概念ベクトルの結束性によるトピックセグメンテーション精度の評価

別所克人　大附克年　松永昭一　林良彦
日本電信電話株式会社　NTTサイバースペース研究所
bessho.katsuji@lab.ntt.co.jp

映像コンテンツのメタデータ自動生成の重要な要素技術であるトピックセグメンテーション技術の手法として、テキストを特異値分解によって得られる概念ベクトルの系列ととらえ、概念ベクトルの結束性に基づいてトピック境界を決定する手法を研究している。すでに、新聞記事テキストやニュース音声の認識結果テキストを入力としたとき、概念ベクトルを用いない従来手法と比べ高精度であることを確認している。本稿では、概念ベクトル生成時の条件がセグメンテーション精度に及ぼす影響について評価実験した結果を報告する。

1. はじめに

ネットワーク上に分散する大容量の映像コンテンツをブロードバンドを通して容易に配信可能となっている現在、目的とするコンテンツに高速に的確にアクセスできるようにするために、コンテンツのメタデータ技術が盛んに研究されている。メタデータの作成には多大の労力・時間を要するので、自動・半自動生成できることが望まれる。ニュース映像コンテンツは一般に複数のトピックから構成されていることが多く、検索の単位としてトピック単位に分割されている方がユーザにとっては利便性が高い。メタデータ自動生成にあたっては、コンテンツ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技術が重要な一要素技術となる。

筆者らはトピックセグメンテーション技術の手法として、テキストを単語の意味表現の一つである概念ベクトルの系列に変換し、概念ベクトルの結束性に基づいてトピック境界を決定する概念ベクトル結束度法を提案してきた[1][2]。すでに、新聞記事テキストやニュース音声の認識結果テキストを入力としたとき、Hearstの方法[3][4]のような単語表記を用いる手法と比べ高精度であることを確認している。概念ベクトル結束度法では、あらかじめ学習用コーパスを基に、特異値分解によって単語にベクトルを割り当てるが、特異値分解を行う際の条件（共起頻度窓幅、圧縮次元数、コーパス量等）により、生成される概念ベクトルが異なってくるため、これらの条件のセグメンテーション精度への影響が考えられる。本稿では、まず概念ベクトル結束度法について説明した後、これらの条件を変化させた場合のセグメンテーション精度の評価実験結果を述べ、セグメンテーション精度への影響度について考察する。

2. 概念ベクトル結束度法

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、学習用コーパス中の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。

表1　共起行列の例

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

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、自立語以外の単語を除去し、各自立語間の一定の窓幅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）。ここで各行に対応する単語が概念ベースに登録される単語（概念語と呼ぶ）であり、各列に対応する単語が概念語と共起する単語（文脈生成単語と呼ぶ）である。共起行列の各行をベクトルと見立てると、各概念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。概念語も文脈生成単語もともに高頻度語を一定数とるが、文脈生成単語の選定においては、あまりにも頻度数が多い単語はコーパス中に普遍的に存在し情報量があまりないと考えられるので、頻度順位が上位の何十個かの単語は除くこともある。

共起行列には一般にデータスパースネスがあり、また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は低いと考えられる。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、[5][6]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。

$n \times p$ の行列 X を共起行列としたとき、特異値分解により共起行列 X は、以下のように分解できる。

$$\underset{n\times p}{X} = \underset{n\times a}{U}\ \underset{a\times a}{S}\ \underset{a\times p}{V^t} \quad (1)$$

ここで、$a = \mathrm{rank}\, X \le \min(n,p)$、$U^tU = V^tV = I$（$I$：単位行列）であり、$S = (c_{ij})$としたとき、$c_{ii} \ge c_{jj} > 0$ $(1 \le^{\forall} i \le^{\forall} j \le a)$である。$c_{ii}$ $(1 \le i \le a)$をXの特異値と呼ぶ。$1 \le^{\forall} b \le a$に対し、

$$\underset{n\times p}{X'} = \underset{n\times b}{U'}\ \underset{b\times b}{S'}\ \underset{b\times p}{V'^t} \quad (2)$$

