E-023

セグメントレベルを考慮した話し言葉のセグメンテーション
Segmentation of Spontaneous Speech Considering Segment Level

福田 雅志† 佐藤 和† 延澤 志保‡ 太原 育夫‡
Masashi Fukuda Kazu Sato Shiho Nobesawa Ikuo Tahara

1 まえがき

近年，新聞記事や論文を対象にした重要文抽出や要約の研究が数多く行われている．そのような研究においては，段落や章などの文書構造を利用した位置情報が有効であるということが知られている．しかし書き言葉と異なり，話し言葉においては，基本単位となる文や段落の境界が定まっていない [1].

このような位置情報を付与する技術として，文書を意味的なまとまりで分割するセグメンテーション技術がある．文書を話題ごとの集まり（セグメント）の集合として分割することで，話し言葉における文書要約や重要文抽出技術の精度向上やインデックス付けの技術が期待できる [2] [3]．しかし，従来の書き言葉を対象としたセグメンテーションは目的としているセグメントサイズが大きく，話し言葉にそのまま適用できない．

そこで本稿では，TextTiling 法の求める大きさと同じセグメントが得られる部分に着目し，最小のサイズから徐々に大きくしていき，境界を求めていくことで，テキストを求める数に沿ったセグメントに分割する方法を提案し，話し言葉へ適用する．本手法を講演録を用いて，複数の被験者により作成された境界位置を正解として評価する．

2 セグメンテーションの手法と目的

文書に複数の話題が存在する場合，文書中の各話題に対応して，同一語の連続，シソーラス上の同一概念に属する語の連続，共起しやすい語の連続などの意味的なつながりをもった語彙的連鎖が存在する [4]．一般に語彙的連鎖は文書中に複数存在し，１つの連鎖が出現している範囲では，その連鎖を構成する語に関する話題が述べられていると考えることができる．このような語彙的連鎖を認識し，各語彙的連鎖ごとに文書を分割することがテキストセグメンテーションである．テキストセグメンテーションの手法としては，併合型・分割型の２種の手法がある．

- 併合型 文書内の文や単語を最小単位とし，隣接する単位を結合する手法
- 分割型 分割されていない文書から話題分割のための境界を探索する手法

†東京理科大学大学院理工学研究科情報科学専攻
‡東京理科大学理工学部

書き言葉に対するテキストセグメンテーション手法はいくつか提案されている．しかし，書き言葉と話し言葉のセグメンテーションでは目的が異なる．書き言葉のセグメンテーションの目的は，筆者が明示的に与えた段落や章に捉われないような意味的なまとまりに分割することであり，得られるセグメントは大きい．それに対して，話し言葉のセグメンテーションの目的は，文や段落，章などの明示的な境界が無いテキストに，その構造を用いるような自然言語処理を適用するための前処理としてテキストを分割することである．

しかし，テキストの構造には，章や節，そして段落などの様々なレベルが存在し，目的に沿ったレベルで分割することが必要である．例えば，話し言葉のインデックス付けの前処理として行う場合には，どのくらいの数のインデックスを付与するのかで，求めるセグメントの数が決定され，要約の前処理として行う場合には，要約率毎に求めるセグメントのレベルが異なる．そこで本稿では，認定するセグメントのレベルを考慮したセグメンテーション手法について述べる．

3 セグメントレベルを考慮したセグメンテーション手法

話し言葉には語以外の最小単位が確かでなく，そのような構造を利用したセグメンテーションが実行できない．語彙的結束性のみでセグメンテーションを行う方法として TextTiling 法 [5] がある．ここでは，TextTiling 法をベースとして，セグメントレベルを考慮したセグメンテーションを行う．本節では，まず TextTiling 法について説明し，その上でセグメントレベルを考慮したセグメンテーション手法について提案する．

3.1 TextTiling 法

TextTiling 法は，まず文書を単語列に分割する．そして文書中のある単語列間の境界を基準点として，その左右に同数の単語列を包含した窓を設ける．左右の窓の類似度（結束度）を求め，基準点を一定間隔でずらしながら類似度の変化に着目し，グラフにおける類似度の極小点をセグメントの境界と推定する手法である (図 1).

