，音声認識処理，そして自然言語処理など複数の技術（機能モジュールと呼ぶ）を利用して番組を解析し，得られた情報を組み合わせてメタデータを制作することを目指している．

従来，スポーツ番組に対してメタデータを付与する研究は，映像解析によるアプローチが数多く行われてきた．サッカー番組を対象とする研究として，色情報を用いて選手，審判などの位置を抽出するもの[7]，パス，シュートなどのボールの動きは，選手から放射状に連続して検出されるという規則性を用いて，ボールの軌跡を追う報告がある[8]．試合中の特定イベントの検出方法としては，選手の動きとボールの位置を画像から抽出し，組み合わせることによりコーナーキックを検出するもの[9]，そのほか複数の選手による長い時間にわたる複雑な動作パターンを，各々の選手の位置と動作により局所的な特徴シンボルの組み合わせとして記述する報告もある[44]．また，試合の解析に重要であるチームワークやプレッシャーを定量的に評価する手法の報告もある[45][46]．しかし，映像情報を解析する手法は，処理時間や精度に問題が残されている．

そこで本章では，自然言語処理を用いて生放送スポーツ中継番組を解析する手法を提案する．生中継番組の一つであるサッカー中継番組を対象とし，サッカー番組のアナウンスコメントからサッカーの試合で発生するシュートやゴールなどのイベントを構成する区間（以後，イベント発生区間と呼ぶ）を抽出し，各イベント発生区間で発生しているイベントとその主関与者となるイベント動作主を抽出する手法について説明する．このイベント発生区間，イベント名，イベント動作主名が，サッカー番組に対するメタデータとなる．また，サッカー中継番組以外の中継番組への応用も可能な手法であり，放送局が所有する大量のコンテンツ管理にも有益と考えられる．イベント発生区間を抽出してからイベントを特定する処理を行うことにより，キーワードマッチングなどの既存の手法と比べて高精度なイベント検出が可能となる．

4.2 では，処理対象とするコメントの特徴について説明する．スポーツ中継番組におけるアナウンサーと解説者のコメントの特徴を，放送されたサッカー番組と情報系番組，ニュース番組を比較することにより調査した．4.3 では，コメントを解析して映像のメタデータとなる情報を抽出する提案手法の詳細と，提案手法による実験，評価結果を説明する．4.4 では，生放送スポーツ中継番組に対して，自然言語処理に加え，映像解析処理，音声認識処理，音響解析処理などにより抽出したメタデータの基となる情報を統合し，より信頼できるメタデータを生成するアプローチについて言及する．4.5 では，生成したメタデータによりスポーツ中継番組から重要なシーンを抜き出した要約を生成し，テレビ番組を記述できるテキストベースの言語 TVML[47]を利用することにより，CG キャスターによる解説付きの要約番組を自動生成するアプリケーションを紹介する．

4.2 サッカー中継番組におけるコメントの特徴調査

サッカー番組におけるアナウンサーと解説者のコメントの特徴を，放送されたサッカー番組(1 試合)を対象として調査した．比較対象としてバラエティ番組(「ためしてガッテン」)のコメント 1 番組分，ニュース番組 1 日分を利用した．文節数，異なり単語数，発話者に関する調査の結果は下記の通りである．

1 文の平均文節数

サッカー	4.6 文節
バラエティ	6.1 文節
ニュース	13.3 文節

異なり単語数（1000 語の名詞・動詞中）

サッカー	136 語
バラエティ	213 語
ニュース	308 語

サッカー番組では，1 文の文節数が他の番組と比較して少ないため，他の番組より複文になる可能性が低いことがわかる．また，サッカー番組では異なり単語数も少なく，ある程度決まった単語が使われると考えられる．統計処理による学習では，単語の表記そのものを素性として利用すると，学習データに出現しない単語がテストデータに頻出するスパースネスの問題が課題となることが多い．しかし，サッカー番組では使われる単語数が少ないため，単語の表記を素性として利用してもスパースネスの問題が少ないと考えられる．提案手法では，学習のための素性を選択する際に，これらの特徴を利用する．

コメントと番組内容との関連を考えると，サッカー番組中の発話は，その内容により以下の 2 種類に分類できる．

試合記述文 : ボールタッチしている選手を中心に，試合の流れや発生したイベントについて実況しているコメント

解説文 : 試合の流れとは直接関係しない補足的な解説で，終わったばかりのイベントや他の試合について解説しているコメント

各発話者のコメントが，「試合記述文」と「解説文」のいずれであるかを人手により判定した結果を表 4.1 に示す.

表 4.1 から，解説者のコメントは解説文であることが圧倒的に多く，発話者の情報は，コメントが試合記述文と解説文のどちらであるかを判定するための素性として有効であると考えられる.

表 4.1　各話者に対するコメント種類の割合

	試合記述文	解説文
アナウンサー	978/1699 (57.6%)	721/1699 (42.4%)
解説者	52/598 (8.7%)	546/598 (91.3%)
その他 (レポーターなど)	6/22 (27.3%)	16/22 (72.7%)

表 4.2　アナウンスコメントの書き起こしデータ(一部)

開始点	終了点	発話者	アナウンスコメント
15074	15134	ア	前半はまもなく 6 分になります
15194	15254	ア	川口です
15224	15284	解	いいですね，川口
15314	15404	ア	川口からのそして早いフィード
15374	15434	ア	エジミウソン
15404	15464	ア	スピードある
15464	15524	ア	相馬だけしかディフェンスはいない
15524	15584	ア	ただ攻撃の枚数がちょっと足りなかったが
15554	15614	解	いやー右いっぱい空いてますよ
15584	15614	ア	シュート打った
15644	15764	ア	あーここはミドルレンジからシュートを狙って来ましたユサンチョル
15794	15884	ア	まあ 4 試合連続ゴール中のフォワードのシュートですから
15884	15974	解	まあとりあえずに挨拶代わりに打っとこうという感じですよね

(発話者欄の「ア」はアナウンサー,「解」は解説者)

提案手法で処理対象とするデータを，表 4.2 に示す．このデータは，コメントの他にその話者（アナウンサー，解説者，レポーターなど），コメントの開始時間，終了時間（放送開始からのフレーム数．30 フレーム＝1 秒）からなる．4.3.3 で示す実験では，これらのデータは人手により作成し，コメントの発話間隔が一定時間あいた場合をコメントの区切れと判断して一文としている．

現在，音声認識に基づく字幕放送も一部行われているので，音声認識手法を実験で用いる方法もあるが，100%の精度が得られないため，提案手法の評価には適さないと考え，今回は人手による書き起こしコメントを処理対象とする．

生放送字幕では，番組音声を聞きながら言い直した音声を認識するリスピーク方式[13]による音声認識システムも利用されている．また，メタデータ生成のためのスポーツ中継番組の音声認識処理の研究も進められており[14]，サッカー中継の実況マイクの出力を対象にした音声認識実験では，絶叫する場面などが含まれる厳しい状況ながら単語認識率82.0%の精度が得られているため，そう遠くない段階で，完全なテキストデータが得られるものと考える．

話者についても，それぞれに接話マイクが用意されているため，その出力情報により容易に識別できる．

4.3　サッカーで発生するイベントに関するメタデータ生成

4.2 では，サッカー番組のコメントの特徴について説明した．この特徴を利用して，サッカー番組からイベント発生区間を抽出し，各イベント発生区間で起きたイベントとそのイ

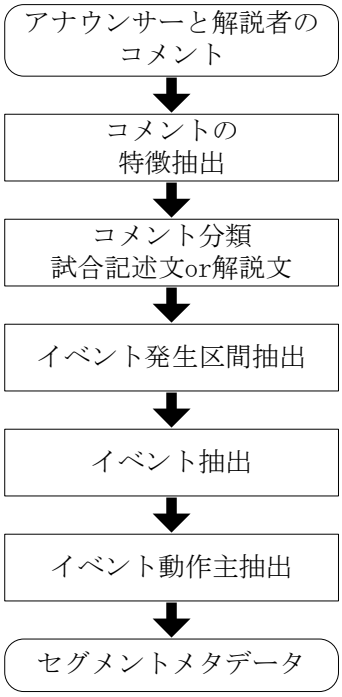


図 4.1　メタデータ生成処理手順

ベント動作主を抽出することにより，メタデータを抽出する手法を提案する．手順を図 4.1 に示す．

まず，コメントの各文から特徴を抽出して，その文が前節で言及した「試合記述文」と「解説文」のいずれであるかを判定する．判定には，2 つのクラスに分類するための機械学習手法であるサポートベクターマシン[28]（以後，SVM）を利用する．SVM による学習は，サポートベクターとなる強力な素性が 2 つのクラスに分類する超平面決定に寄与するため，少量の学習データでも良い特徴が見つかれば精度の良い学習器を生成することができる．実験で利用したアナウンスコメントは 13 試合と少量であったため，提案手法では SVM を利用している．

あらかじめ学習データを与えて学習を行い，その結果に基づき処理対象とするコメントの一つの文が試合記述文と解説文のいずれであるかを判定する．そしてイベントを構成する区間であるイベント発生区間を抽出する．最後に，イベント発生区間中に出現したパスやシュートなどのイベントとそのイベント動作主を抽出し，映像内容を記述するメタデータとする．以下，各処理について説明する．

4.3.1 コメント分類

アナウンサーと解説者のコメントを，SVM を利用して「試合記述文」と「解説文」のいずれであるかを判定するために，コメントの特徴を抽出し，素性とする．この処理では，まず，コメントを単語ごとに分かち書きにする形態素解析と，単語間の係り受け関係を抽出する構文解析を行う．その結果を利用して，一つの文から表 4.3 に示す特徴を抽出する．

統計ベースによる言語モデルの学習では，スパースネスの問題を防ぐために，単語の表記そのものを利用しないで，その上位概念や品詞などを利用することが多い．しかし，4.2 における調査により，サッカー中継番組のコメントは，使用される単語の異なり数がニュースなどに比べて少ないことが解った．そこで本手法では，特徴 1 と特徴 5，そして特徴 7 において表記そのものを特徴として利用する．試合開始前に出場登録された選手名とアナウンサー名，解説者名が予めわかるため，特徴 3〜特徴 6，特徴 17 ではその情報を利用する．また，選手名とチーム名に限り，表記そのものではなく上位概念を利用し，全ての選手名とチーム名をそれぞれ同一の記号列に置き換える．サッカー中継番組のコメントには複文が少ないという特徴もあるため，従属節は考慮せず，主節にある用言とその格要素を素性として選択している．主節に用言が無いことも考えられるため，その場合は最終文節の名詞を特徴 1 で選択する．

図 4.2 に，特徴抽出例を示す．図中の表の各列は，表 4.3 に示した特徴 1〜特徴 18 に対応する．例えば，コメント 2 からは，最終文節「シュート」，最終文節を修飾する人名以外の表記「チャンス」，接続詞が「存在する(1)」，そして，それ以外の項目が「存在しない(ϕ)」という特徴が抽出されている．これらの特徴を素性として，手作業で与えた正解データにより式(2.18)の判定関数を学習する．式(2.18)では，入力データ x と学習データ x_i 間の内積をとっているが，「シュート」や「チャンス」などの表記そのものが入る特徴は，その表記

表 4.3　抽出する特徴

開始点	アナウンスコメント
特徴 1	最終文節の動詞（または名詞）の表記
特徴 2	最終文節の格
特徴 3	最終文節を修飾する人名の有無
特徴 4	最終文節を修飾する人名の格
特徴 5	最終文節を修飾する人名以外の表記
特徴 6	最終文節を修飾する人名以外の格
特徴 7	最終文節を修飾する動詞句の表記
特徴 8	接続詞の有無
特徴 9	助動詞（丁寧）の有無(最終句)
特徴 10	間投助詞の有無(最終句)
特徴 11	助動詞（完了）の有無(最終句)
特徴 12	助動詞（過去）の有無(最終句)
特徴 13	接続助詞の有無
特徴 14	終助詞の有無(最終句)
特徴 15	助動詞（断定）の有無(最終句)
特徴 16	助動詞（打消）の有無(最終句)
特徴 17	解説者・アナウンサーの名前の出現の有無
特徴 18	発話者(アナウンサー:0, それ以外:1)

アナウンスコメント１：今野とも奪い合いましたが，最後のシュートは浮いてしまいました.

アナウンスコメント 2：さあまた岡野がまたスピードを生かす，抜ける，抜けた，チャンス，シュート.

特徴	1	2	3	4	5	6	7	8	9	10
	11	12	13	14	15	16	17	18		
アナコメ 1	**浮く**	ϕ	ϕ	ϕ	**シュート**	**は**	**奪い合う**	ϕ	1	ϕ
	ϕ	1	1	ϕ	ϕ	ϕ	ϕ	ϕ		
アナコメ 2	**シュート**	ϕ	ϕ	ϕ	**チャンス**	ϕ	ϕ	1	ϕ	ϕ
	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ		

図 4.2　アナウンスコメントからの特徴点抽出例

が完全に一致したもののみ，その要素間の内積の値を 1 として計算する．判定関数の変数に各文に含まれる特徴を代入することにより，コメントが「試合記述文」と「解説文」のいずれであるかを判定する．

4.3.2　メタデータ生成処理

イベント発生区間抽出

サッカーのダイジェスト番組を制作する場合,「シュート」シーンでは「シュート」に到るまでの経緯も重要と考えられる．そこで，サッカーで発生する「シュート」などの重要イベントの前後で発生した「パス」などの関連イベントも含めてイベント発生区間として抽出する．提案手法では，以下の２項目を手掛かりとして，サッカーの試合におけるイベントの切れ目を判定し，イベント発生区間を抽出する．

(a)　試合実況ではない発話
(b)　一定時間コメント無し

試合実況ではないコメントとは「解説文」に該当するので，(a)は解説文か否かにより判断する．ここで，発生しているイベントに対するコメントの途中で短時間の解説が出現することがあるため, 2 文以上連続して発話された解説文をイベントの切れ目とする. (b)は，コメントの開始時間，終了時間を利用して判断する．

この処理により，時系列に並ぶコメント列に対して，そのイベントの切れ目が挿入される．このイベントの切れ目の間にある試合記述文集合に対応する番組映像を，イベント発生区間と判定して抽出する．イベント発生区間抽出例を図 4.3 に示す.

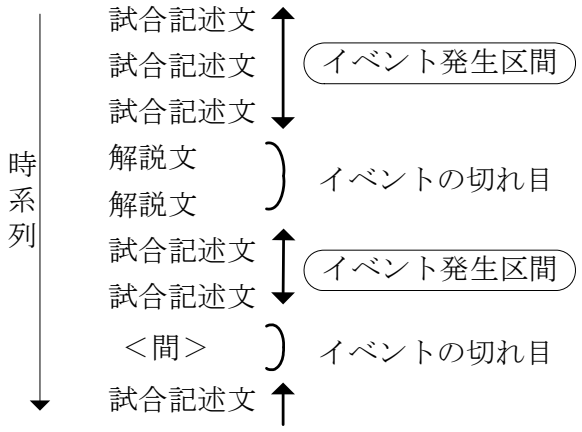


図 4.3　イベント発生区間抽出例

イベント抽出

抽出されたイベント発生区間から，各イベント発生区間で起きた重要なイベントを抽出する．今回,「シュート」「ゴール」「ファウル」「イエローカード・レッドカード」「フリーキック」「コーナーキック」「オフサイド」のイベントを処理対象とした．ここで，一つのイベント発生区間に対して，複数のイベントの抽出を許可している．例えば,「コーナーキック」「シュート」「ゴール」が連続して発生した場合には，同じイベント発生区間にこれ

ら 3 つのイベントが含まれる. また, レッドカードは発生回数が少ないと予想されるため, イエローカードとともに扱う．従来手法では，コメントとこれらのイベントを表現する単語とのキーワードマッチングを行い，イベント候補を抽出している[16][17][18]が，これらの単語が，コメント中に存在しても，実際にそのイベントが映像中に発生しないことも多い．例えば,「上野にはミドルシュートがあります」というコメントがあっても，試合で実際にシュートは打たれていない. 提案手法では, 抽出したイベント発生区間を構成する「試合記述文」に対してキーワードマッチングを行い，一致したイベントを，該当するイベント発生区間のイベントとする．このため，実際に発生しているイベントのみを精度良く抽出できる．

各イベントには複数の表現が存在する．例えば,「ゴール」イベントでは,「ゴール」の他に,「得点」,「先制点」,「逆転」,「決める」などアナウンサーや試合状況によって表現方法が異なる．そこで，再現率の向上のために，一つのイベントに対して複数の関連する単語もマッチングさせる. イベントに関連する単語の一覧を表 4.4 に示す. 選択した単語は, イベント名が含まれる単語，またはイベント名の言い換え表現である.「ゴール」イベントに関しては，アナウンサーがイベント発生時に一般的に良く使うと考えられる単語も追加している．

表 4.4　イベントに関連する単語

イベント名	イベントに関連する単語
シュート	シュート,ループシュート,ミドルシュート, ロングシュート,ヘッディングシュート, ボレーシュート,バナナシュート, ドライブシュート
ゴール	ゴール,得点,先制,逆転,同点,決める,決まる
イエローカード・ レッドカード	イエローカード,イエロー,警告,レッドカード, 退場
フリーキック	フリーキック,間接フリーキック, 直接フリーキック
コーナーキック	コーナーキック,ショートコーナー，コーナー
オフサイド	オフサイド
ファウル	ファウル

イベント動作主抽出

各イベントに対して，その主関与者となるイベント動作主を推定する．この処理では，以下の優先順位により，動作主の選手名を抽出する．

1. 選手名が直接イベントを修飾する場合
   「小笠原のシュート」

2. イベントが含まれる文の主語が選手名であり，選手名が修飾する動詞の目的語がイベントである場合
   「上野がシュート狙って来た」
3. イベントの後部に選手名がある場合
   「シュートを打って来たエジミウソン」
4. イベントが含まれる文に選手名がある場合
   「三浦，中に入って，ミドルシュート打って来た」
5. イベントが含まれる文の一つ後ろの文に，選手名があり，対象イベント以外のイベントが存在しない場合
   「シュート」
   「上野の左足は強烈ですね」
6. イベントが含まれる文の一つ前の文に，選手名があり，対象イベント以外のイベントが存在しない場合
   「ユサンチョルが持った」
   「シュート」

これらの項目の何れにも該当しない場合は，動作主不明とする．

4.3.3 メタデータ生成実験

提案手法の有効性を検証するために，NHK で放送された 13 試合のサッカー番組のコメントの書き起こしテキストデータ(24245 文）を利用して，実験を行った．対象とした 13 試合を実況したアナウンサーは計 6 名(4 試合担当 1 名, 2 試合担当 4 名, 1 試合担当 1 名), 解説者は計 9 名であった．以下にコメント分類実験，イベント抽出実験，そしてイベント動作主抽出実験の詳細を記す．

コメント分類実験

コメントが試合記述文と解説文のいずれであるかを判定する処理を検証するため，人手により, 対象とする全てのコメントに対して正解を与える. 判定処理は SVM^{light} [48]を用い，学習データを 12 試合，テストデータを 1 試合として 13 通りの交差検定を行った．表 4.5 に，判定された結果で正しかった割合を示す適合率と，正解データから正しい結果が得られた割合を示す再現率を示す．

表 4.5 の結果を検証するために，従来手法として有名なベクトル空間モデルを利用した手法[35]により同様の実験を行った．この手法では，人手により与えられた試合記述文と解説文，そしてテストデータに含まれる単語に対して TFIDF 値[39]により重み付けをしたベクトル空間モデルを作成し，テストデータから生成したベクトルと，試合記述文集合，解説文集合から生成したベクトルとの余弦を指標として，テストデータが何れに属するかを決定する．判定結果を表 4.6 に示す．提案手法の結果は，全ての項目でベクトル空間モデルを利用した手法の結果を上回り，手法の有効性が確認できる．

素性として利用した 18 種の特徴の中で, どの特徴がコメントの分類に貢献していたかを

表 4.5　提案手法による「試合記述文」「解説文」判定結果

	適合率	再現率
試合記述文	8161/9104 (89.6%)	8161/9622 (84.8%)
解説文	13680/15141 (90.4%)	13680/14623 (93.6%)

表 4.6　ベクトル空間モデルを利用した手法による
「試合記述文」「解説文」判定結果

	適合率	再現率
試合記述文	7734/11234 (68.8%)	7734/9622 (80.4%)
解説文	11123/13011 (85.5%)	11123/14623 (76.1%)

判定するため, 特徴を一つだけ除いて 17 種の特徴を使った素性によるコメント分類実験を行なった．結果を評価し，その適合率と再現率の調和平均(F 値)，F 値のマクロ平均とマイクロ平均を算出した．マクロ平均とは，試合記述文と解説文に含まれる事例数と無関係に２つの F 値を単純平均した値であり，式(4.1)で定義される．

$$F_{\text{macro}} = \frac{F_{\text{試合記述文}} + F_{\text{解説文}}}{2} \tag{4.1}$$

一方，マイクロ平均は事例数を考慮した平均であり，試合記述文と解説文の２つの分類結果を合わせた適合率と再現率から式(4.2)により定義される．

$$F_{\text{macro}} = \frac{F_{\text{試合記述文}} \times N(\text{試合記述文}) + F_{\text{解説文}} \times N(\text{解説文})}{N(\text{試合記述文}) + N(\text{解説文})} \tag{4.2}$$

2 つの F 値を指標に用いることにより，試合記述文と解説文の分類性能の違いを吸収することができる．特徴の有効性比較を行った結果を表 4.7 に示す．表 4.7 では F 値が小さいものほど, 除いた特徴がコメント分類に有効であったと判断できる. 表 4.7 の結果では, 最終文節の動詞または名詞の表記の特徴を除いた特徴 1 以外を利用した場合の F 値のマクロ平均とマイクロ平均がともに最低であり，文の最終文節の動詞または名詞の表記が最も有効な特徴であることが分かる．逆に，特徴 4(最終文節を修飾する人名の格)や特徴 16(助動詞打消しの有無)を除いた結果は，F 値のマクロ平均とマイクロ平均が全てを利用したものより上昇しており，有効な特徴でないと判断できる．現在の試合状況を説明する「試合記述文」は現在形，過去の事象などを説明する「解説文」は過去形や過去完了形が多用されやすいと考えられるが，特徴 11(助動詞「完了」の有無)や特徴 12(助動詞「過去」の有無)なども，F 値のマクロ平均とマイクロ平均は全てを利用したものとほぼ同じであった．動

表 4.7　特徴の有効性比較

利用した特徴	F 値 試合記述文	F 値 解説文	F 値 マクロ平均	F 値 マイクロ平均
ALL	0.8716	0.9192	0.8954	0.9008
特徴 1 以外	0.7761	0.8269	0.8015	0.8047
特徴 2 以外	0.8713	0.9191	0.8903	0.8959
特徴 3 以外	0.8654	0.9151	0.8903	0.8959
特徴 4 以外	0.8723	0.9194	0.8959	0.9012
特徴 5 以外	0.8572	0.9121	0.8846	0.8912
特徴 6 以外	0.8692	0.9180	0.8936	0.8992
特徴 7 以外	0.8657	0.9166	0.8912	0.8971
特徴 8 以外	0.8707	0.9188	0.8947	0.9002
特徴 9 以外	0.8699	0.9177	0.8938	0.8992
特徴 10 以外	0.8676	0.9171	0.8924	0.8981
特徴 11 以外	0.8715	0.9192	0.8953	0.9008
特徴 12 以外	0.8713	0.9190	0.8952	0.9006
特徴 13 以外	0.8708	0.9189	0.8948	0.9003
特徴 14 以外	0.8694	0.9178	0.8936	0.8991
特徴 15 以外	0.8695	0.9178	0.8937	0.8991
特徴 16 以外	0.8719	0.9193	0.8956	0.9010
特徴 17 以外	0.8716	0.9192	0.8954	0.9008
特徴 18 以外	0.8657	0.9150	0.8904	0.8959

詞の過去形や完了形といった情報は，「試合記述文」と「解説文」の判定に，大きく貢献している訳ではないことがわかる．

イベント抽出実験

全ての特徴を利用したコメントの分類結果を利用して，イベント発生区間を抽出した．抽出されたイベント発生区間は 1 試合平均 113.5 個，1 区間の平均継続時間は 20.1 秒であった．さらに，各イベント発生区間で起きたイベントを抽出した．一つの試合を対象とした抽出結果の一部を表 4.8 に示す．表 4.8 における時間は番組開始からのフレーム数（30 フレームで 1 秒）を示す．実験で利用した 13 試合のうち，映像が入手できた 5 試合(アナウンサー計 4 名)を対象として，手作業により正解イベントを抽出し，対象イベントの抽出結果を検証した．この際，正解イベントは発生時刻で与え，抽出したイベントが含まれるイベント発生区間と正解イベントの発生時刻の比較を行う．手作業により与えられたイベント発生時刻が，抽出された同じイベント名のイベント発生区間に含まれる場合を正解とした．結果を表 4.9 に示す．

表 4.8　イベント抽出結果（一部）

時間(フレーム数)	イベント	動作主
11598f～12887f	フリーキック	ビスマルク
14685f～15524f	コーナーキック	小村
15584f～15764f	シュート	ユサンチョル
22447f～22867f	シュート	上野
29460f～29610f	シュート	三浦
29970f～30239f	ファウル	中村
33566f～33716f	シュート	小笠原
34525f～34705f	シュート	ビスマルク
37012f～37582f	ファウル	中田
37012f～37582f	イエローカード	中田
38481f～39631f	フリーキック	中村
39710f～40069f	ファウル	小村
44535f～45644f	イエローカード	小村
49900f～50199f	シュート	小笠原
52687f～53646f	フリーキック	小村
52687f～53646f	ファウル	波戸
61348f～61738f	コーナーキック	小笠原
73186f～73726f	フリーキック	波戸
81388f～81728f	シュート	中田
85504f～85923f	シュート	上野
87722f～88681f	シュート	ユサンチョル
89920f～89400f	ファウル	平瀬

適合率は，抽出したイベントの含まれるイベント発生区間に，対応する正解イベントが含まれる割合を示し，「シュート」が 90.3%，「ファウル」が 94.1%，「イエローカード・レッドカード」が 90.5%，「オフサイド」が 100%と，これらの 4 つのイベントに対しては良好な結果が得られた．「ゴール」に対する適合率は 38.5%と低い．イベントに関連する単語「決める」などの動詞は，再現率向上には有効であるが，適合率を低下させてしまった．イベント抽出時にキーワードマッチングだけでなく，構文構造によりイベントの有無を判定する処理が効果的と考えられる．また，「フリーキック」と「コーナーキック」の適合率も他のイベントと比べて低い．これらは権利獲得型のイベントであり，「フリーキック」や「コーナーキック」の権利が与えられたときにアナウンサーは説明することが多くみられた．この際の誤検出が多いため，権利獲得型のイベントはその後の区間もイベント発生区間に含めるなど，イベントの種類に応じた抽出処理により精度向上が期待できる．

表 4.9　提案手法によるイベント抽出検証結果

	適合率	全イベントに対する再現率	コメント上の再現率	F 値
シュート	93/103 (90.3%)	93/151 (61.6%)	93/117 (79.5%)	0.845
ゴール	5/13 (38.5%)	5/16 (31.3%)	5/16 (31.3%)	0.345
ファウル(オフサイド以外)	48/51 (94.1%)	48/188 (25.5%)	48/63 (76.2%)	0.809
イエローカード・レッドカード	19/21 (90.5%)	19/26 (73.1%)	19/26 (73.1%)	0.556
フリーキック	15/22 (68.2%)	15/219 (6.8%)	15/32 (49.6%)	0.556
コーナーキック	22/37 (59.5%)	22/41 (53.7%)	22/35 (62.9%)	0.612
オフサイド	21/21 (100%)	21/33 (63.6)	21/26 (80.8%)	0.894
total	223/268 (83.2%)	223/672 (33.2%)	223/315 (70.8%)	0.765

