キャプションと記事テキストの文字列照合による報道番組と新聞記事との対応づけの自動化

角田達彦[†] 大石巧[††]
渡辺靖彦[†††] 長尾眞[††]

本稿では，ニュースの各報道に対応する新聞記事をキャプションと記事の文字列照合により特定する手法を提案する．キャプションと記事中に共通に現れた文字列の長さと出現位置，出現頻度により重みづけし，類似度を計算する．そして類似度が最大で閾値以上のものを選ぶ．学習サンプルによって各パラメータの値を決定した結果，学習サンプルで再現率100%，適合率93.2%，約2週間後のテストサンプルで再現率98.0%，適合率77.8%（閾値のみ決め直した場合，再現率98.0%，適合率84.5%），約7カ月後のテストサンプルで再現率97.1%，適合率79.5%（閾値のみ決め直した場合，再現率94.1%，適合率85.3%），という精度が得られた．また事例を検討し，長い文字列に重みを与えすぎることの弊害を明確にした．

Automatic Alignment between TV News and Newspaper Articles by String Matching between Captions and Article Texts

TATSUHIKO TSUNODA,[†] TAKUMI OOISHI,[††] YASUHIKO WATANABE[†††] and MAKOTO NAGAO[††]

We propose a method of automatic alignment of newspaper articles with corresponding TV news. The method extracts maximum length strings matched between the articles and the caption texts. Then it calculates similarity and picks up the nearest article if the similarity exceeds a given threshold. The similarity is based on the information of the string, i.e. its length, position and frequency in the text. By adjusting the weighting values, our method achieved 100% recall and 93.2% precision for learning samples, 98.0% recall and 77.8% (84.5% if the threshold was readjusted) precision for unseen samples which is about two weeks from the learning samples, and 97.1% recall and 79.5% (94.1% recall and 85.3% precision if the threshold was readjusted) precision for unseen samples which is about seven months from the learning samples. We also clarified the problem of excessive weighting for long strings.

1. はじめに

TVのニュースでの報道と新聞記事は，製作過程はまったく独立であるが，それらの内容は，その日の大きな事件など同一の情報源に基づいていることが多い．すなわち，同一の出来事が，新聞のような文書中心の媒体とTVのような映像を中心とした媒体との，複数の媒体によって表現されている．逆に，それらの対応をとらえられれば，1つの事柄を複数の観点から多角的にとらえ直すことになる．場合によっては完全に対応はしなくても，密接に関係するものを抽出しておけば，互いの情報の不足部分を補い，より柔軟な検索を行える．たとえば，ニュース映像は画像や音声の情報が豊富であるのに対して，新聞は言語的情報が豊富であり，自然言語処理技術を用いて深い情報の抽出を言語的に行うことができる．そして，新聞の関連記事の抽出の技術[1),2)]などを用いて，間接的にニュース番組の同時的関連および時間的つながりをとらえ，分類し，映像データベースを構築できる可能性がある．そのようなデータベースは，たとえば電子図書館[3),4)]のようにネットワークなどを通じてさまざまな情報収集がなされる場合に必要となる．そこでは映像ライブラリという形で，画像，文書，音声などの複合メディア

† 東京大学医科学研究所
The Institute of Medical Science, University of Tokyo
†† 京都大学工学研究科電子通信工学
Department of Electronics and Communication, Kyoto University
††† 龍谷大学理工学部電子情報学科
Department of Electronics and Informatics, Ryukoku University

によって情報を表現するために，双方向のメディア変換がなされるのに十分な情報が統合されている必要がある．

映像の検索のために，その内容に応じて索引づけをするための要素技術の１つとして，カラービデオ映像のカット変わりを検出する手法があげられる[5),6)]．色のヒストグラムの局所的情報を基に，フレームの不連続性を検出し，カット変わりを検出する．そして各区間での最初のフレームや最初の短時間の映像を検索時の参照キーとする．検索時にも色のヒストグラムを用い，検索キーと類似度の高い参照フレームのある区間を特定する．これらの研究は，画像情報のみで内容を分割し，画像をキーとして検索を行っている．しかし，検索要求は自然言語によってなされる場合も多い．そのような場合には，これらの手法だけでは不十分である．

映像の意味的構造を言葉によって記述することを目指したものの１つとして，柴田の研究[7)]がある．放送番組の場合，編集のために，映像制作者が被写体の名称や動作などをシナリオや取材メモにキーワード的に記述しておくことが多い．これらの内容記述の中のキーワードの一致の程度によってドキュメントの各箇所を階層的にクラスタリングし，時間依存的な階層構造を自動的に構築していく．その結果をデータベースとし，映像ブラウザとして用いる方法が示されている．また井手らの研究[8)]は，各場面の画像中の内容を記述する場合の人間の記憶量を反映した量と記述形式を，心理学的に考察している．これらの研究は，専門家が画像を見て相当な量のキーワードを付けることを前提としているが，そのようなデータは一般に利用できるものとはいい難い．一般に入手できる情報源から何らかの手段で自動的に獲得できることが望ましい．

この問題を解決する手段の１つとして，放送された番組やビデオデータなどの映像に付与された音声を認識し，言語情報とすることが考えられる．有木の研究[9)]は，ニュースの報道を対象にし，音声・ビデオデータの自己組織化を目指している．画像処理によって動画のシーンカットを検出し，かつカメラワークのパラメータを抽出する．次にニュースのナレーションの音声をHMMによって認識し，ニュースキーワードを抽出する．それらのキーワードに基づき，上で分割された各映像の分野を推定する．また課題の１つとして映像中の文字も認識し情報として用いることが示唆されている．

またカーネギーメロン大学（CMU）では，Informedia Project[10)]という，デジタルライブラリを構築するプロジェクトが進められている．その一環として，ビデオデータから内容を抽出する研究が行われいている．まず，画像処理と音声信号の変化などを利用して，映像を2～5分程度のセグメント（ビデオパラグラフという）に区切る．次に映像の音声を認識してテキストにし，その中から取り出したキーワードを各ビデオパラグラフに割り付けてライブラリ化する．検索時は，ユーザの質問を自然言語処理によって拡張し，様々な情報の類似度を計算することによって，完全に照合しなくとも，内容が適切であれば検索結果として返せる．画像処理，音声認識，自然言語処理の各要素技術の精度が高いうえにそれぞれの情報を相補的に利用する形で組み合わせている点で，大変価値の高いシステムである．

しかし，映像に付随する音声言語は，どうしても映像の説明という観点が多くなることと，話し言葉によって視聴者に理解してもらうことのために，情報の密度や全体の情報量が低くなりがちであるし，非文も多い（これについては「考察」の章で詳しく述べる）．また，音声認識率は単語単位で９割と高いが，文は平均20個程度の単語を含むことを考えると，文全体を完全に認識することはほとんど不可能であり，文あたり２個程度の誤りが必ず現れる．これを現状の形態素解析・構文解析システムに入力したとき，一部の誤りが文全体に影響することが多く，構造的な情報の抽出ができなくなる．したがってこのCMUのシステムでもキーワードベースの記述という形をとっている．CMUのシステムは英語であるが，日本語の場合は漢字に変換する必要があり，ニュースぐらいの広範囲の語彙を扱うもので，漢字変換の曖昧さを満足に扱えるものはない．また，日々更新される話題の中で，新語・未登録語は必ず現れるが，ビデオで閉じた情報の中では，そのような新語・未登録語を獲得する手段は画像中のキャプションなどの限られた情報を利用せざるをえなく，実現は難しいと思われる．

これに対し新聞記事は，漢字などの表意文字を読者が目で見て，さらにその他の機能語によって情報を立体的に組み上げることができるように作られている．その文章中には，事件自体，その背景，詳細な内容，世論，記者の意見，将来予測，事件の社会に及ぼす影響などがかなり決まった順番と表現で述べられている．そして個々の文はよく推敲されており，非文も少ない．現在の技術では99%以上の精度で形態素解析ができ，95%程度の構文解析ができる．それを利用した情報抽出の精度も大変高いところまできている．上のように完全な音声認識ができない間は，新聞記事のように完

全な形の文が他にあるならば，それを利用することも1つの方法と思われる．また，音声で必ず問題となる新語・未登録語，特に漢字や片仮名による造語を獲得するための情報源とすることもできる．

