TR−SLT−0015

ニュース記事単位のトピックに基づいた テキストセグメンテーション

Text segmentation based on the topic of a news article unit

西脇 正通
Masamichi Nishiwaki

2002 年 7 月 16 日

概要

ATR が保有するデータベースの内、「あすを読む」書き起こし(日)の各文に対し、NHK ニュース(日)、日経新聞(日)から関連記事を抽出するとともに、その情報を利用してニュース記事単位のトピックに基づいたテキストセグメンテーションを「あすを読む」書き起こし(日)の各番組で行った。実験では簡易的に文脈を考慮する手法も採用しながら、比較的小規模なコーパスに対し大規模コーパスの情報を利用することでスパースネスを解決する手法を提案する。実験の結果、提案手法で行ったセグメンテーションの簡易的な精度が 43.5%となり、先行研究の 36.2%とくらべ向上したことが分かった。

(株) 国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619−0288「けいはんな学研都市」光台二丁目 2 番地 2 TEL：0774−95−1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City" 619-0288, Japan
Telephone:+81-774-95-1301
Fax :+81-774-95-1308

目次

図一覧

表一覧

1．はじめに

ATR では、独話ニュース解説番組「あすを読む」の書き起こしコーパスを構築している。このコーパスは放送ニュース原稿や新聞記事と比べひとつの記事が長く、ニュース解説の性格として番組内では過去の複数のニュースを引用し、さらに話者独自の見解も含んでいる。

今回は、簡易的に文脈を考慮しつつ、「あすを読む」のある1文に対して、異なるコーパスであるNHK放送ニュース汎用原稿コーパスと日経新聞コーパスそれぞれの記事から、関連性の高い記事を抽出した。その結果を利用して、ニュース記事単位を基準に、トピックに基づいた文単位のテキストセグメンテーションを、「あすを読む」の書き起こし原稿に対して行ったので、報告する。

2．単語の分布に基づくテキストセグメンテーション

話題の境界を検出し一連のテキストをセグメンテーションする手法には、固有表現を手掛かりにする方法や、テキスト内での単語の出現分布からベクトル空間モデルを作成し、ベクトルの類似性を利用する方法などがある。

単語の出現分布によるベクトル空間モデルを利用してのセグメンテーションは、長いテキストを大きなセグメントに分ける場合に向いた手法である。「あすを読む」の書き起こしのように、1 番組中に複数の話題がある比較的短いテキストにこの手法を適用する場合、対象領域内で観測したデーターのスパースネスが顕著に影響する。このため、シソーラス辞書の利用や、共起情報による単語のクラスタリングを併用しなければならない。また、現在構築中の「あすを読む」の書き起こしコーパスは、現時点で単語のクラスタリングを学習できるほどの十分な量はない。

小規模のコーパス中のテキストを扱うために、単語のクラスタリングを他のコーパスから学習する手法があるが、今回はこのような比較的短いテキストに自動で頑健なセグメンテーションを行うため、他のコーパスから単語のクラスタリングを学習せずに、そのコーパスの記事単位をベクトル空間モデルで直接利用した。

今回の手法は、ベクトルの各次元が大規模コーパスの各記事と1対1で対応しているため、コーパスから得た情報の正確さや性質、セグメンテーションの根拠を把握しやすいことも長所である。

3．文書ベクトルの計算

ベクトル空間モデルの作成では、まず「あすを読む」のある1文中に出現する単語集合を用い、その文の特徴をあらわすこととした。それぞれの単語には重み付けを行い、索引語を各次元とする文書ベクトルを作成する。

今回の重み付けでは、3つのコーパスそれぞれにおける索引語の出現文書数(DF)から取得した各索引語の残差 IDF(RIDF)と、「あすを読む」の対象となる 1 文から取得した索引語頻度(TF)を使用した。ただし、計算の簡略化のために、「あすを読む」コーパスについても1番組を1文書として計算した DF を採用した。

ある索引語の生起確率が独立しているときの IDF は、各事象が独立して生起することを仮定したポアソン分布により推定できる。ポアソン分布は、観測データーが十分な量である場合に二項分布から近似して導かれ、文書集合全体での索引語の出現頻度とコーパスの文書数から、1文書内にk回生起する確率を求めることができる。

生起確率が独立していない索引語、つまり文書中に一度出現すると何度も出現する内容語(content word)の IDF は、non content word と仮定してポアソン分布で推定した IDF より大きくなる。観測した IDF と推定した IDF の差をとることで、RIDF は索引語がどの程度重要な内容語であるかを示している。

（索引語重み）＝（TF）・（RIDF）
（TF）$= \log(1+f_i)$
（RIDF）＝（IDF）－（ポアソン分布により推定されたIDF）
$= \log(n/n_i) - \log(n/n(1-P(0;\lambda_i)))$
$= \log(n/n_i) + \log(1-e^{-\lambda i})$
ポアソン分布 $P(k;\lambda_i) = e^{-\lambda i}\lambda_i^k/k!$
期待値 $\lambda_i = F_i/n$
F_i＝（索引語 w_i の文書集合全体での出現頻度）
f_i＝（索引語 w_i の文書中における出現頻度）
n＝（コーパスの文書数）
n_i＝（索引語 w_i の出現文書数）
k＝（仮定する文書集合全体での出現頻度）

