自 然 言 語 処 理 152－12
（２００２．１１．１２）

関連記事を利用したテキストセグメンテーション

西脇 正通† 田中 英輝
ATR音声言語コミュニケーション研究所

比較的小規模なニュース解説番組の書起こしテキスト各文に対して、放送ニュースなどの異なる大規模コーパスから関連記事を抽出し、その情報を利用してテキストセグメンテーションをする手法について述べる。本手法は対象とするテキストの各文の文書ベクトルに関連記事との関連度を使用したので、大規模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比較的把握しやすい特徴がある。

Text segmentation using the related articles

Masamichi Nishiwaki　Hideki Tanaka
ATR Spoken Language Translation Research Laboratories

This paper proposes a text segmentation method for a news commentary TV program. Our method defines a cohesion score between sentences by document vectors which are calculated from relation scores of another large-scale news corpus. Because it is easy to confirm the information obtained from another corpus, the ground of segmentation can be understood easily.

†現在は NHK 放送技術研究所ヒューマンサイエンスに所属。

1. はじめに

著者らは異種コーパスの統合利用技術を研究している。そのひとつとして、あるテキストを処理する際に異種コーパス中の関連したテキストを利用し、データスパースネスを解消することを考えている。

本稿では、ニュース解説テレビ番組の講演テキストを関連記事を利用してテキストセグメンテーションする手法について報告する。

テキスト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[1]や、テキスト内での単語の出現分布を利用する方法[2]がある。本稿では、単語の出現分布によりベクトル空間モデルを作成し、ベクトルの類似性を利用して境界を検出した。この手法は、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。本稿で扱う講演テキストは、1番組中に複数の話題がある比較的短いテキストであるため、この手法を適用すると対象領域内で十分な単語の出現分布が得られない[3]。このため、シソーラスの利用や、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。しかし、われわれの講演テキストは、単語のクラスタリングをできるほどの十分な量はない。

本手法ではこのようなテキストに自動でセグメンテーションを行うため、関連する比較的大規模なコーパスを利用している。本手法は、単語のクラスタの学習をせずに、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することで、頑健なセグメンテーションが期待できる。また、ベクトルの各次元が関連記事と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比較的把握しやすい。

2. 提案手法の概要

テキストセグメンテーションをするには文間のつながり(**結束度**)を計算する必要がある。本手法では文間の結束度を、その文に関連する記事(**関連記事**)の集合(**関連記事集合**)の類似度で定義した。関連記事集合の類似度は、関連記事の関連度を並べた文書ベクトル[1]を使ってベクトル空間モデルで計算した。この文書ベクトルを**関連度文書ベクトル**と呼ぶ。

関連記事は、索引語頻度(TF)と残差 IDF (residual IDF; **RIDF**)[4]で重み付けした索引語[2]の集合の類似度を**記事関連度**と定義し、文に対して記事関連度が高い記事を大規模コーパスから収集

[1] 文書ベクトル[4]は、文書の内容を表わす要素をベクトルの各次元に割り当てることで、文書の特徴を表わしている。本稿での文書とは、物理的に完結したテキストの意味ではなく、特徴を計算するときの対象領域の意味で用いている。

[2] 本稿での索引語は、単語の標準形と品詞の組みである。また、RIDF を利用して自動的に内容語に対する重み付けを大きくしているため、索引語はテキストの語彙そのものとなる。

した。索引語の集合の類似度は、索引語の TF・RIDF の値を並べた文書ベクトルを使ってベクトル空間モデルで計算した。

RIDF は、セグメンテーションの対象となるテキストを含むコーパスと、関連記事を含むコーパス、それぞれで計算した。

TF は、セグメンテーションの対象となるテキスト各文の**補正した TF** と、関連記事内の TF を用いた。

セグメンテーションの対象となるテキストでは、前後の文脈を考慮できるよう**ハニング窓**を使用したフレーム化を行って補正した TF を求めることで、各文の**補正した文書ベクトル**を得た。

これらの詳細については以降で述べる。

2.1. 文書ベクトルの計算

記事関連度を計算するベクトル空間モデルの作成では、まず講演テキスト中の各文で索引語集合を用いた文書ベクトルを作成することで、その文の特徴をあらわす。それぞれの索引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。