この他，新聞記事そのものに，テキストの情報検索の対象としての需要がある．ニュース報道と新聞記事のように，形のうえでは独立しているが同じ内容の情報や，関連する情報を結び付けたい（ハイパテキスト化）というユーザの要求がまずある．また，新聞では，関連する記事を抽出する精度の高い技術などが確立されている．本研究のようにニュースと新聞記事を結び付けることができれば，関連する新聞記事どうしを介することによって，独立に存在していたビデオ情報も内容的に関連するものを結び付けることができる．

これらをふまえ，本稿では，TVのニュースでの報道に対応する新聞記事を自動的に対応づけする手法を提案する．ニュース映像による報道の内容を知る手がかりとして，画像中に現れるキャプションを用いる．これは報道番組の製作過程で，画像に対して補足的な説明を加えているものである．たとえば，各報道の冒頭では，報道を簡潔明瞭に示すタイトルが現れる．また人名や地名などが該当する映像に加えられる．これらのキャプションを映像から抽出する技術は確立されており，かなり信頼性が高い[11),12)]．このようなキャプションのテキストと新聞記事テキストとの間で文字列照合を行い，類似度を計算し，閾値以上で最大の類似度を持つものどうしを対応づける．

テキスト間の対応づけを行う技術は，情報検索の分野でさまざまな研究が行われているが，各メディア内で対応づけをするものがほとんどである．その中で本研究に関係のあるものは，関連する新聞記事どうしを対応づけるものである．奥らの研究[1)]では，各紙のその日の「～面に関連記事」という記述から紙面数を特定した後の，該当記事をより細かく推定するための情報として，見出し間の文字の一致した数を用いることが提案されている．1面トップ記事とその関連記事とでは，見出しの一部が類似する傾向を利用したものである．だが，この方法は各紙のその日の記事どうしに利用が限られる．また見出しのみの情報を用い，全体の文字の一致した数のみで，各文字列の長さは考慮していない．新谷らの研究[2)]では，その日に限らず関連記事を抽出する方法が提案されている．名詞，サ変名詞，未定義語などの単語を取り出し，その記事内の出現位置（タイトル内か，リード文内か，本文内かなど），記事内の出現回数，そしてその単語が稀な単語かありふれた単語かを判定し，重みづけを変化させることによってより内容を反映した記事間の類似度を求める．

新谷らの研究で用いられているような，単語の文章中での出現位置や出現頻度などの情報は比較的容易に利用できる．それらはキーワードの判定の研究[13)]や，検索質問に対して各文書に重要度を付与する研究[14)]などにも用いられており，特に新聞などのようにある程度形式の定まった文書の場合には有効である．本研究でも，文字列の重みを決定し類似度を求めるための1つの情報としてキャプションと新聞記事のそれぞれのテキスト内での位置と頻度を利用している．キャプションと新聞記事の類似度は，一致した文字列の長さ，頻度，文字列の新聞記事内での位置，文字列の現れたキャプションの性質（タイトルか否か），キャプションの文字数，新聞記事の文字数を考慮する．本研究はより簡便で高速な処理を目指したことと，テキスト間での一致文字列の長さと関連性との相関について考察することを目的としたため，形態素解析は行っていない．名詞などを形態素解析によって取り出し重みづけする研究[15)]も同時に行っているが，その詳細については別稿にゆずることにする．類似度の計算方法は，キャプションや新聞記事の性質に大きく依存する．このため本研究では重みづけのパラメータと閾値を学習サンプルによって決定するとともに，キャプションと新聞記事とに共通に現れる文字列の性質と，それが関連性にどのように結び付くか，このタスクの限界点は何かを明らかにする．そして決定したパラメータをテストサンプルに適用し，手法の有効性を確かめる．

以下，2章でキャプションと新聞記事の特徴について述べ，3章で本稿で提案する対応づけの手法について述べる．4章で学習コーパスによるパラメータの調整の結果とテストコーパスでの有効性について示し，5章で実験結果の考察，本手法の限界，そして得られた知見についてまとめる．

2. キャプションと新聞記事の特徴

2.1 キャプションの特徴

TV番組は一般に画像と音声を主な媒体として内容を伝える．ニュースではさらに，画像の中に文字の情報（キャプション，テロップ）を入れこむことによって，より正確に短時間で複雑な情報を伝達するように工夫することが多い．たとえば事件名や人物名などを画像中の適切な箇所に漢字で入れれば，音声のみよりも曖昧さが少ない．逆にキャプションの方に注目した場合，番組の内容を示す情報の多くを含む☆．

ニュースのキャプションは，内容によっておおまかに区別すると，次のようになる．

表1 キャプションの典型的な例とキャプションの型
Table 1 Typical example of set of captions and its contents.

日銀 支店長会議	タイトル
“景気は緩やかながら回復”	(続き)
日銀支店長会議	状況説明
日銀 松下総裁	人名+所属・職名
“景気は緩やかながら回復”	発言内容のまとめ
大和総研 秋本 投資調査部長	人名+所属・職名
所得 2%強の伸び	発言内容のまとめ
設備投資 サービス・通信・運輸に	発言内容のまとめ
“96 年度 2%台半ばの成長”	発言内容のまとめ
あさひ銀行 大阪 調査部長	人名+所属・職名
個人消費 低迷	発言内容のまとめ
民間設備投資 更新・補修が中心	発言内容のまとめ
能力増強=本格的投資なし	発言内容のまとめ
“公共投資切れで景気低迷懸念”	発言内容のまとめ

(1) タイトル：報道内容全体を示す事件名，話題など
(2) タイトル以外
  (a) 個々の映像の状況説明，事態の簡潔な説明など
  (b) 人名と敬称・職名・地位など
  (c) 現場の地名や映像の撮影時刻など
  (d) 発言内容そのもの（外国語を翻訳した場合など）
  (e) 発言内容を番組製作で編集し，簡潔にしたもの
  (f) その他番組製作に関する情報（報道記者名など）

典型的なキャプションの例と，その個々のキャプションを分類した結果を**表1**に示す．この表から，タイトルの他に人名や所属・職名，発言内容のまとめにも，人間が報道の内容を判断する多くの情報が含まれていることが分かる．上のように分類したキャプションの型は，画面上の位置や飾りなどによって解析できる場合が多いが，必ずしも一意に分類できるとは限らないことと，今回は簡単な処理を行うことを考えたため，タイトルとタイトル以外の部分とに大きく分けるだけにし，その中では均一なコーパスとして扱うことにした．タイトルは，ほとんどの報道番組で下線などの分かりやすい飾りを付けているため，特定が容易である．本研究では，画像処理などによってタイトル部とそれ以外の区別のみされているキャプションのコーパスが

☆ ただし，時刻表示や，他の番組の案内や開始時刻の通知，報道記者名など番組の製作側に関する内容など，直接内容に関係ないものを示す場合もある．これらは映像のレイアウトなどを利用して除外するようにしているが，構造的に区別がつかないこともある．

表2 新聞記事の典型的な例と構造
Table 2 Typical example of newspaper article and its structure.