窓間の類似度は，次に示す cosine measure で定義さ

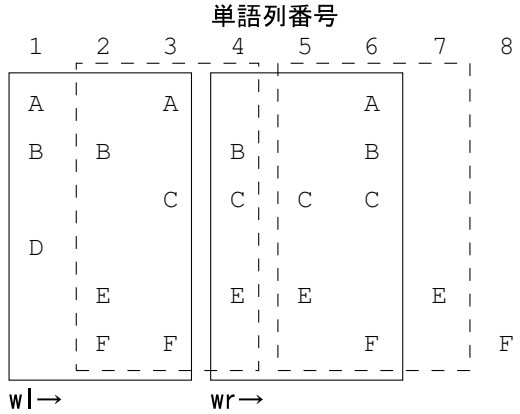

図 1: 類似度の計算

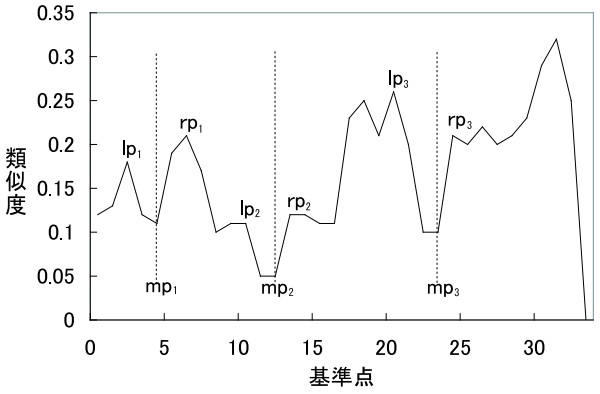

図 2: セグメントの検出

れている．

$$sim(wl, wr) = \frac{\sum_t f_{wl}(t) f_{wr}(t)}{\sqrt{\sum_t f_{wl}(t)^2 \sum_t f_{wr}(t)^2}} \quad (1)$$

ここで，wl と wr は，それぞれ左窓と右窓であり，$f_{wl}(t)$ と $f_{wr}(t)$ は，それぞれ，単語 t の左窓，右窓における出現頻度である．図 1 の例において，単語列 3 と 4 の境界を基準点として左右の窓における単語A,…, F の出現頻度による類似度は，式 (1) より 0.77 と計算される．式 (1) を用いて，基準点を文書の先頭から末尾に向かって一定間隔で移動しながら各基準点における左右の窓の類似度をプロットすると図 2 に示すようなグラフになる．ここで，類似度が極小値をとる基準点，すなわち，左右の窓の結束性が極小となる位置をセグメントの境界とする．

ただし，類似度の微妙な揺れを無視するため，極小点 mp の類似度 S_{mp} と，左側の極大点 lp における類似度 S_{lp}，右側の極大点 rp における類似度 S_{rp} の差を考慮し，以下の式で depth score（以下，d）を求め，d がしきい値 d_{th} を越えた場合にセグメントの境界とする．$\overline{S}$ は類似度の平均，σ は類似度の分散である．

$$d = (S_{lp} - S_m) + (S_{rp} - S_{mp}) \quad (2)$$

$$d_{th} = \overline{S} - \sigma/2 \quad (3)$$

3.2 セグメントの数を考慮したセグメンテーション手法

TextTiling 法には大きい窓幅を使うと大きい話題の切れ目が認識でき，小さい窓幅を使うと小さな話題の切れ目が認識できるという特徴がある [8]．この特徴を利用することにより，窓幅の大きさを変更することで，求めるセグメントのレベルを変更することが可能となる．

そこで本手法は、求める境界の大体の数を入力することでセグメントレベルを設定している．対象とするテキストを 2〜3 つに分割するセグメンテーションとテキストを 10 以上にも分割するセグメンテーションではセグメントのレベルが異なる．そこでセグメントのレベルとして数を設定し，その数に近いセグメントに分割することで，レベルに基づいたセグメンテーションを行う．