全イベントに対する再現率は，試合で発生したすべてのイベントに対する抽出できたイベントの割合を示す．この再現率の平均は 33.2%と低い結果であった．これは，実際にイベントが発生した際に，そのイベントを表す単語がコメントに出現する割合が低いことが原因となっている．「シュート」と「ファウル」を対象として，これらのイベントがコメントに出現した割合を人手により評価した結果,「シュート」イベントが発生していても，コメント中には言及されていないケースが 22.5%,「ファウル」については 66.5%も存在した．「間接フリーキック」となるような軽微な反則は，コメントでは，ほとんど言及されていない．コメントに出現する割合が，本手法の全イベントに対する再現率の限界値となるため, コメントとして取り上げられにくいような重要度が低いイベントは, 精度が低くなる．

本手法の言語処理的な精度を評価するため，人間によって番組映像を参照しないでイベントが発生していると推測できるコメントとイベント名を特定してもらい，そのコメントが持つ開始時間から終了時間までをイベント発生区間とした正解データを作成した．この正解イベント発生区間が，提案手法により抽出した該当イベントの発生した区間に含まれる割合を示す再現率(表 4.9 コメント上の再現率)を調査した. 主要イベントの平均は 70.8%という結果が得られた．

提案手法の有効性を検証するため，従来研究で行われているキーワードマッチングのみによるイベント抽出手法との比較実験を行った．キーワードマッチング手法では，各イベントに対して複数のキーワードを対象とし，そのキーワード出現の有無により，イベントの発生の有無を判定する．キーワードは，提案手法のイベント抽出処理で利用した表 4.4

に示すイベントに関連する単語を利用し，同一イベントに対するキーワードが 5 秒以内に抽出された場合に限り，同じイベントを指すと判断した．キーワードマッチングによるイベント抽出を検証した結果を表 4.10 に示す．発生イベントを網羅的に抽出するため再現率

表 4.10　キーワードマッチングによるイベント抽出検証結果

	適合率	全イベントに対する再現率	コメント上の再現率	F 値
シュート	116/231 (50.2%)	116/151 (76.8%)	116/117 (99.1%)	0.667
ゴール	11/181 (6.1%)	11/16 (68.8%)	11/16 (68.8%)	0.112
ファウル(オフサイド以外)	61/89 (68.5%)	61/188 (32.4%)	61/63 (96.8%)	0.803
イエローカード・レッドカード	24/69 (34.8%)	24/26 (92.3%)	24/26 (92.3%)	0.505
フリーキック	23/72 (31.9%)	23/219 (10.5%)	23/32 (71.9%)	0.442
コーナーキック	24/55 (43.6%)	24/41 (78.8%)	26/35 (68.6%)	0.533
オフサイド	26/56 (46.4%)	26/33 (78.8%)	26/26 (100%)	0.634
total	285/753 (37.8%)	285/672 (42.4%)	285/315 (90.5%)	0.533

表 4.11　イベント動作主抽出結果

イベント	適合率	再現率
シュート	81/84 (96.4%)	81/93 (87.1%)
ゴール	2/3 (66.7%)	2/5 (40.0%)
ファウル(オフサイド以外)	31/44 (70.5%)	31/48 (64.6%)
イエローカード・レッドカード	19/19 (100%)	19/19 (100%)
フリーキック	10/12 (83.3%)	10/15 (66.7%)
コーナーキック	7/17 (41.2%)	7/22 (31.8%)
オフサイド	11/17 (64.7%)	11/21 (52.4%)
提案手法 total	161/196 (82.1%)	161/223 (72.2%)
ベースライン手法	120/128 (93.7%)	120/233 (53.8%)

は高いが適合率は低い．適合率とコメント上の再現率による調和平均を計算したところ，すべてのイベントに対する F 値は 0.533 であった．提案手法では 0.765 であり，キーワードマッチングのみを利用する手法に比べて高精度にイベント抽出が行われていることがわかる．

イベント動作主抽出実験

イベント抽出に成功したものに対して，イベントの主関与者となるイベント動作主抽出の検証を行った．抽出された動作主の正解率を示す適合率と，対象イベントに対して正解動作主の抽出できた割合を示す再現率を表 4.11 に示す．

比較のためのベースライン手法として，構文構造によりイベントを表す単語の動作主体を判定して，イベント動作主と判断する実験を行った．この処理では，イベント動作主を，イベントを表す単語と修飾関係にある選手名，または，イベントを表す単語と同じ動詞の格要素にある選手名としている．ベースライン手法の結果も表 4.11 に示す．

適合率と再現率の調和平均を表す F 値は，提案手法が 0.768，ベースライン手法が 0.684 であり，構文構造による手法に比べて有効性が確認できた．

「シュート」と「イエローカード・レッドカード」は，高精度にイベント動作主が特定できた．これは，重要なイベントであり，かつイベント動作主が一人に特定できるイベントであったためと考えられる．ここで，「イエローカード・レッドカード」のイベント動作主とは，イエローカードまたはレッドカードを受けた選手を示す．

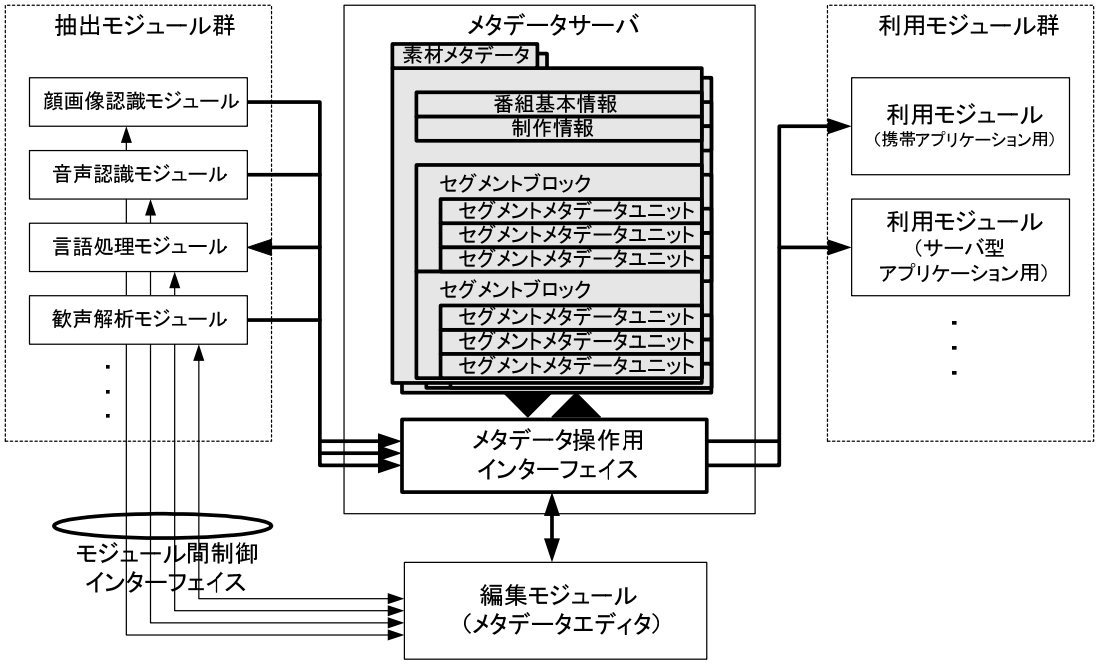
「フリーキック」や「コーナーキック」は，イベント動作主より，ゴール前に居る選手を誤検出することが多く見られた．

また，「ファウル」や「オフサイド」は，2 人の選手が関係するイベントである．今回は，「ファウル」はファウルを与えた選手，「オフサイド」はボールを受けた選手としたが，ファウルを受けた選手やボールを出した選手が誤って抽出されたため，精度が低くなっていた．

「ゴール」は，その表現の多様性のために，誤検出が多くなっていた．精度向上のためには，イベントごとに異なるルールを適応するなどの対処策が考えられる．

4.4 他の情報から生成したメタデータとの統合

前節までに説明した手法により，サッカーの中継番組における特定イベントが発生した区間と，そのイベント名，イベント動作主を抽出することができる．表 4.9 に示す抽出実験結果では，平均適合率は 83.2%，コメント上の再現率は 70.8%であり，従来手法と比較して良好な結果ではあったが，この処理結果を放送局におけるメタデータとして利用するためには，さらなる精度向上が望まれる．そこで，画像，音声，言語などの複数の情報を解析して高精度にメタデータを制作するメタデータ制作システムを試作した．メタデータ制作システムの構成を図 4.4 に示す．



構成要素	機能概要
メタデータサーバ	番組メタデータの蓄積管理 メタデータ操作（作成、変更、削除、検索）
抽出モジュール	シーン切り出し機能、内容記述機能
編集モジュール	セグメントメタデータの読み出し、編集、結合処理、関係記述
利用モジュール	メタデータの検索、統合処理、他の形式への変換など

図 4.4　メタデータ制作システムの構成

```
<?xml version="1.0" encoding="Shift_JIS" ?>
<Soccer>
  <Game>
    <Contents>
      <AudioVisual>
        <SceneCollection>
          <Structure type="1ndHalf">
            <Scene id=S20000805KAvsYM_1stHarf_009>
              <Narration>シュート打った。あーここはミドルレンジからシュートを狙って来ました
                        ユサンチョル。</Narration>
              <Content>
                <Who>ユサンチョル</Who>
                <WhatAction>シュート</WhatAction>
              </Content>
              <MediaTime>
                <MediaRelTimePoint>15584</MediaRelTimePoint>
                <MediaDuration>180</MediaDuration>
              </MediaTime>
            </Scene>
        </SceneCollection>
      </AudioVisual>
    </Contents>
  </Game>
</Soccer>
```

図 4.5　生成メタデータ例

メタデータ制作システムは,「メタデータサーバ」,「抽出モジュール」,「編集モジュール」,そして,「利用モジュール」から構成される.「メタデータサーバ」はデータベース機能を持ち,「メタデータ抽出モジュール」からの処理結果を蓄積,管理する.

「メタデータ抽出モジュール」は,画像解析,音声認識,言語処理などにより番組を解析し,シーンを切り出すセグメンテーションや,切り出された各セグメントに対して内容を記述するラベリング機能を持つ.4.3 で提案した言語情報からメタデータを生成する処理は,このメタデータ抽出モジュールの一つとなる.抽出手法ごとに独立したコンピュータに実装し,ネットワークを利用したモジュール間インターフェイスによりメタデータサーバ,編集モジュールと接続する.今回試作したメタデータ制作システムには,NHK で開発した表 4.12 に示す 8 種の技術[3][6][12][15][49][50][51]を,モジュール化して組み込んだ.メタデータは,「シーンの時間情報」(例:12〜18 秒)と,「内容記述情報」(例:得点場面)で構成される.異なる機能の抽出手法を組み合わせることで,メタデータの区間と内容を得ることができる.

「編集モジュール」では,抽出モジュールで生成されメタデータサーバに集められたメタデータを,メタデータエディタと呼ぶ編集システムで一覧しつつ,人手により,複数の

表 4.12　イベント動作主抽出結果

モジュール名	モジュール機能種別	処理概要
映像解析によるスロー映像区間抽出	セグメンテーション	スロー区間の映像の繰り返し提示パターンを周波数軸から抽出し,スロー映像区間を抽出する
映像解析によるシーン種別判定	セグメンテーション	画像中の特定の色の領域の大きさを閾値判定し,シーンの種別判定を行う(今回は緑色領域に注目し,グランドのシーンを抽出する).
シーン解析によるカット点検出	セグメンテーション	ショットを矩形領域に分割し,パターン化した情報によって表される構図の変化に注目してカット点検出処理を行う
顔画像認識による選手名抽出	ラベリング	顔の特徴点を複数の解像度でウエーブレット特徴による可変テンプレートマッチングを用いた顔画像認識を行う
オブジェクト認識による選手位置検出	ラベリング	オブジェクト認識,追跡技術を利用した,サッカー選手位置検出システムの位置情報を利用してプレイのイベントを抽出
歓声解析による重要イベント区間抽出	セグメンテーション	観客音声収録マイクの音声を利用し,歓声の音声パワーが急に変化した部分を抽出する
音声認識によるアナウンスコメント抽出	ラベリング	ニュース生字幕制作用の音声認識システムを元に作成したもので,スポーツ中継時のアナウンサーマイクからの音声の音響モデル,言語モデルに適用させた音声認識を行う
言語処理によるイベント,イベント動作主抽出	セグメンテーション,ラベリング	音声認識結果を素に,文型情報を利用してアナウンス文を試合記述文と解説文に識別し,試合記述文中に含まれるイベントおよびそのイベントに関連する動作主を抽出する

メディアから得られた情報を統合し，1 つの手法では欠落する情報の補完や，シーンの時間情報と内容記述情報といった相補的な情報を統合して，有効なメタデータを抽出することができる．この結果を人手で確認することによって，放送局における信頼できるメタデータの制作を可能とする．生成されたメタデータは，図 4.5 のような XML 形式で保存される．生成されたメタデータは，「利用モジュール」である VOD のアプリケーションなどに，メタデータ操作用のインターフェイスを介して渡すことができる．

4.5 生成したメタデータの応用

メタデータとユーザプロファイルを利用して個人に適合したテレビ番組を自動生成するシステム TV4U(TV for you)の研究が進められている[52]．TV4U は，ユーザプロファイルをもとに，ユーザ個人の所望する映像やテキスト等の素材を，メタデータを参照して収集し，好みの演出を付加して番組を生成するシステムである．本章で提案した手法により生成されるメタデータを，TV4U における映像素材収集時に利用することにより，1 試合分のサッカー映像をカスタマイズ視聴できるアプリケーションを構築した．

あらかじめユーザは，図 4.6 に示すユーザーインターフェイスを利用し，好きなチーム名，好みの演出，視聴したいイベントなどのユーザ情報を入力する．TV4U は入力されたユーザ情報により，自動生成されたメタデータを使って，ユーザの所望する条件に適合した番組（この場合，サッカーの試合をダイジェストで見られる番組）の生成のために必要な映像素材を，イベントに対応するシーンの映像として検索する．図 4.7 は，演出としてニュース形式が選ばれた例を示す．TV4U により，ユーザが希望したイベントが発生している映像を提示するとともに，CG のキャスターが，その映像に対して説明を加えるといった，サッカーのダイジェスト番組が自動生成される．

ここで，CG のキャスターが話す映像の説明文は，アナウンスコメントの「試合記述文」と「解説文」の判定結果を利用して生成する．検索されたシーンには，試合記述文の集合が対応している．その試合記述文の集合と次のシーンに対応する試合記述文の集合の間は，必ず解説文の集合が存在している．これらの解説文集合は，検索されたシーンで発生したイベントに関する解説であることが多い．そこで，検索されたシーンの直後の解説文集合から，検索されたシーンで発生したイベント名，またはその動作主が含まれている文を取り出し，抽出されたシーンに発話時刻が近い解説文を CG のキャスターが話す映像の説明文とすることができる．例えば，表 4.2 のコメントを解析すると，開始点 15584 フレームのコメントから「シュート」が発生したイベントとして抽出される．イベント「シュート」を受けて CG のキャスターが話す説明文は，「シュート」が含まれるシーンの次の解説文集合中にある開始点 15794 フレームから始まる文「まあ 4 試合連続ゴール中のフォワードのシュートですから」を選ぶことができる．

図 4.6　ＴＶ４Ｕユーザーインターフェイス（個人の嗜好の設定）

図 4.7　ＴＶ４Ｕ番組自動生成例

4.6　おわりに

本章では，サッカー番組のアナウンサーと解説者のコメントを統計処理により解析することにより，サッカー番組に対して番組内容を時間ごとに説明するメタデータを付与する

手法を提案した．コメントを「試合記述文」と「解説文」の 2 種類に分類することにより，イベント発生区間を抽出し，各区間で起きたイベントとイベント動作主を抽出した．コメントの分類実験では, F 値が試合記述文で 0.871, 解説文で 0.919 と良好な結果が得られた．イベント抽出実験では，特定イベントに対しての平均適合率は 83.2%，コメント上の再現率は 70.8%であり，従来手法のキーワードマッチングのみを利用した手法と比べ有効性が認められた．イベント動作主抽出実験では，適合率 82.1%，再現率 72.2%であった．

試合で発生したイベントに対する平均再現率は 33.2%と低い値であった．これは，実際にイベントが発生した際に，そのイベントを表す単語がコメントに出現する割合が低いことが原因となっていた．このため，コメント以外からの情報補間も考慮に入れたメタデータ生成処理が効果的と考えられる．このため，自然言語処理に加え，映像解析処理，音声認識処理，音響解析処理などにより抽出したメタデータの基となる情報を統合し，より信頼できるメタデータを生成するためのメタデータエディタを試作した．

さらに，自動抽出したメタデータを利用して，TV4U における映像素材収集時に利用することにより，1 試合分のサッカー映像をカスタマイズ視聴できるアプリケーションを構築し，提案手法の効果的な活用法を示した．

本章ではサッカー中継番組を前提としたが，ラグビーやアメリカンフットボールなどのスポーツ中継番組でも，コメントは試合記述文と解説文に分類できるので，本手法を適用できると考えている．サッカー番組だけでなく，他のスポーツ番組への適用範囲の拡大を図ることにより，放送される多くの番組への実用的なメタデータ付与支援が可能となる．

5 章　ニュース番組を対象としたメタデータ自動生成

本章では，ニュース番組で使われるテキスト（以後，ニュース記事と呼ぶ）を入力として，一定期間中に発生した話題を抽出し，抽出した話題を要約することにより，ニュース番組に対するメタデータを自動生成する手法について述べる．

5.1　はじめに

ニュース番組には社会の情勢や最新の流行など，豊富な情報が含まれているため，効率的な管理，活用が特に求められている．現在 NHK では，ニュース映像に対して人手により映像の被写体や出来事についての説明文を付与することにより，ニュース映像の管理を行っている．この説明文と，ユーザが入力した検索キーワードとのキーワードマッチングを行うことにより，ニュース番組で利用した映像を検索している．しかし，大量のニュース番組の映像に対して人手により映像説明文を付与する作業には，多大な労力を要し，さらには，放送局では既に膨大な量のニュース映像を蓄積しているため，キーワードマッチングによる検索では，的確な検索は難しくユーザの欲する映像を探す負担が大きい．

ニュースには，政治，経済，社会，スポーツなどの変化しない静的なジャンルの他に，時期とともに変化する動的なジャンルが存在する．この動的なジャンルを，ここでは話題とする．具体的な例では「W杯サッカー」「サミット沖縄会合」「テロ対策基本法案の国会審議」などが話題となる．さらには，各話題の中に，その話題の遷移を表現する代表的な出来事が存在する．例えば，話題「テロ対策基本法案の国会審議」では，「法案の国会提出」から始まり，「衆議院での審議」，「参議院での審議」，「法案の修正」，「法案の可決」などの項目が存在する．このような話題と，話題中の代表的な項目は，インデックスとして放送局のニュース番組管理に利用できる上に，具体的な内容を表現しているため，視聴者が番組を選択するときの重要な鍵となる．

そこで，ニュース記事を解析して，一定期間中に発生した話題を抽出し，抽出した話題を要約する手法について提案する．この処理により得られた話題名，同じ話題に属するニュース記事，話題中の代表的な項目が，ニュース番組に対するメタデータとなる．

従来，複数ニュース記事を対象とした内容分析に関する研究として，テンプレートを利用して，そのスロットを埋めていく手法が提案されている[53]．しかしこの手法では，あらかじめ手作業によりニュースの種類ごとのテンプレートを決めておく必要があり，さらにテンプレートのスロットごとに抽出ルールを生成しなければならない．この作業には大変な労力を要する．また，一つの話題を構成する記事集合をあらかじめクラスタリングして要約を生成する手法が提案されている[54]．しかし，話題中の重要な一つの出来事が，一つのクラスタを形成しなければならないため，重要な出来事の記事数が少ない話題には適さない．本章で提案する要約手法では，話題に特有な係り受け構造を利用して，テンプレートを必要としない要約を実現する．また，クラスタリング処理を行わないため，複数記事が出現しないような内容も要約結果として取り出すことができる．

まず 5.2 では，処理対象とするニュース記事について説明し，その特徴を紹介する．5.3 ではニュース記事を，出現単語の時系列の変化を考慮した手法により分類して，一定期間中に発生した話題を抽出する手法を説明する．5.4 では，抽出した各話題を，話題における単語や統語構造の定型性を評価することにより話題中の代表的な項目を抽出し，複数ニュース記事を要約する手法を提案する．

5.2 ニュース記事の特徴調査

NHK の放送の読み原稿として利用されるニュース記事を処理対象とする．ニュース記事の例を表 5.1 に示す．ニュース記事には，ニュース記事を特定する ID，ニュース記事タイトル，ニュース記事を作成した部局名，ニュース記事テキストなどの情報が含まれる．ニュース記事テキストは，番組のクローズドキャプションとほぼ同じテキストである．ニュース記事は，1 日当たり約 200 記事が作成されている．我々が所有するデータベースには，13 年分のニュース約 98 万記事（約 555 万文）もの大量のテキストデータが蓄積されている．ニュース記事の第一文は"リード"と呼ばれ，伝えるべき 5W1H(when, where, who, what, why, how)などの内容が具体的に記述されている[55]．そのため，リードはニュース記事内容の全貌を説明する事が多い[56]．そこで，本章で提案する手法は，ニュース記事のリードにあたる第 1 文のみを処理対象とする．

このリードでは，主観的な表現，あいまいな形容詞，副詞は使われない．例えば，「良い天気」といった表現は避けられ，「雲ひとつ無い天気」という表現が使われる．

また，使われる動詞にも一定のルールがある．ニュース番組では聞き返しができないため，そのニュース記事中ではできるだけやさしい言葉が求められる．そこで，漢語表現は避けられる．例えば「判明する」の代わりに「わかる」に，「示唆する」の代わりに「ほのめかす」が使われる．さらに，汎用的な動詞表現も避けられる．例えば，「行われる」という動詞表現はできるだけ避けられ，主格が「議論」の場合は「行われる」の代わりに「交わされる」，「会合」の場合は「開かれる」が使われる．他にも，接続助詞「が」を使わない，接続詞も避けるという特徴がある．

さらに，類似する話題では，その話題特有の単語や統語構造が使われることが多い．実際に，「選挙」と「国会審議」に関連する話題に属するニュース記事と，ニュース記事を無作為抽出した「任意の話題」に属するニュース記事の，それぞれ 9227 記事を対象とし，そこに出現する単語の種類を調査した．結果を表 5.2 に示す．任意の話題に出現した名詞は 14,102 種類であったのに対し，「選挙」に関連する話題は 7,292 種類，「国会審議」は 6,521 種類と，同数の記事中で使われる名詞の種類は半数程度であった．動詞は，「任意の話題」が 2,633 種類，「選挙」が 1,780 種類，「国会審議」が 1791 種類と，同様に特定の話題で使われる種類が少なかった．話題特有の単語が存在し，その単語が多用されるために，特定の話題に出現する単語の種類が少ないと考えられる．

同様に，人名，地名，組織名の各話題における出現のべ数を調査した．この結果から，

表 5.1　処理対象とするニュース記事例

X|199806171668
N|再差替・円安問題・日米で協議へ
Z|sakuseibukyoku_name=経済
S|大蔵省は，急激に進んでいる円安・ドル高を食い止めるため，あすにも来日するアメリカのサマーズ財務副長官との間で，市場介入を含めた協調体制を確認したい考えです.
S|このところ急激に進んでいる円安・ドル高については，アメリカ政府も強い懸念を示しており，明日にもサマーズ財務副長官を日本に派遣して，松永大蔵大臣などと対応策を協議することにしています.
S|この中で，大蔵省は，きょう成立した平成十年度の補正予算などを通じて，景気のテコ入れに力を入れていることや，金融の安定に向けて不良債権問題の抜本的な解決策のとりまとめを進めていることなど，円安防止に向けた日本の姿勢を示す方針です.
S|その上で，大蔵省は，現在の為替水準が，行き過ぎた円安になっているという点で，日米が共通の認識を示すことができるものと見ており，こうした動きを食い止めるために，日米が協力して市場介入を行なう協調体制についても確認したい考えです.
S|しかし，アメリカ政府部内には，市場への効果が限られているなどとして，協調介入に慎重な見方もあり，日米両国が円安の流れに対して，どこまで強力な姿勢が示せるかが焦点になっています.

X|199806170847
N|株価大引け
Z|sakuseibukyoku_name=経済
S|きょうの東京株式市場は，円相場が値を戻していることを受けて朝方はこれまで値下がりしていた銘柄を買い戻す動きが出ましたが，午後からは銀行株を中心に逆に売り注文が増え平均株価はきのうの終値とほぼ同じ水準で取引を終えました.
S|主要銘柄の平均株価の終値は，きのうの終値より五円ちょうど安い，一万四千七百十五円三十八銭でした.
S|すべての銘柄の値動きを示すトピックス・東証株価指数は，逆に〇点三七上がって，千百五十六点八四，一日の出来高は四億四百万株でした.
S|市場関係者は「円相場が値を戻しているのを受けて朝方はこのところ値下がりが目立っていた銘柄を買い戻す動きが強まった．しかし，市場では国内の景気が回復傾向を示さない限りは円相場や東京市場の株価の本格的な上昇は望めないという見方が強いため買い注文を出す動きは一時的なものにとどまり，アジア市場の株価が高くなっている事にも特に反応を示さなかった」と話しています.

表 5.2　各話題に属するニュース記事に出現した単語数

話題の種類	名詞（種類）	動詞（種類）	人名（のべ数）	地名（のべ数）	組織名（のべ数）
選挙	7,292	1,780	7,403	13,300	3,470
国会審議	6,521	1,791	6,992	7,995	6,234
任意の話題	14,102	2,633	4,519	16,832	3,298

人名は,「選挙」と「国会審議」では,「任意の話題」に比べて圧倒的に多く出現しており，この話題には欠かせない要素だと推定できる.「選挙」と「国会審議」では,「任意の話題」より地名の出現が少なく，組織名は「国会審議」に頻出しているという特徴が伺える．このような話題に特有な特徴は，ニュース記事のクラスタリングや，話題の要約処理において有益な情報となる.

5.3　一定期間のニュース記事からの話題抽出法

前節で述べたように，ニュース記事は，その第一文はニュース内容の全貌を説明することが多く，これに対して，第二文以降は話題抽出処理では不要な要素が多い．そのため，解析処理では，ニュース記事の第一文のみを処理対象として利用する．まず，話題抽出の対象とする期間のニュース記事の形態素解析を行い，ニュース記事に含まれる名詞を全て抽出する．次に，その抽出した単語の時期ごとに変化する重要度を，統計値を利用して定義する．その重要度をもとに，ニュース記事のクラスタリングを行う．類似する記事が集まったクラスタがニュース記事における話題となる．次に，抽出された全てのクラスタに対して，そのクラスタを特徴付ける名詞句を抽出する．最後に，抽出した話題を，対象とした前後の期間の話題と比較し，話題のトラッキングを行なう．以下にニュース記事からの話題抽出処理，クラスタとして表現された話題を代表する名詞句抽出処理，そして，話題のトラッキング処理について説明する.

5.3.1　ニュース記事からの話題の抽出

ニュース記事のクラスタリングを行うために，まず，ニュース記事に含まれる単語に対して，χ^2 値と IDF 値を利用して，時期ごとに変化する重要度を定義する．この重要度は，対象とする時期の話題に関係が深い単語ほど大きな値が付けられる.