景気の回復改めて確認——日銀支店長会議	見出し
「金利維持」触れず	
(写真，写真説明ともになし)	写真説明
日本銀行の全国支店長会議が八日，日銀本店で開かれ，松下康雄総裁のあいさつなどで「景気は緩やかながら回復しつつある」との認識があらためて示された．当面の金融政策について，総裁は「景気回復の基盤を万全とすることに重点を置き，展開を注意深く点検していく」と述べたが，前回一月の支店長会議での「現在の金融緩和姿勢を維持する」という言葉はなく，市場関係者の一部には，超低金利政策から次の金利水準を模索しているのではないか，との見方が強い．	第一段落
各地の支店長は，公共投資や住宅投資の増加や，企業収益の改善を指摘し，「自動車，工作機械が操業度を引き上げている」(名古屋) など，景気回復の明るい兆しを報告した．同時に「絶好調だった半導体などの電子部品の生産・輸出が今年に入って減少するなど，懸念材料もある．景気は回復しているが，極めて緩やか」(大阪) といった慎重な見方も示された．	第二段落

電子的に入手可能である[11),12)]という状況を仮定する．

2.2 新聞記事の特徴

新聞記事の典型的なものは，見出し，リード，本文，写真などの説明文などの構造を持っている．その一例を**表2**に示す．それぞれの内容の違いをまとめると，次のようになる．

(1) 見出し：記事の内容全体を簡潔明瞭に伝える．話題，当事者，重要な国名・地名などの小見出し．
(2) リード文：記事の主要な事柄を，文章で伝える．事件など自体の説明が1文目で，そして直接の背景となる事柄が2文目などで説明されることが多い．
(3) 本文：扱う事項を筋道立てて詳しく説明する．時間的な経緯や論理的なつながり，また他の関連する事柄などを網羅的に述べている場合が多い．
(4) 写真などの説明文：記事に写真がつけられている場合，現場，人物，該当物体などの写真の表す内容そのものと，記事の内容との関係が述べられる．記事内容の中心的な人名や地名などの固有名詞が含まれることが多い．

新聞記事は，上の順番に従って読むことが想定されているが，今回のタスクでは，記事内容の中心的な部分を探すことに力点が置かれるため，写真の説明文も見出しと同程度に重要である．

表2に示した新聞記事の例はインターネットより入手可能なもので，前述のニュースの例（表1）に内容が対応している．この例では写真はなかった．また

リード文は本文と明示的に区別されていないが，一般の新聞記事のリード文に相当するものは本文の第一段落であると判断した．

今回用いたネットワーク上にあるコーパスも，一般の新聞記事に基づくコーパスも，上の構造をとらえることは大変容易であるため☆，これらの情報を積極的に利用した．

3. ニュースの報道と新聞記事の対応づけの手法

本稿では，ニュースの報道のそれぞれに対して，対応する新聞記事を1つ選択するというタスクを目的とする．選択の対象とする新聞記事の範囲は，ニュースと新聞記事のそれぞれの製作工程の長さなどを考慮して決める必要がある．具体的対象は実験の章で述べる．

対応する新聞記事を選択する手法は以下の手順に従う．

(1) 現在対象としている報道のキャプションと，新聞記事群の中から取り出した1つの記事の間で，類似度を計算する．類似度計算は，次の手順で行う．
   (a) キャプションと記事とで文字列の照合を行い，一致する文字列（各箇所で最長）を列挙する．
   (b) 文字列の長さや出現位置，テキストのサイズなどを考慮して重みづけをし，足し上げた得点を類似度とする．

   これを対象となる新聞記事すべてに対し計算する．

(2) 対象となる新聞記事すべての中から最大の類似度を持つものを求め，それがあらかじめ決められた閾値よりも大きければ，それを解として出力する．小さければ，「解なし」と出力する．

以下の節で，一致文字列の取り出し方，重みづけと得点化について詳しく説明する．

3.1 最長一致文字列の抽出

キャプションのテキストと新聞記事のテキストを見比べ，より長い文字列がより多く一致していれば，それらはより類似していると考えられる．そこで，そのような文字列を見つけることを考える．

たとえば**図1**に示すように，キャプションに「日銀支店長会議」という文字列があったとする．そして新聞記事の方に「日銀で支店長会議が行われた」という

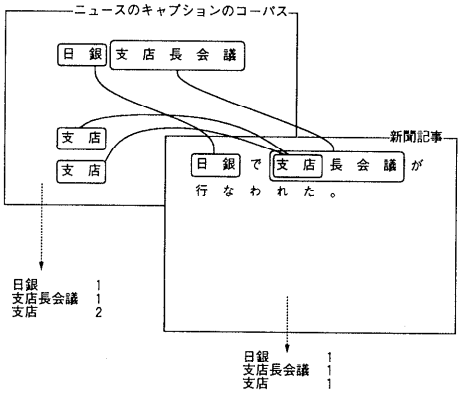


図1 各箇所での最長一致文字列の抽出の様子と頻度の数え方
Fig. 1 The method of extracting matched strings and counting their frequency.

文字列があったときは，各部分での最長一致文字列は，「日銀」という2文字列と「支店長会議」という5文字列である．これを一致文字列の側から見て，「日銀」という文字列はキャプションと新聞記事にそれぞれ1回現れ，そして「支店長会議」もそれぞれのコーパスに1回現れるという数え方をする．

キャプション側にさらに「支店」という文字列が2つあったとすると，上に加えて独立に，「支店」という文字列がキャプションに2回，新聞記事に1回現れたと見なす．つまり，この例では新聞記事の側の「支店長会議」の部分は，キャプションの「支店長会議」とも「支店」とも独立に照合する．そのときの照合に応じて一致範囲をなるべく大きくとる☆☆．

ただし，平仮名はすべてコーパスから削除し，文字列の境界であると見なした．平仮名のほとんどは助詞など，テキストの内容を端的に示す言葉にはなりにくく，ノイズとなりやすいと判断したためである．

3.2 文字列の特徴による重みづけと得点化

前節のように取り出した文字列に対し，その長さと位置に応じて重みづけをし，出現頻度に比例させて得点に加算していく．計算方法は次の式(1)に従う．

$$Score = \sum_{i,j,k} W_l(|k|) \cdot \left(\frac{W_{news}(i) \cdot n_{news}(k,i)}{S_{news}(i)} \right) \cdot \left(\frac{W_{arti}(j) \cdot n_{arti}(k,j)}{S_{arti}(j)} \right) \quad (1)$$

☆ HTMLのタグづけがされている．一般誌の場合にも少なくとも段落分けがされている

☆☆ 本稿では，これを最長一致とよぶことにする．

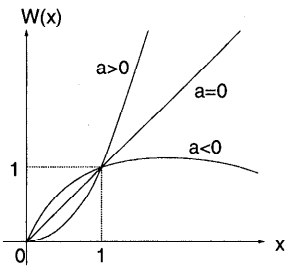


図2 重み関数 W (x) の振舞い
Fig. 2 Behavior of weighting function W (x).

$W_l(\|k\|)$	文字列 k の長さに応じた重み
$W_{news}(i)$	文字列のキャプション内の位置による重みづけ
	$W_{news}(1) \equiv 1$
$S_{news}(i)$	キャプションの各箇所の文字数
$n_{news}(k,i)$	キャプションの各箇所での文字列 k の出現頻度
	i = 1: タイトル, i = 2: タイトル以外
$W_{arti}(j)$	文字列の新聞記事内の位置による重みづけ
	$W_{arti}(1) \equiv W_{arti}(2) \equiv 1$
$S_{arti}(j)$	新聞記事の各箇所の文字数
$n_{arti}(k,j)$	新聞記事の各箇所での文字列 k の出現頻度
	j = 1: 見出し, j = 2: 写真説明
	j = 3: リード文, j = 4: 本文

上の (1) 式中の，キャプションのコーパスと新聞記事のコーパスとの間で照合した文字列の，長さによる重みづけの計算には，次の関数を用いる．

$$W_l(x) = x \cdot 2^{a(x-1)} \qquad (2)$$

ただし，x は文字列の長さとする．式中のパラメータ a によって，文字列の長さによる重みづけの度合を調整する．a を 0 にとれば，文字列の長さに比例した重みづけになる（**図2**）．正にとれば，指数関数的に増加する．a を大きくとればとるほど，文字列の長さの増加による重みづけの度合が大きくなる．逆に負にとれば，いったん増加するものの，途中より減少を始め，0 に漸近的に近づくため，長い文字列の重みづけは大きくしない結果になる．

$W_{news}(i)$ と $W_{arti}(j)$ はそれぞれ，上の文字列が現れた位置による重みであり，他の重みに積算される．特にニュースの側の i=1 はタイトル内にある場合を，そして新聞記事の側の j=1 は見出し内にある場合を示す．さらに新聞記事の側の j=2 は写真の説明文の中にある場合を示す．これらには報道や記事の内容を直接表す言葉が多く含まれていることがほとんどである．そこで，これらの位置での重みを 1 と設定し ($W_{news}(1) = 1, W_{arti}(1) = 1, W_{arti}(2) = 1$)，他の位置での相対的な重みを学習サンプルによって決定することにする．

$S_{news}(i)$ はキャプションのタイトル部分 (i = 1) とタイトル以外の部分 (i = 2) の，それぞれの文字数を表す．同様に $S_{arti}(j)$ は新聞記事の見出し (j = 1)，写真説明 (j = 2)，リード文 (j = 3)，本文 (j = 4) の，それぞれの文字数を表す．これらは各位置での文字列の頻度をテキストのサイズによって正規化するために用いられる．

