F-4　概念ベクトルによるトピックセグメンテーションのニュース音声への適用
Conceptual-Vectors-Based Topic Segmentation for Broadcast News Speech

別所 克人†　大附 克年†　松永 昭一†　林 良彦†
Katsuji Bessho　Katsutoshi Ohtsuki　Shoichi Matsunaga　Yoshihiko Hayashi

1．まえがき

広帯域アクセス回線の普及とともに、巨大なデータ量のサウンドデータや動画データをネットワークを経由して配信するストリーミングが広がりつつある。膨大なマルチメディアコンテンツが日々生成される中で、ユーザが求める情報に高速にアクセスできるように、マルチメディアコンテンツを編集・構造化し、メタデータを自動生成する技術が重要となっている。ニュース等のコンテンツにおいては一つの番組内に一般に複数のトピックが入っており、時間順に見ていたのでは所望のトピックにすぐにアクセスできない。所望のトピックに即時にアクセスできるように、ストリームをトピック単位に切り出し、索引付けしておく必要がある。ストリームをトピック単位に分割するトピックセグメンテーションはメタデータ生成のための必須技術であり、自動・半自動化できればメタデータ生成のためのコストが大幅に削減されることが期待できる。

このような背景から現在、筆者らはニュース等の映像コンテンツのメタデータ自動生成を目的として、映像コンテンツのトピックセグメンテーションの研究を進めている。映像コンテンツには、映像、音声、テロップ、BGM 等の情報が入っているが、ニュース等の場合、特に音声の、セグメンテーションに果たす役割が大きいと考えられる。筆者らはこれまでテキストデータを対象として、単語の意味表現の一つである概念ベクトルを用いたトピックセグメンテーションの研究を行ってきた[1]。今回、本手法をニュース音声の認識結果に対して適用した評価実験を行い、映像コンテンツのセグメンテーションに有効である見通しを得たので報告する。

2．提案手法の概要

2.1　概念ベース生成

本手法では、あらかじめ学習用コーパスを基に、各単語にその共起パターンをベクトル化して得られる意味表現（概念ベクトルと呼ぶ）を対応付け、単語とその概念ベクトルの対の集合である概念ベースを生成しておく。ある 2 単語に対応するベクトル値が近ければ、共起パターンが似ているので、この 2 単語は意味的に近いということが推測される。

概念ベースの生成では、まず学習用コーパスを形態素解析した後、各自立語間の 1 文中に共起する頻度をカウントした共起行列を作成する（表 1 参照）。共起行列の各行をベクトルと見立てると、各自立語にその共起パターンを表すベクトルが対応付けられる。但し、このままではデータのスパースネス性が生じることを始めとして、テキストデータから抽出される単語の情報には常に欠落があると予想されるため、ベクトル間の類似度の精度が低下すると考えられる。また、一般にベクトルの次元数は非常に大きなものとなるため、計算量も無視できないものとなる。このため、[2][3]の手法に従い、共起行列を特異値分解により、次元数を縮退させた行列に変換する。変換後、長さ 1 に正規化された各ベクトルが概念ベクトルであり、単語とその概念ベクトルの対の集合が概念ベースである。

表 1　共起行列の例

	・・・	国会	・・・	園芸	・・・
・・・	・・・	・・・	・・・	・・・	・・・
選挙	・・・	287	・・・	3	・・・
・・・	・・・	・・・	・・・	・・・	・・・
苗木	・・・	2	・・・	93	・・・
・・・	・・・	・・・	・・・	・・・	・・・

2.2　セグメンテーションアルゴリズム

セグメント対象テキスト中の各単語に、概念ベース中のベクトルを対応付けて得られるベクトル列は、単語の意味の変遷を表していると考えられるので、このベクトル列の変化を利用してテキストの分割が行えることが期待できる。

本手法では、セグメント対象テキストを形態素解析して単語に分割し、得られた各単語のうち、自立語のみに概念ベース中の対応するベクトルを付与する。

単語間の境界位置の前後に、一定の単語数の窓を設定し、各窓ごとに、その窓に含まれる単語のベクトルの重心を計算する。各窓に対応する重心ベクトルは、窓の意味を表現するベクトルと見なせる。前後の重心ベクトルの余弦測度を、この境界位置の結束度として計算する。

次に、結束度の微弱な振動を除去するため、各境界位置の結束度を、当該境界位置とその前後一定数の境界位置の結束度の平均に変換する（結束度の平滑化）。

トピック境界では、結束度が極小となっていると期待される。結束度が極小となる境界位置（極小点と呼ぶ）を i，極小点の左側で単調減少している部分の開始位置を l，右側で単調増加している部分の終了位置を r とし、それぞれの結束度を C_i，C_l，C_r としたとき、極小点 i に対し、谷の深さを示す以下の depth score と呼ばれる値 d_i を計算する。

$$d_i = (C_l - C_i) + (C_r - C_i)$$

depth score の大きい極小点から、境界候補として出力する。

セグメンテーションの従来手法の一つである Hearst 法[4][5]は、前後の各窓ごとに、その窓に含まれる単語の出現頻度ベクトルをとり、そのベクトル間の余弦測度を結束度とするものである。

†日本電信電話株式会社　NTT サイバースペース研究所
NTT Cyber Space Laboratories, NTT Corporation
1-1 Hikarinooka Yokosuka-Shi Kanagawa 239-0847 Japan

