自然言語処理 115－11
情報学基礎 43－3
（1996. 9. 12）

キャプションと記事テキストの最長一致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

京都大学　工学研究科　電子通信工学

〒606-01 京都市左京区吉田本町

[+]龍谷大学 理工学部 電子情報学科

〒520-21 大津市瀬田大江町横谷 1-5

e-mail: tsunoda@kuee.kyoto-u.ac.jp

要旨

本稿では，ニュースのキャプションと記事の文字列の照合により各報道に対応する新聞記事を特定する手法を提案する．一致文字列の長さ，キャプションと記事中での出現位置により重みづけし，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種パラメータを決定した結果，学習サンプルで再現率 100 %，適合率 93.2 %，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 %) 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過ぎることの弊害を明確にした．

キーワード　情報検索，ニュース，キャプション，新聞，最長一致文字列照合，対応づけ

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

Tatsuhiko TSUNODA　Takumi OOISHI　Yasuhiko WATANABE†　Makoto NAGAO

Department of Electronics and Communication, Kyoto University

Yoshida-honmachi, Sakyo, Kyoto 606-01, Japan

†Department of Electronics and Informatics, Ryukoku University

Seta, Otsu, Shiga 520-21, Japan

e-mail: tsunoda@kuee.kyoto-u.ac.jp

Abstract

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, and 98.0 % recall and 77.8 % (84.5 % if the threshold was readjusted) precision for unseen samples. We also clarified the problem of excessive weighting for long strings.

Keywords　Information Retrieval, News, Caption, Newspaper, Maximum Length String Match, Alignment

1 はじめに

TV のニュースでの報道と，新聞記事は，製作過程は全く独立である．しかし，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も多い．すなわち，新聞のように文書を中心とした媒体と，TV のように画像・音声などの映像を中心とした媒体との，複数の媒体により同一の出来事が表現されているととらえることができる．

逆に，それらの対応をとらえられれば，一つの事柄を複数の観点から多角的にとらえ直すことができる．また，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，一連の事柄の中の探索を可能にし，より柔軟な検索を行なえる．そして，新聞の関連記事の抽出の技術 [1] 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することができる．

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，あるいは密接に関連する新聞記事を自動的に対応づけする手法を提案する．TV ニュース映像による報道の内容を知る一つの手がかりとして，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．例えば，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 [2, 3]．このようなキャプションと，新聞記事内の文字列の照合を行ない，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 [4]．

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．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．本稿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つくか，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．

表 1: キャプションの典型的な例とキャプションの型

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

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されていることが多い．例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[1]．

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．

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，

[1] 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

景気の回復改めて確認――日銀支店長会議　「金利維持」触れず	見出し
（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

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わかる．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な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わかりやすい飾りをつけているため，特定が容易である．

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わす内容そのものと，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[2]，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一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とりだした一つの記事の間で，類似度を計算する．類似度計算は，次の手順で行なう．
   (a) キャプションと記事とで文字列の照合を行ない，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．

[2] 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

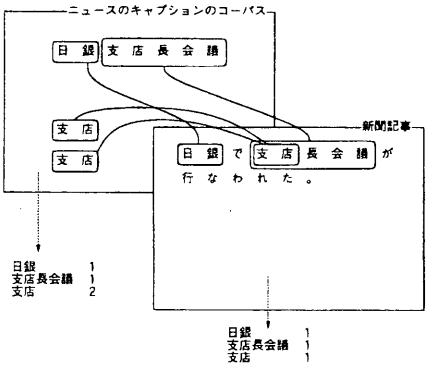

図 1: 各箇所での最長一致文字列の抽出の様子と頻度の数え方

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，それが予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.

例えば図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なわれた」という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一回現れ，そして「支店長会議」もそれぞれのコーパスに一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二回，新聞記事に一回現れたとみなす．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[3].

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みなした．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.

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

$W_l(\|k\|)$	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.

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

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.

[3] 本稿では，これを最長一致とよぶことにする.

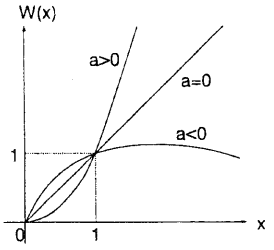

図 2: 重み関数 W(x) の振舞い

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わ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わす．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わす．これらは各位置での文字列の頻度を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) で表わされるように，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．

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，記事数を表わす．

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

どちらもスポーツニュースを除いてある[4]．3 月の 19 日分を学習コーパス，4 月の 13 日分をテストコーパスとした．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一つ選択し，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なう．

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

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

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

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．位置による重みだ

[4] スポーツはどちらの場合も特殊な作り方をしているため，今回の対象外とした．

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．$W_i(x)$ 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なった．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．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．

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わかる．

次に，指数部のパラメータは $a = 0$ よりも，$a = 1$ の方がよい．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．例えば，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，関係がないのに関わらず，「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わかる．

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

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 %

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

$W_{news}(2)$	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 %

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わかる. $W_{arti}(3)$ 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と, 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わかる.

上の結果をみ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とりだ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.

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に示す. 調整の結果, 学習サンプルで閾値を決めた場合は再現率 98.0%, 適合率 77.8% となった. 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,

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

$W_{arti}(3)$	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

$W_{arti}(4)$	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 %

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わかる. すなわち, 重みの設定などは精度の向上につながっている. また, 初期値を直観で設定したため, もともと汎用的に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

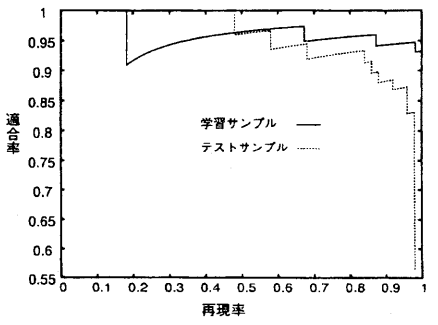

図 3: 学習サンプルで決定したパラメータによる対応づけの評価. 学習サンプルとテストサンプルに適用.

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に適用した結果

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	2.5	87.8	98.0 %	77.8 %	91.2	98.0 %	84.5 %

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

これらの文字列が全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．

5 全体の考察

直観的には，長い文字列は直接対応する記事にしか現れなく，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），関係のない記事にも現れることがあるからである．むしろ，短い文字列でも，出現回数で効くことが多い．この事は関連性の要求の度合にも強く依存する．例えば「アトランタオリンピック」などの極めて長い語は，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．

一つの解決方法は，キャプションの内容を調べ，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，形態素解析を行ない体言止めでないものを除く [4] など，重要でない発言などの部分を削除することが考えられる．

今回は，予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つの分類を行なったが，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つきが強く，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．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．

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．テストサンプルでは，そのパラメータと閾値を使ったとき，再現率 98.0%，適合率 77.8%，閾値のみ決め直したときは再現率 98.0%，適合率 84.5% になった．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] など多くの応用事例に適用可能と思われる．

参考文献

(1) 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関連記事の検索手法, 電子情報通信学会技術研究報告 NLC96-1, (1996), pp. 1–8.

(2) 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, (1996), pp. 3–9.

(3) Sakai, T., A History and Evolution of Document Information Processing, *2nd International Conference on Document Analysis and Recognition*, (1993).

(4) 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).