χ^2 値は，観測値と期待値がどの程度一致しているかを測る指標である．この値を単語出現頻度の時期変化に対して適用することにより，単語の出現の時期ごとの偏り，つまり話題性を評価できる．母集団を対象とする月から前の一年間とし，対象月 m の単語 t の出現頻度を $n(t,m)$，その出現期待値を $e(t)$とした時，単語 t の対象月における話題性を表す χ^2 は式(5.1)となる.

$$\chi^2(t,m) = \frac{(n(t,m) - e(t))^2}{e(t)} \tag{5.1}$$

出現頻度が期待値より小さいときも，大きい場合と同じ正の値をとってしまうため，式(5.2)の値を利用する．

$$\overline{\chi}^2(t,m) = \begin{cases} \dfrac{(n(t,m) - e(t))^2}{e(t)} & n(t,m) \geq e(t) \\ 0 & otherwise \end{cases} \tag{5.2}$$

IDF 値は，ニュース記事中に頻繁に出現する単語ほど一般的な単語と見なし小さな値をとる．対象月 m の一ヶ月のニュース記事の総数を $N(m)$，一ヶ月のニュース記事中で単語 t が出現するニュース記事数を $df(t,m)$としたとき，IDF(t,m)は式(5.3)となる．

$$\mathrm{IDF}(t,m) = \log \frac{N(m)}{df(t,m)} \tag{5.3}$$

χ^2 値により月ごとに変化する単語の話題性を評価できる．また，IDF 値により一般的な単語を処理対象から除外できる．この２つの値を相乗的に利用して，月ごとに変化する対象月 m の単語 t の重要度 $weight(t,m)$を式(5.4)で定義する．

$$weight(t,m) = \overline{\chi}^2(t,m) \times \mathrm{IDF}(t,m) \tag{5.4}$$

1999 年 9 月のニュース記事に含まれる単語(延べ約 36 万単語)の出現頻度と重要度 $weight(t,m)$の上位を表 5.3 に示す．頻度では「日本」「事件」など，この月の話題分類に不適当な単語も上位に多く出現しているが，重要度 $weight(t,m)$では「ティモール」「キルギス」「多国籍軍」など，この月の話題をより具体的に説明する単語に大きな値が付けられ，良好な結果が得られている．

次に，ニュース記事に含まれる単語の重要度を利用して，ニュース記事のクラスタリングを行い，類似ニュース記事の集合である話題を抽出する．クラスタリングは，一ヶ月毎のニュース記事を対象としてベクトル空間法[35]を利用して行う．各ニュース記事は，必ず一つのクラスタに属すると仮定した．ニュース記事は，記事中に含まれる単語をベクトルの要素に，その単語の重要度をベクトルの要素の値としたベクトルで表現し，ニュース記事間の類似度を, 2 つのニュース記事のベクトルの内積とした. クラスタリング処理は，以下の手順で行う．

1. 全てのニュース記事を，一つのニュース記事から成るクラスタとする．
2. 全てのクラスタ間の類似度を計算し，最も類似した２つのクラスタを統合する．
3. 2 の処理を, すべてのクラスタ間の類似度がしきい値以下になるまで繰り返し行う．

表 5.3　単語の出現頻度と重要度の上位 25 単語（1999 年 9 月）

単語の出現頻度		単語の重要度	
日本	1136	台風	12630.70
台風	736	ティモール	10976.65
事件	615	キルギス	8992.53
東京	604	台湾	8803.52
党	559	秋場所	6372.21
アメリカ	504	地震	5970.14
警察	459	武装	5703.43
人	459	多国籍軍	5082.90
大臣	454	東	4650.57
政府	432	高潮	4640.72
東	425	勢力	4518.42
影響	415	中部	4495.72
線	402	治安	4395.39
選手	393	不知火町	3945.48
台湾	391	円高	3444.57
雨	370	域	3168.68
ティモール	354	暴風	3218.85
問題	338	ニュージーランド	3036.37
自民	331	中央アジア	3017.05
男	324	オークランド	2861.57
事故	324	不祥事	2756.20
本部	323	維持	2655.18
考え	314	次男	2606.51
ＪＲ	312	ディリ	2471.16
会社	311	ウズベキスタン	2412.75

5.3.2　類似記事集合(話題)を説明するラベルとなる名詞句抽出

クラスタリング処理により，内容が似たニュース記事は同じクラスタに分類される．ここで生成された全てのクラスタに対して，クラスタを特徴付ける名詞句を抽出し，クラスタを説明するラベルとする．

まず，各クラスタに含まれるニュース記事群から，そのクラスタを代表するニュース記事を抽出する．クラスタ中のニュース記事に含まれる単語のクラスタへの寄与度を，そのクラスタでの単語の出現率と単語の重要度の積とし，記事に含まれる単語の寄与度の合計がクラスタ中で最大のものを，そのクラスタの代表ニュース記事とする．

次に，この代表ニュース記事に含まれる全ての名詞を対象に，連続する名詞群，助詞「の」

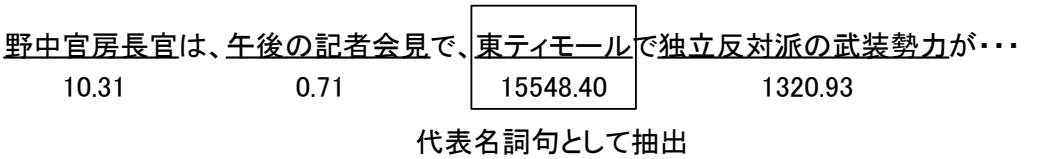


図 5.1　代表ニュース記事から代表名詞句の抽出例（数字は単語の寄与度の合計）

で接続した名詞群を抽出する．抽出した名詞群の中で，そのクラスタにおける寄与度の和が最大のものを，クラスタを代表する名詞句とする．図 5.1 の例では，代表記事に含まれる名詞群の中で，その寄与度の和が最大の「東ティモール」が代表する名詞句として抽出されている．この作業により，すべてのクラスタに，そのクラスタを特徴付ける名詞句が付与される．

最後に，対象とした期間に抽出されたクラスタを，話題性の大きさによって順位付けする．ここで，クラスタに含まれる記事数と重心ベクトルの大きさとの積をそのクラスタの重要度とする．重心ベクトルは，クラスタを構成する記事群のベクトルの和をその記事数で割り生成する．重要度が大きなクラスタに付けられた名詞句ほど話題性が大きいと判断し, クラスタの重要度の降順に, そのクラスタに付けられた名詞句を話題として提示する．

1999 年 9 月のニュース記事 7,482 個に対してクラスタリング処理を行い，話題を説明するラベルとなる名詞句を抽出した結果の上位 8 項目を，表 5.4 に示す．

表 5.4　話題を説明するラベルとなる名詞句を抽出した結果の上位 8 項目（1999 年 9 月）

話題を説明するラベル	ニュース記事数	クラスタの重要度
台風十八号	700 個	1.28×10^7
東ティモール	295 個	7.57×10^6
台湾中部の南投市	299 個	5.22×10^6
中央アジアのキルギス	123 個	3.17×10^6
先進七カ国の蔵相・中央銀行総裁会議	199 個	1.27×10^6
横綱・若乃花	104 個	8.02×10^5
野球のシドニーオリンピック・アジア地区予選	119 個	5.08×10^5
トルコ西部の大地震	49 個	4.64×10^5

5.3.3　話題のトラッキング

5.3.2 では月単位の話題を抽出する手法を説明した．しかし，一つの話題は複数の月に渡って継続することがある．特に世間で騒がれるような大きな話題ほど長く継続することが

多い．実際に，提案した手法により抽出された各月ごとの話題を対象として，その前後に関連した話題があるかを手作業により調査した．その結果，1999 年 1 月～12 月の上位 8 項目の話題では，29.2%がその前後の月に関連した話題が存在した．ここでは，隣接する月の話題間の類似性を評価して，話題のトラッキングを行う．

ニュース記事から抽出した話題は，出現する単語をベクトルの要素とする重心ベクトルで特徴付けられている．この重心ベクトルと隣接する月の話題が構成する重心ベクトルとの類似度を,クラスタリング処理時と同様に２つのニュース記事のベクトルの内積とする．この類似度が経験的に決定したしきい値より大きい時に関連していると判断する．この処理を，対象期間を広げて，話題のトラッキングを行なった．数ヶ月に渡り騒がれるような話題は，その期間中，毎月関連する話題が出現し，それらの話題をすべて関連付けることにより，大きな一つの話題として認識できる．

5.4 ニューステキストの話題要約

話題のトラッキング処理の結果，まとめられた話題には，出来事の連鎖，生起に関する特有のモデルが存在し，各々の出来事に対してはゆらぎの抑制された単語や統語構造による表現がなされる．以下の例，話題「ガイドラインの関連法案の審議」におけるニュース記事では，下線部「～法案は，～で採決され，～賛成多数で可決されました.」が，特有な統語構造と考えられる．

例）
日米防衛協力の指針・いわゆるガイドラインの関連法案は，先程，参議院の特別委員会で採決され，自民，自由，公明の三党などの賛成多数で可決されました．

このような表現は，話題「ガイドラインの関連法案の審議」だけでなく，話題「テロ対策基本法案の審議」など，他の「国会審議」に関するニュース記事にも見られる．本論文では，このような，類似する話題の集合を「分野」と呼ぶ.「分野」には，複数の話題が含まれ，分野に属する話題を構成するニュース記事には，例で示したような分野特有の定型的な表現が多く出現する．本論文における話題と分野の関係の例を図 5.2 に示す.

図 5.2 の例において,「選挙」という分野は話題「第十九回参議院選挙」,「富山市長選挙」,「都知事選挙」などから構成されており，各話題には，「選挙の公示」,「立候補の表明」,「選挙運動」,「投票」,「開票」などの分野に共通する出来事が存在する．さらには，各出来事には,「選挙名」や「投票日」などの共通する詳細情報が存在する．ニュース記事において，分野に共通する出来事を表現する際に，先の例で示したような分野特有の定型的な表現が利用される．分野特有の定型的な表現が多く含まれるニュース記事は，話題を特徴づける重要なニュース記事であると考えられる．そこで，この定型部分から話題要約文を生成する手法を提案する．定型部分抽出のために，同一話題に属する記事集合における単

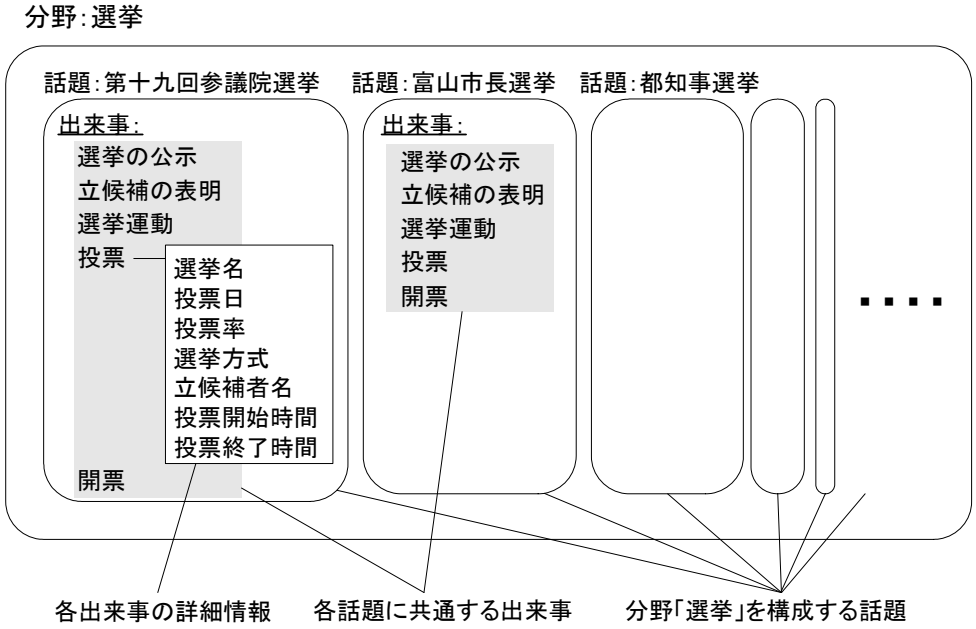


図 5.2　話題と分野の関係の例

語間の係り受け関係に注目する.

係り受け関係を利用しないアプローチでは，前述の「～法案は，～で採決され，～賛成多数で可決されました.」のように文中の離れた場所にある特徴を評価することが難しい. 例えば，連続する語を特徴として与える単語 n グラムモデルを利用する場合,「法案は，先程, 参議院の特別委員会で採決され」の部分だけでも 10 以上の形態素が存在するため,「法案」と「採決」の両方の形態素を考慮した特徴を抽出することは難しい．また，形態素の出現順で特徴付ける手法も考えられるが，形態素数のべき乗の order の計算量となり，計算効率を考慮すると，係り受け関係を利用する手法が現実的と考えられる．そのため，本手法では係り受け関係を特徴として，その定型性を統計的に評価し定量化する.

以下に，係り元の自立語と係り元の文節の付属語となる助詞，そして係り先の自立語からなる 3 項組からなる係り受け関係の定型性評価手法と，係り受け関係の定型性を利用した話題要約処理について記す.

5.4.1　係り受け関係の定型性評価

分野特有の定型的な表現を評価するため，ある対象分野に属する話題を構成する全てのニュース記事を構文解析し，係り受け関係を持つ 2 つの文節の自立語と係り元の文節の付属語となる助詞（助詞を介さず直接係る場合は φ）の 3 項組を抽出する．そして，対象分野における 3 項組の特異性を，観測値と期待値がどの程度一致しているかを測る指標である χ^2 値[57]を利用して評価する．係り元の自立語 w_1，係り元の文節の付属語となる助詞 w_2，係り先の自立語 w_3 としたとき，3 項組(w_1,w_2,w_3)の対象分野における出現頻度を $n(w_1,w_2,w_3)$, その期待値を $e(w_1,w_2,w_3)$としたとき, $\chi^2(w_1,w_2,w_3)$は式(5.5), $e(w_1,w_2,w_3)$は式(5.6)

で与えられる．

$$\chi^2(w_1,w_2,w_3)=\frac{(n(w_1,w_2,w_3)-e(w_1,w_2,w_3))^2}{e(w_1,w_2,w_3)} \tag{5.5}$$

$$e(w_1,w_2,w_3)=\frac{N(topic)\times n_{all}(w_1,w_2,w_3)}{N} \tag{5.6}$$

ここで N は全ニュース記事数，$N(topic)$は対象分野 $topic$ におけるニュース記事数，$n_{all}(w_1,w_2,w_3)$は，3 項組(w_1,w_2,w_3)の全ニュース記事における出現頻度を示す．

観測値 $n(w_1,w_2,w_3)$が期待値 $e(w_1,w_2,w_3)$より小さいときも，大きい場合と同じ正の値をとってしまうため，実際には式(5.7)の値を利用した．

$$\overline{\chi}^2(w_1,w_2,w_3)=\begin{cases}\chi^2(w_1,w_2,w_3) & n(w_1,w_2,w_3)\geq e(w_1,w_2,w_3)\\ 0 & otherwise\end{cases} \tag{5.7}$$

このとき，単語の属性が人名，組織名，地名である場合は，抽象化した属性名を利用し，例えば「自民党の政策」と「民主党の政策」は，共に「"組織名"の政策」として χ^2 値を計算する．χ^2 値が大きい 3 項組ほど，その分野に特異に出現していると言える．

同一話題の多くの記事中に出現する 3 項組は，個別の記事が伝える出来事の内容を特定するための弁別能力に乏しい．そこで，そのような 3 項組の定型性評価値を制限するために，IDF 値を利用した．要約対象とする一つの話題を構成するニュース記事の総数を N, そのニュース記事中で 3 項組 (w_1,w_2,w_3)が出現した記事数を $DF(w_1,w_2,w_3)$としたとき，この 3 項組の IDF 値，$IDF(w_1,w_2,w_3)$は式(5.8)で与えられる．

$$IDF(w_1,w_2,w_3)=\log\frac{N}{DF(w_1,w_2,w_3)} \tag{5.8}$$

さらに，3 項組の品詞の組み合わせにより，定型性を評価する重みに制限を与える．品詞組み合わせによる重み $C(w_1,w_2,w_3)$は，表 5.5 に示す値とした．(w_1,w_2,w_3)=(名詞，助詞，

表 5.5　3 項組の品詞組み合わせによる重み付け
(ϕ は空集合で，その要素が無いことを示す．)

w_1,w_2,w_3	$C(w_1,w_2,w_3)$
名詞, 助詞, 動詞	1.0
名詞, 助詞, 名詞	0.2
動詞, φ, 動詞	0.1
その他の組み合わせ	0.05

動詞)の組み合わせは，動詞とその格構造の情報となり，出来事を表現する基本構造と考えられるため最重要としている．

χ^2 値，IDF 値，さらに品詞による制限値を相乗的に考慮することにより，話題要約文の構成要素候補を抽出するための 3 項組の定型性を評価する重み *weight*(w_1 ,w_2 ,w_3)を以下の

表5.6　3項組の定型性を評価する重み計算結果（上位30組）

weight	3 項組
5721.2	賛成多数 / で / 可決される
3865.1	参議院 / に / 送られる
3305.3	衆議院選挙 / に / 向ける
2922.8	次 / の / 衆議院選挙
2417.4	衆議院本会議 / で / 可決される
2346.0	参議院本会議 / で / 可決・成立する
2031.5	国会内 / で / 述べる
1468.4	政府 / は / 提出する
1078.9	考え / を / 示す
988.0	国会対策委員長会談 / が / 開かれる
918.1	参考人質疑 / が / 行われる
828.6	採決 / が / 行われる
822.9	国会対策委員長 / が / 会談する
810.0	修正案 / が / 可決される
799.3	予算案 / が / 通過する
786.9	衆議院 / を / 通過する
785.4	今 / の / 国会
753.1	成立 / を / 目指す
712.1	質疑 / が / 行われる
676.1	賛成多数 / で / 可決・成立する
654.4	総括質疑 / が / 始まる
593.8	審議 / が / 始まる
573.8	通常国会 / に / 提出される
570.3	衆議院予算委員会 / で / 関連する
562.2	施政方針演説 / に / 対する
547.9	考え / を / 強調する
535.7	成立 / を / 図る
487.3	調整 / が / 続く
484.3	衆議院 / の / 解散・総選挙
476.3	早期成立 / に / 向ける

ように定義する.

$$weight(w_1, w_2, w_3) = C(w_1, w_2, w_3) \times \chi^2(w_1, w_2, w_3) \times \mathrm{IDF}(w_1, w_2, w_3) \quad (5.9)$$

表 5.6 に「テロ対策基本法案の国会審議」に関するニュース記事に出現した 3 項組の定型性を評価する重みの計算結果の上位 30 組を示す.「賛成多数で可決される」「参議院に送られる」といった，国会審議に関するニュース記事の型にはまった表現が上位にある．逆に「経済の問題」といった国会審議に特有の表現でない 3 項組の $weight(w_1,w_2,w_3)$の値は 0 であった.

5.4.2 係り受け関係の定型性を利用した要約処理

5.4.1 では，特定の話題を構成するニュース記事の定型的な表現を評価する 3 項組の定型性を評価する重み定義した．この値を利用して，話題を構成する各ニュース記事文を要約し，各文から生成された要約文の定型性の度合いを評価する．ニュース記事は，これから発生する出来事について触れる場合など，実際に発生した事実・出来事について記述されていない可能性もある．そこで，ニュース記事に含まれる動詞に着目し，実際に発生した事実・出来事についての言及（以後,「確定事項についての言及」と呼ぶ）であるか否かを判定する．これらの結果を利用して，最終的に話題を構成するニュース記事群から要約を生成する．以下に各処理について記す.

各ニュース記事の要約

係り受け関係の定型性を利用したニュース記事 1 文の要約処理例を図 5.3 に示す. まず，あらかじめクラスタリングによって同定されている，対象の話題を構成するニュース記事から，その話題における 3 項組の定型性を評価する重みを計算する．係り受け関係にある 3 項組が対象の話題に特異的に出現する場合は，その定型性を評価する重みは 0 より大きな値をとる．そこで，単一記事に含まれる 3 項組の中で，定型性を評価する重みがゼロでない組を取り出す．図 5.3 の例では,「日米防衛協力の指針・いわゆるガイドライン関連法案は，きょうの衆議院本会議で，自民，自由両党と公明党・改革クラブの三会派による修正の上，賛成多数で可決され，参議院に送られました.」というニュース記事から,「衆議院本会議／で／可決される」,「賛成多数／で／可決される」,「可決され／φ／送られる」,「参議院／に／送られる」の 4 つの 3 項組が取り出されている.

次に，取り出された 3 項組において，共通の文節から取り出された項を持つ 3 項組を統合して単一記事の要約文を生成する．図 5.3 の例では,「可決される」と「送られる」が共通な文節となり，これらを含む 3 項組を，文節の出現順で統合することにより,「衆議院本会議で賛成多数で可決され参議院に送られる」という文が生成されている．このとき，統合した 3 項組が持つ定型性を評価する重みの合計が，文の重要度の指標とする．図 5.3 の例では，統合された 3 項組が持つ定型性を評価する重みの合計 12327.7 が，このニュース

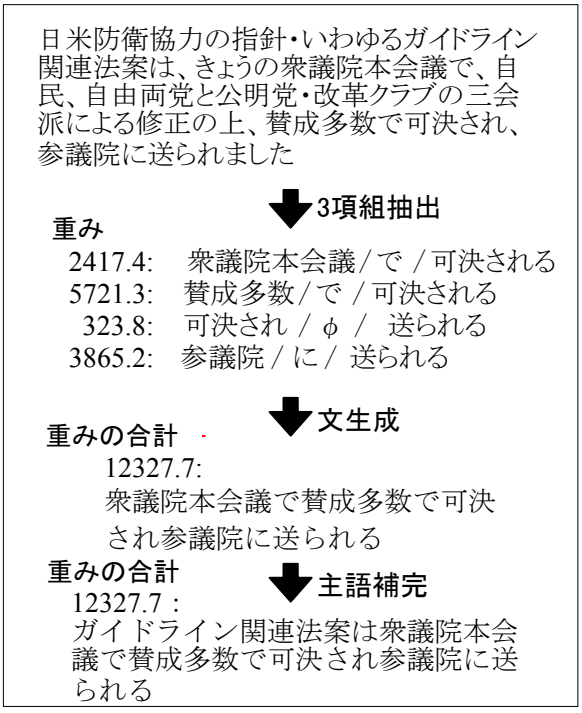


図5.3　係り受け関係の定型性を利用したニュース記事1文の要約処理例
(図中の数値は 3 項組みの定型性を評価する重みを示す)

記事に対する定型性評価の重みとなる．

3 項組を統合することにより生成した文は，その分野における定型的な係り受け関係のみを抽出して繋げている．図 5.2 の例で示した各出来事の詳細情報などの項目は話題ごとに異なり，各出来事の詳細情報が含まれる 3 項組の定型性を評価する重みは小さくなることが多い．図 5.3 の例では，「ガイドライン関連法案／は／可決される」という 3 項組は，他の国会審議の話題では出現しないため，国会審議の分野における 3 項組の定型性を評価する重みは 0 となってしまう．そこで，生成された文に含まれる動詞に対して，その主語と目的語を補完する．この処理では，構文解析結果を利用した．主語と目的語を補完して最終的に生成された文を，ここでは定型文と呼ぶ．この処理により，図 5.3 の例では生成された文「衆議院本会議で賛成多数で可決され参議院に送られる」に対して，その主語である「ガイドライン関連法案は」が補完され，最終的に「ガイドライン関連法案は衆議院本会議で賛成多数で可決され参議院に送られる」という定型文が生成される．

確定事項の判定

各ニュース記事の要約処理により，特定の話題に属するニュース記事集合から定型文を抽出できた．しかし，ニュース記事では，全てが既に実施した事実・出来事を述べた確定事項を表現しているとは限らない．例えば，次のニュース記事には「審議する」「開く」「求める」「行う」「入る」「決める」の 6 つの動詞が出現している．

例）

ガイドライン関連法案を審議している参議院の特別委員会は，きょう理事懇談会を開き，来月十日に，小渕総理大臣と全ての閣僚の出席を求めて総括質疑を行い，審議に入ることを決めました．

ここで，「開く」「決める」は既に実施した事実を述べた確定事項だが，「審議する」「求める」「行う」「入る」は，実施中，もしくは，これから実施される予定の未確定事項である．本手法では，既に実施された話題の構成要素を抽出することを目的とし，確定事項のみを対象とする．

この処理では，事態の確実性を表す名詞[58]（「こと」「考え」「方針」「意向」「見通し」など）以外の名詞を修飾する動詞を，文の主題とは無関係と判断し，確定・未確定の判定処理の対象から除いた．上記の例では，「審議する」の判定処理は行わない．

確定・未確定の判定処理は，動詞の時制を利用する．動詞が「過去形」の場合は確定，「現在形」の場合は未確定とした．日本語の場合，過去にあった出来事を表す動詞は，通常，過去を表す助動詞を動詞の語尾に伴うことにより過去形となる．しかし，動詞が複数出現する複文では，最後に出現する動詞以外は，過去を表す助動詞が省略されてしまう．また，条件文でも例外が生じる．そこで，以下の場合は，例外処理を行い判断した．

- 条件を表す名詞が存在する場合

  過去を表す助動詞を伴っても，未確定とする

  例）「日本に武力攻撃が加えられた場合は，・・・」→「加えられた」は「未確定」と判定

- 連用修飾節の動詞の場合

  係り先の連用節と同じ時制として判定する

  例）「・・・と述べ，・・・ことを示しました．」→「述べ」は「示しました」と同じ時制「過去」として「確定」と判定

この処理を話題「ガイドライン関連法案の審議」を構成する 331 個のニュース記事に対して行い，手作業による結果と比較検証した．その結果を表 5.7 に示す．出現した 929 個の動詞中，810 個(87.2%)の動詞に対して正解が与えられ，ある程度，良好な結果が得られている．発生した出来事を未発生と誤判定してしまった原因の多くは，連用修飾節におけ

表5.7　3項組の定型性を評価する重み計算結果（上位30組）

	確定事項	未確定事項
確定と判定	354(95.7%)	16(4.3%)
未確定と判定	103(18.4%)	456(81.6%)

る係り受け解析の失敗によるものであった.

話題の要約生成

各ニュース記事の要約の結果得られる定型文と，ニュース記事に対する定型性評価の重み，そして，動詞の確定・未確定の判定結果を利用して，複数ニュース記事から形成される話題の要約を行う．ここで，各ニュース記事から生成した定型文の中で，文末の動詞が「発表語」で，その前に「こと」以外の「事態の確実性を表す名詞」がある場合は，その前に述べられた行為の確定性が低いことが判っている[58]．そのため本手法では，「考えを表明する」などが含まれる定型文を処理から除いた.

さらに，同一内容について述べたニュース記事も数多く存在するため，類似内容の定型文も複数抽出してしまう．そこで重複する定型文を削除する処理を行う．この処理では，以下の 2 つの条件を満たす場合に重複した定型文と判断し，定型性評価の重みが低い文を削除する.

- 一定値（本実験では 0）より大きい定型性を評価する重みを持つ 3 項組の係り受け関係で，その内容に不整合（二項が同じで一項のみ異なる組み合わせ）が存在しない
- 共通である 3 項組の定型性を評価する重みの合計が一定値以上（本実験では，{ min (2 文の定型性を評価する重みの和) / 2 }以上）

例えば，抽出された定型文の「衆議院本会議で可決される(定型性を評価する重み 2417.4)」と「衆議院本会議で，賛成多数で可決され，参議院に送られる(定型性を評価する重み 12327.7)」は上記の条件を満たすため，文の定型性を評価する重みが低い「衆議院本会議で可決される」は削除される.

確定事項を表現すると判定された動詞を文末に持つ定型文で，その定型性を評価する重みが一定値（本実験では 500）以上の文から，行為の確定性が低い文と，重複した定型文を削除することにより，話題結果とした．全ての定型文に対して，その重要性を表す定型性を評価する重みを与えたため，ニュース記事の要約率は，このしきい値を変化させることにより，容易に変更できる.

話題「ガイドライン関連法案の審議」に関するニュース 331 記事を要約した結果を表 5.8 に示す．衆議院本会議での趣旨説明，特別委員会の参考人質疑，衆議院本会議の可決，参議院特別委員会の可決，参議院本会議での可決成立など，法案審議の話題の遷移を把握するために重要と考えられる要素が短文で抽出されている.