関連記事を含むコーパスでも、記事単位で同様に文書ベクトルを作成し、文と記事の関連度の計算に使用する。

索引語の重み付けに使用した RIDF はそれぞれのコーパスにおける索引語の出現文書数(DF)から取得した。TF は、講演テキストの対象となる文の補正した TF および、関連記事から取得した TF を用いた。ここでは、講演テキストについても1番組を1文書として計算した DF と RIDF を採用している。

索引語 t_i の IDF を $IDF(t_i)$、索引語 t_i の生起が独立しているとしてポアソン分布により推定した IDF を $PIDF(t_i)$とすると、索引語 t_i の $RIDF(t_i)$は以下の式から求められる。

$$\begin{aligned} RIDF(t_i) &= IDF(t_i) - PIDF(t_i) \\ &= \log(n/n_i) - \log(n/n(1-P(0;\lambda_i))) \\ &= \log(n/n_i) + \log(1- e^{-\lambda i}) \end{aligned} \quad (1)$$

$IDF(t_i) = \log(n/n_i)$

$PIDF(t_i) = \log(n/n(1-P(0;\lambda_i)))$

ポアソン分布 $P(k;\lambda_i) = e^{-\lambda i}\lambda_i^k/k!$

期待値 $\lambda_i = c_i/n$

k =（**仮定した TF**）

c_i =（**索引語 t_i の文書集合全体での出現数**）

n =（**文書集合の文書数**）

n_i =（**索引語 t_i の DF**）

2.2. ハニング窓を使用したフレーム化による文脈の考慮

講演テキストの各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、関連文書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領域を複数文とすることで、文脈を考慮するとよいと考えた。そのために、ハニング窓を使用したフレーム化(図 1)を行い、索引語の出現位置に応じた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。

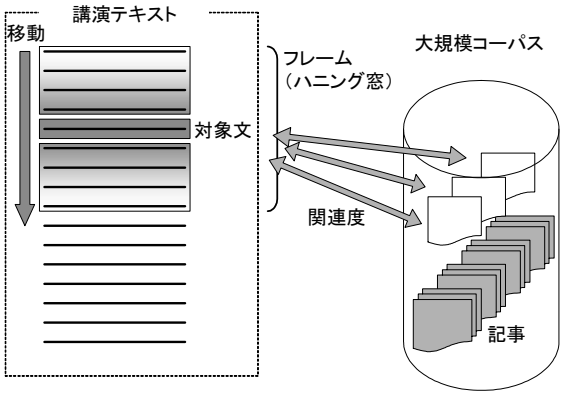


図 1 フレームを利用した関連記事の抽出

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、講演テキスト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、つまり計算ごとの移動量は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・RIDF 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。記事関連度の計算では、この補正した文書ベクトルを使用する。

索引語が出現する位置の文番号を l、フレーム中央の文番号を j、フレーム長を L とすると、ハニング窓関数 $h_j(l)$は次のようになる。

$$h_j(l) = (1+\cos(2\pi(l-j)/L))/2 \quad (\ |l-j| \leqq L/2\ )$$

フレーム長を9文としたときの、フレーム内の各文に対するハニング窓関数の値は[0.095 0.345 0.655 0.905 1 0.905 0.655 0.345 0.095]となる。

以上から関連度を計算するときに使用する補正した文書ベクトルと関連記事の文書ベクトルは次のように求められる。

講演テキストの文 S_j の補正した文書ベクトル(S_jを中央にするフレーム F_j の文書ベクトル)