4．簡易的な文脈の考慮

「あすを読む」の各1文はそれ単独で見た場合、その1文中に出現する単語だけでは、関連記事との対応付けに十分な情報が含まれていないことがある。そこで、文書ベクトル作成時の対象範囲を複数文とする必要がある。また、「あすを読む」の書き起こしコーパスは番組単位で分割されているが、1番組中での明確な段落分けは行われていない。しかし、1番組には複数の事件などが述べられており、ある文書ベクトルを作成する対象範囲に話題の境界があると、適切な文書ベクトルを取得できない。そこで、今回はハニング窓を使用したフレーム化（図 1）を行い、索引語の出現位置から計算されるハニング窓関数の値を乗じて補正した TF をフレーム毎に求めた。TF の補正には単語単位、文字単位で細かく変化するハニング窓関数を使用することも考えられるが、ここでは、「あすを読む」の内容が通常の話し言葉に比べより推敲された話し言葉であることから、1文内での話題転換が少ないと考え、文単位を基準とした。また、フレームの間隔、つまり最初の文から最後の文までスイープするときの移動単位は1文とした。フレーム長は対応付けを行う記事の長さ、取得するベクトルデーターの規模などを考慮し9文とした。フレームの中央の文にはハニング窓関数の最大値1を適用する。得られる TF・IDF を各次元の値とする文書ベクトルは、この中央の文の文脈を考慮した文書ベクトルといえる。窓関数に分解能を考慮した適切な関数を採用することで、フレーム長をさらに長く取ることも考えられる。

ハニング窓関数 $h_l(i) = 0.5(1+\cos 2\pi((i-l)/W))$ （ $|i-l| \leqq W/2$ ）
W＝（フレーム長）
l＝（フレームの中央の文番号）
i＝（対象の文番号）

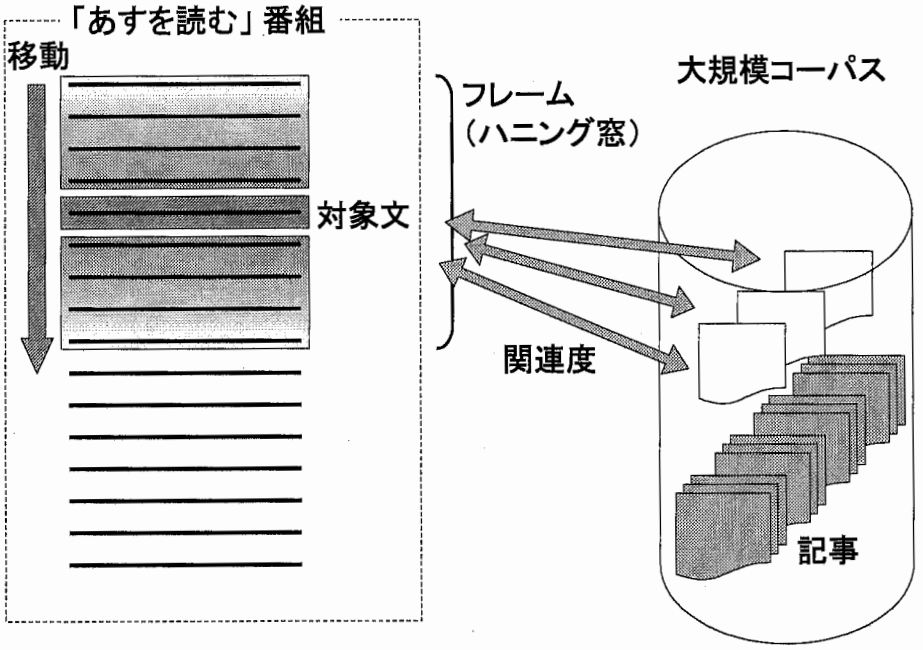


図1　フレームを利用した関連記事の抽出

5．実験

以下の手順で実験を行った。

(1)　あらかじめ形態素解析した表1に示す3つのコーパスについて、標準形と品詞の組を索引語とした索引語頻度(TF)と文書頻度(DF)を、それぞれ求めた。各コーパスには、日本語と英語の2言語が含まれるが、今回は日本語を対象とした。

Corpus	Period	Articles (K)	Sentences (K)	Words (K)	Words/Sent.
NHKコーパス 放送ニュース汎用原稿	4/1995 - 3/2000	287	1,593	75,602	47.45
日経新聞コーパス	1/1995 - 12/2000	1,758	18,620	499,405	26.82
「あすを読む」 書き起こしコーパス	11/1999 - 1/2001	0.305	18	539.3	29.07

表1　各コーパスの概要

(2)　TFとDFから、RIDFとIDFを求め、RIDFの性質を観測した。

(3)　ハニング窓を使用したフレーム化により、「あすを読む」コーパス各文の文書ベクトルを求めた。また、NHKコーパス、日経新聞コーパスの各記事の文書ベクトルを求めた。今回は、各コーパスと時期が重なっている「あすを読む」コーパスの最初の30番組を対象とした。(表2)

「あすを読む」の文 S_j を中央にするフレーム F_j の文書ベクトル

$$f_j = w_1 s_{j-4} + w_2 s_{j-3} + w_3 s_{j-2} + w_4 s_{j-1} + w_5 s_j + w_6 s_{j+1} + w_7 s_{j+2} + w_8 s_{j+3} + w_9 s_{j+4}$$

ハニング窓

$$W = [w_1\ w_2\ w_3\ w_4\ w_5\ w_6\ w_7\ w_8\ w_9] = [0.095\ 0.345\ 0.655\ 0.905\ 1\ 0.905\ 0.655\ 0.345\ 0.095]$$

