分析区間長を可変としたテキスト分割手法

内海　慶　　藤井　敦　　田中　和世
筑波大学大学院図書館情報メディア研究科

1　はじめに

複数の話題を含む文書に対して話題の境界を検出し, 話題ごとに文書を分割するために「テキスト分割」が研究されている.

既存の手法 [1-6] では, テキスト内の語彙的な結束性を利用してテキスト分割を行う. 単語分布の類似度を結束性として利用する手法 [2, 4] や意味ネットワーク上での活性伝播に基づく結束性を利用する手法 [5] がある.

これらの手法では, テキストを一定の長さを持つ「分析区間」の単位に分割して, 分析区間ごとの結束性を計算する. しかし, 分析区間の長さが最適でないと種々の問題が生じる. 分析区間が短い場合は, 分析区間どうしで共通する単語が少なくなり, 「過分割」が起こる. 分析区間が長い場合は, 話題の境界が分析区間に内包されてしまい, 「検出漏れ」が起こる.

本研究は, 分析区間を可変長とすることで上記の問題を解消する.

2　提案するテキスト分割の手法

2.1　概要

本研究で提案するテキスト分割手法は, テキストを分析区間へ分割し, 隣接する分析区間どうしの類似度を用いて話題の境界を検出する.

本手法の特長は, 分析区間を可変長にしてテキスト分割を行い, 異なる分析区間長で得られる複数の結果を統合する点にある.

分析区間の長さを変化させると, 複数の「分割パターン」が得られる. 多くの分割パターンで支持される境界位置は分割の確信度が高い. そこで, 本研究では「投票方式」によって分割精度を向上させる.

また, 投票方式によって, 先行研究のように分析区間の長さを事前に最適化する必要がなくなるという利点がある.

本研究では, 「文」を分析区間の最小単位とする. ここで, 話題の境界は文と文の間にあるとする. 初期状態では, 全ての文と文の間が境界候補である.

提案するテキスト分割の手順は次の通りである.

1. 図 1 のように各境界候補（a〜d）から前後 N 文までを分析区間とする.
   図 1 において A〜E は文である. 図 1 は分析区間長を $N = 1, 2$ とした時に, 境界候補点をずらしながら分析区間を作成する様子を表している.
2. 分析区間に含まれる単語の重要度を計算し, 分析区間のベクトル（トピックベクトル）を作成する.
3. 2 つの分析区間についてトピックベクトルの類似度を計算する.
4. 境界候補を 1 文ずつずらしながら, 全ての境界候補について手順 1.〜3. の処理を繰り返す.
5. 手順 3. で計算した類似度の変化に基づいて, 話題の境界位置を検出する.
6. 分析区間の長さ N を 1 から 1 つずつ増やしながら, 手順 1.〜5. を繰り返す.
7. 図 2 のように, 各境界候補について, 6. までで得られた各分析区間長の分割パターンを用いて投票を行う.
   図 2 で,「○」は境界候補（a〜d）が境界として検出されたことを示し,「×」は境界候補が境界として検出されなかったことを示す. 例えば, 候補 b は 3 つの分割パターンから支持されており, 他の候補よりも境界としての尤度が高い.
8. ある境界候補が得た票数が閾値を越えた場合, その境界候補を境界と決定する.

先行研究に対する本研究の違いは, 手順 1.〜6. において分析区間の長さを変化させて複数の分割パターンを得ている点と, 得られた複数の分割パターンに対して, 手順 7. と 8. で投票を行って結果を統合する点である. 手順 1.〜5. は従来手法と同じである.

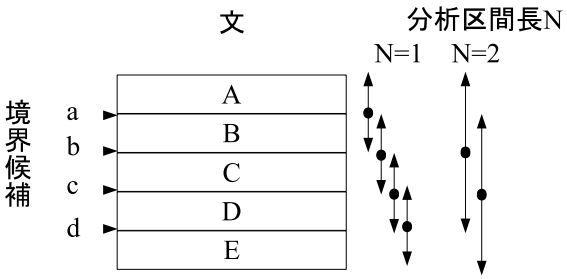

図 1　可変長の分析区間を作成する例

図 2　投票による境界検出の例

2.2　分析区間の作成

分析区間は, 文と文の間から前後 N 文を含むように構成する. 基準となる文と文の間が境界候補である. 基準点をずらしながら, 全ての境界候補について前後に N 文の長さの分析区間を作成する. N の値は 1 から始まり, 事前に設定した上限まで 1 ずつ増やす.