そのようなアルゴリズムは以下のようになる．

Step.1: 目的に沿ったセグメントの数の範囲を設定する．

Step.2: 最初は，窓枠に含まれる単語列を 1 列として，TextTiling 法を行い，テキストの境界を推定する．

Step.3: 推定された境界数が，設定した数を満たさない場合は Step.2 に戻り，類似度計算に使用する単語列を増やす．満たす場合はセグメンテーションを終了する．

このようなアルゴリズムにより，初めは最小単位の窓幅を用いて小さな段落を検出し，それが目的の数や大きさより細かすぎる場合には，段々と窓幅を大きくしていくことで，より大きなセグメントを認定し，セグメントの数を考慮したセグメンテーションが可能になる．

4 話し言葉の特徴を利用したセグメンテーションの精度向上

TextTiling 法を話し言葉に適用した場合，フィラーや言い直しなど，話し言葉特有の語が類似度の計算に悪影響を及ぼすという問題がある．話し言葉における「えー」などのフィラーは，意味を持たず文書構造に関係なく任意の位置に挿入される．特にフィラーの出現頻度の高い箇所では類似度を下げる要因となる．この理由は，左右の窓の類似度に対する影響であり，0 となる基準点が連続すると，境界を認定するための閾値に悪影響を及ぼし，段落検出の精度が下がり，正確なセグメント境界が認定できなくなるためである [7]．

また，単語列の境目を，そのままセグメントの境界とできないという問題がある．TextTiling 法は単語列間の境界を検出した後，その近くの形式段落をセグメントの切れ目としていた．しかし，話し言葉は形式段

落が存在しないので，単語列間の境界に近い，段落になりそうな箇所を改めて探さなければならない．

4.1 類似度の計算に用いる語の選別

話し言葉では類似度計算に不要な語が増え，さらに窓が小さくなると左右の窓における類似度が顕著に低くなるため，左右の窓の正確な類似度が計算できないという問題がある．本手法では，使用する語の種類を増やし，窓枠の類似度を全体的に上げることにより解決する．

不要な語に関しては，形態素解析の結果で得られた品詞情報により，フィラーを取り除くことは容易であり，言い直しも連続で繰り返した単語を 1 つにまとめることにより，類似度計算への影響を減らせる．

窓幅が小さい時の問題に対しては，単純に類似度に使用する語を増やす程，類似度において 0 になる部分が少なくなり，小さな窓幅での類似度の計算が可能になる．しかし，全ての語を使用することにより類似度が上がっても，それはセグメント境界を示すための意味的な指針とはならず，類似度が平らになり，逆にセグメント境界を求めるだけの十分な差が検出できなくなる．

そこで，本手法はフィラーなどの話し言葉特有の類似度計算に関係のない単語と共に,「が」や「は」などの助詞,「です」や「ます」などの助動詞を取り除く．セグメント境界を求めるのに意味のある単語を全て使用することにより，正確な類似度の計算を可能にした．

4.2 セグメント境界に適した単語列への分割

単語列間の境界をそのままセグメントの境界にできないという問題に対して，本手法では単語列をセグメントの境界に適した箇所で分割することで解決する．

本稿では，話し言葉における句読点の挿入手法 [1] を参考とし，句点を挿入する箇所でテキストを単語列に区切る．具体的には，話し言葉に対して形態素解析を行い,「です」「ます」を検出し，そこまでを単語列として扱う.「〜と」「で〜」など，他にも句点が挿入される箇所も存在するが，段落の区切りとはなりにくいので，今回は境界として明確であり，誤差も少ない「です」と「ます」だけを対象としている．このようにテキストを単語列に区切ることで，セグメントの境界として適切な箇所の類似度計算を行い，そのままセグメント境界としての認定を可能にする．