文 S_j の文書ベクトル

$s_j = {}^t[s_{1j}\ s_{2j}\ \cdots s_{nj}]$　（nは索引語数）

索引語 w_i の文 S_j における重み

s_{ij} = （S_j での w_i の TF）・（w_i の「あすを読む」での RIDF）

NHKコーパス、日経新聞コーパスの記事 A_i の文書ベクトル

$a_i = {}^t[a_{1i}\ a_{2i}\ \cdots a_{ni}]$　（nは索引語数）

a_{ji} = （A_i での w_j の TF）・（w_j の各コーパスでの RIDF）

番組番号	話者 ID	放送日	タイトル
1	02	11/29/1999	死刑適用の基準
2	07	11/30/1999	消費者契約法制定へ
3	40	12/01/1999	雇用確保のジレンマ
4	25	12/02/1999	波乱の WTO 農業交渉
5	04	12/03/1999	村山訪朝団と日朝関係
6	02	12/06/1999	定期借家制度導入へ
7	08	12/07/1999	2000 年問題どこまで備えるのか
8	01	12/08/1999	インターネットが変える経済
9	15	12/09/1999	イスラム過激派との対決
10	12	12/10/1999	警察再生への道
11	31	12/13/1999	マカオ返還と台湾
12	14	12/14/1999	ロシア政治事情
13	03	12/15/1999	国会閉会と政局
14	17	12/16/1999	来年度の税制改正
15	13	12/17/1999	難航する医療保険改革
16	17	12/20/1999	どうなる景気と財政予算・大蔵原案
17	01	12/21/1999	迷走するペイオフ論議
18	10	12/22/1999	放射線被ばくと医療
19	01	01/03/2000	失われた 10 年の教訓
20	03	01/05/2000	2000 年政治の選択肢
21	14	01/06/2000	エリツィン退陣後のロシア
22	13	01/07/2000	団塊世代と高齢社会
23	18	01/11/2000	震災 5 年これからの課題
24	32	01/12/2000	ゴラン高原返還交渉
25	02	01/13/2000	何のための春闘か
26	13	01/17/2000	震災 5 年　これからの心のケア
27	12	01/18/2000	社会保障と年金
28	11	01/19/2000	米大統領選挙
29	03	01/20/2000	動きだす憲法調査会
30	01	01/21/2000	新たな摩擦？NTT の接続料

表 2　テキストセグメンテーションを行った番組

(4)　NHKコーパスと「あすを読む」コーパス、日経新聞コーパスと「あすを読む」コーパス、それぞれの文書ベクトル間のコサインを求めて「あすを読む」の各文に対しての記事の関連度 r とした。

「あすを読む」の文 S_j と記事 A_i の関連度
$r_{ij} = \cos(a_i, f_j) = (a_i \cdot f_j) / (\|a_i\| \|f_j\|)$
f_j は文 S_j を中央にするフレーム F_j の文書ベクトル
a_i は記事 A_i の文書ベクトル

(5)　NHKコーパスについて、関連度 r の高い記事の内容を確認した。
(6)　関連度 r の分布を確認した。
(7)　「あすを読む」コーパスの各文 S_j について、NHKコーパスおよび日経新聞コーパスの各記事との関連度 r の集合を、S_j の文書ベクトル d_j とした。次に、隣り合う文の文書ベクトル d_j、d_{j+1} のコサインを求めて文の結束度 c_j とし、この増減を番組内で求めた。ここで、2 つのコーパスに対するコサイン値 r の集合を混合してひとつの文書ベクトルにすると精度が向上すると予測されるが、混合による弊害の可能性もあるので今回は行わなかった。

文 S_j と文 S_{j+1} の結束度 $c_j = \cos(d_j, d_{j+1})$
文 S_j の文書ベクトル $d_j = {}^t[r_{1j}\ r_{2j}\ r_{3j} \ldots r_{mj}]$　（m は記事数）

(8)　一定の閾値以上の結束度 c_j をとる文 S_j、S_{j+1} の組を抽出し、共通のトピックを持つセグメントとした。それぞれのセグメントが適当であるか、基準を設け人手で確認した。

6．実験結果と考察

6．1．各コーパスの索引語数

各コーパスの索引語数は表 3 のようになった。

コーパス	索引語数
NHKコーパス 放送ニュース汎用原稿	141,705
日経新聞コーパス	712,198
「あすを読む」書き起こしコーパス	15,745

表 3　各コーパスの索引語数

6．2．IDF と RIDF の比較

表 4 に「あすを読む」コーパスで RIDF の大きいものから順にソートした索引語の例を示す。

上位の傾向としては、大量の DF=1 の索引語に対しても段階的にスコア付けされており、DF=2 以上の索引語に対しても高いスコアを付与している場合がある。「高まる」など DF が比較的低い索引語でも、出現頻度が低いと RIDF も低くなっている。「から」のようにほとんどの文書に出現する索引語は、IDF と同様に RIDF も低い。

「今晩は」「それでは今晩はこれで失礼をいたします。」などに使われる「それでは」「失礼」「今晩」が、RIDF では最下位に来ているところも興味深い。特に「それでは」「失礼」は全文書数 305 と比べ DF が十分小さい値である。

ポアソン分布より推定した IDF が実際の IDF よりも高い場合もある。予測した分布よりも実際の生起の分布の方がより分散する傾向が強いことを示しており、この場合 RIDF は負の値をとる。本実験のその後の計算では、負の値もすべて 0 とした。ただし、表 4 の例からもわかるように意図的に多くの文に分散して含まれているとも解釈できる。このような極端な値はNHKコーパス、日経コーパスには無かった。

索引語	出現頻度	DF	IDF	RIDF
古墳/名詞-一般	48	1	5.720311777	3.793544254
傍受/名詞-サ変接続	33	1	5.720311777	3.442896925
マカオ/名詞-固有名詞-地域-一般	31	1	5.720311777	3.383597935
野球/名詞-一般	63	2	5.027164596	3.348485974
フッ素/名詞-一般	28	1	5.720311777	3.286654006
トキ/名詞-一般	28	1	5.720311777	3.286654006
雪印/名詞-固有名詞-組織	25	1	5.720311777	3.178172145
タリバン/未知語	25	1	5.720311777	3.178172145
エーアールエフ/未知語	25	1	5.720311777	3.178172145
胚/名詞-一般	793		4.621699488	3.144121201
…				
れる/動詞-接尾	3446	304	0.003284	0.003272
高まる/動詞-自立	56	51	1.788486	0.003127
以前/名詞-副詞可能	43	40	2.031432	0.002657
から/助詞-格助詞-一般	2218	304	0.003284	0.002589
。/記号-句点	18550	305	0	0
ます/助動詞	17105	305	0	0
…				
それでは/接続詞	206	163	0.626562	-0.084643
失礼/名詞-サ変接続	150	145	0.743578	-0.201942
今晩/名詞-副詞可能	411	296	0.029952	-0.270989