以上のように，ある報道とある記事の類似度は，式 (1) で表されるように，文字列の長さによる重み，キャプション中での位置による重み，そして新聞記事での位置による重みを一致文字列ごとに積算し，文字列の頻度に比例して得点化することによって求められる．そしてある閾値以上の，最大の類似度を持つ記事を，キャプションに対応づける．その重みづけのためのいくつかのパラメータと閾値を，学習サンプルによって決定する．

4. 学習コーパスとテストコーパスに対する実験

4.1 学習コーパスとテストコーパス

実験に用いたニュースは，NHK の 9 時のニュースの全国版の部分である．1 つのニュースあたり 5〜10 報道程度が放送されている．また新聞記事は朝日新聞のインターネット版で，朝刊と夕刊の部分を用いた．どちらもスポーツニュースを除いてある☆．3 月の 12 日分を学習コーパス，学習コーパスから約 2 週間後の 4 月の 10 日分をテストコーパス A，そして学習コーパスから約 7 カ月後の 10，11 月の 12 日分をテストコーパス B とした．その内訳を**表3**に示す．

ニュースの各報道に対し，その日の夕刊と翌日の朝刊の中（約 30 ないし 70 記事）から対応する記事があれば 1 つ選択し，なければ「解なし」と答えることが今回の目的である．

4.2 評価値の定義

本稿では，次に定義される再現率と適合率，そして評価値をもとに評価を行う．

$$\text{再現率} = \frac{\text{システム出力のうちの正解数}}{\text{対応する記事のあるキャプションの数}}$$

$$\text{適合率} = \frac{\text{システム出力のうちの正解数}}{\text{システムの出力の数}}$$

$$\text{評価値} = \max_{\text{閾値}} \left[\frac{\text{再現率} + \text{適合率}}{2} \times 100 \right]$$

評価値は，閾値を変化させたときの，再現率と適合率

☆ スポーツニュース，スポーツ記事は特殊な作り方がされているため，今回の対象外とした．いずれも装飾やタグなどによって構造的に容易に区別できる．

表3 学習コーパスとテストコーパス.( ) 内はそれぞれの報道数,記事数を表す

Table 3 Corpus for learning and testing. Each number between a set of braces shows the number of articles.

学習コーパス		テストコーパス A		テストコーパス B	
ニュース	新聞記事	ニュース	新聞記事	ニュース	新聞記事
3/11 (5)	3/12 朝 (29)	4/1 (9)	4/1 夕 (31)	10/23 (9)	10/23 夕 (30)
3/13 (6)	3/14 朝 (30)		4/2 朝 (32)		10/24 朝 (32)
3/14 (5)	3/15 朝 (30)	4/2 (10)	4/2 夕 (41)	10/24 (10)	10/24 夕 (37)
3/15 (6)	3/16 朝 (30)		4/3 朝 (30)		10/25 朝 (27)
3/18 (6)	3/18 夕 (33)	4/3 (9)	4/4 朝 (31)	10/25 (9)	10/25 夕 (37)
	3/19 朝 (29)	4/4 (10)	4/4 夕 (36)		10/26 朝 (34)
3/19 (6)	3/20 朝 (29)		4/5 朝 (30)	10/28 (12)	10/28 夕 (37)
3/21 (5)	3/22 朝 (29)	4/5 (10)	4/5 夕 (32)		10/29 朝 (25)
3/22 (6)	3/22 夕 (39)		4/6 朝 (32)	10/29 (13)	10/29 夕 (36)
	3/23 朝 (30)	4/8 (11)	4/8 夕 (31)		10/30 朝 (37)
3/26 (6)	3/27 朝 (31)		4/9 朝 (32)	10/30 (9)	10/30 夕 (30)
3/27 (7)	3/27 夕 (37)	4/9 (11)	4/9 夕 (42)		10/31 朝 (27)
	3/28 朝 (34)		4/10 朝 (27)	10/31 (8)	10/31 夕 (30)
3/28 (6)	3/28 夕 (33)	4/11 (10)	4/11 夕 (35)		11/1 朝 (41)
	3/29 朝 (10)		4/12 朝 (10)	11/1 (10)	11/1 夕 (36)
3/29 (9)	3/29 夕 (31)	4/12 (3)	4/12 夕 (37)		11/2 朝 (35)
	3/30 朝 (30)		4/13 朝 (30)	11/4 (9)	11/4 夕 (14)
		4/13 (4)	4/13 夕 (41)		11/5 朝 (24)
			4/14 朝 (25)	11/5 (6)	11/5 夕 (43)
					11/6 朝 (27)
				11/6 (1)	11/6 夕 (29)
					11/7 朝 (41)
				11/7 (6)	11/7 夕 (31)
					11/8 朝 (21)
(73 報道)	(514 記事)	(87 報道)	(605 記事)	(102 報道)	(761 記事)

の百分率の平均の最大値である.

4.3 学習コーパスによるパラメータ値の決定

学習サンプルによってパラメータを決定するとき,複数あるパラメータを同時に決定するのは,解空間が大変大きくなり,効率が悪い.そこで,妥当そうな初期値を最初に設定し,精度を大きく左右すると思われるパラメータから順に変更し直すという方法をとる.

4.3.1 初期値の設定

文字列長に関するパラメータは最初に値を探索するので,初期値の設定は不要である.位置による重みであるが,タイトル部全体の文字数と,それ以外の文字数の比率から,文字列あたり 10 倍程度の重みがすでにタイトル部にかかっていると解釈できる.同様に,新聞記事に関しても,見出し部にリード文や本文の 10 倍から 20 倍程度の重みがかかっている.そこでこれらを補正するため,タイトル以外,見出し以外の重みの初期値をやや大きめに,$W_{news}(2) = 1.5$, $W_{arti}(3) = 2.5$, $W_{arti}(4) = 2.5$ と設定した.$W_l(x)$ のパラメータ a はこれらの重みを用いて次節で決定する.それをもとにさらに上の重みを順に調整し直す.

4.3.2 文字列長による重みづけのパラメータの決定

文字列の長さに応じた重みづけを決定する.重みづけは前章の式 (2) を用い,指数部のパラメータ a の値を $0, 1, 2, 3$ にしてそれぞれ実験してみた.同時に,文字列の長さの下限を 1 文字と 2 文字のそれぞれで変えて実験を行った.その理由は,日本語は漢字 2 文字の言葉が大変多く,内容との関係が大変大きいのに対し,1 文字では曖昧さが多く,ノイズとなる可能性もあると判断したからである.しかし実験の結果,**表 4** のように,$a = 1$ かつ文字列長の下限が 1 文字の場合が最も評価値が高く,95.5 となった.

表4 文字列長による重みづけの関数のパラメータ a と用いる文字列の長さの下限に対する評価

Table 4 Evaluation for the parameter of weighting function according to string length and the threshold of the length.

a の値	長さ下限	評価値	再現率	適合率
0	1	89.1	89.1%	89.1%
0	2	94.9	98.2%	91.7%
1	1	95.5	96.4%	94.6%
1	2	94.9	98.2%	91.5%
2	1	94.7	96.4%	93.0%
2	2	94.7	96.4%	93.0%
3	1	91.3	94.5%	88.1%
3	2	91.3	94.5%	88.1%

まず,この表の上から 4 行目までを比較検討すると,長さに比例する重みづけ ($a = 0$) の場合には,1 文字一致のものはノイズとなりうるが,指数部を適切に設定すれば ($a = 1$),1 文字一致のものも弁別の助けになることが分かる.

次に,指数部のパラメータは $a = 0$ よりも,$a = 1$ の方が良い.これは一致文字列の長いものに大きな重みを与えた方が良いという直観に合う.だが,$a > 1$ では逆に悪くなるのは,長い一致文字列でも,関係のない記事に現れることがありうるからである.たとえば,米大統領選挙の「ドール候補 圧勝」のニュースと,「反テロへ国際連帯訴え」の新聞記事は,関係がないのにもかかわらず,「クリントン大統領」,「クリントン」,「大統領」という文字列が共通して現れる.