5.4.3 考察

ガイドライン関連法案の国会審議に関するニュースの出現数の推移を図 5.4 に示す．ニュース記事が多く出現した 4 月 27 日は衆議院本会議で可決，5 月 24 日は参議院特別委員

表5.8　話題「ガイドライン関連法案の審議」に関するニュース記事の要約結果

日付	要約結果	定型性
1999/2/12	理事会で指針いわゆるガイドラインの関連法案を審議する特別委員会を衆議院本会議で設置することを決める	645.2
1999/3/12	ガイドライン関連法案は衆議院本会議で趣旨説明と質疑が行なわれ法案の早期成立に野党側の協力を求める	1361.9
1999/3/19	両党の間で今の国会での成立を目指すことを確認する	1592.7
1999/3/29	国会対策委員長が会談し第三週に通過を目指すことを確認する	832.4
1999/4/1	衆議院の特別委員会は理事会で自民・自由両党が法案の採決を行うよう求める	507.7
1999/4/7	衆議院の特別委員会で，参考人質疑が行なわれ，四人の参考人が，意見を述べる	1447.2
1999/4/26	衆議院特別委員会で修正案が可決される	810.0
1999/4/26	小渕総理大臣は今の国会での成立に向けて協力を要請する	1218.4
1999/4/27	採決が行われ賛成多数で可決される	6590.2
1999/4/27	ガイドライン関連法案は衆議院本会議で賛成多数で可決され参議院に送られる	12327.7
1999/4/27	小渕総理大臣はガイドライン関連法案が衆議院を通過したことについて国会内で記者団に対し述べる	3250.1
1999/4/28	参議院の特別委員会は理事懇談会を開き出席を求めて総括質疑を行い審議に入ることを決める	1044.8
1999/5/9	参議院本会議で質疑が行われる	1093.8
1999/5/13	参議院の特別委員会で参考人質疑が行われ有職者三人が意見を述べる	936.4
1999/5/20	初会合を開き今の国会で法案の成立を目指す方針を確認する	1979.1
1999/5/24	関連法案は参議院の特別委員会で採決され三党などの賛成多数で可決される	5989.3
1999/5/24	関連法は参議院本会議で採決され三党などの賛成多数で可決され成立する	6462.0

会，本会議で可決されたことが一覧できる．このニュース記事の出現数と，話題におる出来事の重要さは，必ずしも一致しない．話題「テロ対策法案の審議」に関するニュース 191 記事を要約した結果を表 5.9 に，その出現頻度の変化を図 5.5 に示す．法案が参議院で可決された 10 月 26 日には，「テロ対策法案の審議」に関するニュースは 6 個しか出現していない．このような重要な出来事に対するニュース記事が少ない話題は，クラスタリングや，

表5.9　話題「テロ対策法案」に関するニュース記事の要約結果

日付	要約結果	定型性
2001/10/4	衆議院議院運営委員会は理事会で「テロ対策特別法案」などを審議するため特別委員会を設置することを決める	620.4
2001/10/5	政府は「テロ対策特別法案」を決定し国会に提出する	6879.0
2001/10/5	法案をまとめ衆議院に提出する	1039.2
2001/10/8	早期成立に向けて協力を求めることを決める	952.5
2001/10/9	特別委員会は出席を求めて質疑を行なうことで与野党が合意する	1124.8
2001/10/10	十九日までに法案の成立を目指すことで一致する	988.4
2001/10/10	参考人質疑を十五日には一般質疑を行なうことを決める	1169.2
2001/10/15	小泉総理大臣は野党各党の党首と個別に会談し早期成立に向けて協力を要請する	936.8
2001/10/15	修正を行なう考えを示して賛成を求める	1607.4
2001/10/16	与党三党などの賛成多数で可決される	6201.0
2001/10/16	テロ対策特別法案が衆議院の特別委員会で可決されたことについて小泉総理大臣は総理大臣官邸で記者団に対し述べる	641.7
2001/10/18	「テロ対策特別法案」が衆議院本会議で与党三党などの賛成多数で可決され参議院に送られる	12203.9
2001/10/18	テロ対策特別法案が衆議院を通過したことについて中谷防衛庁長官は記者団に対し述べる	1553.8
2001/10/18	小泉総理大臣は「テロ対策特別法案」が衆議院本会議で可決されたことについて述べる	2150.7
2001/10/19	合同理事会が開かれ審議日程について出席を求めて質疑を行うことを決める	752.9
2001/10/26	「テロ対策特別法案」は参議院外交防衛委員会で与党三党の賛成多数で可決される	6187.2

単語の変化点を用いる従来手法では，要約することが難しい．しかし，本手法では，法案の国会への提出，衆議院特別委員会での可決，衆議院本会議での可決，参議院外交防衛委員会での可決といった項目が抽出されていることがわかる．

分野「国会審議」に属する 34 個の話題を対象に，その話題を構成する複数ニュース記事を要約した結果を検証した．その結果，衆議院／参議院本会議における法案可決に関する記述は，再現率 90.7%で抽出され，良好な結果が得られた．抽出されなかった理由の 71.0%が，動詞の確定・未確定判定の誤りにあり，残りが定型性を評価する重みのしきい値によ

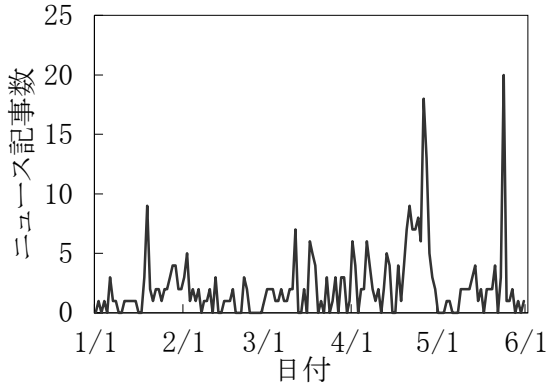


図5.4　「ガイドライン関連法案の審議」に関するニュースの出現記事数

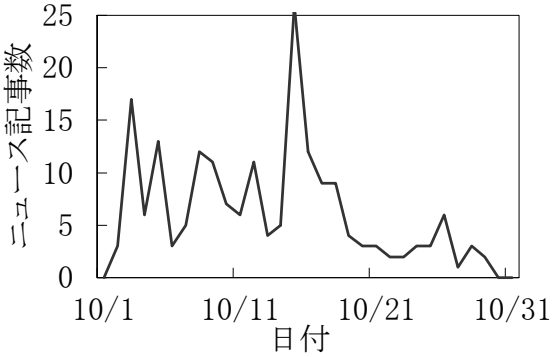


図5.5　「テロ対策法案の審議」に関するニュースの出現記事数

る問題であった．今後，過去を表す助動詞以外の要素も取り入れた確定・未確定判定処理が必要と考えられる．

また，法案可決の要約結果の 22.4%は，場所格が抜けていたため，どこで可決されたか，という重要な要素が落ちていた．表 5.9 でも，10 月 16 日に「与党三党の賛成多数で可決される」という定型文が生成されている．この定型文は，「衆議院の特別委員会で」という場所格が抜けている．ニュース記事では，冗長な部分を少なくするため，複文の構造を持つ場合は，動詞の共通の格は省略されてしまう．そこで，要約時に省略された場所格を補完する必要が生じる．これは，従来から提案されている主語補完の技術[59]を応用することにより解決可能と考えられる．

5.5　おわりに

本章では，単語の出現頻度の時間軸における変化量に着目した指標により，話題性に基づいた時期ごとに変化する単語の重要度を定義し，大量のニュース記事から特定期間に発生した話題を抽出する手法を提案した．さらに，複数の期間にまたがり継続する話題のトラッキングを行なう処理も行うことにより，長期に継続する話題を大きな一つの話題とし

て認識する実験を行った．また，抽出した話題に対して，話題が属する分野特有の定型的な表現を評価することによって，話題の要約を生成する手法を提案した．この処理ではニュース記事の出現数からでは判断できない話題の主要な出来事を抽出する要約を実現している．話題要約手法は，国会審議以外の話題でも，基本要素をテンプレートで表現できるような話題に関する複数ニュース記事であれば適応可能と考えられる．

話題の要約処理では，前処理となる話題のクラスタリング処理，トラッキング処理の結果により，出力が大きく変わる可能性がある．しかし，提案した手法では，統計的に係り受け関係の定型性を評価しているため，話題の集合である分野に一定数のニュース記事が含まれていれば，話題に含まれるニュース記事数には大きく影響を受けずに要約が可能と考えられる．

要約結果の評価は，難しい課題として知られている．要約技術の向上のため，近年，自動要約評価型のワークショップ[60]も開催されており，これらのワークショップでは，主催者が特定のタスクに対する要約結果の正解を人手で与え，計算機により要約した結果との類似性により評価している．今後，提案手法をそのようなタスクへ応用することにより，生成された要約結果のさらなる検証を行い，より良いメタデータ生成へと進める必要がある．

ニュース記事には，国語辞典などの辞書に載っていない単語が大量に存在する．ニュース記事を解析する際，このような辞書に載っていない単語の扱いが問題となる．また，人は，ニュース記事を読めば，その内容を的確に把握することができる．人間が持つ一般的な常識や，世間の話題に対する背景知識が，ニュース記事を理解するための助けとなっていると考えられる．6 章では，辞書に載っていない単語の扱いと，人間が持つ知識を計算機上で扱う手法について論じる．

6 章　テキストからの知識獲得

単語間の関係などの語彙に関する知識を，大量のテキストを解析することにより自動獲得する技術について論じる．これらの知識は，放送コンテンツ中のテキストデータ解析精度の向上と，放送コンテンツを利用したアプリケーション生成に効果的となる．

6.1　はじめに

3 章から 5 章までは，テレビ番組に対してメタデータを生成する手法を論じた．これらのメタデータ生成処理では,テキスト中に出現する単語の表記を特徴として利用している．しかし，テキストを表層的に比較する単語表記の特徴だけでは，人間が行うような深い考えに基づく処理はできない．より深い処理を行うためには，人間が持つ一般常識や語彙に関する知識が効果的と考えられる．また，メタデータを利用して放送された番組を効果的に二次利用するアプリケーションにおいても，このような知識は有益と考えられる．そこで，大量のテキストを解析して語彙に関する知識を自動獲得する手法を提案する．本章において獲得対象とする知識を図 6.1 に示す．

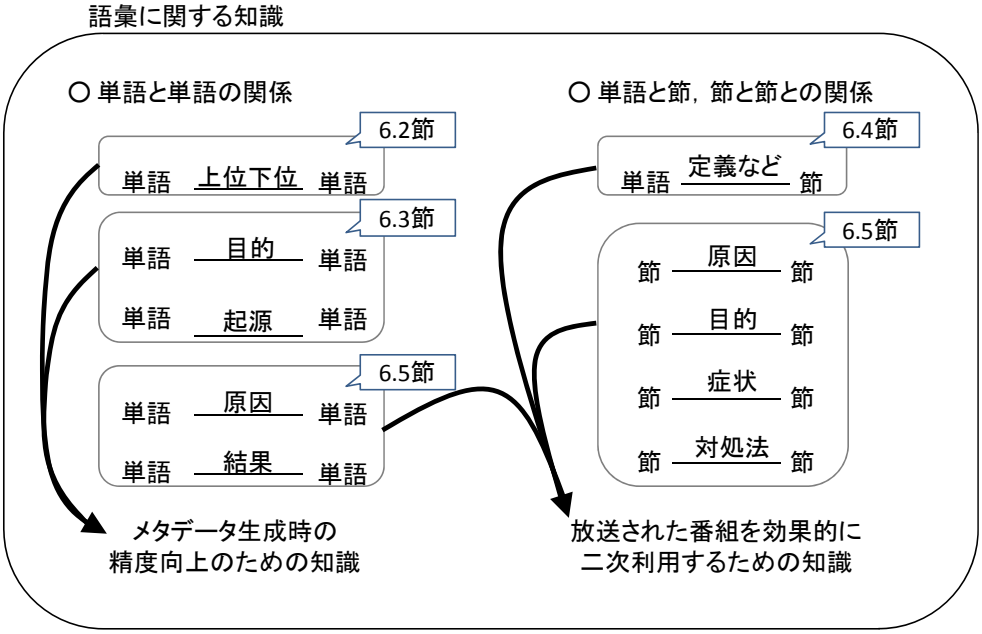


図6.1　　獲得対象の知識

単語と単語の関係は，計算機が扱う知識として有効と考えられ，シソーラスと呼ばれる単語の上位下位関係などを記した辞書が数多く存在する[61][62][63]．これらのシソーラスは，人手により生成されているため，単語の網羅性が問題となり，さらには新しい語を追加する更新作業にも，大きな労力を要してしまう．そこで，単語と単語の関係を，大規模なコーパスから抽出する手法を提案する．まず 6.2 では，大量のテキストデータを解析する際に問題となる未知語処理について述べる．未知語とは，辞書に登録されていない単語の事を言い，コンピュータ上でテキストデータを解析する場合，未知語の存在が問題とな

る. 頑健な解析システムを構築するためには, 未知語処理が必要となる. そこで, 6.2 では, 未知語の特徴を明らかにし，名詞未知語の属性推定のための知識について整理する．この処理で扱う属性は，単語の上位語に相当する．6.3 では，コーパスから単語に対する語彙情報を獲得する手法について説明する．従来，語彙的知識獲得に関する研究は，単語間の「上位－下位」「部分－全体」といった関係を対象としたものが多く存在する[64][65][66][67]. ここでは，従来行われていない「目的」と「起源」の関係を推定対象とする．6.2，6.3 では英語を対象としているが，同様の処理は日本語でも可能と考えられ，このような知識を獲得することにより，クローズドキャプションを解析してメタデータを自動生成する処理の精度向上につながると考えられる．例えば，「衆議院」と「参議院」は，国会を構成する議院である，という知識があれば，この 2 つの単語には関連があることが分かり，提案手法における話題のクラスタリングやトラッキング処理において，より高精度な解析が可能となる．

6.4 ではニュースに関するテキストを対象として，単語と，その説明を行っている節を自動獲得して，単語と節の関係を推定する手法を提案する．従来，テキストから用語の定義を抽出する研究として，「α とは β である」，「α は β」といった，表層パターンマッチングに基づいた手法が提案されている[68]．提案手法では，従来手法で使われたパターンとは異なる表現も対象とし，その説明文の単語に対する意味的役割の推定も行う．6.5 では，健康に関する番組のクローズドキャプションを対象として単語間の因果関係を抽出する手法について説明する．従来手法では，「ため」などの明示的に因果関係を表す表現を利用していた[69]が，ここでは，このような手掛かり語も自動獲得する手法を提案する．さらに，単語間だけではなく，節間の関係を推定する手法も提案する．6.4，6.5 の処理で獲得される知識は，言葉の定義や原因，理由など，物事を調べる際に利用できる知識である．そこで，獲得した知識を教育分野において利用するためのマルチメディア教育支援システムを紹介する．これらの知識の獲得元となるテキストには映像とのリンクも存在するため，3.4 で言及したマルチメディア百科事典などのアプリケーションにおいても有益となる．

6.2 未知語処理

6.2.1 未知語の特徴調査

近年，機械で扱える辞書は大規模で充実したものとなってきており，登録単語数が数十万語を超えるものも珍しくはない．しかし一方で，毎年非常に多くの新語が作り出されており，こういった日々増え続ける新語を，逐次，辞書に登録することは困難である．そのため，大規模コーパスや，新聞社・放送局におけるテキスト，Web テキストを解析する際には，どうしても辞書に登録されていない未知語が出現してしまう．実用的な言語処理システムの構築のためには，このような未知語が存在する場合でも，構文解析，情報抽出処理などにおいて，可能な限り未知語の品詞やその意味属性（上位語）の推定を行い，正常に解析を終了させることが好ましい．未知語の品詞や意味属性の推定のために，まずは未

知語が出現する際の特徴について英文コーパスを用いて考察する．

本論文における未知語の定義は以下とする．

（定義）その語自体が辞書に含まれていないすべての単語．（複数形，過去形，比較級，最上級などの語尾変化によるものは未知語に含めない）

ただし，ハイフンを使用した明らかな複合語の場合は，切り出された語の中に一つでも未知語がある場合のみ，その複合語を未知語とする．

未知語の特徴を調査するため, 一般的な英文が集められている LOB CORPUS を用いた．このコーパスには，約 100 万単語，約 5 万 4 千文が含まれている．また，品詞と単語の属性情報が単語ごとに付けられている．LOB CORPUS の文例を表 6.1 に示す．

未知語であるか否かを判定する辞書として，EDR 電子化辞書と UNIX のスペルチェッカを併用した．これらには，併せて 12 万 6 千単語が登録されている．コーパスに対して辞書に含まれているか否かを照合し，未知語を抽出した．未知語の数を表 6.2 に示す．コーパスは 54,297 文からなっていたので，一文あたりの平均単語数は約 18 個となり，平均 4 文に 1 個の未知語が存在することになる．

表6.1　LOB CORPUSの文例

```
D01   2    ^ with_IN so_QL many_AP problems_NNS to_TO solve_VB ,_, it_PP3
D01   2 would_MD be_BE a_AT great_JJ help_NN to_TO
D01   3 select_VB some_DTI one_CD1 problem_NN which_WDTR might_MD be_BE the_ATI
D01   3 key_NN to_IN all_ABN the_ATI others_APS ,_, and_CC
D01   4 begin_VB there_RN ._. ^ if_CS there_EX is_BEZ any_DTI such_ABL
D01   4 key-problem_NN ,_, then_RN it_PP3 is_BEZ undoubtedly_RB
D01   5 the_ATI problem_NN of_IN the_ATI unity_NN of_IN the_ATI gospel_NN ._.
D01   5 there_EX are_BER three_CD views_NNS of_IN the_ATI
D01   6 fourth_OD gospel_NN which_WDTR have_HV been_BEN held_VBN ._.
```

表 6.2　未知語の出現数

全単語数	1,013851 語
未知語数	14,892 語

表 6.3　未知語の分類

未知語の種類	単語種類数	のべ数
大文字で始まる単語 例：Herold, Lytton, Nato	4971 種	11,151 個 (83%)
小文字で始まる単語 例：asphyxia, caesium,negro	948 種	1,776 個 (13%)
省略形の単語 例：approx., i.r.s., r.m.s.	64 種	212 個 (2%)
ハイフンによる複合語 例：avant-garde, co-existence	214 種	336 個 (2%)
全体	6,197 種	13,475 個

抽出した未知語を，大文字で始まる単語，小文字で始まる単語，省略形の単語，ハイフンによる複合語の 4 種類に分類し，その数を評価した．結果を表 6.3 に示す．なお，このコーパスでは，文の先頭も小文字で記述されている．表 6.3 では大文字で始まる未知語が多いことが目立つ．この 4,971 種類の単語を調べたところ，そのうち 4,786 種類(96%)が固有名詞であった．固有名詞は，その意味が，人物，場所，組織，施設，生産物，理論などに限定される．また，省略形の単語の未知語では，64 種類中 32 種類(50%)が単位を意味する単語であった．このような固有名詞や単位表現を文書から抽出するワークショップは，これまでに英文, 日本語文などを対象に開催され[70][71], この成果を利用することにより，これらの単語の意味属性の推定は高精度な処理が可能である．

ハイフンを含む複合語の未知語は，214 種類中 88 種類(41%)が接頭辞(co-, neo-, pre-, un-など)によるものであった．これらの接頭辞は未知語の品詞，意味属性の手掛かりとなり得る．

小文字のみの未知語を対象として，その品詞を調査した．結果を表 6.4 に示す．ここでは，英語以外の単語については品詞を考えず，外来語として分類している．表 6.4 の結果より，名詞，形容詞，外来語の頻度が高いことがわかる．外来語が使われる際には明示的にイタリック体が使われることが多く，この語の分類は容易である．また，その単語の意味を必要とすることも少ないと考えられる．

小文字のみからなる未知語を対象に，語幹が登録語であるかを調査した．結果を表 6.5 に示す．表 6.5 では，小文字のみからなる形容詞と副詞の未知語の大部分は，登録語に接頭辞，接尾辞を付加した単語，つまり，登録語からの派生語であることがわかる．形容詞と副詞の未知語は，派生に関する処理が最も有効で，また，この処理だけで，ほぼ十分で

表 6.4　小文字で始まる未知語の品詞調査結果

品詞	単語種類数	のべ数
名詞	337 種	636 個
動詞	55 種	91 個
形容詞	153 種	233 個
副詞	58 種	214 個
外来語	326 種	571 個
その他	19 種	31 個
全体	948 種	1,776 個

表 6.5　小文字で始まる未知語が派生語である割合

品詞	単語種類数	登録語からの派生の未知語種類数	派生ではない未知語種類数
名詞	337 種	127 種	210 種
動詞	55 種	19 種	36 種
形容詞	153 種	147 種	6 種
副詞	58 種	48 種	10 種
全体	603 種	341 種	262 種

あることが分かる．名詞と動詞の未知語も，登録語からの派生語が多く，派生に関する処理は有効な一手段であるといえる．

また，表 6.5 の結果から，登録語からの派生でない未知語 262 種類中の 246 種類(93.9%)は名詞か動詞に含まれるので，新たな意味の単語を作り出すときは，名詞か動詞になりやすいことがわかる．その絶対数が多い名詞には，言葉の新造能力があり，新語は名詞が多いということが，このことからわかる．

6.2.2 名詞未知語の意味属性（上位語）推定のための知識

名詞の意味（上位語）を考えるためには，意味属性の分類が重要な課題となる．意味属性として，本論文では旧科学技術庁の Mu プロジェクト[72]において行われた分類を利用する（表 6.6）．これは 12 個のファセット（上位概念）と 48 個の意味マーカ（下位概念）からなる．名詞未知語がこれらのいずれに属するかを決定するアプローチとして，形態素レベル, 句レベル, 文レベルの三つのレベルが考えられる. 以下に各レベルについて述べ, その有効性について論じる．

表 6.6　名詞意味マーカ体系

ファセット	意味マーカ
国・機関・組織	
生物	人，動物，植物，その他
無生物	自然物，部品および材料，生産物，施設，その他
知的抽象物	理論・法則・学問，知的抽象的道具・方法，知的抽象的材料，知的抽象的生産物，その他
部分	部分・要素，生物の器官および構成要素，その他
属性	属性名，関係，形態，状態，構成，特徴，その他
現象	自然現象，力・エネルギー，生理的現象，社会的現象，物象，制度・慣習，その他
心得	感覚・反応，認知・思考，その他
行動	行為，動き，その他
測度	数，数量名，基準標準，単位，その他
場所・空間	
時間	時点，時間間隔，所要時間，その他

形態素レベルの意味属性推定

通常，接頭辞は，単語に意味の変化を与える．しかし，表 6.6 の意味マーカは荒い分類であるため，接頭辞が付くことによってその範囲を超えて意味が変化することは少ない．従って，接頭辞と接頭辞を除いた語幹が登録語であるなら，意味マーカは，その接頭辞を除いた語幹と同じと推定できる．例えば，未知語“unawareness”は，接頭辞“un”を除いた“awareness”と同じ意味マーカ「認知・思考」と推定できる．

接尾辞は意味の変化を与えない．主として，品詞の変化を与えるだけである．しかし，

接尾辞は，接続できる語幹との間に意味的な制限があるため[73]，この特徴を用いることにより，その単語の意味を限定することができる．名詞をつくる 26 個の接尾辞に対して，接尾辞を除いた語幹の品詞により単語の意味を分類し，この知識を利用する．表 6.7 にその一部を示す．例えば未知語“expressionism”は，接尾辞が“ism”， 接尾辞を除いた語幹の品詞が名詞であることから，その意味マーカは「行為」，「生理現象」，「理論」のいずれかであると推定できる．

表 6.7　接尾辞による意味属性の推定

接尾辞	接尾辞を除いた語幹の品詞	意味属性
-(a)cy	名詞	国・機関・組織，生産物，行為
	動詞	行為
	形容詞	状態，特徴
-(e)ry	名詞	施設
	動詞	機能
	形容詞	状態，特徴
-ism	名詞	行為，生理的現象，理論
	形容詞	状態，特徴，理論

未知語が明白な複合語である場合，複合語中の後部の部分にある名詞と同じ意味マーカであると推定できる．例えば未知語“applecake”は“cake”と同じ意味マーカ「生産物」と推定できる．複合語には，この例のような内心複合語と，複合語中の後部の部分にある名詞と同じ意味とはならない外心複合語がある[74]．後者はこの手法では誤った推定がなされる．しかし，未知語となるような，新たに作りだされる単語は，ほとんど前者に含まれる．実際に無作為に抽出した未知語 57 種類中，56 種類が内心複合語であった．

句レベルの意味属性推定

句は，まとまって一つの意味をなす．形態素レベルで推定できなくても，あるまとまった句の中で，その前後関係等から意味を限定できることがある．「名詞句 1+of+名詞句 2」の型の名詞句では，名詞句 1 と名詞句 2 の間には，「属性－対象」，「所有－対象」，「部分－全体」などの関係があると考えられるため，表 6.6 で示した意味マーカに対してこれらの関係となり得る意味マーカを整理することにより，ある程度の意味マーカの限定ができる．例えば “... pains of the dromozoa”という文において未知語“dromozoa”は，“pain”の意味マーカである「感覚・反応」を属性や所有，部分に持つ意味マーカに限定できる．

統語的に並列の関係にある二つの単語は，それぞれ同じ上位概念を持つ．したがって，未知語と並列関係にある語が登録語である場合，未知語の意味は，その登録語が持つ上位ファセットに限定できる．例えば，“... from pre-existing heart disease or from almost pure asphyxia”という文では，未知語“asphyxia”は，“disease”の意味マーカ「生理的現象」の上位

ファセット「現象」に限定できる.

前置詞句中の名詞は，その前置詞により意味属性が限定できる場合がある．例えば，“All the birds in my birdroom appeared ...”という文では，未知語“birdroom”は“in”の可能な前置詞目的語「場所」,「時間」,「道具」を提供する意味マーカに限定できる.

文レベルの意味属性推定

文の動詞が抽象的関係を意味する状態動詞（be 動詞など）である場合，主体あるいは対象のいずれか一方から，他方の意味属性を推定できる．例えば，“... drowning is not a simple asphyxia ...”という文では，be 動詞の主体と対象の関係は「同一属性」,「上位－下位」,「対象－属性」などがあるため，未知語“asphyxia”は主体“drowning”の意味マーカ「行為」と「同一属性」,「上位－下位」,「対象－属性」の関係にある意味マーカに限定できる.

また, 各動詞の格フレーム情報を利用することも有効な手段となる. 動詞に対する格は，近年，大量のコーパスを解析した結果がいくつか公開されている[75]．例えば，“... the dancers began the calinda, ...”という文では，未知語“calinda”は，動詞”begin”の対象格に入ることができる意味マーカに限定できる.

各レベルにおける意味属性推定の有効性調査

形態素レベルではコーパスから抽出した未知語 337 種，句レベル，文レベルでは小文字の未知語 100 種（出現文数 277 文）を対象として，意味属性推定の有効性についての調査を行った．調査は，まず，知識の有効性を限定可能な意味マーカ数により以下の四つのランクに分類する.

1. 得られた意味マーカが 4 個以下
2. 得られた意味マーカが 5 個以上 20 個以下
3. 得られた意味マーカが 21 個以上 48 個以下
4. 意味マーカ限定不能

各レベルにおける調査結果を表 6.8 に示す．この表では，例えば，句レベルの知識だけを用いた場合，全体の 22%の未知語が 4 個以下，23%が 5 個以上 20 個以下，8%が 21 個以上 48 個以下の意味マーカに限定可能で，残りの 47%が句レベルによる意味マーカ限定ができなかったことを示す.

調査の段階で，未知語を含む文には，次の三つの特徴が見られた.

(1) 未知語と並列関係にある名詞が同じ文中に出現していることが多い
(2) 未知語が be 動詞の主体または対象の位置にあることが多い
(3) 「名詞 1+of+名詞 2」の型の名詞句において，未知語は名詞 1 より名詞 2 に位置することが多い.