表 4 「あすを読む」索引語の IDF と RIDF

索引語	出現頻度	DF	IDF	RIDF
ピブーン/未知語	14	1	12. 568334	2.639033
タトゥー/未知語	14	1	12. 568334	2.639033
木樋/名詞-一般	11	1	12. 568334	2.397876
…				
真須美/名詞-固有名詞-人名-名	2196	464	6.428450	1.550689
…				
ダイアナ/名詞-固有名詞-人名-名	1120	241	7.083538	1.534339
…				
行き/名詞-接尾-一般	14004	2993	4.564303	1.518795
…				
潮位/名詞-一般	201	99	7.973215	0.707835
円/名詞-接尾-助数詞	110860	45322	1.846787	0.707747
初日の出/名詞-一般	136	67	8.363642	0.707726
…				
ます/助動詞	1494405	286080	0.004308	-0.001217
ところで/接続詞	1593	1591	5.196216	-0.001515
。/記号-句点	1676098	287230	0.000296	-0.002636
トピックス/名詞-一般	2471	2471	4.755956	-0.004297

表 5 NHKコーパス索引語の IDF と RIDF

表 5 のNHKコーパスを見ると最下位は、株式市場の記事で使用される「トピックス」となっている。経済記事に多い「円」は、コーパス中に経済記事も多いため IDF はあまり大きくないが、RIDF は大きな値をとっている。RIDF に 0.7 以上をとる索引語は 9,401 語で索引語数の約 6.6%である。日経コーパスで RIDF に 0.7 以上をとる索引語は 17,780 語(2.5%)、「あすを読む」コーパスで RIDF に 0.7 以上をとる索引語は 1,653 語(10.5%)である。

表 6 の日経コーパスでは、丁寧語に使用する「ます」の RIDF が非常に大きくなっている。新聞では丁寧語を一般の記事に使わないことが影響している。今回は「あすを読む」コーパスでの RIDF が 0 のため影響が無かったが、新聞記事のクラスタリングには有効だと考えられる。句点の RIDF が上位 130,053(18.3%)にあった。句点を手掛かりに新聞記事をクラスタリングできるといえるが、これは今回使用した日経新聞コーパスが他のコーパスに比べ、表などの扱いにおいて未整備であることが原因と考えられる。

索引語	出現頻度	DF	IDF	RIDF
精励/名詞-サ変接続	1528	51	10.264969	3.399368
ｔｒａｄｅｒ/記号-アルファベット	29	1	14.196795	3.367286
チチタケ/未知語	25	1	3.218867	3.218867
…				
ます/助動詞	172578	48546	3.406528	1.209982
…				
真須美/名詞-固有名詞-人名-名	960	349	8.341723	1.011533
…				
ダイアナ/名詞-固有名詞-人名-名	1073	551	7.885060	0.666113
…				
行き/名詞-接尾-一般	3766	2996	6.191762	0.227450
…				
区画/名詞-サ変接続	7182	4052	5.889829	0.569915
円/名詞-接尾-助数詞	1532310	537319	1.002448	0.569894
濱陽/名詞-固有名詞-地域-一般	810	458	8.069926	0.569888
…				
。/記号-句点	13798346	1276074	0.137496	0.137415
…				

表 6　日経コーパス索引語の IDF と RIDF

6．3．関連記事の抽出結果

NHKコーパスから抽出された関連記事のうち上位10記事について、小規模であるが精度を観測した。観測には「その原稿を元にしていると認められる」「ほぼ同じ内容」「同じ話題」「内容を補足している」「話題がかけ離れている」の 5 段階の評価基準を使った。「あすを読む」の番組中には、解説者の考え、番組進行上必要な説明などが含まれており、そのような個所では精度が低かった。特に解説者の考えは、政治家のコメントや政府の発表、実際の状況などと反対の内容が多くなりがちなので、今回の評価基準では低い精度となった。逆に、特に番組の導入部では、状況の説明が多くなるので、精度は高くなる傾向が見られた。このように番組内でテキストセグメンテーションに向いた個所と向かない個所があり、向かない個所のベクトルモデルは安定しないことが確認できた。

評価が高かった対応関係は前後数ヶ月以内など、作成された時期が互いに近いものが多かったが、約1年間時期がずれている記事でも正しく対応付けされていた。さらに時期がずれている対応関係でも、「内容を補足している」と評価できるものが多く見受けられた。

索引語の中には行事名など対応付けの上で非常に強力であると思われるキーワードがあるが、異コーパス間での対応付けでは、例えば「WTO」と「ダブルティーオー」などのように表記の方針が違うため共通する索引語として使用できないものもあった。その他にも「エヌジーオー」が「エヌジー」と「オー」に分割されてしまうような形態素解析での誤りや、「広い範囲」と「包括的」のような換言もあった。しかし、その場合でも他の比較的一般的な内容語の組み合わせによって精度の高い対応付けが得られた例もあり、文脈を適切に考慮したことで、未知語が重要なキーワードになる場合の大幅な精度劣化を避けることができたと考えられる。

6．4．各記事に対する関連度の分布

「あすを読む」コーパスとNHKコーパスとの間で求めた文書ベクトルのコサイン値 r(関連度)について、その分布を図 2 に示す。ここでは、4 番組についてNHKコーパスの全記事に対する関連度の度数平均を求めたが、関連度が 0.1 以上の記事は「あすを読む」1 番組中に少ない記事で 1.39%、多い記事で 2.29%、平均 1.82%であった。番組によって分布に若干の違いがあるが、その割合と対応付けの精度の間に明らかな関連性は確認できなかった。