これらを合わせ考えると,文字列長は短いが多く現れた文字列と,重みの大きい長い文字列との,それぞれの誤り率に応じたバランスをとる必要があり,長い文字列に重みを与え過ぎてはならないことが分かる.

4.3.3 キャプション中の位置による重みづけの調整

キャプション中の位置による重みづけの調整に際しては,上の結果から,文字列の長さに関しては 1 文字より考慮し,文字列長による重みに関するパラメータは $a = 1$ に固定することにする.また新聞記事中の位置による重みは初期値のままとする.

文字列がキャプションのタイトル以外の場合の,タイトル中の場合に対する相対的な重み $W_{news}(2)$ を変化させた結果,**表 5** のように,1.0 のときが最も評価値が高くなった.初期値の 1.5 のときに比べ良くなっているのは,ニュースの「アイヌの人たちに新立法を」(タイトル部分) という報道では,タイトルにしか「アイヌ」という語が現れないが,それが相対的に大きく

表5 文字列がキャプションのタイトル以外に現れるときの重みづけの評価
Table 5 Evaluation of weight for strings which appear in non-title section in news captions.

$W_{news}(2)$	評価値	再現率	適合率
0.0	86.2	85.4%	87.0%
0.5	95.6	98.2%	93.1%
1.0	96.5	98.2%	94.7%
1.5	95.5	96.4%	94.6%
2.0	95.5	96.4%	94.6%
2.5	94.5	94.5%	94.5%
3.0	94.5	94.5%	94.5%

重みづけられたため，対応する新聞記事を特定することができたためである．だが，タイトル以外の部分の重みを過度に小さくすると，人名や地名など，タイトルに現れない語を加味することができなくなり，結果が悪くなる．

ところでこの $W_{news}(2) = 1.0$ という値だが，タイトル部，タイトル以外の部分とも各々の文字数で正規化するため，個々の一致文字列を見れば，すでにタイトル部に10倍程度の重みがかかっている．このため，その重みを保持するとみれば自然な値である．

4.3.4 新聞記事中の位置による重みづけの調整

上の結果を用い（$a = 1$，1文字以上，$W_{news}(2) = 1.0$），文字列の新聞記事中の位置による重みづけ，すなわち $W_{arti}(3)$ と $W_{arti}(4)$ を順に調整する．まず，$W_{arti}(4) = 2.5$ のままで $W_{arti}(3)$ を0から5まで変化させる．すると**表6**のように，3.5から4.5付近が評価値が高いことが分かる．$W_{arti}(3)$ を小さくすることは，新聞記事のリード部を無視することによって相対的にリード部以後を大きくみる悪い点と，見出しの部分を相対的に大きくする効果があり，その関係はトレードオフにある．これを0にしたときに生じた誤り例は，文字列長による重みの決め方で問題になった米大統領選挙の事例で，これは「クリントン大統領」などの文字列がリード部より後にあるためである．リード部の重みを小さくすると，相対的に他の場所の重みが大きくなるからである．また他の例として，「春闘電機8,800円台で決着へ」というニュースに対して，関係のない株価の記事が出力された．株価の見出しに「円」と「円台」という文字列が含まれていたのと，記事全体の大きさが極端に小さいために，これらが強調されてしまったのが原因である．

以上から，リード部は無視できず，見出し部分などとの重みのバランスをとる必要があることが分かる．

上の結果を見てプラトーの真中にある $W_{arti}(3) = 4.0$ を用い，最後に $W_{arti}(4)$ を変化させた．その結果，**表7**のように，初期値である2.5と2.0で評価値が最大となった．その前後よりも評価値が良いのは，リード部以後の本文は，ある程度弁別に寄与するが，あまり大きく重みをつけると，関連はするが対応はしない記事を取り出す影響もあることを示している．実際に，この重みが3.0のときは，「少し関連する記事」が取り出され，誤りとなる事例が見られた．

表6 文字列が新聞記事のリード文に現れるときの重みづけの評価
Table 6 Evaluation of weight for strings which appear in leading section in newspapers.

$W_{arti}(3)$	評価値	再現率	適合率
0.0	93.9	96.4%	91.4%
0.5	93.9	96.4%	91.4%
1.0	94.7	96.4%	93.0%
1.5	94.7	96.4%	93.0%
2.0	96.5	98.2%	94.7%
2.5	96.5	98.2%	94.7%
3.0	96.5	98.2%	94.7%
3.5	96.6	100.0%	93.2%
4.0	96.6	100.0%	93.2%
4.5	96.6	100.0%	93.2%
5.0	94.9	98.2%	91.5%

表7 文字列が新聞記事の本文に現れるときの重みづけの評価
Table 7 Evaluation of weight for strings which appear in main section in newspapers.

$W_{arti}(4)$	評価値	再現率	適合率
0.0	93.1	96.4%	89.8%
0.5	93.1	96.4%	89.8%
1.0	94.9	98.2%	91.5%
1.5	94.9	98.2%	91.5%
2.0	96.6	100.0%	93.2%
2.5	96.6	100.0%	93.2%
3.0	95.8	100.0%	91.6%

4.4 テストコーパスでの評価

学習サンプルをもとに各段階で決めたパラメータを，それぞれテストコーパスで適用した結果を**表8**と**表9**に示す．調整の結果，学習サンプルで閾値を決めた場合は，約2週間後のテストコーパスAでは再現率98.0%，適合率77.8%に，約7カ月後のテストコーパスBでは再現率97.1%，適合率79.5%になった．この結果は初期値での結果に比べて必ずしも良い値でないが，それは，閾値の設定が個々の事例に依存するためで，テストサンプル自体で閾値を決め直すと，良くなっているのが分かる．閾値のみ決め直した場合は，テストコーパスAでは再現率98.0%，適合率84.5%に，テストコーパスBでは再現率94.1%，適合率85.3%になった．すなわち，重みの設定自体はある程度精度の向上につながっているものと思われる．また，初期値を直観で設定したため，もともと汎用的に

表8 学習コーパスに基づき各段階で決定したパラメータ値を用いて本手法をテストコーパスAに適用した結果
Table 8 Experimental results of our method which is applied to the corpus for testing. The set of value of parameters was determined with the corpus for learning.

a	長さ	W_{news}	W_{arti}	W_{arti}	学習コーパスによる閾値を使用			テストコーパスAによる閾値を使用		
値	下限	(2)	(3)	(4)	評価値	再現率	適合率	評価値	再現率	適合率
1	1	1.5	2.5	2.5	89.4	96.0%	82.8%	89.4	96.0%	82.8%
1	1	1.0	2.5	2.5	88.0	96.0%	80.0%	89.7	94.0%	85.5%
1	1	1.0	4.0	2.5	84.9	96.0%	73.8%	89.4	96.0%	82.8%
1	1	1.0	4.0	2.0	87.8	98.0%	77.8%	91.2	98.0%	84.5%

表9 学習コーパスに基づき各段階で決定したパラメータ値を用いて本手法をテストコーパスBに適用した結果
Table 9 Experimental results of our method which is applied to the corpus for testing. The set of value of parameters was determined with the corpus B for learning.

a	長さ	W_{news}	W_{arti}	W_{arti}	学習コーパスによる閾値を使用			テストコーパスBによる閾値を使用		
値	下限	(2)	(3)	(4)	評価値	再現率	適合率	評価値	再現率	適合率
1	1	1.5	2.5	2.5	88.3	92.6%	84.0%	88.3	92.6%	84.0%
1	1	1.0	2.5	2.5	88.1	94.1%	82.1%	88.8	86.8%	90.8%
1	1	1.0	4.0	2.5	87.4	97.0%	77.6%	90.1	92.6%	87.5%
1	1	1.0	4.0	2.0	88.3	97.1%	79.5%	89.7	94.1%	85.3%

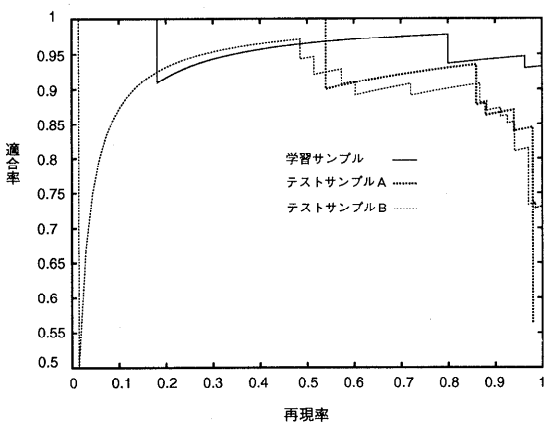


図3 学習サンプルで決定したパラメータによる対応づけの評価. 学習サンプルとテストサンプルに適用
Fig. 3 Experimental results according to threshold value.