5 評価実験

5.1 実験コーパス

実験の対象として，国立国語学研究所から提供されている『話し言葉コーパス（以下，CSJ)』を使用した．CSJ は，学会講演など，主にモノローグを対象として収集，構築されているコーパスである．そのうち談話境界を与えられている 40 テキストを用いた．40 テキストの内訳は，模擬講演 25 講演と学会講演 15 講演である．模擬講演とは，いくつかの大きなテーマを与え，そのテーマに従って講演者に 10〜15 分程度講演してもらったものである．談話境界の認定作業は，まず 3 人の分析者が，10 個程度（5〜15 個）の意味的なまとまりに分割し，その 3 人の結果をすりあわせることにより，談話境界を認定している．

実験の目的は，人手で与えられた談話境界をシステムで検出できるか検証することである．そこで，談話境界が与えられたそれぞれのテキスト毎に約 10 個のセグメントを得るセグメンテーションを行う．形態素と品詞情報を得るためには，日本語形態素解析システム「茶筌」 [9] を使用した．

そして，単語列への分割と類似度の計算に使用する語の選別を行った．その結果「です」「ます」で区切られた箇所が，段落の境界と一致していた割合は 96.4 %であり，殆どの段落境界が認定可能である．

5.2 セグメントレベルを考慮したセグメンテーションの実験結果

評価は，フィラー,「が」や「けれど」などの接続詞や接続助詞のような単語によるずれを取り除き，本手法が認定するセグメント境界と人手で認定された境界が，完全に一致した場合のみを正解とする．その理由は，書き言葉のセグメンテーションと異なり，話し言葉のセグメンテーションはそれ自体が指針となるべきだからである．

評価としては，適合率と再現率を用いた．適合率，再現率は以下の式で求める．

$$\text{適合率} = \frac{\text{出力結果に含まれる正解境界数}}{\text{出力境界数}}$$

$$\text{再現率} = \frac{\text{出力結果に含まれる正解境界数}}{\text{正解境界数}}$$

さらに上記の適合率 P と再現率 R を用いて，F 値でもあわせて評価する．

$$F \text{ 値} = \frac{2PR}{P+R}$$

窓幅を固定してセグメンテーションとセグメントの数を設定したセグメンテーションの精度の差について表 1 に示す．

表 1: 窓幅によるセグメントの精度の比較

窓幅	適合率	再現率	F 値
単語列 1 で固定	19.3	28.6	21.9
単語列 2 で固定	24.0	11.6	13.9
単語列 3 で固定	5.7	1.3	2.1
セグメント数で変更	31.6	27.6	29.4

セグメントの数を考慮する本手法の結果は，書き言葉に対して用いられていた窓枠を固定する既存の手法をそのまま用いた結果に対して，再現率で若干落ちている箇所はあるが，適合率や F 値は全体的に上回っている．

5.3 話し言葉に対して精度を向上させたセグメンテーションの実験結果

また，TextTiling 法を話し言葉への適用した場合の精度の差を，先程と同じように適合率，再現率，F 値で表 2 に示す．

表 2: 話し言葉へ適用したセグメント精度の比較

精度向上	窓枠	適合率	再現率	F 値
無し	単語列 1 で固定	19.3	28.6	21.9
無し	単語列 2 で固定	24.0	11.6	13.9
無し	単語列 3 で固定	5.7	1.3	2.1
無し	セグメント数で変更	31.6	27.6	29.4
有り	セグメント数で変更	30.9	36.3	31.2

話し言葉に対して精度を向上させることにより，若干落ちた再現率が戻り，窓枠を固定する既存の手法に対して，適合率と再現率，F 値の全てで上回った．

6 考察

6.1 セグメントレベルを考慮したセグメンテーションの考察

表 1 の結果より，本手法は，既存のセグメントのレベルを考慮せず固定する手法に対して，ほぼ全ての結果において上回っている．この原因は，既存手法のようにレベルを固定した場合には，各々のテキストに対して，どのレベルでセグメンテーションを行えば良いかが判明しないためである．仮にあるテキストで，特定のレベルのセグメンテーションが有効であると判明しても，それを他のテキストにそのまま用いて有効であるとは限らない．しかし，本手法ではセグメントの数というレベルを設定することにより，各々のテキストに対して，レベルに応じた処理が可能である．それにより全体的に精度が向上している．これは目的に沿ったセグメンテーションにはセグメントレベルの考慮が必要であるという仮定を裏付ける結果となっている．