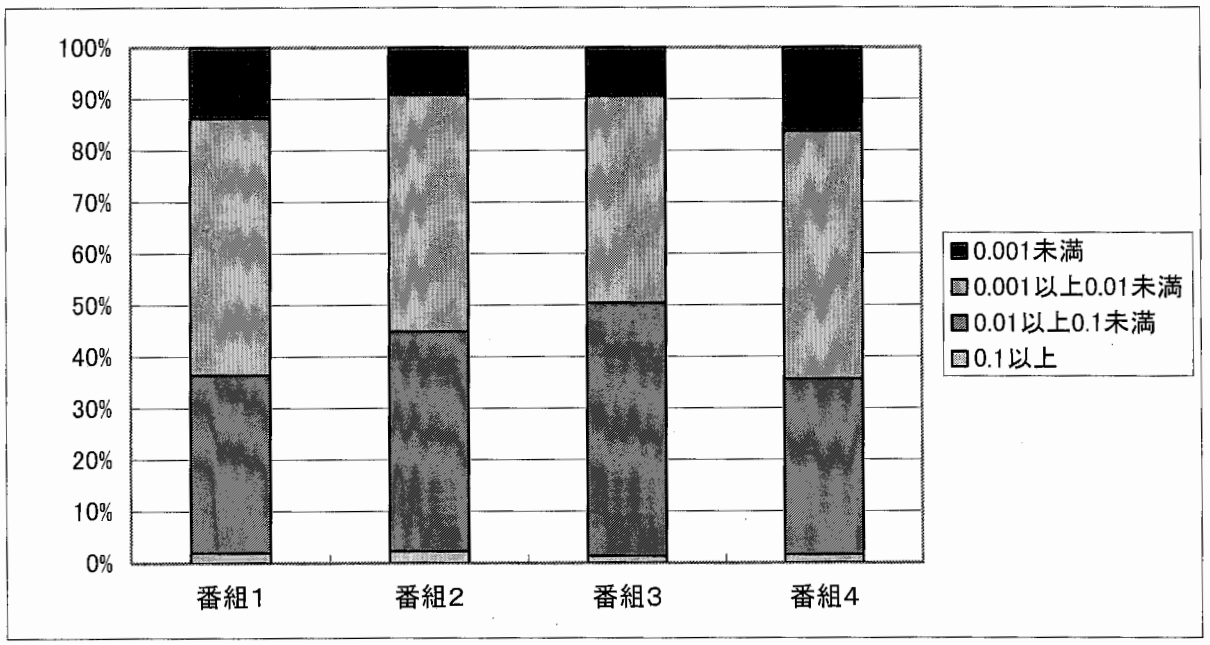


図 2 「あすを読む」コーパスとNHKコーパス 文書ベクトル間の関連度の分布

6．5．隣り合う文の類似度

「あすを読む」の隣り合う文間で、関連記事を各次元とした文書ベクトル間のコサインを計算し、文の結束度とした。その変化をグラフにして目視により観測した。(図 3)

NHKコーパスを使用した場合と日経新聞コーパスを使用した場合を比較すると、結束度の番組内での変化の傾向は類似していることが確認できた。これは、両コーパスの記事単位の内容が似通っているためと考えられる。

比較的傾向が似ていないものとして「警察再生への道」をタイトルにする番組が挙げられる。この番組で結束度の傾向が違う個所は、1 文 1 文がニュース記事としては比較的独立しており、日経新聞コーパスを使用した場合の方が安定していた。(図 4)これは、日経新聞コーパスで一連の話題をまとめた記事が多かったと予測できる。同一コーパスでも性質が違う記事があると、利用目的によっては得られる結果に影響を与えるので、話題だけではなく記事の用途などの性質によっても分別し