をとるとき、Xの第i行目の行ベクトルは、p次元空間中のあるb次元部分空間にある、$U'S'$の第i行目の行ベクトルに射影変換される。$U'S'$の各行ベクトルは、U'の対応する行ベクトルを、各座標ごとに対応する特異値の割合で伸縮したものなので、本研究では、変換後のベクトルを$U'S'$ではなく、U'の行ベクトルをその長さで割って単位ベクトルに正規化したものとする。変換後の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。

本研究では、ＮＴＴが開発した発想誘導型情報検索システム[7][8]を用いて概念ベース生成を行っている。

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の変化は、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、まずセグメント対象テキストを形態素解析した後、自立語以外の単語を除去する。残った自立語で概念ベースを検索し、対応するベクトルを付与する。

単語（ベクトルを割り当てられた自立語即ち概念語に限定する）間の境界位置（一定の単語数間隔でとってよい）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、極小点を直近の文境界に変換した上で、境界候補として出力する。

3. 評価実験

3.1. 実験データ

同一の学習用コーパスを用いても、概念ベース生成の際の条件を変えることにより、選定される概念語や概念語に割り当てられる概念ベクトルは異なってくる可能性がある。特に重要と思われる各種条件とセグメンテーション精度との関係についての評価実験を3.2.以降で順次述べていく。

精度を比較する際の基準となる概念ベースを以下のように生成した。学習用コーパスとしてはＣＤ－毎日新聞2000年版1年分の約106,000記事の見出しと本文の部分を用いた。概念語として高頻度語30,000語をとり、文脈生成単語として頻度順位が上位51番目以降の1,500語をとった。共起頻度をとる際の窓幅を1文とし、圧縮後の次元数を100とした。

セグメント対象テキストとしては、RWCP検索・要約用ニュース音声データベースにある、一人のプロアナウンサー男性がニュース放送原稿を読み上げて得られた音声データを、ポーズで区切られた区間ごとに書き起こしたテキストを用いた。この音声データ及び書き起こしテキストは1原稿につきそれぞれ1ファイルある。書き起こしテキストファイル41個を連接したものをセグメント対象テキストとした。また、対応する音声データファイル41個を連接したものを音声認識エンジンVoiceRex[9]を用いて認識させた結果得られるテキストも用いた。認識精度は単語誤り率7.8%である。書き起こしテキストも音声認識結果テキストも、ともに連接前の各ファイルを正解のトピック区間と仮定した。各セグメント対象テキストの情報は表2のとおりである。ここでは、ポーズで区切られた区間を文と呼んでいる。ベースラインとは、ランダムに選んだ文境界が正解トピック境界である確率を意味する。

セグメンテーション時の窓幅は、1トピックあたり平均のべ概念語数の3分の1とし、結束度は各単語境界に対し算出した。テキストの前の方と後の方で、前後いずれかの窓幅が規定の長さに満たない場合でも結束度を算出した。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平均で行った。セグメンテーションの出力結果としては、正解トピック境界数分だけの文境界を出力した。

表2　セグメント対象テキストの情報

	書き起こしテキスト	音声認識結果テキスト
トピック数	41	41
文数	877	1203
ベースライン	4.6%	3.3%
のべ自立語数	2775	2707

出力結果に対し、精度は以下のように算出される。

再現率＝正解の出力境界数/正解境界数

適合率＝正解の出力境界数/出力境界数

F値＝(2×再現率×適合率)/(再現率＋適合率)

3.2. 共起頻度窓幅

基準となる概念ベースに対し、共起行列作成時の共起頻度をカウントする窓幅を変化させたときの精度を検証する。学習用コーパスにおいて、1文あたり平均のべ自立語数が8.4個、1記事あたり平均のべ自立語数が105.6個であることを鑑み、窓幅を適宜変化させたときの書き起こしテキストに対するセグメンテーション精度のF値をプロットしたものが図1である。図1において±27単語などとあるのは、注目している概念語の前後27個の自立語をとり(但し記事の範囲を越えない)、その範囲内で概念語と文脈生成単語との共起頻度をとったことを意味している。±0文、±1文とは正解とする正解トピック境界からの範囲を意味する。