$$f_j = \sum_{l=0}^{8} h_j(j-4+l)\, s_{j-4+l}$$

文 S_j の文書ベクトル

$s_j = {}^t[s_{1j}\ s_{2j}\ \ldots s_{Ij}]$　　（I は索引語数）

文 S_j における索引語 t_i の重み

$s_{ij} = TF_j(t_i) \cdot RIDF(t_i)$

文 S_j での索引語 t_i の TF

$TF_j(t_i) = \log(1+C_j(t_i))$

$C_j(t_i)$は索引語 t_i の文 S_j における出現頻度

関連記事 A_m の文書ベクトル

$a_m = {}^t[a_{1m}\ a_{2m}\ \ldots a_{Im}]$　　（I は索引語数）

索引語 t_i の記事 A_m における重み

$a_{im} = TF_m(t_i) \cdot RIDF(t_i)$

記事 A_m での索引語 t_i の TF

$TF_m(t_i) = \log(1+C_m(t_i))$

$C_m(t_i)$は索引語 t_i の記事 A_m における出現頻度

2.3. 記事関連度の計算

講演テキストの各文の補正した文書ベクトルと関連記事の文書ベクトルがなす角のコサインを求めて、講演テキストの各文に対しての記事の関連度 r とする。

講演テキストの文 S_j と記事 A_m の文書ベクトルがなす角を θ (a_m, f_j)とすると、関連度 r_{mj} は以下のように求められる。

$$r_{mj} = \cos\theta(a_m, f_j) = (a_m \cdot f_j) / (\| a_m \| \| f_j \|)$$

2.4. 結束度の計算

講演テキストの各文 S_j について、各記事 A_m との関連度 r_{mj} を計算しそれを並べたものを、結束度を計算するときの S_j の関連度文書ベクトル d_j とした。次に、隣り合う文の関連記事集合の類似度として、d_j、d_{j+1} がなす角のコサインを求め、これを文の結束度 u_j とし、この増減を各番組内で求めた。

関連度文書ベクトル d_j、d_{j+1} がなす角を θ (d_j, d_{j+1}) とすると、文 S_j と文 S_{j+1} の結束度 u_j は以下のように求められる。

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）

結束度 u_j

$$u_j = \cos\theta(d_j, d_{j+1}) = (d_j \cdot d_{j+1}) / (\| d_j \| \| d_{j+1} \|)$$

3. 実験

以下の手順で実験を行った。

3.1. 各コーパスでの RIDF

表 1 に示すあらかじめ形態素解析[3]した 3 つのコーパス[4]を使用し、標準形と品詞の組を索引語とした RIDF、IDF をそれぞれのコーパスで求め比較した。

講演テキストには、NHK番組「あすを読む」の書起こしテキストを使用し、関連記事を含む他のコーパスと時期が重なっている最初の 30 番組をテキストセグメンテーションの対象とした。このテキストは放送ニュース原稿や新聞記事と比べひとつのテキストが長く、テキスト内では過去の複数のニュースを引用している。

この講演テキストの各文に対して、大規模なニュースコーパスである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から、関連するニュース記事を得て関連記事とする。

表 1　各コーパスの概要

コーパス	期間	文章数（千）	文数（千）	1 文章内平均文数	1 文内平均単語数
ＮＨＫ放送ニュース汎用原稿	4/1995 - 3/2000	287	1,593	5.55	47.45
日経新聞	1/1995 - 12/2000	1,758	18,620	10.59	26.82
「あすを読む」書起こし	11/1999 - 1/2001	0.305	18	59.02	29.07

各コーパスの索引語数は表 2 のようになった。

表 2　各コーパスの索引語数

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

表 3 に「あすを読む」書起こしコーパスで RIDF の大きいものから順にソートした索引語の例を示す。

コーパス中では「以前お伝えしました」などに使われることが多い「以前」などは、IDF が高く RIDF は低い。また、「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。」などの番組中の挨拶に使われる「それでは」「失礼」「今晩」が、RIDF では 0 以下となっている[5]。

これらの RIDF が低い索引語は、内容語とならないものとして自動的に計算でき、内容と関係が低いコーパス特有のスタイルも関連記事との関連度に影響しにくいと考えられる。

表 4 にNHK放送ニュース汎用原稿コーパスでの RIDF を示す。株式市場の記事で使用される「トピッ

[3] 形態素解析には茶筌[5]を使用した。品詞は詳細品詞である。

[4] 各コーパスには、日本語と英語の 2 言語が含まれるが、本稿では日本語を対象とした。

[5] RIDF を求める式(1)において、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、RIDF は負の値をとる。記事関連度の計算では、負の値もすべて 0 とした。