て使用する必要性がある。また、この例のような違いが原因となり、コーパスを混合した場合の弊害が予想される。

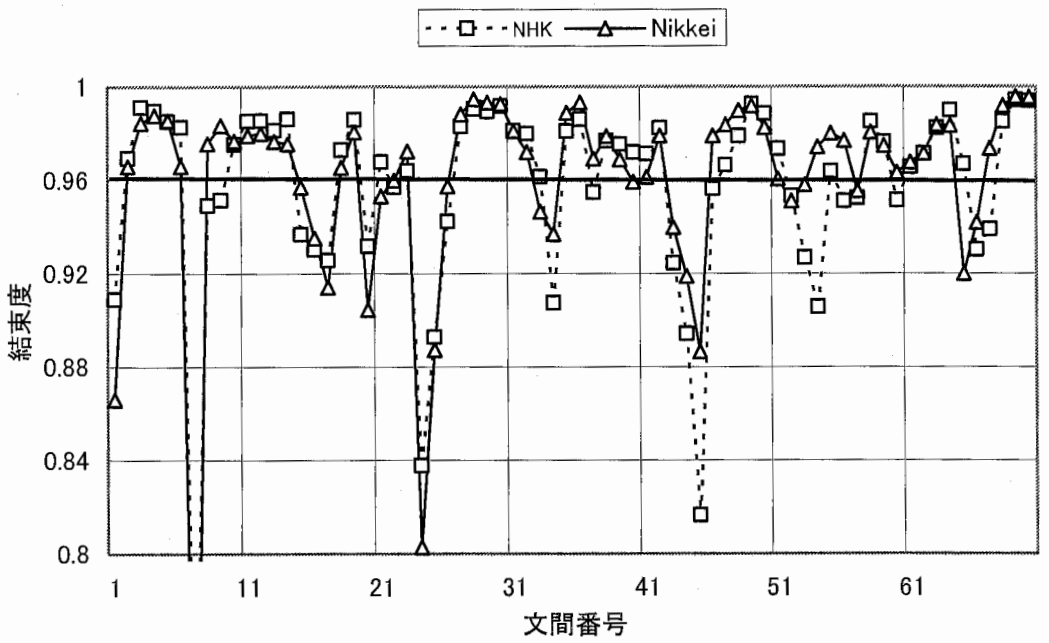


図 3 「あすを読む」の文の結束度の変化例（番組1）

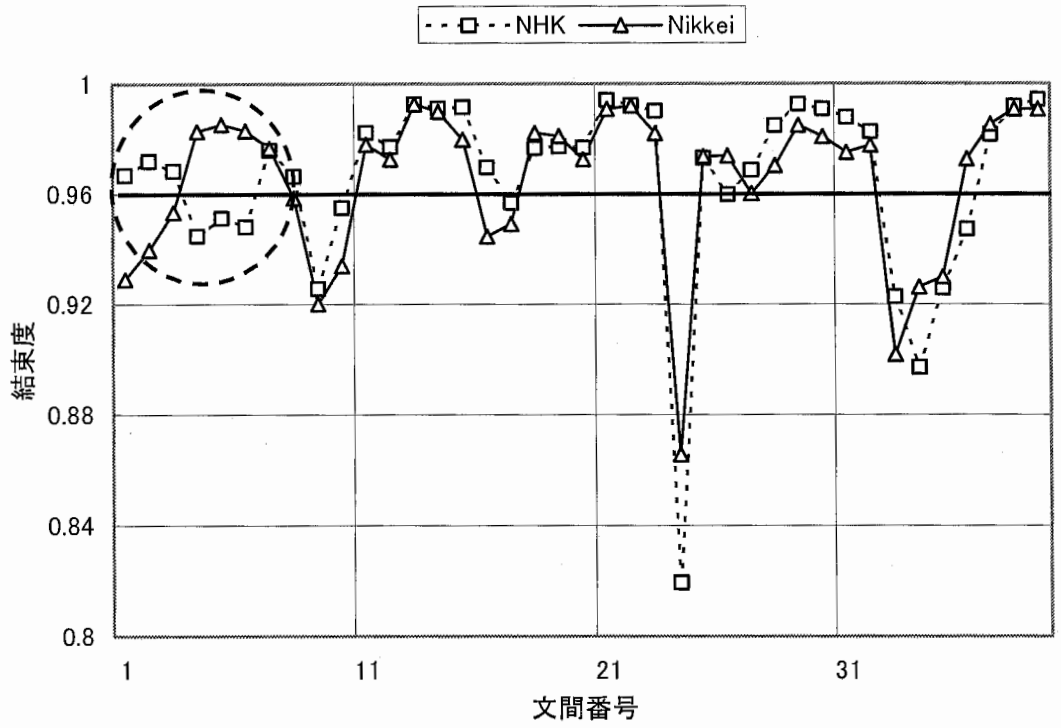


図 4 「あすを読む」の文の結束度の変化例（番組 13）

6．6．共通のトピックを持つセグメントの抽出

得られた文の結束度のグラフを目視して、適当な値として結束度が 0.96 以上の値をとる場合に共通のトピックを持つと仮定した。この閾値以上を連続してとる範囲を、共通のトピックを持つセグメントとして抽出した。結束度は日経新聞コーパスで得られた結束度を採用した。

抽出結果の評価には、共通のトピックを持つセグメントの先頭が「しかし」「また」「ところで」「そして」「さらに」「例えば」「では」「これに対して」「まず」などの話題転換の機能を持つフレーズで始まる割合を求めた。各番組の1文目から始まるセグメントは、真であるとしてカウントした。得られた値は、抽出したセグメントの開始位置の精度といえる。各番組についての精度を図 5 に示す。

「国会閉会と政局」をタイトルにする番組では抽出した5つのセグメント、「2000年政治の選択肢」をタイトルにする番組では抽出した9つのセグメント、それぞれすべてが正解であった。ただし、番組によっては接続詞やそれに相当する表現が少ないものもあったので、得られた値は実際の精度よりも低い値になっていると考えられる。

30 番組を通しては抽出されたセグメントが 296 個、うち正解が 135 個で、平均正解率は 45.6%であった。番組中で先頭のセグメントは無条件に真としたので、これを除くとセグメント数 266 個、うち正解が 105 個で、平均正解率は 39.4%となった。

日経新聞コーパスから得た単語共起知識を利用した先行研究[i] [1]の実験結果を、最初の 23 番組について同一の方法で評価した結果、抽出されたセグメントが 152 個、うち正解が 55 個で、平均正解率は 36.2%であった。番組中で先頭のセグメントを除くとセグメント数 129 個、うち正解が 32 個で、平均正解率は 24.8%となった。

同様に今回の手法で最初の 23 番組を対象にすると抽出されたセグメントが 233 個、うち正解が 97 個で、平均正解率は 43.5%であった。番組中で先頭のセグメントを除くとセグメント数 210 個、うち正解が 74 個で、平均正解率は 35.2%となった。(表 7)

	先頭のセグメントを除かないときの平均正解率		先頭のセグメントを除いたときの平均正解率	
	関連記事を利用したセグメンテーション	単語共起知識を利用したセグメンテーション	関連記事を利用したセグメンテーション	単語共起知識を利用したセグメンテーション
平均正解率	43.5%	36.2%	35.2%	24.8%