良い精度で求められると期待されるが，学習サンプルによって学習事例の特性に合わせてしまった面もあると思われる．**図3**に，学習サンプルで決めたパラメータを用い，閾値を変化させたときの，再現率，適合率の変化を示した．

テストサンプルでの誤りは，対応する記事がないとき,「かなり密接に関連する記事」を出力したものが半数を占めた．残りの半数の誤りの原因をあげる．

- 同じ人物が関わる：橋本首相，米大統領
- 同じ場所が関わる：北朝鮮，シェラレオネ
- 社会面用語：逮捕，容疑者
- 省庁：大蔵省
- 一文字一致：日，ガ，リ，数字

これらの文字列がまったく関係ない事件などに同時に現れたため，ノイズとなった．この傾向はテストサンプルA，B間でほとんど違いがなかった．詳しい考察を次章で述べる．

5. 全体の考察

5.1 関連しない組での一致文字列の特徴

直観的には，長い文字列は直接対応する記事にしか現れず，大きな重みをつければ弁別の精度が向上すると思われている．しかし，今回の検討により，それは必ずしも正しくないことが明らかになった．人名，地名，一般用語にも長い文字列があり（特に片仮名で），対応しない組，ほとんど関係のない組にも共通して現れることがあるからである．

表10に，対応する組合せからほとんど関係のない組合せまでの，すべての組合せに対して一致する文字列を調べた結果を示す．表中の数は，それぞれの関連性に対するすべての組に現れる一致文字列の数を，文字列の長さごとに数えたものである．この表の一番下に示されている組数でそれぞれの数を割れば，1つの組あたりの一致文字列の数が求まる．これらの値から，文字列が長いものほど頻度が少なく，また関連性が低くなるほど頻度が少ない．これは自然である．

だが，ほとんど関連しない組合せ（表10中の「非関連」の部分）にも長い文字列の一致がいくつか見られる．これらの中で4文字以上の一致文字列を調べ，分類した結果，**表11**のように，人名，組織名，地名などの固有名詞や，分野に依存した普通名詞などがほ

表10 関連性による文字列一致の頻度の違い
Table 10 Difference of frequency of the matched strings according to relevance.

文字列長	学習コーパス				テストコーパスA			
	ほぼ対応	密接関連	少し関連	非関連	ほぼ対応	密接関連	少し関連	非関連
1	7578	1583	995	117576	6819	3682	1171	142539
2	3589	499	245	5573	2724	1062	232	6814
3	892	101	31	462	688	360	42	688
4	458	39	18	93	388	122	6	105
5	229	24	6	18	121	26	3	27
6	97	2	8	4	70	3	4	8
7	42	2			24			
8	7			1	1	2		
9	2				9			
10	9							
組数	79	26	20	2621	76	61	22	4465

表11 ほとんど関連しない組に現れる長い一致文字列の傾向
Table 11 Characteristics of long strings which are matched between non-relevant pairs of captions and newspapers.

字数	文字列の種類	事例数 学習	事例数 テスト	文字列の例	例の出現したニュース・記事のそれぞれの内容 ニュース側	新聞記事側
8	人名+肩書	1	0	クリントン大統領	大統領選挙．ドール候補発言内．	反テロ首脳会議．議長声明．
6	組織名	4	4	オウム真理教	オウム真理教破産	予算案審議．オウムは一問題
	地名・地域名	0	4	シェラレオネ	シエラレオネで邦人殺害か．	リベリア，首都での戦闘激化．
5	人名	1	9	クリントン	大統領選挙．ドール候補発言内．	反テロ首脳会議．議長声明．
	地名・地域名	0	7	ワシントン	普天間基地返還で基本合意	日米保険協議．
	組織名	4	5	医道審議会	安楽死医師有罪判決．	診療科名意見書提出．
	固有名詞	0	1	衆院予算委	金融関連法 先送りも．	会社更生法の適用は困難．
	普通名詞	5	3	不動産会社	木津信乱脈融資．	弁護士弁護料着服．
	数+助数詞	8	2	000万円	長時間労働自殺賠償請求．	先生の制服．購入費予算．
4	人名+肩書	7	12	橋本首相	薬害エイズ和解．	国連常任理事国入り支持．
	肩書	2	6	事務次官	薬害エイズ．	国民所得統計速報．
	組織名	12	12	大阪府警	木津信本店捜査．	ニュートラム暴走事故．
	国名	2	2	アメリカ	テロ対策首脳会議．	アメリカマナティー死亡．
	国名+肩書	1	0	台湾総統	台湾総督選挙．	中国，米下院決議を非難．
	地名	1	0	台湾海峡	台湾総督選挙．	中国合同演習活発化．
	固有名詞	1	1	首脳会議	テロ対策首脳会議．	ミャンマー観光年始動
	普通名詞	51	60	負担割合	薬害エイズ和解．	介護保険連合見解．
	数+助数詞	12	12	年12月	チョン大統領起訴事実否認．	ごみ処分場問題．
	その他	4	0	東京，大	エイズ薬害訴訟．	太平洋銀破綻．

とんどであることが分かった．これらを個別に見ると，以下のような特徴がある．

(1) 固有名詞の一致が見られる．人名（+肩書），組織名，地名・地域名など，一般には出来事に固有と思われている固有名詞が，ほとんど関係のない組合せにも共通して現れることがある．その典型的な例は，すでに述べたように，「クリントン大統領」が米大統領選挙の記事にも反テロ首脳会議にも現れた事例である．より詳細な関係を扱うことが必要になった場合には，これらの固有名詞を中心とした観点からの検索をするということも考えられるが，今回のタスクではノイズとなる．これは後で考察する関連性の範囲の問題と密接に関わる．

(2) 普通名詞の一致が見られる．特に政治や経済，社会などの分野に依存した普通名詞が一致することが多い．これらはおおまかな分野を特定する場合には役立つが，細かく区別するには妨げとなる．また，「記者会見」などのように，異なる分野にまたがって共通に現れるものもある．

(3) 実際には固有名詞か普通名詞かが区別しにくいものも多い．たとえば「事務次官」や「大統領」が固有名詞か普通名詞かは状況に依存するし，1度現れた固有名詞の一部が普通名詞によって表現されることもある．すなわちこれらは名詞の指示性とも密接に関係する．また，メディアの片方では固有名詞，もう一方では普通名詞の場合もある．

(4) 上の問題に加え，さらに部分一致の問題がある．以下のようなさまざまな場合の部分一致が起こり，本稿で提案するような手法に対してノイズとなりうる．
- 組織名が肩書に含まれることがある（例：新王子製紙会長）．
- 異なる組織名の一部が一致することがある（例：日本住宅金融と住宅金融専門会社）．
- 組織名の一部と普通名詞が一致することがある（例：コスモ信用組合と信用組合）．
- 国名と普通名詞の一部が一致することがある（アメリカマナティー）．
- 国名と組織名の一部が一致することがある（アメリカオンライン）．
- 異なる固有名詞の一部が一致することがある（テロ対策首脳会議と ASEAN 首脳会議）．
- 異なる普通名詞の一部が一致することがある（与党首脳会談と日米首脳会談）．
- カタカナの普通名詞の一部が一致することがある（ペースメーカーとメーカー）．

(5) たとえば「メーカー」のように，カタカナの普通名詞は長いため，漢字の場合よりも一致する長さが長くなる傾向が見られる．

(6) 数+助数詞：経済の分野での処理を考慮して，今回は含めたが，長くなる傾向があり，ノイズになりやすい．今後，内容を解析して除くことも考えられる．

(7) その他，句読点やコーパス特有の記号などによって，期せずして長い文字列を作る可能性がある．今回，特殊記号などをすべて除く前処理を行った．

これらの特徴から，長い一致文字列は関連性の低いものに対しても本質的に現れうるものであることが分かる．

5.2 関連性の問題

上の考察は関連性の要求の度合にも強く依存する．たとえば「アトランタオリンピック」などのきわめて長い語は，特により細かい種目まで弁別する場合などでは，著しく類似度や閾値に影響し妨げとなる．