(1)(2)の特徴を明確にするため, コーパスから無作為抽出した同条件にある登録語（100

表 6.8　未知語意味推定処理の各レベルにおける有効性

レベル	ランク 1	ランク 2	ランク 3	ランク 4
形態素レベル	49%	6%	0%	45%
句レベル	22%	23%	8%	47%
文レベル	9%	20%	24%	47%

表 6.9　未知語の特徴調査

	未知語	登録語
並列関係が出現する割合	44/100	24/100
be 動詞の主体または対象として出現する割合	18/100	8/100

種 277 文）との比較調査を行った．その結果を表 6.9 に示す．並列関係には，未知語は登録語の 1.83 倍(未知語 44 種，登録語 24 種)， be 動詞の主体または対象には，未知語は登録語の 2.25 倍(未知語 18 種，登録語 8 種)出現しており，(1)(2)の特徴が確認できる．また，「名詞 1+of+名詞 2」の型の名詞句中に出現した 35 種の未知語に対して調査を行ったところ，名詞句 1 の位置には 35 種中 8 種，名詞句 2 の位置には 27 種が出現していた．

表 6.8 では，低いレベルほど有効な情報が得られていることがわかる．形態素レベルが有効であるという事実は，未知語は登録語を派生させて新しく作り出されることが多いことを示している．

句レベルが有効である理由の一つとして，多くの文において，未知語と並列関係にある名詞が出現しているということがあげられる．未知語は登録語と比較すると，並列な関係にある語が 2 倍近く出現していることがわかる．辞書に登録されていないような難しい単語を使用する場合，人が容易に理解できるように，並列構造などにより意味を補っているものと考えられる．

また，be 動詞文の主体または対象の位置にある未知語の出現回数も，登録語と比較すると 2 倍以上であった．主動詞が be 動詞である文は，一般的に陳述文であるので，これも，未知語の意味を人が容易に理解できるようにしていると考えられる．

文レベルにおいては，be 動詞による陳述文以外からは多くの情報は得られなかった．この理由として，動詞の多義性が挙げられる．

また，「名詞 1+of+名詞 2」の型の名詞句中の名詞句において，名詞 2 が未登録語であることは名詞 1 の 3 倍以上多かった．この理由は，名詞 1，名詞 2 に入りやすい名詞の意味に偏りがあるからと考えられる．例えば，名詞 1 には「属性」を意味する単語が入りやすい．このことより，未知語の意味属性には偏りがあるという仮説が提案できる．この特徴は，未知語の意味属性推定の一つの手掛かりとして利用できる．

6.3　単語の語彙体系知識獲得

6.3.1　語彙体系知識の概要

テキストデータに出現する単語の意味を表現するために，その語の意味的な役割となるQualia Structure[76]を定義するという手法がある．Qualia structure は，欧州における１２の言語で相互参照できる語彙の意味体系を構築することを目的としたプロジェクトであるSIMPLE (Semantic Information for Multifunctional Plurilingual LExicons[77])で採用された語彙体系であり，語の指示対象（概念）の意味記述の要素として，以下の４つの役割を定義している．

A) formal role : 語の指示対象と他を区別する情報で，一般に上位の概念を表す
B) constitutive role : 指示対象が持つ内部的な性質・構成要素
C) telic role : 指示対象が持つ典型的な機能・目的
D) agentive role: 指示対象の起源，指示対象が引き起こす事象

例えば，名詞「本」では，「出版物」が formal role,「テキスト」が constitutive role,「読む」が telic role,「書く」が agentive role にあたる．

これまでに，formal role に該当する名詞の上位概念をテキストから抽出する手法や，constitutive role に該当する単語の「部分－全体」の関係を抽出する手法は，いくつか提案されている[64][65][66][67]．また，formal role や constitutive role の関係は，既存の辞書である WordNet の hypernym, meronym の関係に該当する．一方，telic role と agentive role の関係を抽出する研究は，あまり行われてきていない．

telic role や agentive role は, 単語の意味を拡張して用いるような場合の換喩表現の解析時に有用な情報である．例えば，“Mary finished her beer.”という文では，単語“finish”の目的語となりにくい単語“beer”があり, 解釈が困難である．このような場合, 単語“beer”の telic role の関係にある“drink”を利用することにより，“Mary finished drinking her beer.”と文を補え，解釈が可能となる．

qualia structure において, 名詞の telic role と agentive role は複数の動詞により表現できる．例えば，名詞“book”の典型的な telic role は“read”，agentive role は“write”であるが，telic role として“study”，agentive role として“publish”も当てはまる．各名詞に対して telic role や agentive role の役割を果たす度合いも動詞により異なる．そこで本章では，名詞の telic role と agentive role の役割を果たす度合いに関する優先順位を動詞に対してランク付けする手法を提案する．ランクが上位である動詞は，名詞に対する対象の役割を果たすと考えることができ，この役割関係が名詞に対する語彙知識となる．

本章で提案する手法では，書き言葉のコーパスとして有名なBritish National Corpus (BNC:[78])を利用する．このコーパスは約9千万単語が含まれている．また，係り受け解析としてRobust Accurate Statistical Parsing(RASP:[79])を利用する．RASPでは，最初に，各単語に対してCLAWS-2(the Constituent Likelihood Automatic Word-tagging System)[80]で定められた品詞タグを付与し，そして，係り受け構造を特定する．係り受け構造の関係として，23種のGrammatical relationsが使われている．例えば，“ncmod”というGrammatical relationは名

詞間の係り受け関係を示し，“dobj”というGrammatical relationは，動詞と，その直接目的語の関係にある名詞との関係を示す．以下にRASPの出力例を示す．

例）入力　“airplane tickets”
RASP 出力　ncmod(_, tichet_NN1, airplane_NN2)
入力　“read books”
RASP 出力　dobj(read_VV0, book_NN2,_)

ここで，NN1，NN2，そして VV0 は CLAWS-2 で定められた品詞タグで，それぞれ単数普通名詞，複数普通名詞，動詞基本形を指す．

6.3.2　語彙知識抽出処理

本節では，名詞に対して，Qualia Structure における telic role と agentive role の役割を果たす動詞をランク付けする語彙知識抽出処理について説明する．まず，実験と評価のための名詞と動詞を限定し, 30 個の名詞と, 各名詞に対してそれぞれ 50 個の動詞を選択して，処理対象とした．また英語を母語とする被験者により，各名詞に対する動詞の telic role らしさと agentive role らしさを主観的に評価し，これを Gold-standard data として学習データと評価用データとして利用した．この実験用の名詞－動詞の組み合わせデータに対して，最大エントロピー法[29]を利用する手法と，手作業により作成したテンプレートを利用する手法の２つの手法により，動詞の telic role らしさと agentive role らしさを評価する．

処理対象とする名詞－動詞の組み合わせデータ

以下に示す 30 個の名詞を実験対象とする．これらのうち 10 個は Qualia Structure に関する過去の文献で扱われていた名詞で，残りの 20 個は無作為に選択した．

[処理対象とする単語]

book, car, knife, speech, food, table, door, prisoner, juice, novel,
executive, delegation, phone, clinic, cash, beef, review, letter,
counter, county, sunshine, accounting, register, complexity, gaze,
profession, investigation, imagination, estimate, maturity

さらに各名詞に対して，それぞれ 50 個の動詞を BNC から実験対象として取り出し，50 個の動詞に対して各名詞との各役割の果たしやすさをランク付けする処理を行う．この 50 個の動詞の一部は，名詞との共起頻度の高いものを選択し，残りはランダムに選択した．例えば，名詞“book”に対して選択された動詞を以下に記す．

[名詞“book”に対して選択された動詞]

abandon, add, appear, believe, borrow, bring, browse, buy, call,
compile, dedicate, design, destroy, dispose, end, expect, fill, find,
follow, get, hand, hold, introduce, keep, lay, make, move, need, pack,
plan, prepare, print, provide, publish, read, remove, return, show,
snatch, start, steal, suit, think, throw, thrust, translate, treat,
want, withdraw, write

この作業により，合計 1500 個の名詞－動詞の組み合わせが選択される．ここでは実験用として，後に生成したランクの評価を行うために動詞の数を限定したが，大規模コーパスから telic role と agentive role などの語彙知識を実際に抽出する際には，動詞の限定は行う必要は無い．

Gold-standard data

英語を母語とする 2 人のアノテーターにより，30 個の各名詞に対する動詞の telic role らしさと agentive role らしさを主観的に評価し，0 から 10 までの整数で各動詞に対する点数を付与した．この点数が 10 である場合は，名詞に対する代表的な telic role または agentive role の役割を果たす動詞であると考えられる．この点数が 0 である場合は，名詞に対して，telic role または agentive role の役割を取りえない動詞であると考えられる．2 人のアノテーターによる点数の平均値を，学習データや評価用データとして利用する Gold-standard data とする．

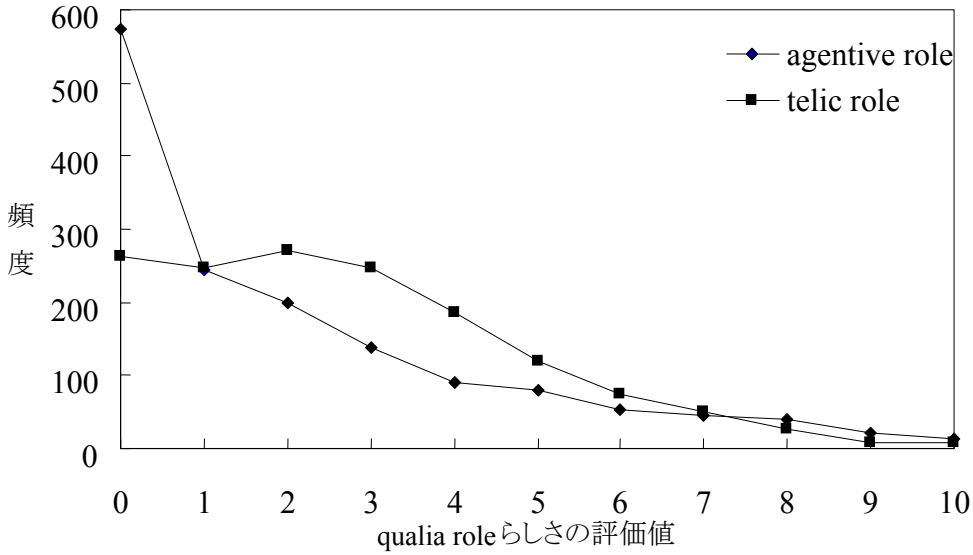


図 6.2　人により付与された qualia role らしさの点数の分布

図 6.2 に人により付与された qualia role らしさの点数の分布を示す．agentive role に対する点数の平均は 2.06，SD(分散)は 5.97 であり，telic role に対する点数の平均は 2.70，SD

は 4.56 であった．agentive role の SD は telic role より大きく，agentive role のほうが，その役割を果たすかを判断しやすいことが伺える．

最大エントロピー法を利用したランキング手法

名詞の qualia role になるような動詞は，特別な構文構造で名詞と共起する傾向がある．そこで提案手法では，このような特別な構文構造を最大エントロピー法により学習する．学習データとして，人手により作成した gold-standard data を用い，その評価点として 7 以上が与えられた動詞を正例，0 であった動詞を負例とする．表 6.10 に名詞“book”に対する正例と負例となる動詞の例を示す.

表 6.10　名詞“book”に対する正例と負例となる動詞の例

Role	正例	負例
agentive role	print, publish, write, make, compile, design, start	abandon, appear, destroy, dispose, follow, hand, hold, keep, lay, pack, remove, return, snatch, suit, throw, thrust, withdraw
telic role	read, browse	call, end, appear, suit

次に BNC から, 正例と負例の名詞－動詞ペアが含まれ, さらに, 名詞－動詞ペアが RASP の解析による Grammatical relations を持つ文を抽出する．全ての名詞－動詞ペアに対して，agentive role では 7810 個の正例と，13780 個の負例が，telic role では 9925 個の正例と 5148 個の負例が抽出された．この学習データから，以下の素性を抽出して最大エントロピー法により学習を行う．

- 対象とする名詞－動詞ペアの Grammatical relations
- 対象とする名詞と，対象動詞以外の単語の Grammatical relations とその品詞タグ
- 対象とする動詞と，対象名詞以外の単語の Grammatical relations とその品詞タグ

例えば，“Can I have a book to read?”という入力文に対して，名詞“book”と動詞“read”を対象として素性を抽出することを考える．入力文を RASP により解析すると，以下の出力が得られる．

(|Can:1_VM| |I:2_PPIS1| |have:3_VH0| |a:4_AT1|
　|**book**:5_NN1| |to:6_TO| |**read**:7_VV0| |?:8_?|)
(|ncsubj| |have:3_VH0| |I:2_PPIS1| _)
(|dobj| |have:3_VH0| |**book**:5_NN1| _)
(|ncsubj| |**read**:7_VV0| |I:2_PPIS1| _)
(|xcomp| |to:6_TO| |**book**:5_NN1| |**read**:7_VV0|)
(|detmod| _ |**book**:5_NN1| |a:4_AT1|)
(|aux| _ |have:3_VH0| |Can:1_VM|)

この RASP による解析結果から，以下の素性が抽出できる．

- “book”, “read”の Grammatical relations
  xcomp.to
- “book”と，“read”以外の単語の Grammatical relations とその品詞タグ
  dtmod, AT　　(単語“a”と“book”から)
  dobj, VH　　(単語“have”と“book”から)
- “read”と，“book”以外の単語の Grammatical relations とその品詞タグ
  ncsubj, PP　　(単語“I”と“read”から)

ここで Grammatical relations の xcomp.to は to 不定詞による述語と，その係り先となる名詞との関係を示し，dtmod は，名詞と限定詞の関係，dobj は述語の目的語，ncsubj は述語と主語の関係を示す．また，品詞タグの AT は冠詞，VH は動詞 have，PP は代名詞を示す．表 6.10 に示すように，動詞“read”は名詞“book”に対して telic role において正例として扱われるため，これらの RASP による解析結果が正例の素性となる．

処理対象名詞を除く 29 個の名詞の正例, 負例に該当する動詞の RASP による解析結果から素性を抽出し，最大エントロピー法により各素性に対する有効性を示すパラメータ λ を学習し，式(2.20)により対象名詞と動詞が文中で共起したときの，その関係が agentive role である条件付き確率 P(rel=agentive role|$f_{n,v}$)と，telic role である条件付き確率 P(rel=telic role|$f_{n,v}$)を計算できる．以下に，文“I always had book to read.”と文“Complete books have been written on this subject.”における処理対象名詞“book”と動詞“read”，“write”の関係の確率値を計算した例を示す．

例文 1）$f1_{\text{book, read}}$：“I always had book to read.”
P(rel=agentive role|$f1_{\text{book, read}}$) = 0.396
P(rel=telic role|$f1_{\text{book, read}}$) = 0.943
例文 2）$f2_{\text{book, write}}$：“Complete books have been written on this subject.”
P(rel=agentive role|$f2_{\text{book, write}}$) = 0.652
P(rel=telic role|$f2_{\text{book, write}}$) = 0.455

P(rel=agentive role|$f_{n,v}$)，P(rel=telic role|$f_{n,v}$)は 0 から 1 の値を取り，0.5 より大きい場合，動詞はその名詞に対して該当する関係を持つと判断できる．上記の例では，動詞“read”は名詞“book”に対して telic role の関係を持ち，動詞“write”は名詞“book”に対して agentive role の関係を持つと判断できる．

名詞と動詞が共起しやすい関係であるほど, agentive role や telic role の関係を持ちやすくなると考えられる．そこで，このような要素を相互情報量[81]により取り入れる．名詞 n

と動詞 v の相互情報量 $MI(n, v)$は式(6.1)で定義される．

$$MI(n,v) = \log \frac{P(n,v)}{P(n)P(v)} \tag{6.1}$$

ここで，$P(n)$は名詞 n が出現する確率，$P(v)$は動詞 v が出現する確率，$P(n, v)$は名詞 n と動詞 v が共起する確率を示す．

対象名詞と動詞がある文で共起したとき該当関係を持つ確率値と相互情報量を利用して，動詞 v が名詞 n に対して agentive role の関係を持つスコア $score_ME_{agentive}(n,v)$と，telic role の関係を持つスコア $score_ME_{telic}(n,v)$を，式(6.2)，式(6.3) に定義する．

$$score\_ME_{agentive}(n,v) = MI(n,v) \times \frac{\sum_{n,v} \left\{ 2 \times P'(rel = agentive \mid f_{n,v}) - 1 \right\}}{DF(n,v)} \tag{6.2}$$

$$score\_ME_{telic}(n,v) = MI(n,v) \times \frac{\sum_{n,v} \left\{ 2 \times P'(rel = telic \mid f_{n,v}) - 1 \right\}}{DF(n,v)} \tag{6.3}$$

$$P'(rel \mid f_{n,v}) = \begin{cases} P(rel \mid f_{n,v}) & if\ P(rel \mid f_{n,v}) \geq 0.5 \\ 0 & otherwise \end{cases} \tag{6.4}$$

ここで，$DF(n,v)$は名詞 n と動詞 v が共起する文数を示す．式(6.2)，式(6.3)では，分子の P'の確率値を 2 倍し，その値から 1 を引くことにより，この値を-1～+1 に正規化している．$score\_ME_{agentive}(n,v)$，$score_ME_{telic}(n,v)$の値が大きいほど，最大エントロピー法を利用する手法において名詞 n と動詞 v は該当する関係を持ちやすいと判断する．

<u>テンプレートを利用したランキング手法</u>

文中で，主語にある名詞と受動態の形態である動詞が存在する場合，その動詞は名詞に対して agentive role の役割を果たしやすい．例えば，名詞“book”に対して agentive role を取ると考えられる動詞“write”は“This book was written by him.”という文などで共起する．同じように，“a N worth V-ing”のような構造がある場合，動詞 V は名詞 N の telic role になりやすいと考えられる．例えば，名詞“book”に対して telic role を取ると考えられる動詞“read”は“a book worth reading.”という文などで共起する．そこで，このような典型的なテンプレートを手作業により生成し，動詞が名詞に対して，agentive role や telic role の役割を果たすかを判定するためのスコアの指標として利用する．表 6.11 に agentive role のテンプレートを，表 6.12 に telic role のテンプレートを示す．

agentive role と telic role において，コーパス中で，名詞 n と動詞 v がテンプレートに当てはまる頻度を $TempF_{agentive}(n,v)$と，$TempF_{telic}(n,v)$としたとき，動詞 v が名詞 n に対して agentive

表 6.11　agentive role 用のテンプレート

テンプレート	例
N BE V[+en]	(the) book was written (by Kim)

表 6.12　telic role 用のテンプレート

テンプレート	例
N (BE | φ) (worth | deserving | meriting) (V[+ing] | V[+nom])	(a) book worth reading
N BE worthy of V[+nom]	(the) book is worth of reading
N (deserves | merits) V[+nom]	(the) book merits reading
Adverb-V[+en] N	(a) well-read book
Adverb V[+en] N	(a) well read book
N BE Adverb-V[ed]	(the) book is well read
V[+ing] N	(I enjoy) reading books
N to V	(a) book to read

N: 名詞，V: 動詞

V[+ing]:動詞 V の現在進行形，V[+ed]:動詞 V の過去分詞形，V[+nom]:動詞 V の名詞化

role の関係を持つスコア $score\_template_{agentive}(n,v)$ と，telic role の関係を持つスコア $score_ME_{telic}(n,v)$を，式(6.5)，式(6.6) に定義する．

$$score\_template_{agentive}(n,v) = \frac{TempF_{agenteive}(n,v)}{DF(n,v)} \tag{6.5}$$

$$score\_template_{telic}(n,v) = \frac{TempF_{telic}(n,v)}{DF(n,v)} \tag{6.6}$$

ここで，$DF(n,v)$は名詞 n と動詞 v が共起する文数を示す．式(6.5)，式(6.6)の値が大きいほど，テンプレートを利用する手法において名詞 n と動詞 v は該当する関係を持ちやすいと判断する．

6.3.3　評価手法の改良：Spearman's rank correlation の改良

提案手法の有効性を測るために，Spearman's rank correlation と呼ばれる 2 つランク付けされたデータの相関を測る指標を用いる．2 つのデータのランクの差を d_x，データ数を n とした場合，Spearman's rank correlation の値 Rs は式(6.7)で与えられる．

$$Rs = 1 - \sum_{x=1}^{n} d_x^2 / E(\sum_{x=1}^{n} d_x^2) = 1 - 6\sum_{x=1}^{n} d_x^2 / n(n-1) \tag{6.7}$$

Spearman's rank correlation の値 *Rs* は，2 つのデータのランク全ての相関を評価しているため, 上位のランクと下位のランクが同等に扱われる. 今回のタスクでは, 名詞の qualia role になるような動詞は少量しか存在せず，残りの qualia role とならない動詞に対してはランク付けした結果に重要な意味を持たない．Gold-standard data の作成で利用した 2 人のアノテーターのランクを，式(6.7)の Spearman's rank correlation の値で評価したところ，telic role では 0.448，agentive role では 0.369 と低い値となった．そこで，Spearman's rank correlation の値を，ランクの上位 *m* 項目に対して評価を行えるよう，式(6.8)のように改良を行った.

$$Rs'(m) = 1 - \sum_{x=1}^{m} d_x^2 / E(\sum_{x=1}^{m} d_x^2)$$
$$= 1 - 6\sum_{x=1}^{m} d_x^2 / m(2m^2 - 3nm + 2n^2 - 1) \quad (6.8)$$

もし，2 つのデータが上位 *m* 個のランクが同じであれば *Rs'*(*m*)=1 となり，上位 m 個のランクが無相関の場合は *Rs'*(*m*)=0 となる. しかし, 2 つのデータに負の相関がある場合は, *Rs'*(*m*)<-1 となり，相関係数としては相応しくない値となってしまう．しかし，今回のタスクでは，負の相関を考慮する必要がないため，問題とならない.

6.3.4 語彙知識抽出実験と評価

名詞の qualia role になる動詞抽出のために，6.3.2 で提案した最大エントロピー法を利用したランキング手法と，テンプレートを利用したランキング手法による実験を行った．30 個の名詞とそれぞれ 50 個の動詞の名詞－動詞の組み合わせデータを処理対象とし, 式(6.2), 式(6.3)，式(6.5)，式(6.6)の値を求めた．最大エントロピー法を利用する手法では，30 個の名詞のうち 1 個を処理対象, 29 個を学習データとして交差検定を合計 30 回行った. 表 6.13 と表 6.14 に, 名詞"book"に対する agentive role と telic role の上位 8 個の動詞を示す. また, 同じ表には，アノテーターによる Gold-standard data の上位も示す．表 6.13 では，2 つの手法ともに，動詞"write"，"publish"，"compile"，"print"など，Gold-standard data の上位にも含まれる agentive role の動詞として相応しい単語が上位にランクされている. また, 表 6.14 では，Gold-standard data で最上位となった動詞"read"が，2 つの手法で最上位にランクされている.

この結果を定量的に評価するため，6.3.3 で提案した Spearman's rank correlation を改良した指標を利用する．agentive role，telic role に対する実験結果を式(6.8)により評価した結果を図 6.3，図 6.4 に示す．図 6.3，図 6.4 では，縦軸を Spearman's rank correlation を改良した指標の値の平均，評価対象とする横軸をトップ *N* 項目としている．例えば，横軸の値が 5 である場合，上位 5 項目を評価対象とする．この図で，Gold-standard data は，2 人のアノテーターによる Spearman's rank correlation を改良した指標の値を示し，この値が各役割を

表 6.13　名詞“book”の agentive role の役割と推定された動詞(上位 8 位まで)

Rank	最大エントロピーを利用した手法 (*score_ME*$_{agentive}$)	テンプレートを利用した手法 (*score_template*$_{agentive}$)	Gold-standard data (2 人のアノテータが付与した点数の平均)
1	dedicate(1.084)	publish(0.157)	write(10.0)
2	publish(0.898)	write(0.102)	publish(8.0)
3	compile(0.651)	read(0.019)	compile(8.0)
4	dispose(0.605)	call(0.015)	print(7.5)
5	write(0.438)	dedicate(0.011)	make(7.5)
6	browse(0.408)	print(0.008)	start(7.0)
7	borrow(0.399)	keep(0.007)	design(7.0)
8	print(0.386)	compile(0.006)	translate(6.0)

表 6.14　名詞“book”の telic role の役割と推定された動詞(上位 8 位まで)

Rank	最大エントロピーを利用した手法 (*score_ME*$_{telic}$)	テンプレートを利用した手法 (*score_template*$_{telic}$)	Gold-standard data (2 人のアノテータが付与した点数の平均)
1	read(2.814)	read(0.316)	read(10.0)
2	write(2.221)	write(0.112)	browse(9.0)
3	compile(2.115)	publish(0.079)	think(6.5)
4	dedicate(1.982)	buy(0.036)	buy(6.0)
5	buy(1.775)	keep(0.016)	provide(6.0)
6	borrow(1.695)	appear(0.015)	borrow(5.5)
7	throw(1.682)	make(0.014)	return(5.5)
8	publish(1.656)	provide(0.014)	start(5.5)

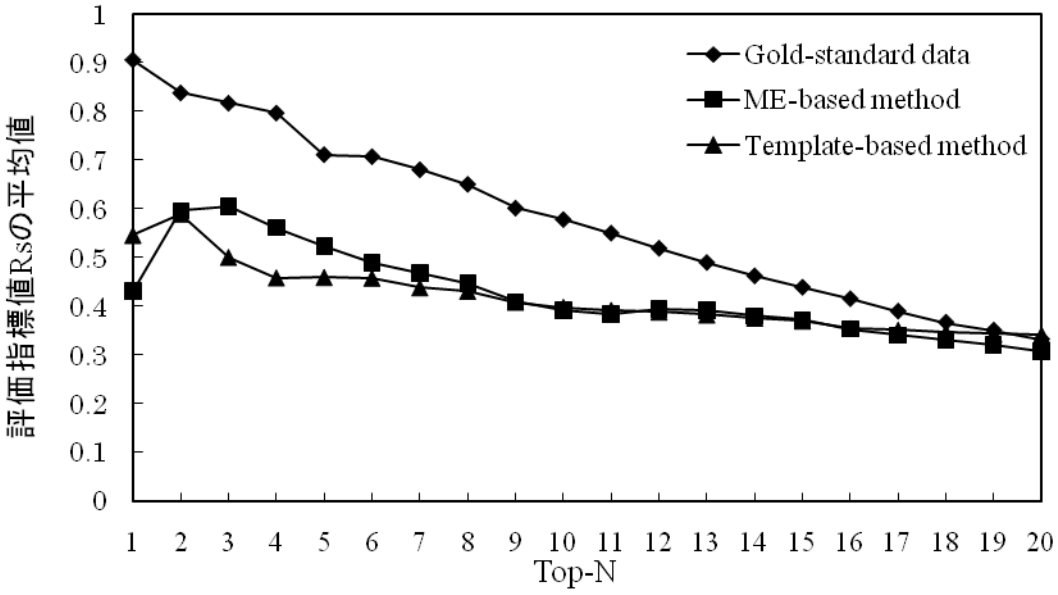


図 6.3　　agentive role における語彙知識抽出実験評価結果

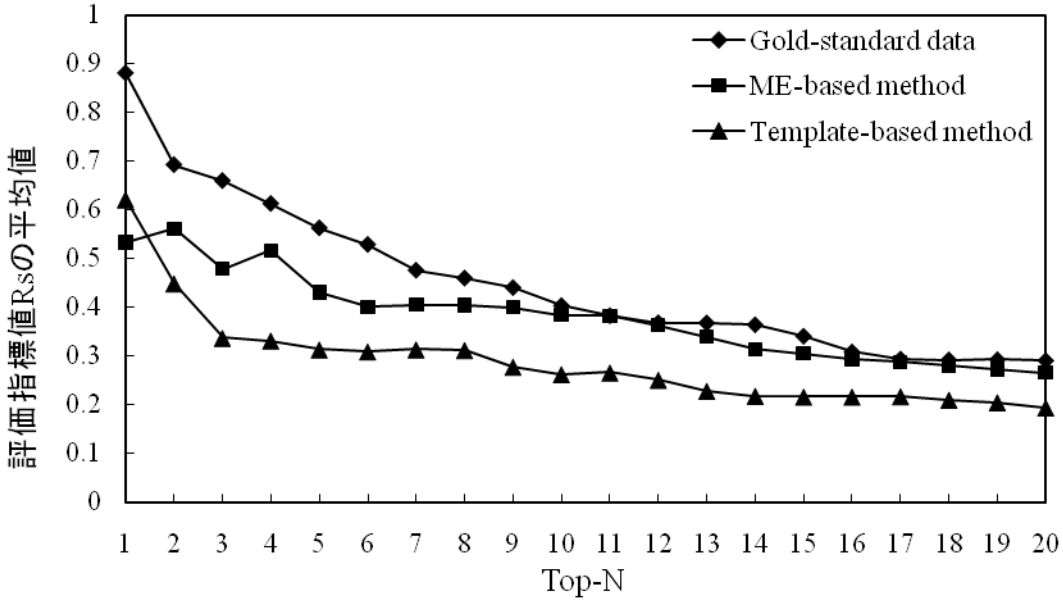


図 6.4　telic role における語彙知識抽出実験評価結果

果たす動詞順位の推定結果の上限値と考えられる．上位項目における Gold-standard data の値は高く，これは，上位項目においては，アノテーターの意見が一致しやすいことを示す．また，telic role における Gold-standard data の値は，agentive role に比べて低く，目的を表す telic role の関係には多くのバリエーションが存在することが予想できる．