クス」のRIDFは0となっている。これは、各記事に1回ずつ出現するためで、RIDF はわずかに負の値となっている。このように、本来記事の特徴を表わすと考えられる内容語の RIDF が低くなるのは、同一単語の生起を単語の独立性の尺度とし、内容語は記事中で何度も使われると仮定しているためで、これを改善するためには独立性を別の尺度で計算する必要があるといえる。

表 3 「あすを読む」書起こし IDF と RIDF

索引語	出現頻度	DF	IDF	RIDF
古墳/名詞-一般	48	1	5.72	3.79
傍受/名詞-サ変接続	33	1	5.72	3.44
マカオ/名詞-固有名詞-地域-一般	31	1	5.72	3.38
野球/名詞-一般	63	2	5.03	3.35
フッ素/名詞-一般	28	1	5.72	3.29
トキ/名詞-一般	28	1	5.72	3.29
雪印/名詞-固有名詞-組織	25	1	5.72	3.18
タリバン/未知語	25	1	5.72	3.18
エーアールエフ/未知語	25	1	5.72	3.18
胚/名詞-一般	79	3	4.62	3.14
踏まえる/動詞-自立	21	20	2.72	0.01
れる/動詞-接尾	3446	304	0	0
高まる/動詞-自立	56	51	1.79	0
以前/名詞-副詞可能	43	40	2.03	0
から/助詞-格助詞-一般	2218	304	0	0
。/記号-句点	18550	305	0	0
ます/助動詞	17105	305	0	0
とりわけ/副詞-一般	14	14	3.08	-0.02
皆さん/名詞-一般	19	19	2.78	-0.03
例える/動詞-自立	21	21	2.68	-0.03
それでは/接続詞	206	163	0.63	-0.08
失礼/名詞-サ変接続	150	145	0.74	-0.20
今晩/名詞-副詞可能	411	296	0.03	-0.27

表 4 NHK放送ニュース IDF と RIDF

索引語	出現頻度	DF	IDF	RIDF
真須美/名詞-固有名詞-人名-名	2196	464	6.43	1.55
ダイアナ/名詞-固有名詞-人名-名	1120	241	7.08	1.53
行き/名詞-接尾-一般	14004	2993	4.56	1.52
潮位/名詞-一般	201	99	7.97	0.71
円/名詞-接尾-助数詞	110860	45322	1.85	0.71
初日の出/名詞-一般	136	67	8.36	0.71
ます/助動詞	1494405	286080	0	0
ところで/接続詞	1593	1591	5.20	0
。/記号-句点	1676098	287230	0	0 (-0.003)
トピックス/名詞-一般	2471	2471	4.76	0 (-0.004)

3.2. 関連記事の抽出

講演テキストの文と関連記事との関連度を計算し、NHK放送ニュース汎用原稿コーパスの関連記事のうち関連度が上位の記事について、内容を確認した。確認には「(5)その原稿を元にしていると認められる」「(4)ほぼ同じ内容」「(3)同じ話題」「(2)内容を補足している」「(1)話題がかけ離れている」の 5 段階基準の評価を行い、その合計をとった。「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評価が低かった。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評価となった。逆に、特に番組の導入部では、状況の説明が多くなるので、評価が高くなる傾向が見られた。

表 5 に示す例の文 16 は「イーユー」「広い範囲」が関連記事では「EU」「包括的」となっていたが、関連度が高い上位 10 記事の評価の合計が 43 となり評価の高い関連記事が多く得られた。

表 5 関連記事上位 10 記事の評価の合計(例 1)

評価	講演テキスト
48	12 ただ農業とサービス分野だけは前回の決着の時の経緯で決まっております。
47	13 いったい各国各地域は今回の交渉に、どんな姿勢で臨もうとしてるんでしょうか。
40	14 これをご覧いただきたいと思いますが、アメリカは農業とサービス部門のみに限定して話し合いをしたいとこう言っております。
34	15 この部門は非常にそのアメリカにとっては競争力が強いところなんですね。
43	16 イーユーはもっと広い範囲で話し合いをしたい、議論をしたいという姿勢です。
44	17 そして日本も同じなんですがやはり多角的にこの交渉の範囲を広げたい、そして同時にですね、この反ダンピング措置の乱用防止こういったことについても、その重要な項目だというふうにしております。
32	18 この交渉の範囲を広げておけば弱い部門だけを攻められることを防げるという思惑もありますし、それから取り引きのカードをいろいろ使うことができるという考え方もあります。