3．評価実験

3.1 実験条件

概念ベース生成用の学習用コーパスとしては、ニュース原稿のテキスト 1,428,900 文を用いた。共起行列の列（ベクトルの各座標）に対応する単語としては、テキスト中の頻度順位が 51 番目から 1,050 番目までの 1,000 個の単語をとり、行に対応する単語としては高頻度語 20,000 語をとった。特異値分解後の次元数は 100 とした。

セグメント対象テキストとして、模擬アナウンサーの音声を音声認識エンジン VoiceRex[6]を用いて認識した結果である 33 トピック分（データ A とする）を用いた。データ A においては、認識結果の NBEST 候補の内、最尤のものを採用した。認識精度は単語誤り率 5.6%である。またこの他に、データ A のポーズで区切られた各音声セグメントごとに正解の書き起こしが記述されたテキスト（データ B）と、データ B で同一文に含まれるテキストが連結され、1 文ごとに書き起こしが記述されたテキスト（データ C）を用いた。データ A, B の各音声セグメントに対応するテキストも便宜上、文と呼ぶことにする。表 2 は各データに関する情報である。ベースラインとは、ランダムに選んだ境界が正解境界である確率である。

セグメント対象テキストに対し、提案手法と Hearst 法によるセグメントを行った。ともに、窓幅は 1 トピックの平均長の 3 分の 1 程度の単語数をとり、結束度の平滑化は、各境界位置とその直前、直後の境界位置の結束度の加重平均で行った。極小点となる境界位置の depth score を求め、depth score の大きな境界位置から、境界位置を直近の文境界に変換した上で出力した。ここで同じ文境界は重複して出力せずに、正解境界数である 32 だけ境界候補を出力した。

表 2　セグメント対象テキストの情報

	A	B	C
トピック数	33	33	33
全文数	387	387	178
1 トピックあたりの文数	11.7	11.7	5.4
ベースライン（完全一致）	8.3%	8.3%	18.1%
ベースライン（1 文ずれ許容）	24.9%	24.9%	54.2%

3.2 実験結果

各データごとの精度は表 3 のようになった。なお、R は再現率（正解の境界候補数／正解境界数）、P は適合率（正解の境界候補数／境界候補数）であり、F は F 値で $F=(2\times R\times P)/(R+P)$ のように算出される。また、E は誤り率で、一定の距離にある 2 単語の組で 1 トピック内にあるものが分断されたり、異なるトピックにあるものが共存している割合であり、低いほど精度がよい。2 単語間の距離として、1 トピックの平均長の半分程度の単語数をとった。

データ A のような認識誤りを含むテキストに対しても、F 値が 83.3%（1 文ずれ許容の場合）にまでいき、本手法が Hearst 法と比較して高精度を出すことが分かる。データ B は、認識精度が 100%の場合であるが、データ A と同程度である。これはデータ A の認識精度が比較的高いことと、データ A に認識誤りの自立語が存在することにより正解境界の結束度が特に低くなったり、データ B にデータ A にはない自立語で正しいセグメンテーションに寄与しないものが存在することにより、かえってデータ B の精度の方が落ちることもあることに起因する。データ C は、そもそもデータ A, B と比べベースラインが高いので精度がより高くなるのは当然であるが、ポーズで区切られた各音声セグメントを真の一文になるように結合する前処理を行うことにより、この精度までいくことが期待できる。

表 3　データごとの実験結果

●データ A

	境界候補	R	P	F	E
提案手法	完全一致	71.9%	71.9%	71.9%	12.3%
	1 文ずれ許容	75.0%	93.8%	83.3%	
Hearst 法	完全一致	50.0%	50.0%	50.0%	30.2%
	1 文ずれ許容	50.0%	78.1%	61.0%	

●データ B

	境界候補	R	P	F	E
提案手法	完全一致	71.9%	71.9%	71.9%	12.8%
	1 文ずれ許容	75.0%	96.9%	84.5%	
Hearst 法	完全一致	46.9%	46.9%	46.9%	30.1%
	1 文ずれ許容	50.0%	78.1%	61.0%	

●データ C

	境界候補	R	P	F	E
提案手法	完全一致	87.5%	87.5%	87.5%	8.7%
	1 文ずれ許容	87.5%	90.6%	89.0%	
Hearst 法	完全一致	62.5%	62.5%	62.5%	27.8%
	1 文ずれ許容	65.6%	81.3%	72.6%	

4．まとめ

概念ベクトルを用いる手法により、認識誤りを含むテキストに対しても高精度の結果が得られることを報告した。本実験では、正解境界数の情報を用いたが実際には正解境界数は不明である。正解境界数を定めるために、depth score の閾値や、過去のニュースデータから求めた 1 ニュースあたりの平均長等の情報を用いることを検討している。また、より認識精度が低い場合にも頑健に高精度を出す方式や、結束度以外の情報を併用する方式等についても検討を進めていく。

参考文献

[1]別所：単語の概念ベクトルを用いたテキストセグメンテーション，情報処理学会論文誌，Vol.42, No.11 (2001)

[2]Schütze, H. : Dimensions of Meaning, Proc. Supercomputing '92, pp.787-796 (1992)

[3]Schütze, H. and Pedersen, J.O. : A Cooccurrence-Based Thesaurus and Two Applications to Information Retrieval, Proc. RIAO '94, pp.266-274 (1994)

[4]Hearst, M.A. : Multi-Paragraph Segmentation of Expository Text, 32nd Annual Meeting of the Association for Computational Linguistics, pp.9-16 (1994)

[5]Hearst, M.A. : TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages, Computational Linguistics, Vol.23, No.1, pp.33-64 (1997)

[6]野田他：音声認識エンジン VoiceRex の開発，音講論，2-1-19, pp.91-92 (1999-9)