最大エントロピー法を利用した手法において，正例として使用した動詞の数は，一つの名詞に対して平均 3.15 個であった．そこで，上位 3 項目を比較対象とする．agentive role に対して，最大エントロピー法を利用した手法では，上位 3 項目における Spearman's rank correlation を改良した指標の値は 0.605，テンプレートを利用した手法では 0.500，Gold-standard data の値は 0.816 であった．また，telic role に対して最大エントロピー法を利用した手法は 0.479, テンプレートを利用した手法では 0.337, Gold-standard data の値は 0.659 であった．最大エントロピー法を利用した手法は，テンプレートを利用した手法より，トップ項目を除き良好な結果が得られた. トップ項目をテンプレートによる手法で取り出し，残りを最大エントロピー法による手法で抽出するなど，2 つの手法を合わせて利用することにより，より Gold-standard data に近い値が期待できる．

今回の実験では評価のために名詞 30 個に対して動詞 50 個の合計 1500 個の名詞－動詞ペアのみを処理対象としたが，提案手法は全ての名詞，動詞に対しても適用可能であり，名詞に対する agentive role や telic role の役割を果たす動詞を，大規模コーパスから自動抽出することができる．構築される語彙知識は，翻訳や文理解など，多くのアプリケーションへの利用が期待できる．

6.2, 6.3 で述べた未知語の意味概念（上位語）に関する知識と，名詞に対する agentive role, telic role の役割に関する知識は，クローズドキャプションを解析する際に有益となる．放

送局が扱うニュースや番組テキストには，数多くの未知語が存在する．この未知語に対する上位語が特定できれば，類似するニュース記事を探してクラスタリングを行う際に効果的となる．また，telic role や agentive role は，要約処理における用語の重み付けにも利用できる．例えば，「衆議院」の telic role として「審議する」「可決する」などの動詞が獲得できていれば，これらの単語を含むニュース記事は重要な項目と判断できる．また，「法案」の agentive role として「作る」「審議する」などの動詞が獲得できれば，同様に，これらの単語を含むニュース記事は重要な項目と判断できる．6.2，6.3 では，英語を対象とした実験を行っているが，処理対象を日本語にすることにより同様の知識が獲得できる．このような知識を大量に集めることにより，より高精度なメタデータ生成処理が可能となると期待できる．

6.4 用語の説明文獲得

6.4.1 説明文の抽出処理

放送用のニュース記事は，最新の社会情勢や一般常識といった有益な情報を含み，映像ともリンクするため，教育用コンテンツ素材としても有用である．例えば，ニュースで使われた用語とその説明を記事から抽出して辞書形式で蓄積し，生徒の質問に答える Q&A システムを構築することが考えられる．この際，毎日のように出現する新たな時事用語とその説明を辞書に自動追加することができれば, Q&A システムを社会の動きと連動させたダイナミックなものとすることが可能になる．ニュース記事で難しい用語や，新語が扱われる場合は，視聴者が容易に理解できるよう，用語の説明を伴うことが多い．そこで本節では，用語とその説明を抽出し，用語の意味属性と，説明に含まれる動詞，助詞，名詞の意味属性などを素性とした最大エントロピー法[29]による学習を利用して，意味関係を分類する手法を提案する．

ニュース記事中で用語を説明する場合，その表現は以下の 3 通りに分類できる．

パターン A)　連体修飾節の係り元に説明，係り先に用語
　例：寝入りばなに怖い夢を見る「入眠時幻覚」は，・・・

パターン B)　連体修飾節の係り元に用語，係り先に説明
　例：「情報家電」と呼ばれる次世代の高速インターネットに対応した家電製品を・・・

パターン C)　文の主部に用語，述部に説明
　例：「クローン規制法」は，クローン技術を使って同じ遺伝子を持つ人間を人工的に作り出すことを禁止する法律です．

例では鍵括弧内が用語，下線部がその説明に対応している．2001 年 6 月のニュース記事を対象として，用語を説明する文を手作業により抽出し，どのパターンに属するか調査した．結果を表 6.15 に示す．

表 6.15　用語を説明するパターンの出現数

パターン	出現数
A	397(74.6%)
B	94(17.7%)
C	41(7.7%)

この結果から，ニュース記事中で用語を説明する場合，圧倒的にパターン A が多いことがわかる．ニュースでは，多くの情報をできるだけ短い時間で伝えることが重要である[55]．そのため，表 6.15 に示すように，１文で用語の説明をするパターン C は避けられ，用語の説明とニュースの主題を同時に記述できるパターン A が多く出現していると考えられる．実際に同じ期間の毎日新聞の記事の第一文と NHK ニュース記事の第一文の文字数を調べたところ，新聞が平均 11.3 文節（122.4 文字）であったのに対し，ニュース記事は，18.2 文節（178.8 文字）と新聞より約 1.6 倍も長い文であり，従属節が多用されていた．そのため，ニュース記事から用語の説明を抽出する処理では，パターン A の解析が重要となる．

また，パターン B では，その説明部分を構成する文節の数が少ないという特徴がある．実際に上記データを調査したところ，平均 2.0 個の文節しか存在しないことが分かった（パターン A は平均 7.9 個，パターン C は平均 9.5 個）．このパターンは，用語を平易な単語（対象用語の上位概念に相当する一般性の高い語）に言い換える目的で使われることが多く，用語自体の意味を十分に表現していない．そこで本手法では，このパターンを用語集に直接引用して利用することはしない．ただし，言い換えられた上位概念語は，用語の説明文生成処理において補完情報として利用する．

パターン C については，従来研究[68]により考察が行われているため，このパターンでは「α は β です」のみに限定して説明部分を抽出する．

提案する手法では，図 6.5 に示す手順により用語集を作成する．まず，ニュース記事集合から，用語集に登録する用語を抽出する．ここでは，強調を意図する鍵括弧で囲まれた名詞句[82]に限定して抽出した．鍵括弧で囲まれていても一般的な語には，その説明を抽出する必要が無いため，ここでは分類語彙表[83]に登録されている語は処理対象から除いている．次に，抽出した用語を含む文について用語説明のパターンを判別し，パターン A とパターン C の文から，用語の説明部分を抽出し説明文を生成する．この際，パターン A では，抽出された連体修飾節と，同一用語に対するパターン B の説明から抽出した上位概念語を用いて説明文を生成する．最後に用語とその説明文の意味的関係を判定して，その結果を用語集データベースに登録する．この時，一つの用語に対して複数の同じ関係の説明文が抽出された場合は，最適な説明文を選択する．以下に各処理について述べる．

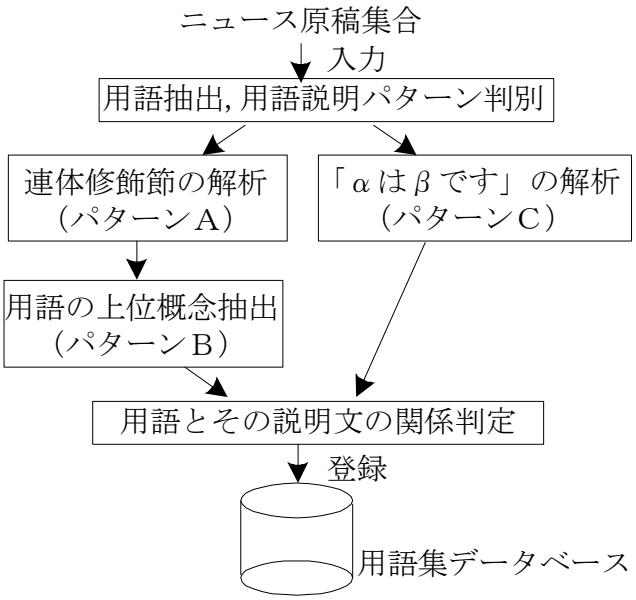


図 6.5　用語集作成手順

用語説明連体修飾節の抽出（パターン A）

用語を修飾する連体修飾節を抽出して，用語の説明として利用する．この処理では，十分な情報量を持つ説明を抽出するために，連体修飾節中に動詞が含まれているもののみを処理対象とする．

また，用語の前に出現する節の，どこまでが用語の説明となるかを判定しなければならない．本手法では，構文解析結果から，直接的または間接的に対象用語に係る節をすべて抽出して，用語の説明とする．しかし，下記のニュース記事では，下線部は用語に直接関係した情報とはならない．

- 群馬県内では，当面，畜産農家が牛を食肉として出荷する際に付ける「生産履歴書」という文書に，・・・

構文解析の結果から判定できそうであるが，複数の構文解析システムによる処理結果では，その全てが，下線部の「群馬県内では，当面,」は「付ける」に係ると解析され，間接的に用語を修飾する節として抽出されてしまう．現状の技術では，この判定は難しく，今後の構文解析システムの課題と考えられる．本手法では，構文解析結果から不要な節を除くために,「係助詞“は”」「代名詞」「時詞」を含む節を除外する処理のみに留めている．

用語の上位概念語の抽出（パターン B，形態素情報，並列構造）

用語を説明するパターン A で抽出した連体修飾節は,動詞の連体形で終わっているため,このままでは，文として不完全である．そこで，用語の上位概念語を抽出して，この連体修飾節と統合することにより，説明文を生成する．

提案手法では，ニュース記事の表層的特徴を基に，用語の上位概念語を抽出する．この

中では，下記の３つのパターンに分類して，抽出を行う．

パターン a)　用語の最終形態素が上位概念語となるもの
　　例）「航空安全法案」
パターン b)　用語の直前に上位概念語があるもの
　　例）通貨「ユーロ」の～
パターン c)　用語の直後に「という」「と呼ばれる」といった定型的な言い換えを意図する句があり，その後に上位概念語があるもの
　　例）「マルス」と呼ばれるコンピュータシステムは～

まず,抽出された用語を形態素解析し,用語の最終形態素が辞書に含まれている場合は,その形態素を上位概念語とする（パターン a)．2002 年 6 月に出現した用語を対象に調査した結果，この処理により，抽出された用語の 54.8%に上位概念語が与えられた．次に，用語が含まれる文を構文解析し，用語と並列関係にある直前の名詞を上位概念語とする（パターン b)．この処理により，抽出された用語の 11.3%に上位概念語が与えられた．

用語の直後に言い換えを意図する句がある場合は，先に説明した用語を説明するパターンＢに該当する（上記ではパターン c)．上記の例では,「マルス」の上位概念語は「コンピュータシステム」と特定できるが，下記の例では，その特定が困難となる．

- ～に溶けていた「ホスゲン」という人体に有毒なガスが，～

この例では，用語「ホスゲン」の上位概念語が「人体」であるか「ガス」であるかを，表層的には判断できない．そこで，用語の直前にある動詞との整合性を調べ（この例では，「人体」＋助詞＋「溶ける」の組み合わせと「ガス」＋助詞＋「溶ける」の組み合わせの過去のニュース記事における出現頻度を調べる)，その相互情報量の大きさを基準として，どちらが上位概念語になりやすいかを判定する．名詞 n と動詞 v の相互情報量 $MI(n,v)$は式(6.1)で与えられる．この処理により，抽出された用語の 9.6%に上位概念語が与えられた．

2002 年 6 月のニュース記事に出現した鍵括弧で囲まれた用語を対象として，パターン a)～パターン c)の 3 つの処理により抽出された上位概念語を検証した結果，適合率は 95.7%と良好な結果が得られた．上位概念語が得られなかった 24.3%の用語は，「もの（こと)」という一般的な単語をその上位概念とした．

抽出された上位概念語を，パターン A で抽出された連体修飾節中の動詞の連体形に繋げて説明文を生成する．

「α は β です」の解析（パターン C）

文献[68]で考察された「α とは β である」というパターンは，国語学では「真の題目」と呼ばれ，用語の説明では有効な表現となる．しかし，2001 年 6 月のニュース記事で，「α とは β である」のパターンは出現しなかった．そこで，本手法では「α とは β である」の関係を包含する表現「α は β である」に着目し，抽出する．（放送用のニュース記事では，語尾に丁寧語が用いられるため，実際には，「α は β です」の表現を抽出する．）

6.4.2 用語と説明文の関係判定処理

抽出された用語と説明文との意味関係には，様々な種類が考えられる．本手法では，6.3で扱ったQualia structure[76]を参考にして分類する．意味記述の役割を，用語の説明部分の役割に当てはめる．下記の例では，下線部の説明が，鍵括弧で囲まれた用語におけるQualia structure(A～D, 6.3.1参照)のそれぞれの役割となる．

A）ビル街や舗装された道路に囲まれた都会の気温が上昇する「ヒートアイランド現象」が，～
B）この他，細長い花弁を水牛の角のように張る「スカホセパラム」の仲間など～
C）森林の整備や緑化の推進を目的とする「緑の募金」を，～
D）ハタミ大統領が提唱する「文明間の対話」～

ニュース記事では5W1Hの情報を的確に伝えることが重要な役割とされている．そこで，5W1Hの関係や特徴となるConstitutive roleについては，さらに細分化された下記の役割を利用して分類する．

E）Time　　：時間
F）Location　：場所
G）Instrument　：道具
H）Contain　　：包含
Ⅰ）Is_a_member_of　：メンバー
J）Is_a_part_of：部分

用語の説明部分の役割が，選択した10個のQualia Structureのいずれに該当するかを推定するために，説明部分と用語の表層的な特徴を手掛かりとする，最大エントロピー法による学習法[29]を利用する．学習における素性は，説明が連体修飾節（パターン A）である場合は，連体修飾節中の用語を直接修飾する動詞の標準表記・時制・態・完了形の有無，動詞の格，その格が含まれる節の自立語表記，自立語の属性，用語の属性とする．ここで自立語と用語の属性は，IREX固有表現抽出タスク[70]により付加された8つの固有表現(組織名，人名，地名，固有物名，日付表現，時間表現，金額表現，割合表現)とし，抽出アルゴリズムは内元らの手法[84]により，あらかじめ判定した結果を利用する．図6.6に素性の例を示す．この例では，“千葉県柏市をホームタウンとする「柏レイソル」～”というニュース記事から取り出された連体修飾節「千葉県柏市をホームタウンとする」と用語「柏レイソル」の関係を推定するために，表記（千葉県柏市，ホームタウン，する），属性（地名，その他，組織名），助詞（を格，と格），動詞の特徴（現在形，能動態，完了形無し）といった特徴を利用している．説明が文の述部（パターン C）にある場合は，断定を表す助動

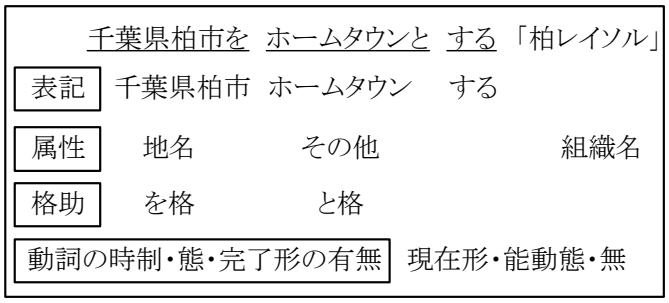


図 6.6　付加する素性例（説明が連体修飾節の場合）

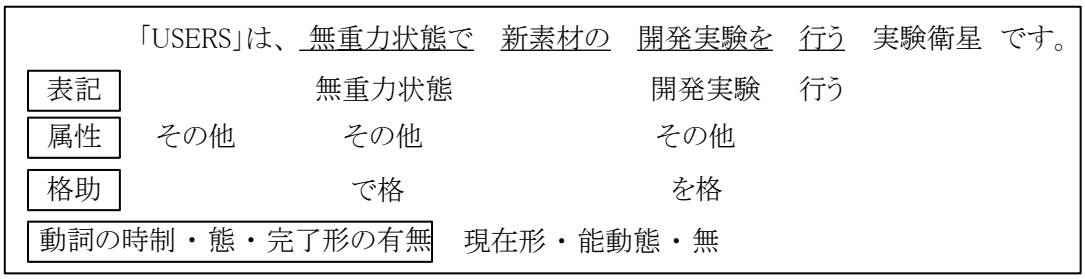


図 6.7　付加する素性例（説明が文の述部の場合）

詞「です」の直前の名詞を修飾する動詞の標準表記・時制・態・完了形の有無，動詞の格，その格が含まれる節の自立語表記，自立語の属性，用語の属性を素性とする．説明が文の述部にある場合の素性の例を図 6.7 に示す．

この素性により，用語とその説明の関係が Qualia Structure のいずれに該当するかを推定する．ニュース記事中に同じ用語が繰り返し出現する場合，同じ関係を持つ用語の説明文も複数抽出される可能性があり，どの説明文が適切か，優先順位をつける必要が生じる．そこで，同じ関係を持つ用語の説明文が複数生成された場合には，説明文の動詞に係る文節数が多く，さらに，新しいニュース記事から生成されたものほど優先順位を高くして順位付けを行う．最後に，用語と，優先順位の高い説明文を，その関係ごとに用語集データベースへ登録する．

6.4.3　説明文抽出・関係判定実験

提案手法を検証するために 2002 年 6 月のニュース記事に出現した 670 組の連体修飾節と用語を対象として，その意味関係を判定する実験を行った．この際，学習データは，2002 年 5 月以前のニュース記事から 3,712 組の連体修飾節と用語の組を無作為に抜き出し，それぞれに Qualia structure の役割分類を手作業で与えて作成した．学習データに少数しか出現しない素性はノイズとなる可能性があるため，出現頻度が 8 回以上観測された素性のみを用いた．推定処理の結果，各役割に所属する確率値が出力される．処理結果の一部を表 6.16 に示す．表中の数値は，各役割に所属する確率値を示す．

この値が一定値（実験では 50%）以上のものを，用語連体修飾節の役割と判定した．評価結果を表 6.17 に示す．

表 6.16　用語の説明文抽出，関係判定実験結果

用語	用語を修飾する連体修飾節 + 上位概念	意味関係	%
メディアパークつくば	茨城県などが出資する第三セクターで江戸時代の街並みを再現した娯楽施設を運営するもの（こと）	Agentive	97.8
虫追い	害虫の駆除と豊作を祈る伝統行事	Telic	54.1
船渡御	「御心霊」が大川を渡るもの（こと）	Constitutive	81.7
安曇野夏季美術講座	「自然と芸術」をテーマに北アルプスのふもと安曇野の自然や世界の芸術について，大学の授業を地域住民に公開する講座	Formal	75.5
エネルギー憲章に関する条約	エネルギー分野の貿易や投資の自由化などを定めた条約	Formal	99.5
早期勧奨退職	小泉総理大臣は，きょうの閣議の後の閣僚懇談会で，官僚のいわゆる天下りを減らすため，定年前に退職を勧めるもの（こと）	Telic	68.1
御芝堂減肥こう嚢	厚生労働省が商品名を公表したもの（こと）	Agentive	97.8
世界最大の恐竜博	恐竜をめぐる最新の研究成果を紹介する博覧会	Formal	84.6
石積み	大規模な堀や大きな石を積んだもの（こと）	Is_a_part_of	98.3
山あげ	イチゴを早く収穫するために夏の間だけ涼しい高原で苗を育てる作業	Formal	99.9
3連複	上位三着の馬の番号を順位に関わらず当てるもの（こと）	Formal	90.6
ほしのこえ	今年三月に発売されたもの（こと）	Time	98.6
ののじ廻し	祭りの見所の一つで円を描くように山車を廻すもの（こと）	Formal	83.8
さかな館	高知県中村市に四万十川の魚など日本と世界の淡水魚を展示したもの（こと）	Is_a_part_of	97.1
景気予測調査	全国の一万社を対象に景気判断を聞く調査	Formal	99.0
改正ハートビル法	一定規模以上の百貨店やホテルなどの建物に対し，お年寄りや障害のある人が利用しやすいようバリアフリー化を義務付ける法律	Telic	78.8
ペリーシチュー	牛肉を使ったもの（こと）	Instrument	66.4
燃料電池車	次世代の低公害車として注目されているもの（こと）	Constitutive	98.6
金魚ちょうちん	江戸時代から山口県柳井市で地元の民芸品として伝わるもの（こと）	Constitutive	72.6
全国滝サミット	滝を観光の柱にしている全国の自治体の担当者が集まるサミット	Is_a_member_of	96.6
びっぐあーす	小豆島を経由して大阪と結ぶ高速艇	Location	91.3

表 6.17　用語と説明文の関係抽出評価結果

役割	適合率	再現率
Formal	0.853	0.858
Telic	0.891	0.661
Agentive	0.792	0.811
Constitutive	0.683	0.545
Time	1	0.286
Location	0.9	0.9
Instrument	1	0.571
Is_a_member_of	0.684	0.65
Is_a_part_of	1	0.5
Contain	0.625	0.5
Total	0.814	0.76

この処理を，日々記事の追加されるニュース記事データベースを対象にして毎日行うことにより，常に最新の用語を含む用語集データベースを作成することができる．約 59 万記事(1991 年～2002 年 9 月)のニュース記事を対象とした実験では，17,662 組の用語とその説明文の組が抽出された．提案手法により，用語に対してその意味関係が明確になった説明文を自動生成することが可能となり，用語に対する知識として利用できる．

6.4.4　生成したメタデータの応用

我々は，放送局が所有するコンテンツを学校教育に活用するマルチメディア教育支援システムの研究を進めている[85]．このシステムでは，ネットワーク上の仮想教室で，遠隔地にいる複数の生徒が話し合ったり，NHK の映像データやテキストデータにアクセスしたりしながらグループ学習を行う．仮想教室には，生徒のグループ学習を支援するエージェントが参加する．このエージェントが持つ情報として作成した用語集データベースを利用するシステムを実装した．この仮想教室システムでは，参加する生徒がチャットを行うように文字をコンピュータに打ち込むことにより，生徒の化身となる CG アバターを通して話し合い，グループ学習を行う．ユーザーインターフェィス画面例を図 6.8 に示す．CG アバターが討論番組のように画面上に現れるため，発言者や対話の流れが視覚的に理解できる．

エージェントは，生徒の会話内容をモニターし，エージェントに対する呼びかけをキーとした質問文章に対して，テンプレートマッチによる解析によりキーワードを抽出して必要な情報の検索を行い応答する[86]．エージェントは意味に関する応答のための情報源として，ニュース記事から自動生成した用語集を利用する．生徒から，「エージェントさん，“用語”って何？」という質問があった場合，エージェントは用語集を検索し，用語の Formal の関係にある説明文を返答する．他にも，「どういった特徴？(Constitutive)」，「何のため？(Telic)」，「いつ？（Time）」，「どこ？(Location)」などの質問に対しても，対応する関係にある説

図 6.8　仮想教室のユーザーインターフェイス

明文を利用して返答できる．ニュース記事には，用語の説明が簡潔に記述されているため，それを利用した Q&A システムでも簡潔に返答でき，このようなシステムに適した情報と考えられる．

エージェントは，仮想教室に参加する生徒の会話状況の中で，一定時間発話が無い場合に，会話が行き詰っていると判断して，その内容に関連する質問を行い会話の活性化を図る．この際，まず，生徒の会話に最も関連が深い話題を，5 章で説明した手法[87]によりニュース記事データベースから抽出した話題群の中から選択する．ニュース記事から抽出した話題は，すべて話題ベクトルで特徴付けられている．仮想教室における生徒間の会話は，会話が途切れた直前の 5 つの名詞をベクトルの要素にした大きさ 1 の単位ベクトルにより特徴付けた．生徒の発言ごとにこのベクトルは更新される．この生徒間の会話内容を特徴付けたベクトルとの内積が最大である話題ベクトルを，生徒の会話に最も関連が深い話題と推定した．

次に，その話題を構成するニュース記事に含まれる用語を，提供する質問の候補とし，この中で，生徒の会話に最も関連が深い用語を選択して質問文を生成する．この処理では，各用語に，Formal と判定された連体修飾節に含まれる単語をベクトルの要素に持つ用語ベクトルを定義する．用語ベクトルの要素の重みは，その用語が抽出されたニュース記事が属する話題の話題ベクトルの要素の重みと同じ値とした．Formal の役割の連体修飾節は，用語を他の用語と区別する説明を与えるため，上記の用語ベクトルは用語を特徴付ける指標となる．

生徒の会話に最も関連が深い話題を構成するニュース記事から抽出された用語の中で，生徒間の会話内容を特徴付けたベクトルとの内積が最大となる用語ベクトルを持つものを，会話に関連した用語であると判断し，質問対象とした．すべての用語ベクトルとの内積が 0 の場合は，用語ベクトルのノルムが最大のものを質問対象用語とした．ここで，会話に既出の用語は，処理対象から除いている．

質問は，あらかじめ決めておいたテンプレートを利用し，質問対象用語について問いかける．例えば，「“ヒートアイランド現象”って何か知ってる？」といった質問文を生成する．

生徒から「知らないから教えて」と言われた場合，Formal と判定された説明を利用して，「ビル街や舗装された道路に囲まれた都会の気温が上昇する現象のことだよ」と応答できる．

提案した用語集システムでは，大量の用語に関する簡潔な説明が蓄積されるため，Q&A システムに適している．日々作成されるニュース記事から新しい用語を抽出してその説明文を生成することにより，用語集を毎日新しいものに自動更新できる．

6.5 因果関係知識獲得

本節では，原因結果の関係を表す因果関係を，大量の日本語テキスト TV 番組「きょうの健康」のクローズドキャプション）から抽出する手法を提案する．

6.5.1 因果関係表現の分類

テキスト中における因果関係の表現は，以下の 3 種類に分類できる．

（1）同一文の名詞ペアに因果関係がある場合
「脳卒中(結果)の原因となる動脈硬化(原因)が促進される．」
（2）同一文の節のペアに因果関係がある場合
「急に運動を始める(原因)と血圧が急上昇します(結果)．」
（3）複数文にまたがって因果関係のある名詞ペア，句のペアが出現する場合
「心臓肥大(原因）が促進される．」
「この結果，心筋梗塞(結果)が起こりやすくなる．」

ここでは，(1)と(2)について論じる． (3)は，接続詞などが手掛かりになるであろうが，難しい問題と考え，ここでは対象外とする．

6.5.2 因果関係抽出処理（同一文の名詞ペアに因果関係がある場合）

これまでに Chang ら[88]は，英文テキストを対象として，因果関係にある単語ペアと構文構造を学習する手法を提案している．この手法では，手掛かり語を特定しないでも因果関係にある単語ペアと構文構造が精度良く抽出でき，並列句中の動詞が共通の目的語を持たない場合にも対処可能であるため，汎用的な手法と考えられる．そこで，Chang らの手法をベースとし，日本語テキストを対象に因果関係を抽出する．名詞ペアと名詞ペア間の構文構造に注目して 2 つの名詞間に因果関係が有るかを判定するために，Naïve Bayes の分類器に EM アルゴリズムを組み合わせた手法[89]を利用する．本手法では，少量のクローズドキャプション中の名詞ペアに，因果関係の有無を判定したラベルを付与し，ラベル無しのクローズドキャプションのラベルを推定する．

まず, 入力されるクローズドキャプションテキストから名詞ペアと名詞ペア間の構文構造を抽出する．対象を分類語彙表[83]により因果関係を表現しやすい名詞に限定し，南瓜[90]による構文解析結果を利用して名詞ペア間がどのような構文構造に位置しているかを抽出する．そして，Preorder String Expression[91]（以後，PSE と呼ぶ）により構文構造の表現とマッチング処理を行なう．ここでは，南瓜により分割された句を自立語と機能語の２つに分割して PSE の構造に利用した．以下に，PSE の例を示す.