表 6 に示す例では、文 78、79 の内容が解説者の考えである傾向が強く、関連記事の評価も低くなった。

表 6 関連記事上位 10 記事の評価の合計(例 2)

評価	講演テキスト
20	76 日本とアメリカでは状況が違うので、アメリカの制度をそのまま持ち込むことはできません。

16	77 私は結局のところ、今ある雇用を守りながら新規の雇用を作り出す、この二つのことをやるためには日本とアメリカのそれぞれの良いところを取りいれるしかないんだろうと思います。
10	78 富士通の例、これはこうした挑戦の一つではないでしょうか。
11	79 それぞれの企業が自分にあった形でどんどん新しい方式に挑戦してもらいたいもんだと思ってます。
17	80 今年の経済白書、日本企業は二百万の余剰人員を抱えていると推計しています。
17	81 この人達が全員仕事を失いますと、失業率は八パーセントにはね上がります。
19	82 企業は体力が落ちているので、これ以上余剰な雇用を抱え込む力はありません。
23	83 終身雇用が崩れるのは避けて通れないのが現実です。
22	84 よらば大樹の陰でやってきたサラリーマンも今や実力主義、実績主義の世界に入ってきたことを自覚して研鑽を積むしかないでしょう。

評価が高かった関連記事は前後数ヶ月以内など、作成された時期が講演テキスト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関連付けされていた。さらに時期がずれている関連記事でも、「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関連付けの上で有効であると思われるキーワードがあるが、異種コーパス間での関連付けでは、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評価が高い関連記事を得た例もあり、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。