2.3　トピックベクトルの作成

トピックベクトルは, 作成された分析区間に含まれる内容語から, 分析区間の特徴を現している語を抽出して作成する. トピックベクトルの作成手順を示す.

1. 作成した分析区間に対して形態素解析を行い, 内容語（名詞と動詞）を抽出する.
2. 分析区間ごとに内容語の重要度を求める.
3. 各分析区間を表すベクトルにおいて, テキストの異なり語数を次元とし, 分析区間に含まれる内容語の重要度を要素にする.

手順 2. において, TF・IDF を用いて内容語の重要度を計算する.

$$TF(w_i, t_k) \cdot IDF(w_i) \qquad (1)$$

$$TF(w_i, t_k) = \log(\text{索引語 } w_i \text{が文書 } t_k \text{に現れる回数}) + 1$$

$$IDF(w_i) = \log\left(\frac{\text{全文書数}}{\text{索引語 } w_i \text{が現れる文書数}}\right) + 1$$

隣接する分析区間の類似度をトピックベクトルどうしのコサインで計算する.

2.4　境界検出

全ての境界候補に対して, 全ての分析区間長で類似度を計算すると, 図 3 のように分析区間長ごとに類似度グラフが描ける.

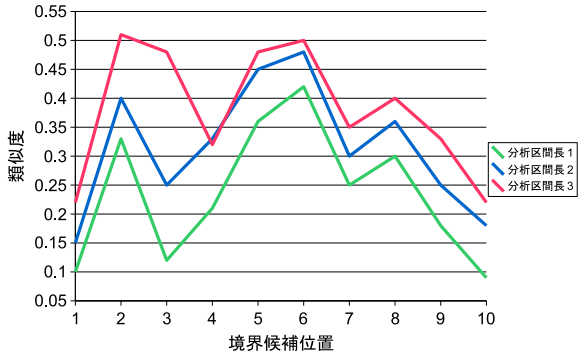

図 3　分析区間長ごとの類似度グラフ

図 3 のような類似度グラフから境界を検出する方法として 2 つの手法が提案されている. 1 つは類似度が閾値以下となる点を境界として検出する手法 [2] である. もう 1 つは類似度が極小となる点を境界として検出する手法 [4] である.

予備実験を行った結果, 後者の精度が高かったため, 本研究で用いた.

極小点で検出すると, 前後と比べて類似度がわずかに低くなった場所が検出される. Hearst の手法 [4] では, そのような極小点を境界として検出しないために, 極小点と前後の極大点を調べて, 極大点と極小点の類似度差が閾値を超えている場合のみ境界として検出する. 閾値は, 類似度の分散を考慮した平均類似度から計算する. しかし, 予備実験では効果的でなかったためこの制約は用いない.

2.5　投票

投票によって, 複数の分析区間長の結果を統合し, 検出された境界の確信度を計算する.

多くの分析区間パターンにおいて境界として検出されている場所は, より正しい境界である可能性が高い. 本研究では, 図 2 のように, 異なる分析区間長で得られた複数の分割パターンによる投票を行い, 票数が閾値を超える境界候補を話題の境界として検出する.

ただし, どの分割パターンでも一律で 1 票とするのではなく, 分割パターンごとに 1 票の重みを変える. 分割パターンごとの票の重みは実験的に設定した.

3　評価実験

3.1　実験データ

実験には, 放送大学の講義音声を人手で転記した 18 件と音声認識で自動的に転記した 5 件を使用した. また, 放送大学の教科書 23 件を使用した. なお, 1 件とは, 1 回分（約 45 分）の放送に対応する.