[入力 1]　名詞 1 が起きると名詞 2 につながります.
[PSE1]　{“つながる”, “と”, “起きる”, “が”, “名詞 1”, 0,0,0,0, “に”, “名詞 2”,0,0,0}
[入力 2]　　名詞 1 が名詞 2 につながる
[PSE2]　{“つながる”, “が”, “名詞 1”, 0,0, “に”, “名詞 2”,0,0,0}

PSE の表現では, 語順と要素”0”により元の構文構造を復元することができる. 2 つの PSE p_1, p_2 に出現する名詞ペア間の構文構造の類似性 $sim(p_1, p_2)$を以下の式(6.9)により評価する.

$$sim(p_1, p_2) = \frac{com(p_1, p_2) \times 2}{wc(p_1) + wc(p_2)} \tag{6.9}$$

ここで，$wc(p_i)$は PSE p_iに出現する名詞 1，名詞 2，0 以外の単語数，$com(p_1, p_2)$はそのうちの PSE の構造を考慮した共通単語数を示す. 例えば上記の例では, 単語“つながる”, “が”, “に”が共通単語になるため，sim(PSE1, PSE2) = 6/8 = 0.75 となる．この類似性評価の値は EM アルゴリズムで利用する.

テキストから抽出した 2 つの名詞は，分類語彙表上での属性が一意に決まる場合はその 5 桁目までの数値を，複数の属性を持つ場合は表記そのものを利用する．2 つの名詞と，その間の構文構造と合わせて 3 項組として扱う．例えば，6.5.1(1)の例は以下のように表現される.

「脳卒中(結果)の原因となる動脈硬化(原因)が促進される．」

↓

<15721, 15721, {名詞 2, “なる”, “と”, “原因”, “の”,名詞 1}>

ここで，”15721”は分類語彙表上での属性番号の上位 5 桁を示す．この 3 項組が因果関係を表すか否かを評価する.

抽出された 3 項組 t_i が因果関係を持つ(c_1)，もしくは持たない(c_0)確率は，以下の式(6.10)で与えられる.

$$P(c_j \mid t_i) = \frac{P(c_j)P(t_i \mid c_j)}{P(t_i)} \qquad i = 1,\ldots,|T|; j = 0,1 \tag{6.10}$$

ここで，|T|は 3 項組 t_i の総数を示す．この値が大きいクラス c_j(c_0 または c_1)を，因果関係の有無の判定結果とする．$P(t_i \mid c_j)$は，以下の式(6.11)とする．

$$P(t_i \mid c_j) = P(CP_{t_i} \mid c_j)P(SP_{t_i} \mid c_j) \tag{6.11}$$

ここで，CP_{ti} は 3 項組 t_i に含まれる２つの名詞間の構文構造を指し，SP_{ti} は 3 項組 t_i に含まれる名詞ペアを指す．この式を利用して，EM アルゴリズムにより $P(c_j \mid t_i)$を推定する．

EM アルゴリズムは，内部状態が不明な不完全データに対して尤度が最大になるような繰り返し学習を行ない，内部状態を推定する手法であり，この場合は教師無しデータが不完全データとなる．まず，すべてのクローズドキャプション集合を対象として，あるクラス c_j のもとで素性となる CP_{ti}，SP_{ti} が発生する確率 $P(CP_{ti} \mid c_j)$，$P(SP_{ti} \mid c_j)$を以下の式(6.12)，式(6.13)により求める(M ステップ).

$$P(CP_{t_i} \mid c_j) = \frac{1 + \sum_{k=1}^{|T|} sim'(CP_{t_i}, CP_{t_k})P(c_j \mid t_k)}{|CP| + \sum_{m=1}^{|CP|}\sum_{k=1}^{|T|} sim'(CP_{t_m}, CP_{t_k})P(c_j \mid t_k)} \tag{6.12}$$

$$P(SP_{t_i} \mid c_j) = \frac{1 + \sum_{k=1}^{|T|} N(SP_{t_i}, t_k)P(c_j \mid t_k)}{|SP| + \sum_{m=1}^{|SP|}\sum_{k=1}^{|T|} N(SP_{t_m}, t_k)P(c_j \mid t_k)} \tag{6.13}$$

ここで，|*CP*|，|*SP*|，|*T*|は，名詞間の構文構造の総数，名詞ペアの総数，3 項組の総数を表し，$N(SP,t_k)$は 3 項組 t_k に名詞ペアが含まれるか否かを表す関数であり，含まれるときだけ 1 の値を取る．$sim'(CP_{ti}, CP_{tk})$は名詞ペア間の構文構造の類似性で，0.5 より大きい場合に $sim(CP_{ti}, CP_{tk})$を，それ以外は 0 を与える．$P(c_j \mid t_k)$の初期値は，因果関係の有無を判定した少量のクローズドキャプション（教師有り訓練データ）を利用して計算する．

次に，Naïve Bayes の式を利用して，$P(c_j \mid t_i)$の期待値を式(6.14)，式(6.15)により計算する（E ステップ）．

$$P(c_j \mid t_i) = \frac{P(c_j)P(CP_{t_i} \mid c_j)P(SP_{t_i} \mid c_j)}{\sum_r P(c_r)P(CP_{t_i} \mid c_r)P(SP_{t_i} \mid c_r)} \tag{6.14}$$

$$P(c_j) = \frac{1 + \sum_{k=1}^{|T|} P(c_j \mid t_k)}{|c| + |T|} \tag{6.15}$$

$|c|$は分類すべきクラスの数を指し，ここでは 2 となる．M ステップと E ステップを繰り返すことにより，クローズドキャプションに出現する 3 項組が因果関係を持つか否かを $P(c_j \mid t_i)$の値により推定できる．さらには，$P(CP_{ti} \mid c_j)$，$P(c_j)$，$P(CP_{ti})$からベイズの定理により $P(c_j \mid CP_{ti})$が計算可能で，因果関係を持つ時の特徴的な構文構造の判定が可能となる．

因果関係抽出処理実験

手法の検証のために，循環器系の話題を取り上げている「きょうの健康」16 番組を対象とし，番組で使われたクローズドキャプション 2180 文から 3 項組 1495 個を抽出して因果関係抽出実験を行なった．無作為に 1 番組を選び，そこから抽出した 3 項組 149 個に対して人手により因果関係の有無をタグ付けして教師有り訓練データとし，残りの 15 番組を教師無しデータとした．繰り返し回数を $P(c_j)$の収束度合を基準として判定したとき（実験では 100 回），因果関係を持つと判定された 3 項組を生成する原文（一部）を表 6.18 に示す．

表 6.18 抽出された因果関係（一部）

$P(c_1 \mid t_i)$	3 項組を生成する原文（括弧内が対象名詞）
0.980	[コレステロール]が高いほど[心筋梗塞]の危険も高くなる
0.978	[コレステロール]が高いと[冠動脈疾患]が起こりやすい
0.973	[動脈硬化]が進んで[虚血性心疾患]を起こす
0.962	[中性脂肪]が高いのは[動脈硬化]の危険信号，と考える
0.955	[脳卒中]の原因となる[動脈硬化]が･･･
0.931	[動脈硬化]を起こすリスク例えば[高血圧]を持つ

表 6.18 の名詞ペアとその間の構文構造は，いずれも訓練データには存在せず，EM アルゴリズムにより獲得できたものである．実験で利用した教師無しデータ中の 1 番組を取り出し適合率を評価した結果，因果関係がある名詞ペアは 73.8%(31/42)，無い名詞ペアは 61.9%(75/121)であった．

6.5.3 因果関係抽出処理（同一文の節ペアに因果関係がある場合）

節間の関係を推定するために，まず，クローズドキャプションから処理対象となる節のペアを抽出する．次に，抽出した節がどのような意味カテゴリーに属すかを手作業で与えたルールにより分類する．適切な意味カテゴリー分類ができれば，節のペアが属する意味カテゴリーの組み合わせにより節間の関係が推定できると仮説を立て，フィッシャーの正確確率検定[92]を行う．検定の結果，節のペアが属する意味カテゴリーの組み合わせが顕著に現れる関係が判明する．この節のペアが属する意味カテゴリーと関係を利用することにより，テストデータから節の関係を推定する．

節ペアの抽出

複文において，節に含まれる述語を表す文節が，別の節に含まれる述語を表す文節を修飾する場合，この 2 つの節は「原因－結果」などの関係を持つ可能性がある．そこで，係り受け関係にある節のペアを抽出する．この時，以下に示す 2 つの条件を満たす節ペアに処理対象を制限した．

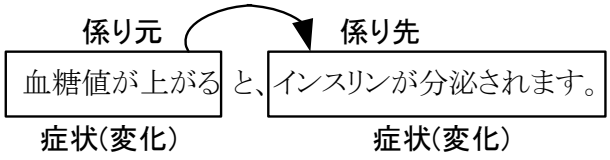
・係り元の述語中の動詞が連用形である，または接続助詞「て」，「と」，「ば」を付属語として伴う

・節ペアのいずれかに健康関連に特有な単語が含まれる

健康関連に特有な単語は，あらかじめ TFIDF 値を計算し，この値の上位を利用した．この処理により，節のペアが大量に抽出される．

節の分類

名詞ペアの関係を判定する処理では，名詞ペアの共通係り先までの単語列が重要な情報となるが，節ペアの場合，直接係り受け関係にあるため節の周辺情報は利用できない．そこで，節に含まれる情報を解析する必要がある．節が意味する事柄によって，その関係が特定できる場合がある．例えば以下の例では，係り元の節と係り先の節がともに「病状の変化」を表す．このようなケースでは係り元の節が係り先の節の原因を表すことが多い．



そこで，処理対象として抽出した節を以下に示す 8 種類の意味カテゴリーに分類する．

【節の意味カテゴリー】

- 症状（状態）　例）血圧が 高い
- 症状（変化）　例）細胞が 障害を 受ける
- 病気（状態）　例）糖尿病が 続く
- 病気（変化）　例）肺炎に なる
- 行為（医療）　例）インスリンを 注射する
- 行為（体の動作）例）ひざを 伸ばす
- 行為（管理）　例）血圧を コントロールする
- その他　例）糖尿病を 含める

分類のために，節に含まれる動詞とその格構造の組み合わせを基とするルールを手作業により作成した．作成したルールの一部を以下に示す．

【ルール例】

[病気]の ＋[*]が ＋　起きる → 病気（変化）

[内臓 or 分泌物]が ＋ 不足する → 症状（状態）

[内臓 or 分泌物]を ＋ 取る → 行為（医療）

[症状]を ＋ 保つ → 行為（管理）

このルールにおいて，[病気]は，「病気」のカテゴリーに属する名詞を示す．このカテゴリーは，あらかじめ「病気」「症状」「行為」「人体属性」「内臓 or 分泌物」「体の部位」「医療品」などに属する名詞を人手により登録したものを利用する．[*]は任意の単語との一致を許す．

節間の関係の分類

番組のクローズドキャプションには，出現する節の間に様々な関係が存在する．提案手法では健康に関する番組について，以下の 4 つの関係とその他を分類対象とする．

- 原因（係り先の節の原因が係り元の節）
  例）インスリンの分泌を増やして，血糖を下げます
- 症状（係り元の節の症状が係り先の節）
  例）糖尿病になると腸管からコレステロールの吸収が増え，…
- 目的（係り元の節の目的が係り先の節）
  例）生活習慣を変えて内臓脂肪や肥満を取る
- 対処法（係り元の節の対処法が係り先の節）
  例）腎不全になり最終的には透析を行う…
- その他（上記 4 つの関係以外）
  例）眼科へ来て，初めて糖尿病が分かるケースが…

節間の関係に特徴的な節ペアが属する意味カテゴリーの組み合わせを判定するために，フィッシャーの正確確率検定を用いる．フィッシャーの正確確率検定は，2 つの変数の間に統計学的に有意な差があるかを判定する検定手法で，近似せずに全ての可能な事象について列挙し，直接有意確率を計算する．節間の関係 x と節の意味カテゴリーの組合せ A との関係を考える場合，表 6.19 に示すような 2×2 分割表を作成する．
この事例が出現する確率 p は以下の式(6.16)で与えられる．

$$p = \frac{{}_{a+b}C_a \times {}_{c+d}C_c}{{}_{a+b+c+d}C_{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{(a+b+c+d)!a!b!c!d!} \tag{6.16}$$

表 6.19　2x2 分割表

	関係 x	関係 x 以外	計
意味カテゴリーA	a	b	$a+b$
意味カテゴリーA 以外	c	d	$c+d$
	$a+c$	$b+d$	$a+b+c+d$

節間の関係 x と節の意味カテゴリーの組み合わせ A に有意な差があるかを片側検定により判定する場合，式(6.17)により頻度 a 以上の確率値の和を求める．

$$\text{有意確率} = \sum_{\alpha \geq a} p(\text{意味カテゴリー}A\text{と関係}x\text{の共起頻度} = \alpha) \tag{6.17}$$

この有意確率が一定値以下の場合，節の意味カテゴリーの組み合わせ A は関係 x を持つ場合と判定できる．

節間の関係推定実験

NHK で放送された「きょうの健康」の糖尿病に関連する 80 番組を処理対象とし，抽出した節の分類実験と，節間の関係推定実験を行った．処理対象のクローズドキャプションを解析して節ペアを抽出し，節の属する意味カテゴリーと，節間の関係の正解を人手により与えた．このうち半分を学習用データ，残りの半分をテスト用データとし，学習用データのみを参照して人手により節を分類するためのルールを作成した．このルールを用いて，テスト用データにある節を 8 つの意味カテゴリーに分類し，その他以外をまとめて評価した結果を表 6.20 に示す．

表 6.20 ルールによる節の意味カテゴリー分類結果

適合率	再現率
98.0%	72.7%
(343/350)	(343/472)

表 6.21 節間の関係に特徴的な節ペアが属する意味カテゴリーの組み合わせ
(有意確率の値が小さい 5 項目)

係り元	係り先	節間の関係	有意確率
症状(変化)	症状(変化)	原因	2.1e-12
病気(変化)	症状(変化)	症状	4.1e-9
行為(医療)	行為(医療)	目的	3.1e-7
行為(体の動作)	症状(変化)	原因	4.8e-4
病気(状態)	行為(管理)	対処法	9.3e-4

節の意味カテゴリー分類結果を利用して，学習用データから節間の関係に特徴的な節ペアが属する意味カテゴリーの組み合わせを抽出した．5%の有意水準による検証を行い，節の意味カテゴリー（係り元と係り先の組み合わせ）と節間の関係 11 組を抽出した．有意確率の値の小さい 5 項目を表 6.21 に示す．

抽出した節の意味カテゴリーと節間の関係を利用し，テスト用データから節間の関係を推定した．表 6.22 に，テキストから抽出した節ペアと，提案手法により推定された節間の関係を示す．また，原因，症状，目的，対処法の 4 つの関係をまとめた評価結果を表 6.23 に示す．節の意味カテゴリー分類結果の再現率が低いため，節間の関係推定実験の再現率は低いが，適合率は 80%を超え，一定の弁別能力があると判断できる．

獲得した因果関係知識は，6.4.4 で言及したマルチメディア教育支援システムなどの教育を目的とするアプリケーションで有用と考えられる．さらに，6.4，6.5 で獲得した知識には，対応する番組映像も存在するため，他のマルチメディアコンテンツの素材としても大

表 6.22 節間の関係推定実験結果

節間の関係	係り元の節	係り先の節
原因	コレステロールが 溜る[症状(変化)]	血管が 狭窄する[症状(変化)]
原因	網膜が 障害される[症状(変化)]	視力が 落ちる[症状(変化)]
原因	ブドウ糖を 飲む[行為(体の動作)]	血糖値が 上がる[症状(変化)]
原因	アルコールを 飲む[行為(体の動作)]	すい 炎に なる[症状(変化)]
原因	食事を 摂る[行為(体の動作)]	血糖値が 上昇する[症状(変化)]
症状	腎症が 進む[病気(変化)]	高血圧が 助長される[症状(変化)]
症状	糖尿病に なる[病気(変化)]	腸管から コレステロールの 吸収が増える[症状(変化)]
症状	腎症神経症を 起こす[病気(変化)]	細い 血管が 変化する[症状(変化)]
目的	簡単に 指先で 血を 採る[行為(医療)]	血糖を 測る[行為(医療)]
目的	血糖値を 下げる[行為(医療)]	インスリンを 注射する[行為(医療)]
対処法	合併症が ある[病気(状態)]	栄養のバランスに気をつける[行為(管理)]
対処法	糖尿病に ならない[病気(状態)]	食事を 考える[行為(管理)]

表 6.23 節間の関係推定実験評価結果

適合率	再現率
81.0% (51/63)	31.3% (51/163)

きな可能性をもつものと考えられる． 3.4 で紹介したマルチメディア百科事典は，メタデータを利用した映像活用の一つのアプリケーションであり，6.4，6.5 で獲得した知識をマルチメディア百科事典へ応用することにより，検索した映像を見せるだけでなく，より高度な調べ学習が可能な百科事典へと拡張可能と考えられる．獲得した知識とメタデータを利用することにより，放送番組を利用した，より効果的なアプリケーションの構築が期待される．

6.6 おわりに

本章では，大量のテキストデータから知識として利用できる有益な情報を自動獲得するための技術について説明した．テキストから抽出可能な基本的な知識として，単語の上位概念や関連単語などの辞書的な語彙的知識が挙げられる．6.2 で述べた未知語処理では，辞書に登録されていない単語が属する上位概念を推定するために，どのような特徴が有益であるかを示した．調査の結果，形態素レベルと句レベルの情報が未知語の意味推定に有効であり，文レベルの情報は動詞の多義性の問題が残されているため，それほど有効では無いことが分かった．6.3 では，辞書に登録されている単語に対して，その単語が持つ典型的な機能や目的を示す telic role と，単語の起源や引き起こす事象を示す agentive role の役割を果たす動詞を自動抽出する手法を提案した．この結果は，単語の意味を拡張して用いるような場合の換喩表現の解析時に有用な情報となる．6.2, 6.3 において獲得された知識は，メタデータ生成処理における解析精度の向上へと応用可能と考えられる．

6.4 では， ニューステキストから難しい用語の説明を抽出し，用語と説明の意味関係を特定する手法を提案した．さらに 6.5 では，健康に関するテキストデータから因果関係を抽出する手法を提案した．この用語に関する知識は，マルチメディア教育支援システムやマルチメディア百科事典などの教育用途のアプリケーションなどで有用となり，番組に付与されたメタデータとの共用により，放送番組の効果的な二次利用を可能にする知識と成り得る．

テキストから知識を獲得する技術は未成熟な分野ではあるが，今後，良質かつ膨大な量のコーパスのさらなる出現などにより，良質な知識を大量に自動獲得することが可能になると考えられる．近い将来，人間が持つ常識などの知識を計算機が持つことにより，計算機が深い言語解析を行い，テキストを理解した上で情報抽出などの処理を行うような環境が整備されることを期待したい．

7 章　結論

本論文では，放送番組で利用されるクローズドキャプションを解析することにより，情報系番組，生放送スポーツ中継番組，ニュース番組に対して，その映像区間，音声区間の内容を詳細に説明するメタデータを自動付与する手法を提案した．さらに，クローズドキャプションや，大規模コーパスなどの世間に流通するテキストデータから，語彙知識を自動獲得する手法について論じた．本論文の要旨をまとめると，以下のようになる．

情報系番組へのメタデータ生成　情報系番組には，特定の事柄を表現するために同じような言い回しが多用される.特定の事柄を表現する定型的な表現区間を抽出することにより，メタデータを生成することができる．この定型的な表現区間を抽出するためには，文章に含まれる大域的な情報まで加味した特徴量が効果的である．紀行番組に対して「場所紹介区間」を表現する定型文章区間を抽出する実験により，複数文にまたがる文節の特徴を統計的に解析する提案手法の有効性を確認した．さらに，大域的な情報まで加味した特徴量をサンプリングすることにより，処理時間の効率化を図る手法も提案し，その有効性を確認した．定型的な表現区間ごとに，人物紹介や場所の紹介などのメタデータを付与することができる．(3 章)

生放送スポーツ番組へのメタデータ生成　生放送スポーツ番組におけるアナウンスコメントには，試合の流れや発生したイベントについて実況するコメント（試合記述文）と，試合の流れとは直接関係しないコメント（解説文）が存在する．統計処理により，アナウンスコメントを試合記述文と解説文に分類することにより，スポーツ番組からイベント発生区間を効果的に抽出することが可能となり，各イベント発生区間における発生イベントはアナウンスコメントから推定できる．サッカー中継番組において，アナウンスコメントを分類し，イベントとイベントの動作主を抽出する実験を行うことにより，その有効性を確認した．スポーツ番組では，このイベントとイベントの動作主が，イベント発生区間に対するメタデータとして利用できる．(4 章)

ニュース番組へのメタデータ生成　ニュース記事において，同じ分野に属するニュースの話題には，各話題に共通する出来事があり，その出来事は話題を特徴付ける重要なニュースである．係り受け関係を持つ 2 つの文節の自立語と係り元の文節の付属語となる助詞の 3 項組の定型性を評価して各話題に共通する出来事を抽出することにより，ニュースの話題の要約を実現できることを，実験により確認した．ニュースの話題とその要約結果を，ニュースに対するメタデータとして利用できる．(5 章)

未知語処理　テキストに出現する辞書に登録されていない単語の上位語を推定する処理では，形態素レベルの情報，句レベルの情報が，有効となる．辞書に登録されていないよう

な難しい単語を使用する場合，人が容易に理解できるように，並列構造などにより意味を補うことや，未知語が be 動詞文の主体または対象の位置にあることが多い．このような構造情報が，未知語の上位語推定において大きな手掛かりとなる．未知語の上位語を推定することができれば，メタデータ付与のための処理をより高精度に実現できる．（6.2）

単語の語彙体系知識獲得　単語の意味を表現するために，単語の目的を表す telic role と起源を表す agentive role がある．名詞に対して，この 2 つの関係に対応する動詞を，大規模なコーパスを解析して，名詞と動詞の文中での構文的な位置を特徴とした統計処理を施すことにより，ランク付けして獲得することができる．このような知識はテキスト要約によるメタデータ生成処理へ応用できる．（6.3）

用語の説明文獲得　ニュース記事に難しい用語が出現する場合，視聴者が容易に理解できるよう，連体修飾節により用語の説明を伴うことが多い．連体修飾節の言葉の特徴を統計的に解析することにより，用語と説明の関係を qualia structure で定義された関係にマッピングできることを確認した．（6.4）

因果関係抽出　放送される番組に付随するテキストには，多くの知識が含まれている．単語間の構造と，単語の意味カテゴリーを特定することにより，因果関係にある少量の単語ペアから，因果関係にある大量の単語ペアを抽出することができる．また，単語間だけでなく，節間の関係も，節の意味カテゴリーを特定することにより，推定することができる．（6.5）

メタデータを利用したアプリケーション　「場所紹介区間」のメタデータを利用して，ユーザが興味のある場所について映像とともに調べることができるマルチメディア百科事典を試作した．サッカーのイベント発生区間と，各区間で起きたイベントとイベント動作主からなるメタデータを利用して，CG キャスターによる解説付きの要約番組を自動生成するアプリケーションを生成した．用語と説明文の関係判定処理の結果を利用して，放送局が所有するコンテンツを学校教育に活用するマルチメディア教育支援システムを構築した．各アプリケーションを通して，メタデータの重要性を再確認した．

以上に示す通り，本研究では，放送される番組に対するメタデータ生成をおもな課題として，番組の種別に応じた適切なテキスト解析手法を考察し，あわせて，メタデータを利用したアプリケーションを開発した．また，大量のテキストから語彙に関する知識を獲得する手法についても考察した．

現在，放送はテレビ受信機だけでなく，パソコンや携帯型受信機でも視聴できるようになり，テレビ番組が持つ役割も大きく変化してきている．放送局でも，「いつでも」，「どこ

でも」テレビ番組を，より楽しく見てもらうことを目標に掲げている．放送番組をテレビ受信機で視聴する形態と，パソコンや携帯型受信機で視聴する形態は，大きく異なる．例えば，携帯型受信機で視聴する場合，画面の解像度が低く，さらには音声も聞き取りにくい環境で視聴している可能性が高い．そのような場合に，テキスト情報で送られるメタデータが大いに役に立つ．また，パソコンから視聴する場合は，放送されている番組ではなく，オンデマンドのサービスにより視聴することも考えられる．オンデマンドによる視聴では，時間的な制約から解放されるため，関連する番組も同時に紹介するなど様々なアプリケーションが考えられる. また, 強力な CPU やメモリを持つパソコンで視聴する場合は，テレビ受信機では処理ができなかった高度な解析も可能となる．テレビ受信機自体も，大容量のハードディスクを持ったものも出現している．今後，放送局では，様々な視聴形態を前提とした新たな放送サービスを始めていかねばならない. 新たなサービスのためには，放送コンテンツの詳細を記したメタデータが重要な役割を果たす．いかに効率的にメタデータを付与し，効果的なアプリケーションを構築していくかが，今後の大きな課題となっている．

また，NHK では，2008 年 12 月からアーカイブスオンデマンドという，放送した番組コンテンツをインターネット経由で配信するサービスが開始された．このようなサービスを多くの視聴者の方々に利用してもらうためにも，大量の番組から所望する番組だけを効果的に検索することを実現するメタデータが重要な鍵となってくる．

本論文の成果が，効率的なメタデータ付与と効果的なアプリケーション構築の一助となれば幸いである．

謝辞

本論文を作成する過程では，大変多くの方々からご指導，ご鞭撻を賜りました．

名古屋大学大学院情報科学研究科メディア科学専攻の大西昇教授には，私が名古屋大学大学院工学研究科に在学時から多大なご指導を頂き，さらには，本論文をまとめるにあたり，終始有意義なご指導，ご鞭撻を頂きました．また，主査として，適切なご助言をして頂いたのみならず，お忙しい中，論文内容に関して他の先生方との議論の機会を設定して頂くなど，多大なご支援を賜りました．心より感謝いたします．

副査であります名古屋大学大学院情報科学研究科メディア科学専攻の村瀬洋教授，名古屋大学大学院情報科学研究科メディア科学専攻の長尾確教授には，専門家として，適切かつ有益なご助言をいただきました．これらのご支援は本論文をまとめるにあたり，特に有益なものとなりました．厚く御礼申し上げます．

名古屋大学名誉教授の杉江昇先生には，名古屋大学大学院工学研究科在学時の担当教授として，多大なご指導，ご鞭撻を賜りました．この時にご教示いただいたことが，現在の研究者としての礎となっております．深く感謝いたします．

愛知県立大学の山村毅准教授，名城大学の佐川雄二教授には，私が自然言語処理の研究に携わるきっかけを与えて頂き，多くのご指導を頂きました．名古屋大学の工藤博章准教授には，名古屋大学大学院工学研究科における同期の卒業生として，良きアドバイスを頂きました．深く感謝いたします．

NHK 放送技術研究所の歴代所長ならびに久保田啓一所長には，本研究の遂行する機会と配慮を頂きました．榎並和雅元放送技術研究所所長（現，NICT けいはんな研究所所長），八木伸行部長には，私が NHK 放送技術研究所に入ったときからの上司として，番組コンテンツ解析の研究の動機を与えて頂き，研究の方向性に対する熱心なご指導を頂きました．柴田正啓部長には，直属の研究指導者として，親身になってご指導，ご助言を頂きました．金淵培氏（現，三星綜合技術院），浦谷則好氏（現，東京工芸大学教授），江原暉将氏（現，山梨英和大学教授）には，自然言語処理に関する様々なアイデアを頂きました．藤井真人氏，田中英輝氏，加藤直人氏，住吉英樹氏，佐野雅規氏，後藤淳氏，小早川健氏，宮崎勝氏，後藤功雄氏，三浦菊佳氏，河合吉彦氏ほか，NHK 放送技術研究所の皆様からは，具体的な研究内容について有益な議論を繰り返し，多大なご助言をいただきました．心から感謝いたします．

有限会社インデキシングの小杉広和氏には，本論文で使用したプログラムを作成して頂き，多くのアドバイスも頂きました．心から感謝いたします．

現在，所属している独立行政法人情報通信研究機構の皆様には，本論文をまとめる機会を与えて頂き，さらには，研究に対する熱心な議論をいただきました．感謝いたします．

メルボルン大学の Timothy Baldwin 氏には，私がスタンフォード大学の滞在研究員として在学していたときから，親身にご指導頂き，多くの知見をご教示いただきました．深謝いたします．

最後に，本論文の作成にあたり，背後から支えてくれた妻・理恵，長男・健太，次男・裕次郎，長女・菜摘に深く感謝いたします．

参考文献

[1] 長坂晃朗, 田中譲, “カラービデオ映像における自動索引付け法と物体探索法,” 情処論, Vol.33, No.4, pp543-550(1992)

[2] 河合吉彦, 住吉英樹, 八木伸行, “可変ブロックサイズのブロックマッチングに基づくカット点検出, ” 電子情報通信学会総合大会講演論文集**,** pp.212(2007)

[3] 望月貴裕, 蓼沼眞, 八木伸行, “フラクタル特徴の変化に基づくカット点検出”, 平 17 信学総全大, D-11-134, p.134, 2005.