表 7　セグメンテーションの精度の比較

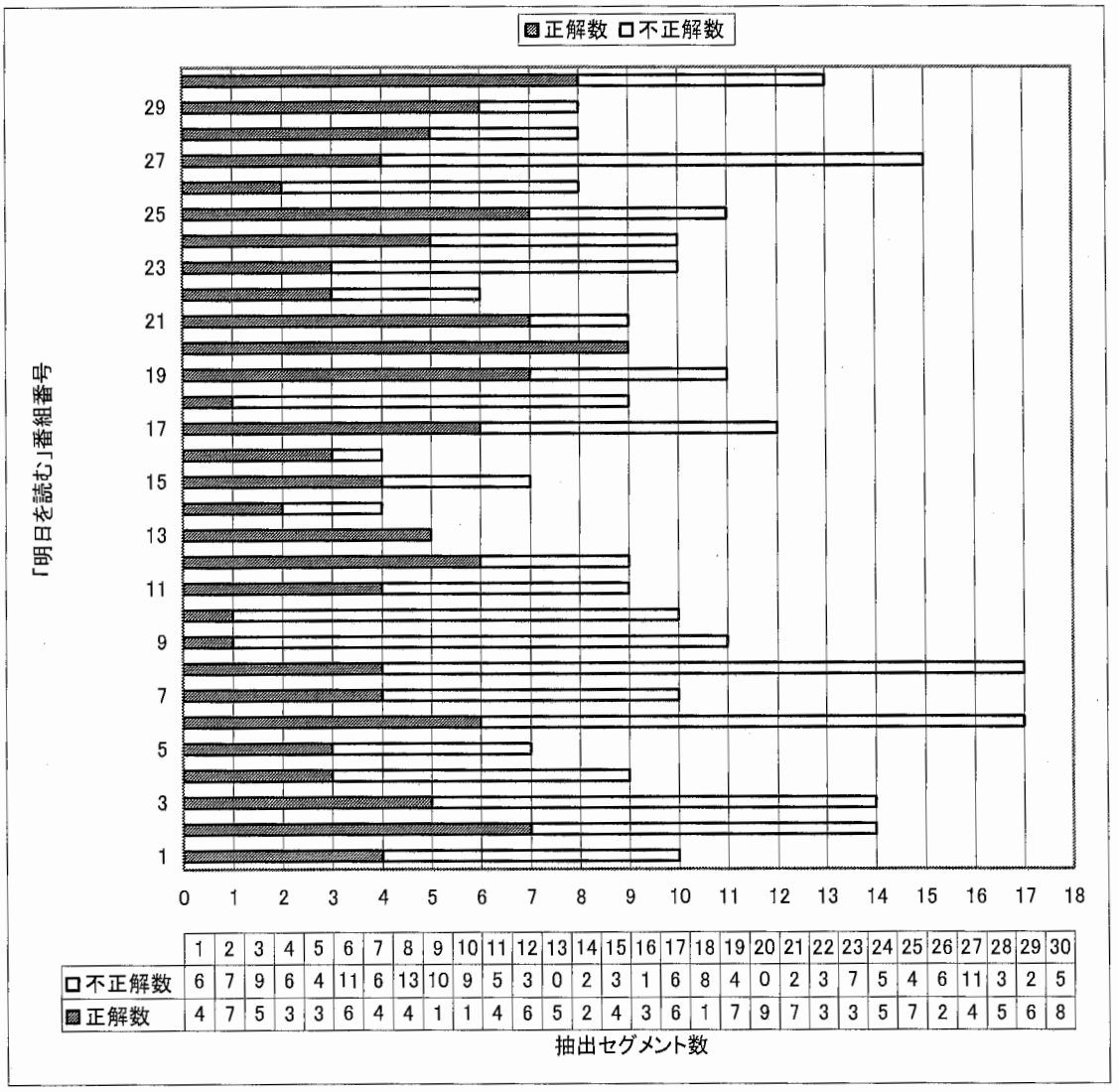
両手法間にはチューニングの違いがあるので単純に精度を比較することはできないが、本手法は比較的高い精度であるといえる。これは、関連する記事の知識を直接利用したため、1対1の単語対の共起情報よりも的確な情報を得られたこと、RIDF を用いることで的確な内容語を選択できたことが主な原因と考えられる。

「あすを読む」コーパスには「これについて具体的に見てみましょう。」など、ニュース記事トピックが含まれない一文が話題の転換点に挿入されていることも多い。フレームに使用したハニング窓関数は、このような文に対して頑健であるが、ベクトル空間モデルで隣り合う文のみをコサイン値で比

---

i) 1995 年1月～1999 年 12 月の日経新聞コーパスを使用して、一定の相互情報量を持つ 8416 対の索引語を抽出し、「あすを読む」コーパスの内容語をその相互情報量にしたがって換言する手法をとっている。日経新聞コーパスの期間が本実験と若干違い、対象とした番組のうち 5 番組は期間に重なりがない。また、フレームの文書ベクトルの計算には内容語の TF を使用しているが、IDF に相当する情報や、窓関数は使用していない。フレーム長は、内容語 50 語であることから、およそ 2 文程度と本実験の 9 文とは大きく違う。しかし、ハニング窓の周辺部の値は低く、実質的なフレーム長は似通っている。フレームの移動単位は内容語 1 語で、本実験よりも細かい。