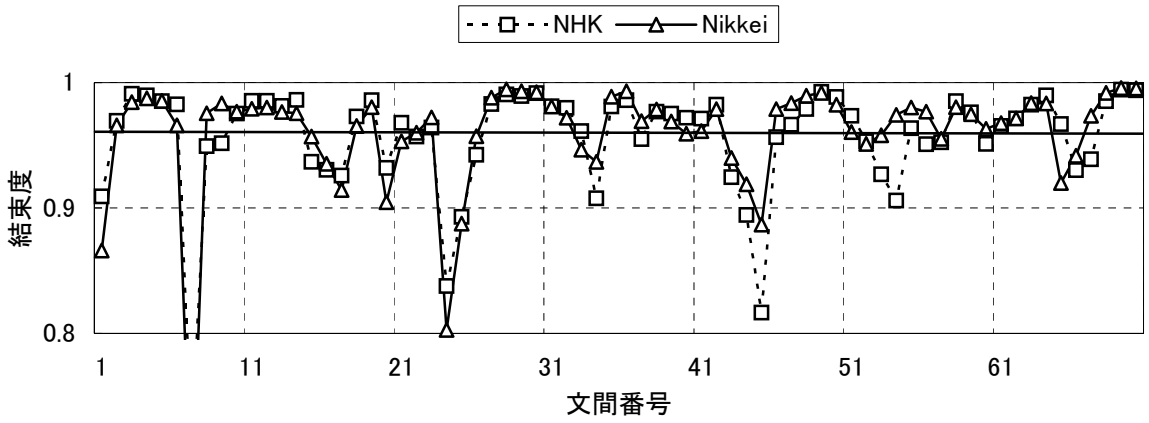


図 2 「あすを読む」の文の結束度の変化例（番組1）

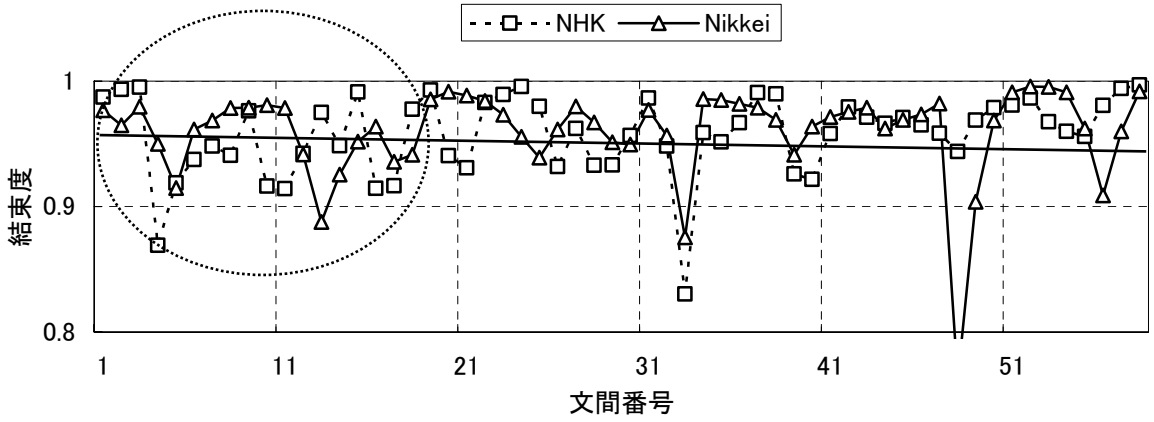


図 3 「あすを読む」の文の結束度の変化例（番組 10）

3.3. 隣り合う文の結束度

講演テキストの隣り合う文の結束度をグラフにして確認した。(図 2)

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 文がニュースのトピック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 3)。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。このことから、コーパス内に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と考えられる。

3.4. セグメンテーション

日経新聞コーパスで得られた文の結束度のグラフをから、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。

得られたセグメントが適当であるかを確認するために、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、簡易的な評価を行った。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なる。

30 番組を通して抽出されたセグメントは 296 個、うち話題転換の機能を持つフレーズで始まるセグメントの割合は 45.6%であった。

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[3]の実験結果を、30 番組に含まれている 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、うち話題転換の機能を持つフレーズで始まるセグメントの割合は 36.2%であった。同様に今回の手法でこの 23 番組を対象にすると、抽出されたセグメントが 233 個、話題転換の機能を持つフレーズで始まるセグメントの割合は 43.5%で、先行研究に比べ 7.3%大きい値であった。

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。

4. おわりに

ニュース解説番組の小規模な書起こしコーパス中のテキストに対して、関連するニュース記事を含む大規模なコーパスを利用し、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。

テキストの特徴を表わすために各コーパスで索引語の RIDF を取得することで、異種コーパス間での表現スタイルの違いを確認するとともに、その影響が少ない内容語の重み付きリストを各コーパスで得ることができた。同時に、RIDF で用いた独立性の尺度が不適切である場合も確認できた。

関連記事の確認では、重要な内容語を欠いて記事関連度が計算されている現象を確認でき、表記基準、文体の違いなどの、異種コーパスを利用する場合の問題点が明らかになった。しかし、本手法ではその場合も他の単語が同時に共起する情報をより多く利用できるため、精度の急激な低下を避けられることも確認できた。

また、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、関連記事の関連度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。

今後は、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、テキストセグメンテーションに使う閾値の決定方法やセグメント境界の検出方法の改善などを行い精度を向上させ、さらに精密な評価を行うとともに、各セグメントの関連記事を直接確認できる特徴を利用して、翻訳支援や機械学習などへの応用も進めたい。

謝辞

本稿は通信・放送機構の研究委託により実施したものである。

発表の機会を与えていただいた NHK に感謝します。

参考文献

[1] 望月源,本田岳夫,奥村学,複数の表層手がかりを統合したテキストセグメンテーション,自然言語処理,Vol.6,No.3,pp.43-58,1999.

[2]Hearst, M. ,Multi-Paragraph Segmentation of Expository Text, Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, pp.9-16,1994.

[3]伊藤山彦,松本賢司,谷田泰郎,柏岡秀紀,浦谷則好,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, 第 61 回情報処理学会全国大会, pp.2-159-160,2001.

[4]北研二,津田和彦,獅々堀正幹, 情報検索アルゴリズム, 共立出版,2002.

[5]茶筌 http://chasen.aist-nara.ac.jp