今回は，あらかじめ人手で正解を作る際，(1) ほぼ完全に対応する，(2) かなり密接に関連する，(3) 少し関連する (4) ほとんど関連しない，の 4 つの分類を行ったが，ほぼ完全に対応する記事を取り出すことのみを目的としたため，大変厳しい評価となった．学習サンプルの最終的な誤りはすべて「かなり密接に関係する記事」だった．新聞の関連記事の抽出でも生じる問題だが，関連性を明確に定義することは難しい．関係の度合が視点・観点に依存することが多いからである．特に，政治と経済の事柄は互いに結び付きが強く，内容の相違が明確でないことが多い．また国際面では主要人物が限られ，同じ人物が異なる記事に現れ弁別を妨げることがある．今後これらの分野ごとの性質の違いを考慮し，精度を向上することが考えられる．

5.3 本手法の特徴と課題

本稿で提案した手法では，形態素解析，構文解析，意味解析をいっさい行っていない．これらの処理はまだまだ時間のかかる処理であるため，現在のところ，これらを使わないことによって処理効率を高めることができる．また辞書の作り方に依存せず，文字列一致による手法の性質を明確にして議論できる．その結果，文字列の長さに応じた重みづけには注意を要することが分かったが，本手法では，短い文字列であっても出現回数が多ければ関連性が高いことを支持することとのバランスをとることによって，全体の精度を高めている．

この文字列の長さの問題の本質は，出現する語がどの程度日常的か，非日常的かに依存している．すなわち，対象となる出来事に完全に固有な語であればノイズとはなりえないが，ありふれた語であれば，ノイズとなりうる．その 1 つの指標として，ここでは一致文字列の長さを用い，その性質を論じた．単語が文書にどの程度固有なものであるかをおおまかに反映した IDF[16] などを考慮に入れることも考えられるが，全体のニュースと記事の組合せが 7000 程度であり，問題となる各文字列の出現する組合せが 1 から 10 程度であることから，IDF を考慮したとしても重みづけの範囲は 1.5 倍程度までにとどまり，文字列の長さの重みづけの程度に対する考察を変えるものにはならない．逆に長い一致文字列は頻度が少ないため上の問題を際だたせることにもなりうるし，稀な語の出現回数と日常的な語の出現回数の比が通常と逆転している場合も多く，問題が繁雑になるため，本稿では考慮の対象から外した．だが，全体の精度に対して若干寄与する可能性は残されており，その調査と考察は今後の課題とする．

もう 1 つの解決方法は，キャプションの内容を調べ，意味まで積極的に踏み込んで目的に合わない部分を削除することである．今回問題になった「クリントン大統領」の長い文字列も，ドール大統領候補の発言の翻訳部分に現れたが，形態素解析を行い体言止めでないものを除く[15] など，重要でない発言の部分などを削除

することが考えられる．また画像中でのキャプションの文字の大きさや位置などによって意味上のより細かい違いをとらえ，ノイズとなるような語を除いていくことも考えられる．

文字列長の1つの意味づけは，複合語の構成要素数を近似していると考えられることである．その例外は，カタカナの文字列で，長さが一定しない問題がある．将来，より高速で，未知語の形態素解析が精度良く行える形態素解析システムが開発されれば，それを利用し，文字列長の代わりに形態素数を数えるなどの対処をすることが考えられる．

本研究では，テストコーパスとして，学習コーパスの約2週間後のデータ（A）と，約7カ月後のデータ（B）を用いてテストを行った．これらのコーパスで扱われた出来事はほとんどが互いに内容が異なるものであった．わずかにオウム真理教による事件に関する報道が重なったが，テストコーパスAでは事件の解明を，そしてテストコーパスBでは裁判の行方を報じたものであり，出現する語句はかなり異なっている．また，国会内の様子などが報道されることもしばしばであるが，テストコーパスAでは「空転国会」が，そしてテストコーパスBでは「社さ三者協議」や「橋本内閣組閣」などが報じられるなど，異なる話題が扱われている．研究当初，2週間ごとにパラメータ値を決め直す方法をとることを想定していたが，これらの2種類のテストコーパスに適用した結果，ほぼ同じ高い正解率が得られたことから，一度パラメータ値を決定すれば，長期間にわたり良い正解率で本手法を適用できることが分かった．

5.4 新聞記事と映像中の音声言語との違い

ニュースの報道の本質的な点は，画像や音声を未加工のまま，あるいは一部編集する程度で情報を視聴者に伝えられることにある．すなわち，その内容は実況的であり，視聴者に現場の状況，発話者の態度や発話内容などの全体のニュアンスなどが直接伝わるという利点がある．このような情報は，現在の技術では，情報検索などの何らかの計算機処理の後，人間が直接閲覧することが望ましい内容を持っている．映像に付与された音声言語はそのために，画像や音響などの様々な情報と統合されて認識する必要のある形になっており，独立した言語としては扱いにくい面がある．表1で示されたキャプションの元となる番組の音声言語を人手によって書き起こしたものを**表12**に示す．これは表2の新聞記事の例と同じ出来事を扱っているものである．この表12の書き起し例から，音声言語は非文が大変多いことが分かる．このような文に対して形態素解析や構文解析を行っても，誤りなく解析することはほぼ不可能であり，それぞれの文からの論理的・構造的な情報はほとんど得られないと考えられる．

表12 TV映像中の音声言語の書き起こし例
Table 12 Example of speech from TV image.

え 次は景気についてのニュースです え 日銀の支店長会議が あ 今日から始まりました え 全国の各支店長から 企業の業績が改善していることなどを背景に景気は緩やかに回復しつつあると いう報告が相次ぎました 会議で 挨拶に立った 日銀の松下総裁は金融と財政面からの 政策効果もあって え 国内の需要は 住宅投資や 公共投資を中心に増加し 生産も緩やかに増加するなど 景気は 緩やかながら回復しつつある という認識を改めて示しました また この後行なわれた 全国の 支店長の報告でも 個人消費や企業の業績が改善していることなどを背景に 景気が緩やかに回復しつつあるという見方が 相次いで 出ました え 今後の景気の見通しについては 現在の景気を支えている うー公共投資の効果が薄れる夏以降が どうなるのかが 一つの ポイントです えー二人の 専門家に話を聞きました え 春闘では まぁー最低昨年並の まぁー 2.8 %から－3 %ぐらいまでのですね まぁーあのベースアップが期待できますし またあの夏冬のボーナスもですね え 少し増えていくでしょうから まぁ 雇用調整 とはいってもですね 所得は まぁ 2 %強の 伸びは十分に期待できると思いますね え 電気中心 大企業中心の 設備投資が だんだんとそのサービス とか え通信 それから 運輸の方へですね 広がりを見せていますし また昨年 10 12月には－あの－資本金1億円未満のですね 中小企業の－ 設備投資が2桁の伸びを示しているわけです ですから 確かに公共投資 住宅投資はですねー 伸び悩むと思いますけども 消費と設備投資で 2 %台半ばのですね 成長は十分に期待できる と思いますね 個人消費はーやはりー底揺れの兆しは一部ありますけどね やはりーまだまだ低迷状態が続いていると それからー民間の設備投資につきましては まぁこれはあのー ここへ来て少し明るくなってきておりますけども やはり その 中身を見てみますとね やはりあのー 更新投資とか 補修投資が中心になって おりまして やはり能力増強投資みたいな そういうものが日本国内で あまり行なわれていない 日本経済の低迷というものが 構造的なところにある 以上 まぁ あのー何と言いますか 公共投資の 支出が続いている限り そのーぅ 経済成長はある程度続くかも しれませんけども それが切れると そういう麻薬が切れると 景気が低迷する というような そういうような あー姿に なりかねないと そういう懸念があるという風に考えているわけで ございますえ このようにーぃ 景気の先行きについては 公共投資が 一段落した後え このまま あ民間が主導する形で 自律的に 回復に向かうのかどうかあ経済専門家の間でも 意見が分かれておりまして え景気は ＊＊＊<不明>難しい局面を向かえている と言えそうです