評価に用いる正解の境界として, 教科書は章と節を用いた. 転記と音声認識結果には, 人手で境界を付与した. 音声認識では, 話者適用と言語モデルの適応を行った [7].

3.2　テキスト分割精度

本手法を再現率, 精度, F 値を用いて評価した. ただし, システムが検出した境界が, 人手で付与した境界の前後 1 文以内であれば正解とした.

分析区間長として 1〜10 文の 10 段階を用意した.

投票に用いる各分析区間の重みを決定するために, データを二等分して交差検定を行った. 重みには, 半分のデータから得られた各分析区間長のテキスト分割精度を使用した. すなわち, 学習データに対する分割精度を向上させた分析区間長は 1 票の重みが大きくなる.

表 1 と 2 に教科書データと転記データを対象とした場合の実験結果をそれぞれ示す.「投票なし」の場合は, 分析区間長を最適化する必要がある. そこで, 複数の分析区間長で得られた結果の中から, F 値が最も高い結果だけを示す.「投票あり」は, F 値が最も高くなった閾値を与えた場合の結果である.

また, 比較対象として, 内山らのテキスト分割ツール [6] に対する再現率, 精度, F 値を表 3 に示す. 表 3 の「教科書」と「人手による転記」に示した数値は表 1 と表 2 の数値とそれぞれ比較可能である.

表 1　教科書データに対する実験結果

投票なし			投票あり		
再現率	精度	F 値	再現率	精度	F 値
0.677	0.292	0.408	0.567	0.315	0.405

表 2　人手による転記データに対する実験結果

投票なし			投票あり		
再現率	精度	F 値	再現率	精度	F 値
0.694	0.420	0.524	0.796	0.422	0.544

表 3　内山ら [6] のツールによる実験結果

教科書			人手による転記		
再現率	精度	F 値	再現率	精度	F 値
0.649	0.131	0.218	0.940	0.173	0.292

表 1 より, 教科書を対象にした場合は,「投票なし」の F 値が 0.408 であり,「投票あり」の F 値は 0.405 だった.

表 2 より, 人手による転記を対象とした場合は,「投票なし」の F 値が 0.524 だったのに対して,「投票あり」の F 値は 0.544 まで向上した.

また, 表 3 では, 教科書と転記の両方で F 値が低くなった. 内山らは, 明らかに話題が異なる複数のテキストを連結させて人工的なテキストを合成し, 実験対象とした. それに対して, 本実験では, 話題が徐々に変化したり, 話題の繰り返しを伴うテキストを使用した. そのため, 内山らの手法は上手く機能しなかったと考えられる.

今回の実験では, 投票の閾値を変化させてテキスト分割実験を行い, その中で F 値が最大となる閾値を選んだ. しかし, 適切な閾値を事前に設定することは難しく, これは今後の研究課題である.

教科書と転記データを用いて, 投票の閾値を変化させた時の再現率と精度と,「投票なし」の再現率と精度を図 4 に示す.

図 4 で,「投票なし」の再現率と精度は, Hearst の手法 [4] における境界検出の閾値を変化させることで再現率と精度を変化させた. これは, 極小点のみで検出を行うと, 各分析区間長に対してテキストの分割パターンは 1 つしか存在せず, 再現率と精度を変化させることが原理的に不可能だからである.

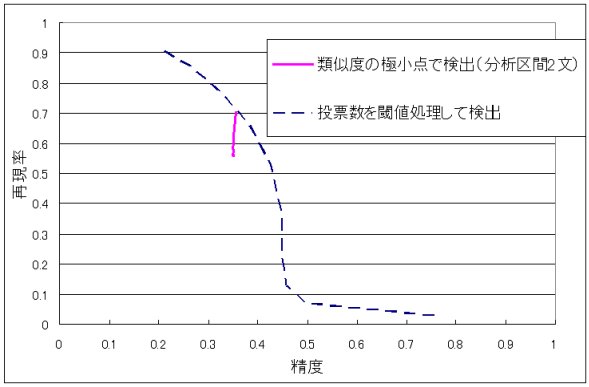

図 4　投票の有無による再現率と精度の比較

投票を行わない場合は, 閾値を変化させても精度の上昇が見られず, 再現率と精度の調整が行えないことが分かる. それに対して, 投票を行った場合は, 投票に対する閾値を調整することで, 再現率と精度の調整が行えることが分かる.