ただし，タスクのための前処理を目的とした場合でなく，何らかの基準により与えられたセグメントの分割のみを目的とした場合には，そのセグメントがどのようなレベルなのかを指定する必要がある．そのためには，他の学習的な手法により，与えられたセグメントのレベルを発見し，それを用いて本手法に適用することが考えられる．

6.2 話し言葉に対して精度を向上させたセグメンテーションの考察

表 2 の結果より，既存手法における最小の窓幅によるセグメンテーションで再現率が高い理由は，最も多くの数のセグメント境界を判定する手段であるので，単純に期待値が高いためである．しかし，これはセグメンテーションの目的とは本質的に異なる．

そのような処理に対しても，本手法を話し言葉に対して適用して精度を向上させることにより，既存手法に対して，適合率と再現率，F 値の全てで上回った．これは，本手法のように様々な窓枠において類似度の計算を行う手法では，小さい窓幅と大きい窓幅のどちらかだけでなく両方において正確な類似度を測定できることにより，精度が向上することを示した．

7 まとめ

本稿では，セグメント数を考慮した話し言葉のセグメンテーション手法を提案した．書き言葉と異なり，話し言葉には文や段落などの境界が与えられていない．そのため，書き言葉のセグメンテーションと異なり，話し言葉のセグメンテーションには，セグメントのレベルを考慮する必要があると考え，TextTiling 法の特徴を用いて，目的に応じてセグメントの数を設定し，その数に即したセグメント境界を認定した．

本手法の結果を，人手で認定した談話境界を用いて，それとの完全一致のみを正解とした評価実験を行った．セグメントのレベルを考慮し，また話し言葉の特徴を利用し精度を向上させ，適合率，再現率のそれぞれで，30.9%，36.3%という結果となった．この結果は，既存の TextTiling 法の結果を上回っている．

今後の課題としては，コーパスを利用し，同一語以外の語彙的関係を使い，より正確な語彙的結束性を数値化することや，学習的な手法と組み合わせることでセグメンテーションの精度を向上させること，そして実際にインデックス付けや要約，重要文抽出などの前処理に使用して，それらの技術に応用することが考えられる．

参考文献

[1] 下岡和也, 南條浩輝, 河原達也,“ 講演の書き起こしに対する統計的手法を用いた文体の整形, ”自然言語処理, Vol.11, No.2, pp.67–83, 2004.

[2] 望月源, 本田岳夫, 奥村学,“ 複数の表層的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.

[3] 新中庸介, 広畑誠, 古井貞熙,“ 講演音声のインデキシングを目的としたセグメンテーション手法の検討, ”日本音響学会 2005 年春季講演論文集, 1-5-4, pp.7–8, 2005.

[4] 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, ”自然言語処理, Vol.6, No.3, pp.101–126, 1999.

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

[6] 平尾努, 北内啓, 木谷強,“ 語彙的結束性と単語重要度に基づくテキストセグメンテーション, ”情報処理学会誌 トランザクション「データベース」Vol.41, No.SIG03–003, pp24–36, 2001.

[7] 福田雅志, 延澤志保, 太原育夫,“ 語彙的結束性に基づく話し言葉のセグメンテーション, ”言語処理学会第 11 回年次大会論文集, pp.620–623, 2005.

[8] 仲尾由雄,“ 語彙的結束性に基づく話題の階層構成の認定, ”自然言語処理, Vol.6, No.6, pp. 83–112, 1999.

[9] 奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座『形態素解析システム《茶筌》 version 2.3.3 使用説明書』, 2003.