1文幅から±159単語幅に至るまで精度は殆ど変わらないが、1記事幅は他の場合と比べ精度が低い。これは記事の中には、のべ自立語数が230個を越えるような非常に長いものが存在するために、概念語と関係の弱い単語との共起頻度数も多く取られることがあり、共起パターンが適切でなくなることが原因であると推察される。より適切な共起パターンを生成するには、概念語からある程度の距離内の単語だけに限定する必要があるといえる。

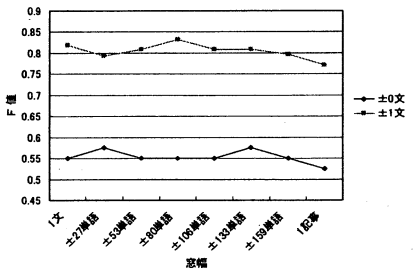

図1　共起頻度窓幅と精度の関係

3.3. 圧縮次元数

3.1.で述べたように基準となる概念ベースにおいて、共起行列におけるベクトル次元数は1,500、特異値分解による圧縮後の次元数は100である。圧縮後の次元数を変化させたときの精度を検証する。図2は、圧縮後の次元数を変化させたときの書き起こしテキストに対するセグメンテーション精度のF値をプロットしたものである。

1,500次元から750次元まで次元数が減るにつれ、精度は上昇している。概念ベクトルが単語の共起パターンを表す上でより適切なものに変化している

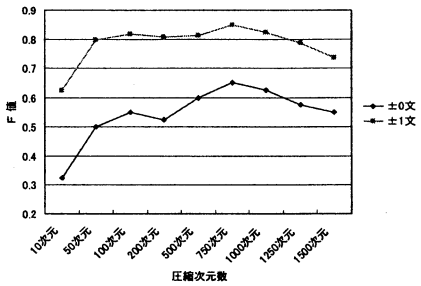

図2　圧縮次元数と精度の関係

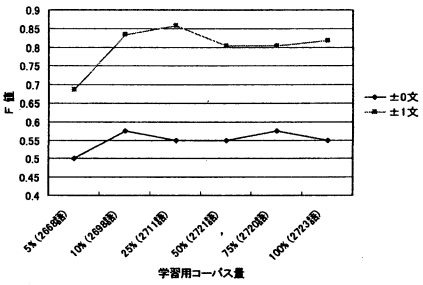

図3　学習用コーパス量と精度の関係

ことを意味しており、特異値分解の効果が現れている。±1文の範囲内なら、1,250次元から50次元に至るまで精度に大きな差はないが、10次元のときは1,500次元のときよりも大きく精度が落ちる。10次元にまで圧縮を行うと、元の共起行列の情報が大きく損なわれるため、概念ベクトルがもはや適切なものでなくなっていると推察される。100次元の場合は精度的にも計算量的にも妥当であるといえる。

3.4. 学習用コーパス量

基準となる概念ベースに対し、元となる学習用コーパスの量を変化させたときの精度を検証する。図3は、基準となる概念ベースの元となる学習用コーパスの量を100%としたときの、コーパス量を変化させたときの書き起こしテキストに対するセグメンテーション精度のF値をプロットしたものである。横軸におけるパーセンテージの後の括弧で囲まれた語数は、書き起こしテキスト中ののべ概念語数である。

100%から10%に至るまで目立った精度の差はないが5%のとき精度が大きく落ちる。これは学習用コーパスの量が少ないために、適切な共起パターンが得られなかったことが一つの原因と考えられる。また、これまで概念語であった単語が学習用コーパスにおいて十分な頻度数がなくなり、概念ベースにおける概念語集合の中に含まれなくなった結果、書き起こしテキスト中の自立語でベクトルを割り当てられない単語が増えたことがもう一つの原因と考えられる。