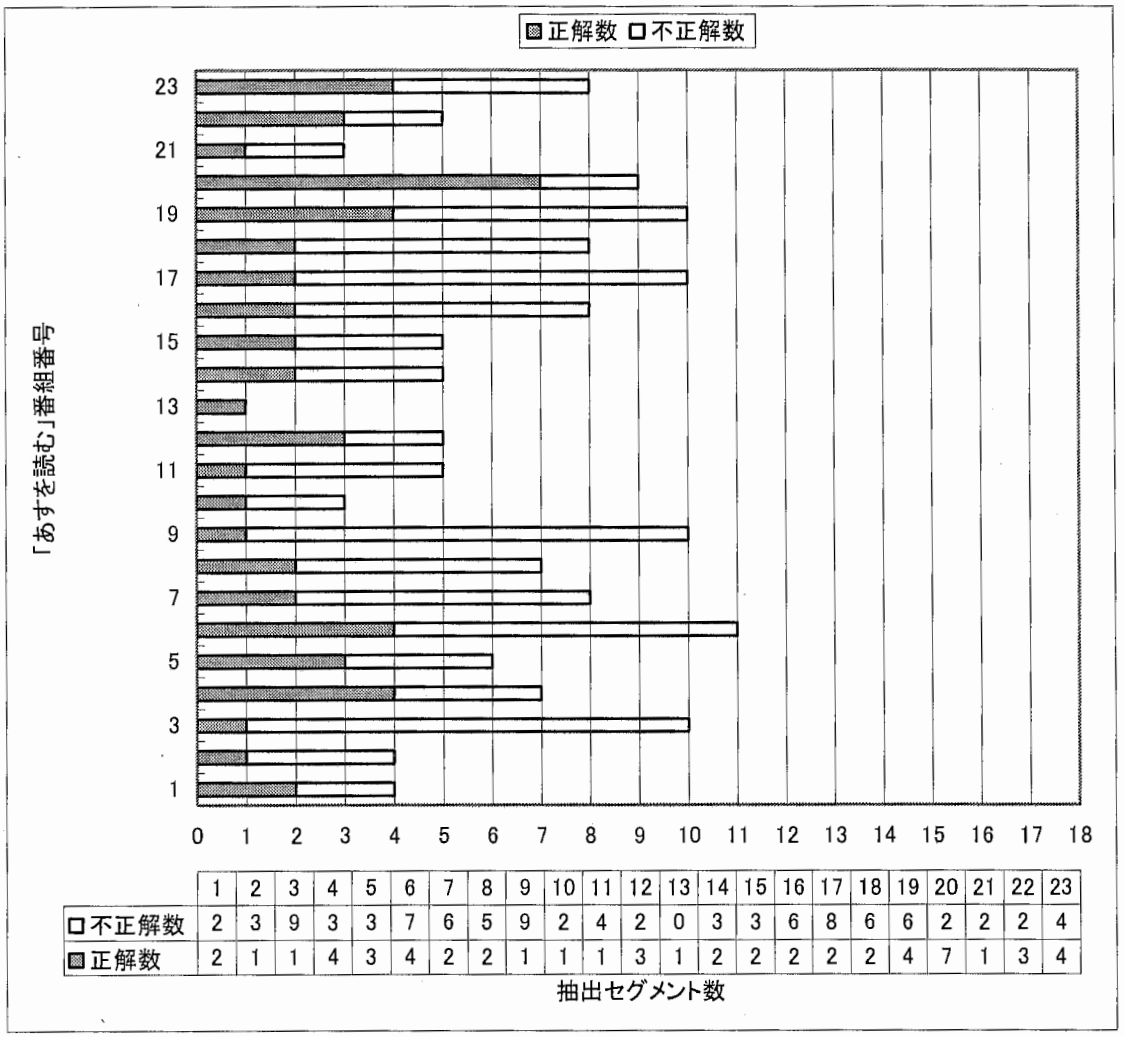
較する場合は、話題の転換を正確に把握できない恐れがある。しかし、今回の実験では、はっきりとしたニュース記事トピックを含まない範囲は、結束度の大きさが不安定になりやすいことが確認できたので、あらかじめそのような文を取り除かなくても、セグメンテーションの精度低下が少ないといえる。



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
不正解数	6	7	9	6	4	11	6	13	10	9	5	3	0	2	3	1	6	8	4	0	2	3	7	5	4	6	11	3	2	5
正解数	4	7	5	3	3	6	4	4	1	1	4	6	5	2	4	3	6	1	7	9	7	3	3	5	7	2	4	5	6	8

抽出セグメント数

図５　抽出したセグメントの開始位置の精度



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
□不正解数	2	3	9	3	3	7	6	5	9	2	4	2	0	3	3	6	8	6	6	2	2	2	4
■正解数	2	1	1	4	3	4	2	2	1	1	1	3	1	2	2	2	2	2	4	7	1	3	4

図6　単語共起知識を利用して抽出したセグメントの開始位置の精度

7．おわりに

小規模の「あすを読む」の書き起こしコーパスに対して、大規模なコーパスであるNHK放送ニュース汎用原稿コーパス、日経新聞コーパスを利用し、ニュース記事トピックに基づく文単位のテキストセグメンテーションを行った。

表記基準、文体の違いなど、複数のコーパスを利用する場合の問題点が明らかになった。それら、各コーパスの性質の違いで、ニュースを扱うコーパス同士でも重要な内容語がコーパス毎に違うことも確認できた。

今回は、コーパス内での1対1の単語共起やそれに類似した単語のクラスタリングを利用せず、各記事に対する関連の強さをベクトル空間モデルの各次元に直接使用することで、他のコーパスから得た情報の正確さや性質、セグメンテーションの根拠を的確に把握できた。それによって、重要な内容語が欠けて計算されている現象を観測できたが、その場合も他の多くの単語が同時に共起する情報を簡単に利用できるため、精度の急激な低下を避けられることも確認できた。

本手法の応用範囲は、多岐にわたる。例えば、「あすを読む」のような解説番組は一種のサマリーとして利用できるので、NHKコーパス、日経新聞コーパスなどの単一コーパス上のクラスタリングでは同一クラスタとならないクラスタ間の関連性を、「あすを読む」コーパスを参照することで計算

できる。また、今回はテキスト内の隣接するセグメント間だけに注目したが、ベクトル空間モデルを使った簡単な作業でこれらのセグメントをクラスタリングすることが可能である。これは、テキストの要約を容易にするだけではなく、今回の手法の長所である、他のコーパスから得た情報の正確さや性質、クラスタリングの根拠を把握し易い特徴も兼ね備えることができ、作成した要約から関連記事を参照するようなアプリケーションの構築も容易である。

今後は、さらに精密な評価を行うとともに、ニュース記事トピックを含まないセグメントの処理方法の改善、DF の計算方法の改善、セグメンテーションに使う閾値の決定方法の改善などを行い、精度を向上させるとともに、関連記事抽出技術の翻訳支援や機械学習、情報抽出への利用、テキストセグメンテーションに利用した情報の意訳を目的とする機械翻訳技術や要約技術への利用を進めたい。

参考文献

[1]伊藤ほか:単語の共起を利用した講演文のテキストセグメンテーション, 情報処理学会第 61 回全国大会, 4T－013, 2001

[2]北ほか:情報検索アルゴリズム, 北研二、津田和彦、獅々堀正幹, 共立出版, 2002