[4] A. Merlino, D. Morey, and M. Maybur, “Broadcast news navigation using story segmentation, ” In *Proceedings of ACM Multimedia*, pp. 381-391(1997)

[5] P. Viola and M. Jones, “Fast and Robust Classification using Asymmetric AdaBoost and a Detector Cascade,” *Advances in Neural Information Processing System* 14, MIT Press, Cambridge, MA, 2002.

[6] A. Matsui, S. Clippingdale, F. Uzawa, and T. Matsumoto, “Bayesian Face Recognition using a Markov Chain Monte Carlo Method,” *Proc. of the 17th International Conference on Pattern Recognition* (*ICPR 2004*), pp.918-921, 2004.

[7] 皆川信司, 川嶋稔夫, 青木由直, “サッカー中継のシーン解析,” 信学’92 春大, D-531, pp.267-273, Mar. 1992.

[8] 大野義典, 三浦純, 白井良明, “サッカーゲームにおける選手とボールの追跡,” 情処学 CVIM 研報, CVIM114-7, pp.49-56, Jan. 1999.

[9] 丸尾二郎, 岩井儀雄, 谷内田正彦, 越後富夫, 飯作俊一, “サッカー映像からの特定映像イベントの抽出,” 信学技報, PRMU99-41, pp.31-38, July 1999.

[10] 望月貴裕, 藤井真人, 八木伸行, 篠田浩一, “投球の次ショットに重きを置いたシーンのパターン化と離散隠れマルコフモデルを用いた野球放送映像の自動イベント分類,” 映像情報メディア学会誌, Vol.61, No.8, pp.1139-1149(2007)

[11] http://www-nlpir.nist.gov/projects/trecvid/

[12] M. Sano, H. Sumiyoshi, N. Yagi, “Generating Metadata from Acoustic and Speech Data in Live Broadcasting,” *Proc. ICASSP-2005*, II-1148, 2005.

[13] 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制作システム,” 信学論(D-II), Vol.J84-D-II, No.6, pp.877-887, Jun 2001.

[14] 佐藤庄衛, 尾上和穂, 小林彰夫, 今井亨, “スポーツ番組用メタデータ制作のための音声認識,” 2004 音響学秋季予稿集, pp.127-128, Sep. 2004.

[15] 佐藤庄衛, 小林彰夫, 尾上和穂, 山田一郎, 佐野雅規, 今井亨, “メタデータ生成のための音声認識の改善, ” 映情学年大, 9-1, 2005.

[16] Y. Chang, W. Zeng, I. Kamel, and R. Alonso, “Integrated Image and Speech Analysis for Content-based Video Indexing,” *Proc. IEEE ICMCS'96*, pp.306-313, June 1996.

[17] M. Lazarescu, S. Venkatesh, G. West, and T.Caelli, “On the Automated Interpretation and Indexing of American Football,” *Proc. IEEE ICMCS'99*, pp.802-806, June 1999.

[18] 新田直子, 馬場口登, “放送型スポーツ映像の意味内容獲得のためのストーリー分割法”, 信学論(D-II), Vol.J86-D-II, No.8, pp.1222-1233, Aug. 2003.

[19] S. Satoh, Y. Nakamura, and T. Kanade, “Name-it:Naming and detecting faces in video by the integration of image and natural language processing,” *Proceedings of15th IJCAI Conference*, Nagoya, Aug. 1997, pp.1488–1493.

[20] 柴田知秀, 黒橋禎夫, “言語情報と映像情報を統合した隠れマルコフモデルに基づくトピック推定,” 情処論, Vol.48, No.6, pp.2129-2139, June 2007.

[21] 奥岡知樹，高橋友和，出口大輔，井手一郎，村瀬洋, "Wikipedia を利用したニュース映像アーカイブへの自動索引付け," 第 5 回デジタルコンテンツシンポジウム, 1-3(2009)

[22] 河合吉彦, 山田一郎, 住吉英樹, 八木伸行, EPG テキストとクローズドキャプション情報を利用した番組スポット候補映像区間の抽出手法,信学技報，Vol.106，No.100, PRMU2006-53, pp.25-30(2006)

[23] 佐野雅規, 山田一郎, 有安香子, 住吉英樹, 柴田正啓, 八木伸行, “メタデータエディタの試作ーメタデータ制作活用のプラットフォームの提案ー,” 信学技報, PRMU04-146, pp.71-76, Apr. 2004.

[24] 山本大介, 増田智樹, 大平茂輝, 長尾確, “映像を話題としたコミュニティ活動支援に基づくアノテーションシステム, “ 情処学論 Vol.48, No.12, pp.3624-3636, Dec. 2007.

[25] K. Nagao, Y. Shirai, and K. Squire, “Semantic annotation and transcoding: Making Web content more accessible,” IEEE MultiMedia Special Issue on Web Engineering, Vol. 8, No. 2, pp. 69-81(2001)

[26] Y. Freund and R.E. Schapire, “A decision theoretic generalization of on-line learning and an application to boosting, “ *Journal of Computer and System Sciences*, Vol.55, No.1, pp.119-139, 1996.

[27] Nakada, Y., Mouri, Y., Hongo, Y. and Matsumoto, T. “Gibbsboost: a Boosting Algorithm using a Sequential Monte Carlo Approach,” *IEEE International Workshop on Machine Learning for Signal Processing*, pp.259-264, 2006.

[28] N. Cristianini, “An Introduction to Support Vector Machines and other kernel-based learning method,” *Cambridge University Pres*s, 2000.

[29] A. L. Berger, P. F. Brown, and V. Della Pietra., A maximum entropy approach to natural language processing., *Computational Linguistics*, 22(1), pp.39-71, 1996.

[30] 北研二, “確率的言語モデル,” 東京大学出版, pp.41-46, 1999.

[31] R. E. Schapire , “ The strength of weak leanability,” *Machine Learning*, 5, pp.197-227, 1990.

[32] H. Szu and R. Hartley, “Fast simulated annealing.” *Physics Letters A*, 122, pp.157-162, 1987.

[33] A. Doucet, N. DeFreitas and N. Gordon, “*Sequential Monte Carlo Methods in Practice*.” Springer, 2001.

[34] S. Della Pietra, V. Della Pietra and Lafferty, J., “Inducing features of random fields,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4), pp.380-393, 1997.

[35] G. Salton, “The Vector Space Model, Automatic Text Processing,” *Addison-Wesley Publishing*, pp.312-325, 1989.

[36] M. Collins and N. Duffy, “Convolution Kernels for Natural Language,” *In Proceedings of NIPS2001*, 2001.

[37] R.E. Schapire and Y. Singer, “BoosTexter: A boosting -based system for text categorization,” *Machine Learning*, 39(2/3), pp.135-168, 2000.

[38] 工藤拓, 松本裕治, “半構造化テキストの分類のためのブースティングアルゴリズム,” 情処学論, Vol.45, No.9, pp.2146-2156, Sept. 2004.

[39] G. Salton and M.J.McGill, “Introduction to Modern Information Retrieval,” *McGraw-Hill Book Company*, 1983.

[40] D. Lewis and M.Ringuette, “A comparison of two learning algorithms for text categorization,” In *Third Annual Symposium on Document Analysis and IR*, 1994.

[41] 山田一郎, “マルチメディア百科事典～膨大な映像資産の有効活用に向けて～,” NHK 技研 R&D, July 2006.

[42] Kikuka Miura, Ichiro Yamada, Hideki Sumiyoshi, Nobuyuki Yagi, “Automatic Generation of a Multimedia Encyclopedia from TV Programs by Using Closed Captions and Detecting Principal Video Objects.” In *Proc. of Eighth IEEE International Symposium on Multimedia*, pp.873-880, 2006.

[43] Kikuka Miura, Ichiro Yamada, Hideki Sumiyoshi, Nobuyuki Yagi, “Identification of names and actions of principal objects in TV program segments using closed captions,” *International Journal of Semantic Computing*, 2008.

[44] 宮森恒, 越後富夫, 飯作俊一, “短時間動作記述を用いた映像のシーン表現と検索方式の検討,” 信学技報, PRMU98-190, pp.107-114, Jan. 1999.

[45] 瀧剛志, 松本貴之, 長谷川純一, 福村晃夫, “サッカー映像からのチームワーク評価方法の検討, “ 信学技報, PRMU96-10, pp.67-74, May 1996.

[46] 菱川剛, 瀧剛志, 長谷川純一, “サッカーにおけるプレッシャー評価方法の改善,” 1999 信学総大, D-12-193, p.366, Mar. 1999.

[47] M. Douke, M. Hayashi, E. Makino, “A Study of Automatic Program Production Using TVML,” In *Proceedings of Eurographics* '99, pp.42-45, 1999.

[48] http://svmlight.joachims.org/

[49] 三須俊彦, 高橋正樹, 蓼沼眞, 八伸行木, “サッカー映像のフォーメーション解析に基づく実時間イベント検出,” 第 4 回情報科学技術フォーラム情報科学技術レターズ (FIT2005), vol.4, LI-003, 2005, p.141-144, 2005.

[50] 山田一郎, 佐野雅規, 住吉英樹, 柴田正啓, 八木伸行, “アナウンスコメントを利用したサッカー番組メタデータ自動生成,” 信学技報, PRMU2004-204, pp.37-42, 2005.

[51] 住吉英樹, 佐野雅規, 八木伸行, “スポーツダイジェスト番組の制作支援を目的としたスロー再生区間検出手法,” 映情メディア冬季大会, 9-6, 2004.

[52] 浜口斉周, 道家守, 林正樹, “TV4U～テレビセット内で作られる自分だけのテレビ番組～, ” 信学技報, PRMU2002-29, vol.102, no.155, pp.63-68, 2002.

[53] McKeown and D. R. Radev: “Generating Summaries of Multiple News Articles”, In *Proc. of the SIGIR-95*,1995.

[54] D.R. Radev, H. Jing, and M. Budzikowska.: “Centroid -based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies”, In *Proc. of the ANLP/NAACL2000 Workshop on Automatic Summarization*, pp.21-30, 2000.

[55] 奥秋義信, “ニュース原稿の書き方～その理論と実際,” 岩崎放送出版社, 1970.

[56] 加藤直人, 浦谷則好, “放送ニュースを対象にした重要文抽出” 言語処理学会第 6 回年次大会論文集, pp.237-240, 2000.

[57] 渡辺靖彦, 竹内雅人, 村田真樹, 長尾真, “χ^2 法を用いた重要漢字の自動抽出と文献の自動分類,” 信学技報, NLC94-25, pp.23-30, 1994.

[58] 木田敦子, 乾裕子, 落谷亮, 西野文人, “情報抽出のための文末表現分析,” 言語処理学会第 6 回年次大会論文集, pp.304-307, 2000.

[59] Yeun-Bae Kim and Terumasa Ehara, “A Method of Partitioning of Long Japanese Sentence with Subject Resolution in J/E Machine Translation,” *Proc. of the 1994 International Conference on Computer Processing of Oriental Languages*, pp.467-473, 1994.

[60] 平尾努, “自動要約評価型ワークショップ : Text Summrization Challenge(TSC)(音声・言語における標準化動向) ,” 信学技報, NLC2004-62, pp.37-42, 2004.

[61] C. Fellbaum, “WordNet: An Electronic Lexical Database,” Cambridge, MA: MIT Press, 1998.

[62] F. Bond, H. Isahara, K. Kanzaki and K. Uchimoto, “Boot-strapping a WordNet Using Multiple Existing WordNets,” In *the 6th International Conference on Language Resources and Evaluation* (LREC), Marrakech, 2008.

[63] 西野文人, 橋本三奈子, 落谷亮, “テキストからの用語とその定義文の抽出,” 言語処理学会第５回年次大会, pp.124-127, 1999.

[64] Marti Hearst, “Automatic acquisition of hyponyms from large text corpora, *In Proc. of the 14th International Conference on Computational Linguistics* (*COLING '92*), Nantes, France, pp.539-545, 1992.

[65] Dominic Widdows and Beate Dorow, “A graph model for unsupervised lexical acquisition,” *In*

Proc. of the 19th International Conference on Computational Linguistics (*COLING 2002*), Taipei, Taiwan, pp.1093-1099, 2002.

[66] Rion Snow, Dan Jurafsky and Andrew Y. Ng, "Semantic taxonomy induction from heterogenous evidence," In Proc. of COLING/ACL 2006, Sydney, Australia, 2006.

[67] Roxanne Girju, Adriana Badulescu, and Dan Moldovan, "Learning semantic constraints for the automatic discovery of part-whole relations," *In Proc. of Human Language Technology Conference - Annual Meeting of the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL*), Edmonton, Canada, pp.1-8, 2003.

[68] 木田敦子, 乾裕子, 落谷亮, 西野文人, "新聞記事からの用語集作成のためのテキスト分析," 情処学 NL 技報, NL134-12, pp.85-92, 1999.

[69] 乾孝司, 乾健太郎, 松本裕治, "接続標識「ため」に基づく文書集合からの因果関係知識の自動獲得, " 情処学論 Vol.45, No.3, pp.919-933, March 2004.

[70] Satoshi Sekine, Hitoshi Isahara , "IREX: IR and IE Evaluation project in Japanese," In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pp.1475-1480, 2000.

[71] http://trec.nist.gov/

[72] 石川徹也, 坂本義行, 佐藤雅之, "Mu プロジェクトにおける意味マーカの概念と体系," 情報学会 NL 技報, 1986-NL-055 , pp.1-12, 1985.

[73] 竝木崇康, "語形成," 大修館書店, 1985.

[74] A. S. Hornby, 伊藤健三(訳), "英語の型と語法," オックスフォード大学出版局, 1977.

[75] Daisuke Kawahara, Sadao Kurohashi, "Case Frame Compilation from the Web using High-Performance Computing," In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 2006.

[76] James Pustejovsky, Peter Anick, and Sabine Bergler, "Lexical semantic techniques for corpus analysis," *Computational Linguistics*, 19(2), pp.331-358, 1993.

[77] http://www.ub.es/gilcub/SIMPLE/simple.html

[78] Lou Burnard, "User Reference Guide for the British National Corpus," *Technical report*, Oxford University Computing Services, 2000.

[79] Ted Briscoe and John Carroll, "Robust accurate statistical annotation of general text," In *Proc. of the3rd International Conference on Language Resources and Evaluation* (*LREC 2002*), Las Palmas, Spain, pp.1499-1504, 2002.

[80] http://www.comp.lancs.ac.uk/ucrel/claws2tags.html

[81] Kenneth W. Church and Patrick Hanks, "Word association norms, mutual information and lexicography," In *Proc. of the 27th Annual Meeting of the ACL*, pp.76--83, 1989.

[82] 後藤功雄, 熊野正, 江原輝将, "かぎ括弧で囲まれた表現の種類の自動判別," 言語処理学会第 6 回年次大会, pp.35-38, 2000.

[83] 中野洋, "分類語彙表形式による語彙分類表（増補版）,"国立国語研究所, 1996

[84] 内元清貴, 馬青, 村田真樹, 小作浩美, 井佐原均, ”最大エントロピーモデルと書き換え規則に基づく固有表現抽出,” 自然言語処理 Vol. 7, No. 2, pp.63-90, Apr. 2000.

[85] 住吉英樹, 山田一郎, 有安香子, 柴田正啓, 八木伸行, “学習コミュニティの対話を支援する仮想教室のシステム化,” 第 1 回情報科学技術フォーラム(FIT2002), Sep. 2002.

[86] 住吉英樹, 望月祐一, 金淵培, 柴田正啓, 井上誠喜, “エージェントを利用した映像検索のためのユーザーインターフェイス,” 信学技報, OFS2000-24, pp.9-14, 2000.

[87] I. Yamada, Y. B. Kim and M. Shibata, “*Topic Event Detection using Japanese News Articles*,” 5th NLPRS1999, pp.375-380, Nov. 1999.

[88] Du-Seong Chang, Key-Sun Choi, “Causal Relation Extraction Using Cue Phrase and Lexical Pair Probabilities,” *IJCNLP 2004*, pp.61-70, 2004.

[89] Kamel Nigam, Andrew Kachites Mccallum, Sebastian Thrun and Tom Mitchell, “Text Classification from Labeled and Unlabeled Document using EM,” *Machine Learning*, Vol.39, No.2/3, pp.103-134, 2000.

[90] 工藤拓, 松本 裕治, “チャンキングの段階適用による係り受け解析,” 情処論, Vol.43, No.6, pp.1834-1842, 2002.

[91] F. Luccio, A. M. Enriquez, P. O. Rieumont and L. Pagli, “Exact Rooted Subtree Matching in Sublinear Time,” *Technical Report*, TR-01-14, 2001.

[92] William L. Hays, Statistics, “Analyzing Qualitative Data,” *Rinehart and Winston Inc*., Chapter18, pp.769-783, 1988.

本研究に関する発表リスト

学術論文

- Ichiro Yamada, Timothy Baldwin, Hideki Sumiyoshi, Masahiro Shibata, Nobuyuki Yagi, “Automatic Acquisition of Qualia Structure from Corpus Data,” *IEICE Transactions on Information and Systems*(ED), Vol.E90-D, No.10, pp.1534-1541, 2007.
- 山田一郎, 三浦菊佳, 河合吉彦, 住吉英樹, 八木伸行, 奥村学, 徳永健伸, "大域的な文章構造の類似性を利用したクローズドキャプション中の定型的な文章区間の抽出", 信学論(D), Vol.J90-D, No.9, pp.2624-2633, 2007.
- 山田一郎, 佐野雅規, 住吉英樹, 柴田正啓, 八木伸行, “アナウンサーと解説者のコメントを利用したサッカー番組セグメントメタデータ自動生成,”信学論(D), Vol.J89-D, No.10, pp.2328-2337, 2006.
- 山田一郎, 山村毅, 佐川雄二, 大西昇, 杉江昇, “英文中に出現する未登録語の特徴分析とその意味推定のための知識の有効性に関する考察,” 情処論, Vol.35, No.12, pp.2725-2733, 1994.

共著学術論文

- Kikuka Miura, Ichiro Yamada, Hideki Sumiyoshi, Nobuyuki Yagi, “Identification of names and actions of principal objects in TV program segments using closed captions,” *International Journal of Semantic Computing*, 2008.
- 三浦菊佳，山田一郎，住吉英樹，八木伸行，奥村　学，徳永健伸, “放送番組を素材としたマルチメディア百科事典の自動構築,” 映像情報メディア学会誌, Vol.62, No.1, pp.110-116, 2008.
- Masanori SANO, Ichiro YAMADA, Hideki SUMIYOSHI and Nobuyuki YAGI, “Automatic Real-time Selection and Annotation of Highlight Scenes in Televised Soccer,” *IEICE TRANSACTIONS on Information and Systems*(ED), Vol.E90-D, No.1, pp.224-232, 2007.
- 金子豊, 中川俊夫, 藤澤俊之, 山田一郎, 南浩樹, 鹿喰善明, 田中豊, “分散処理環境下におけるコンテンツの相互運用のためのオブジェクトモデルの提案,” 信学論(D), Vol.J89-D, No.5, pp.932-942, 2006.
- Kyoko ARIYASU, Ichiro YAMADA, Hideki SUMIYOSHI, Masahiro SHIBATA, Nobuyuki YAGI, “Visualization of Text-Based Dialog in a Virtual Classroom for e-Learning,” *IEICE TRANSACTIONS on Information and Systems*(ED), Vol.E88-D, No.5, pp.836-842, 2005.
- 住吉英樹，山田一郎, 村崎康博，　金淵培，　八木伸行, 柴田正啓, “新しい教育放送サービスのための情報検索システム,” 映像情報メディア学会誌, Vol.57, No.2, pp.253-261, 2003.

国際会議

- Ichiro Yamada and Timothy Baldwin, “Automatic Discovery of Telic and Agentive Roles from Corpus Data,” *Proceedings of The 18th Pacific Asia Conference on Language Information and Computation*(*PACLIC18*), pp.115-126, 2004.
- Ichiro Yamada, Masahiro Shibata, and Yeun-Bae Kim, “Multiple Text Summarization using Fixed Expressions in News Articles,” *IEEE International Symposium on Natural Language Processing and Knowledge Engineering* ,Vol.2, MD2E-4, pp.460-465, 2002.
- Ichiro Yamada, Yeun-Bae Kim, Masahiro Shibata and Noriyoshi Uratani, “Topic Event Detection using Japanese News Articles,” *5th Natural Language Processing Pacific Rim Symposium 1999*, pp.375-380, 1999.
- Ichiro Yamada, Hideki Sumiyoshi, Yeun-Bae Kim and Masahiro Shibata, "TV program skimming method using similarity of scene descriptions,” *SPIE International Symposium Multimedia Strage and Archiving System Ⅲ*,Vol.3527, pp.253-260, 1998.
- Ichiro Yamada, Yeun-Bae Kim and Masahito Shibata, “Video Retrieval based on the Sentences Similarity,” *Proceedings of The National Language Research Institute Fifth International Symposium*, pp.145-153, 1997.

全国大会

- 山田一郎，宮崎勝，三浦菊佳，住吉英樹，柴田正啓，八木伸行, “放送番組のクローズドキャプションを対象とした健康に関する知識獲得へ向けて,” 第 14 回言語処理学会年次大会, E2-5, 2008.
- 山田一郎，宮崎勝，三浦菊佳，住吉英樹，八木伸行, “同一文中に出現する複数の節間における因果関係抽出の検討,” 第 6 回情報科学技術フォーラム(FIT2007), No.2, E-048, pp.251-252, 2007.
- 山田一郎，中田洋平，松井淳，松本隆，三浦菊佳，住吉英樹，八木伸行, “GibbsBoost による類似文章検索の検討,” 第 13 回言語処理学会年次大会, C3-2, pp.538-541, 2007.
- 山田一郎，三浦菊佳，河合吉彦，住吉英樹，八木伸行，奥村学，徳永健伸, “ＥＭアルゴリズムを利用した属性名抽出の検討,” 2007 年電子情報通信学会総合大会，D-5-12, p.52, 2007.
- 山田一郎，三浦菊佳，住吉英樹，八木伸行，奥村学，徳永健伸, “AdaBoost を利用した字幕テキストからの定型表現文章区間抽出の検討,” 第 12 回言語処理学会年次大会発表論文集 D1-4, pp.101-104, 2006.
- 山田一郎，小早川健，三浦菊佳，住吉英樹，八木伸行，崔杞鮮, “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,” 第 5 回情報科学技術フォーラム(FIT2005), No.2, E-001, pp.113-114, 2005.
- 山田一郎，佐野雅規，住吉英樹，柴田正啓, “アナウンスコメントを利用したサッカー番組のメタデータ自動生成の検討,” 第 4 回情報科学技術フォーラム(FIT2004), no.2,

E-030, pp.177-178, 2004.

- 山田一郎, 柴田正啓, 金淵培, Key-Sun Choi, “Web を情報源とした Q&A システムの検討,” 第 9 回言語処理学会年次大会, C7-5, 2003.
- 山田一郎，有安香子，住吉英樹，柴田正啓，八木伸行, “仮想教室の学習コミュニティにおける発想支援エージェントの検討,” 第 2 回情報科学技術フォーラム(FIT2002), E-33, p.p147-148, 2002.
- 山田一郎, 柴田正啓, 金淵培, “ニュース原稿を利用した用語集作成の検討", 第 8 回言語処理学会年次大会, B4-4, pp.507-510, 2002.
- 山田一郎, 柴田正啓, “ニュース原稿を利用した用語集作成の検討,” 情報処理学会第 63 回全国大会, 1L-1, pp.29-30, 2001.
- 山田一郎, 金淵培, 柴田正啓, 浦谷則好, “ニュース記事からの話題構成要素抽出の検討～国会審議に関する話題を対象として～,” 第 7 回言語処理学会年次大会, C4-4, pp.297-300, 2001.
- 山田一郎, 金淵培, 柴田正啓, “ニュース原稿を利用した話題トラッキングの検討,” 情報処理学会第 61 回全国大会, 5U-6, pp.193-194, 2000.
- 山田一郎, 金淵培, 柴田正啓, “ニュース原稿からの話題抽出を利用したテレビ番組選択の検討,” 情報処理学会第 59 回全国大会, 2P-7, pp.71-72, 1999.
- 山田一郎, 金淵培, 柴田正啓, 浦谷則好, “ニュース記事を利用したトピック抽出の検討,” 第 5 回言語処理学会年次大会, A2-3, pp.116-119, 1999.
- 山田一郎, 金淵培, 柴田正啓, 浦谷則好, “ニュース原稿のクラスタリングを用いたトピック抽出,” 1998 年映像情報メディア学会冬季大会, 8-3, p82, 1998.
- 山田一郎, 金淵培, 柴田正啓, “映像シーケンサにおける文の類似性推定についての考察,” 1997 年映像情報メディア学会年次大会, 15-9, pp189-190, 1997.

研究会・シンポジウム

- 山田一郎，中田洋平, 松井淳, 松本隆，三浦菊佳，住吉英樹，八木伸行, “サンプリング技術を利用した文章類似性評価,” 情処学 NL 研報, Vol.2007, No.76, 2007-NL-180, pp.127-132, 2007.
- 山田一郎，三浦菊佳，住吉英樹，八木伸行，奥村学，徳永健伸, “AdaBoost を利用した字幕テキストからの定型表現文章区間抽出,” 情処学 NL 研報, NL174-5, pp.25-30, 2006.
- 山田一郎, 佐野雅規, 住吉英樹, 柴田正啓, 八木伸行, “アナウンスコメントを利用したサッカー番組メタデータ自動生成,” 信学技報, NLC2004-122, pp.37-42, 2005.
- 山田一郎，住吉英樹，柴田正啓, “ニュース記事に出現する用語と説明文の意味関係自動獲得,” 情処学 NL 研報, NL152-21, pp.145-152, 2002.
- 山田一郎, 柴田正啓, “ニュース記事の定型性を利用した話題要約の検討,” 情処学 NL 研報, NL148-7, pp.45-50, 2002.

- 山田一郎, 柴田正啓, “ニュース原稿を利用した話題抽出とニュース話題選択の個人化の検討,” 知識発見のための自然言語処理シンポジウム, 1999.
- 山田一郎, 住吉英樹, 金淵培, 柴田正啓, “シーン記述文の類似性を用いた番組自動要約システム,” 信学技報, PRMU97-254, pp.23-30, 1998

共著解説

- 村崎康博, 山田一郎, 金淵培, “放送と情報処理 : エージェントテレビ - 番組洪水から視聴者を救うテレビ -,” 情報処理, Vol.40, No.7, pp.1-5, 1999.