3.5. セグメント対象テキスト中の概念語数

3.4節でセグメント対象テキスト中の概念語数が少ないことが精度の低下の原因となることを述べた。セグメント対象テキスト中の概念ベクトルを割り当てられない自立語は、セグメンテーションにおいて全く考慮されない。そこでセグメント対象テキスト中の全ての自立語に概念ベクトルを割り当てた場合の精度を検証する。基準となる概念ベースの元となる学習用コーパスとセグメント対象テキストとをマージさせ、学習用コーパス中の高頻度語30,000語と、セグメント対象テキスト中の全自立語をマージしたものを概念語集合とした。この概念語集合において、頻度順位が上位51番目以降の1,500語を文脈生成単語として概念ベースを生成した。

書き起こしテキストと音声認識結果テキストに対し、セグメンテーション精度のF値をプロットしたものが図4、5である。両図において一部自立語不使用とあるのは、学習用コーパスのみから生成した概念ベースを使用した場合の精度を意味し、全自立語使用とあるのは、学習用コーパスとセグメント対象テキストをマージしたものから生成した概念ベースを使用した場合の精度を意味している。2723語のような語数はセグメント対象テキスト中ののべ概念語数である。また、横軸の±0文などとあるのは、正解とする正解トピック境界からの範囲を意味する。

いずれのテキストでも全自立語を使用した場合の方がそうでない場合よりも精度が微増する。セグメント対象テキストにおいて、全ての自立語に概念ベクトルを割り当てた方が、窓における概念ベクトルの重心が窓の意味をより的確に表し、結果として結束度やdepth scoreもより適切なものとなることが理由と考えられる。

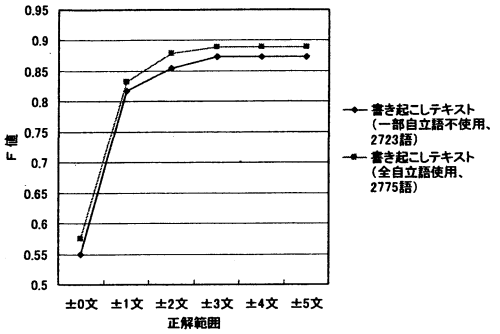

図4 書き起こしテキスト中の概念語数と精度の関係

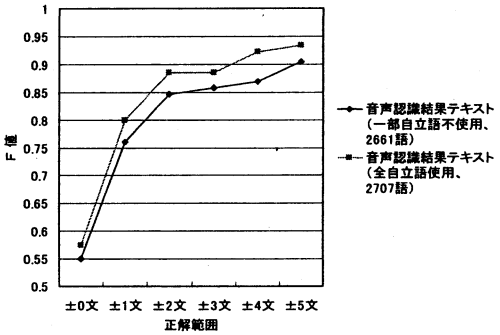

図5 音声認識結果テキスト中の概念語数と精度の関係

4. おわりに

概念ベース生成時の各種条件がセグメンテーション精度に及ぼす影響について評価実験を行い考察を行った。今後、学習用コーパスとセグメント対象テキストとの関係（内容的・時期的に近いか否か等）や、概念ベース生成時とセグメンテーション時のパラメータ間の関係（共起頻度窓幅と結束度算出時の窓幅との関係等）が、セグメンテーション精度に及ぼす影響についても考察していきたい。

参考文献

[1]別所克人：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]別所克人他：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，FIT2002，F-4, pp.201-202.

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

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

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

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

[7]Kato, T., Shimada, S., Kumamoto, M. and Matsuzawa, K.：Idea-Deriving Information Retrieval System, Proc. 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, pp.187-193(1999).

[8]熊本　睦，島田茂夫，加藤恒昭：概念ベースの情報検索への適用—概念ベースを用いた検索の特性評価，情報処理学会研究報告，Vol.SIG-ICS 115, pp.9-16(1999).

[9]野田喜昭他：音声認識エンジンVoiceRexの開発，日本音響学会秋季研究発表会，2-1-19, pp.91-92(1999-9).