前述のCMUのシステムのように，そのような情報を得ることをせず，キーワードを抽出することに処理を絞ることを考えたとき，得られるであろうキーワードは，表1で示されるようなキャプション中に現れる語句程度であろうと思われる．実際に表12の書き起こし文章の中で，表1のキャプションで示された語句を下線で示したが，これを見ると，これらの語句で，文章の内容がほぼ代表されていることが分かる．非文や，論理的に錯綜した部分などの存在を考えると，むしろキャプションの方が内容をより正確に表現している場合も少なくない．

音声言語全体の談話に目を向けると，映像主体であり，音声はその映像を説明する役割が大きいため，音声言語全体の談話は論理的でない部分も多い．また発話も比喩が多かったり，同じことの繰返しもみられる．そして，画像のキャプションなどと結び付けないと話者が特定できないなど，音声言語単独では談話の一貫性が乏しい．内容そのものは，出来事そのものを伝えることが中心となっており，過去の出来事との関連性など，出来事の全体の中での位置づけなどの情報が少ない．

これに対し，新聞記事は一連の事件の中での関連性，一般的な事実とのかかわり，社会に対する影響，未来予測，記者の見解などが必要に応じて盛り込まれ，必要十分に論理的に要約されている．また，だれが言ったかなどの指標が明確であるなど，曖昧さをほとんど含まずメディア単独で情報を伝えることができる．また現状の形態素解析や構文解析を新聞に適用した際の精度の高さから，論理的な情報の抽出が比較的容易であるなどのメリットがある．

これらを考慮し，本研究では映像に対応する新聞記事を自動的に抽出することを目的とした．だが上述のように映像自体にもモーダルなど新聞記事にはない情報も多く含まれている．将来，このようなモーダルを適切に統合する研究が進めば，映像中の音声情報もより積極的に利用していくことが望ましいと考えられる．

6. おわりに

TVニュースの各報道に新聞記事を一致文字列の長さ，位置，頻度などにより対応づける手法を提案し，学習サンプルによってパラメータを決定した結果，学習サンプルに対して再現率100.0%，適合率93.2%が得られた．そのパラメータと閾値を使ったとき，約2週間後のテストサンプルでは再現率98.0%，適合率77.8%に，また約7カ月後のテストサンプルでは再現率97.1%，適合率79.5%になった．それぞれのテストコーパスで閾値のみ決め直したときは，約2週間後のテストサンプルでは再現率98.0%，適合率84.5%に，また約7カ月後のテストサンプルでは再現率94.1%，適合率85.3%になった．この結果から，パラメータの値を調整したサンプルから時期の離れたサンプルに対しても，かなり良い正解率で対応づけをすることができることが分かった．

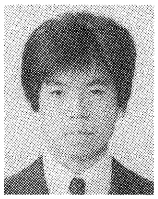
各種パラメータを検討した結果特に，一般の常識に反し，長い一致文字列に重みを与え過ぎてはならないことが明らかになった．本稿で述べた手法と考察は一般性があり，新聞記事どうしの関連記事の抽出[1),2)]など多くの応用事例に適用可能と思われる．

参考文献

1) 奥 雅博，鷲崎誠司，田中智博：関連記事の判定に関する検討，言語処理学会第2回年次大会発表論文集，pp.89–92 (1996).
2) 新谷 研，角田達彦，大石 巧，長尾 眞：形態素の共起頻度と出現位置による新聞の関連記事の検索手法，情報処理学会論文誌，Vol.38, No.4, pp.855–862 (1997).
3) 長尾 眞：電子図書館（岩波科学ライブラリー15），岩波書店 (1994).
4) Schatz, B. and Chen, H.: Building Large-scale Digital Libraries, *IEEE Computer*, No.5, pp.22–26 (1996).
5) 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，情報処理学会論文誌，Vol.33, No.4, pp.543–550 (1992).
6) Tonomura, Y.: Video Handling Based on Structured Information for Hypermedia Systems, *International Conference on Multimedia Information Systems '91*, pp.333–344, McGrow-Hill, Singapore (1991).
7) 柴田正啓：映像の内容記述モデルとその映像構造化への応用，電子情報通信学会論文誌（D-II），Vol.J78-D-II, No.5, pp.754–764 (1995).
8) 井手一郎，田中英彦：マルチメディア・データベースに要求される画像の描写に関する研究，第51回情報処理学会全国大会論文集，Vol.3, pp.303–304 (1995).
9) 有木康雄：音声・ビデオデータの自己組織化方式の研究，自己組織化型情報ベースシステムと利用に関するシンポジウム予稿集，pp.2-2-1–2-2-8 (1993).
10) Wactlar, H., Kanade, T., Smith, M.A. and Stevens, S.M.: Intelligent Access to Digital Video: Informedia Project, *IEEE Computer*, No.5, pp.46–52 (1996).
11) 美濃導彦：知的メディア検索技術の動向，人工知能学会誌，Vol.11, No.1, pp.3–9 (1996).
12) Sakai, T.: A History and Evolution of Document Information, *Proc. 2nd International Conference on Document Analysis and Recognition* (1993).
13) 木本晴夫：日本語新聞記事からのキーワード自動抽出と重要度評価，電子情報通信学会論文誌（D-I），Vol.J74-D-I, No.8, pp.556–566 (1991).
14) Wilkinson, R.: Effective Retrieval of Structured Documents, *Proc. 17th ACM SIGIR*, pp.311–317 (1994).
15) 渡辺靖彦，岡田至弘，角田達彦，長尾 眞：TVニュースと新聞記事の対応づけ，情報処理学会研究報告，NL-96-114 (1996).
16) Sparck J.K.: A Statistical Interpretation of

Term Specificity and Its Application in Retrieval, *Journal of Documentation*, Vol.28, No.1, pp.11–21 (1972).

(平成8年10月24日受付)
(平成9年 4月 3日採録)

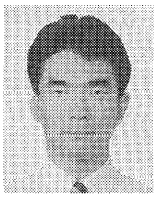


角田 達彦（正会員）

1967年生．1989年東京大学理学部物理学科卒業．1995年東京大学工学系大学院博士課程修了．工学博士．同年京都大学工学研究科助手．1997年東京大学医科学研究所特別研究員，現在に至る．IJCNN'93 Student Award受賞．1994年情報処理学会学術奨励賞受賞．岩波ソフトウエア科学第15巻「自然言語処理」（岩波書店，共著）．言語処理学会，日本神経回路学会，電子情報通信学会，人工知能学会，日本認知科学会各会員．

大石 巧

1971年生．1996年京都大学工学部電気工学第二学科卒業．同年，同大学大学院修士課程進学，電子通信工学専攻在学中．情報検索の研究に従事．言語処理学会会員．



渡辺 靖彦（正会員）

1966年生．1991年京都大学工学部電気第二学科卒業．1993年同大学院工学研究科修士課程電気工学第二専攻修了．1995年同博士課程退学．同年龍谷大学理工学部助手，現在に至る．自然言語処理，言語と画像の統合の研究に従事．言語処理学会，人工知能学会各会員．

長尾 眞（正会員）

1936年生．1959年京都大学工学部電子工学科卒業．1961年同修士課程修了．1965年京都大学工学博士．1961年京都大学工学部助手，1968年京都大学工学部助教授，1973年京都大学工学部教授，現在に至る．1976～94年国立民族学博物館併任教授，1986～90年京都大学大型計算機センター長，1995年～京都大学附属図書館長．専攻は自然言語処理，画像処理，人工知能．情報処理学会論文賞（4回），通産大臣賞，郵政大臣賞，京都新聞文化賞，IEEE Emanuel R. Piore賞など．電子情報通信学会副会長（1993～95年），情報処理学会副会長（1994～96年），言語処理学会会長（1994～96年），日本認知科学会会長（1988～90年），機械翻訳国際連盟初代会長（1991～93年），アジア・太平洋機械翻訳協会会長（1991年～），パターン認識国際委員会副会長（1986～88年）．著書：論理と意味（岩波書店），知識と推論（岩波書店），機械翻訳はどこまで可能か（岩波書店），電子図書館（岩波書店），岩波情報科学辞典（岩波書店），自然言語処理（岩波書店），画像認識論（コロナ社），パターン情報処理（コロナ社），言語工学（昭晃堂）ほか．