テキスト中の話題には詳細度がある. ニュース番組における個々のニュースは明らかに話題が異なる. それに対して, 放送大学のような講義では, より

細かな単位で話題が徐々に変化する. 本手法は, 閾値の設定によって分割すべき話題の詳細度に対応することが可能である.

音声認識と人手による転記に対して投票によるテキスト分割を行った場合の F 値を表 4 に示す. 実験では, 音声認識と転記それぞれに最適な閾値を与えた.

表 4　音声認識と転記によるテキスト分割の F 値

文書	単語誤り率	F 値	
		音声認識	転記
法と裁判 10	12.84	0.519	0.612
家族法 2	53.53	0.351	0.495
食物とからだ 2	39.86	0.492	0.585
太陽系の科学 2	48.77	0.423	0.482
古典古代の歴史 2	37.86	0.580	0.634

表 4 では, 転記の F 値が 0.5〜1.5 ほど高くなった. しかし, 単語誤り率と F 値の変化には特に関連がなかった. 今後は実験データを増やしてさらに検討する必要がある.

音声認識結果の 5 件と, 同講義の転記 5 件それぞれについて, 投票を行わなかった場合と行った場合の再現率, 精度, F 値を表 5 と 6 に示す.「投票なし」と「投票あり」の手法は表 2 と同じである.

表 5　音声認識結果での比較

投票なし			投票あり		
再現率	精度	F 値	再現率	精度	F 値
0.754	0.363	0.490	0.589	0.358	0.446

表 6　転記での比較

投票なし			投票あり		
再現率	精度	F 値	再現率	精度	F 値
0.764	0.497	0.602	0.761	0.439	0.557

表 5 と 6 で,「投票なし」と「投票あり」を比較すると,「投票なし」の方が F 値が 0.05 ほど高くなった. また,「投票なし」と「投票あり」で, 表 5 と表 6 を比較すると, どちらも表 6 の方が F 値が 0.11 ほど高くなった. このことから, 投票方式は音声認識の結果に対しては有効でなかった. ただし, 投票方式は音声認識結果に対して, 分析区間長を固定した場合のテキスト分割と同程度の頑健さを持つことが分かった.

4　おわりに

本研究は, 既存のテキスト分割手法における分析区間の長さに起因する問題を解決するために, 複数の分析区間パターンによる投票方式を提案した. また, 投票方式によるテキスト分割の評価実験結果について報告した.

今後の研究課題は, 投票数に関する閾値を自動的に最適化することと,「Web ページ」や「特許」などの様々なジャンルで実験を行うことである.

参考文献

[1] 望月源, 岩山真, 奥村学, 語彙的連鎖に基づくパッセージ検索, 言語処理, vol.6, No.3, pp. 101–126, 1999. pp. 33–64, 1997.

[2] 緒方淳, 山本夏夫, 有木康雄, 講義データを対象とした音声認識と構造化の検討, 情報処理学会研究報告, SLP37-14, pp. 79–84, 2001.

[3] 山本夏夫, 緒方淳, 有木康雄, トピックセグメンテーションに基づく講義ビデオの構造化の検討, 音声言語情報処理, 42–10, pp.59–64, 2002.

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

[5] 小嶋秀樹, 古郡廷治, 単語の結束性にもとづいてテキストを場面に分割する試み, 情報処理学会研究報告, 93-NL-95, pp. 49–56, 1993.

[6] 内山将夫, 井佐原均, 統計的手法による分野非依存のテキスト分割, 自然言語処理, vol.8, No.4, pp.19–36, 2001.
`http://www2.nict.go.jp/jt/a132/members/mutiyama/software.html#textseg`

[7] Atsushi Fujii, Katunobu Itou, Tetsuya Ishikawa, LODEM: A system for on-demand video lectures, Speech Communication. (To appear)