語彙的結束性に基づく話し言葉のテキストセグメンテーション

福田 雅志　　延澤 志保　　太原 育夫
東京理科大学

1　まえがき

テキストを対象とした重要文抽出や要約の研究が数多く行われている．新聞記事や論文を対象にした研究においては，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．しかし論文と異なり，講演などの書き起こしにおいては，境界がはっきりと決まらないだけでなく，係り受け関係の交差やフィラー，そして倒置表現があるなど問題があり [1]，既存の手法を利用するのは困難である．

このような問題を解決する技術として，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）の集合として考えることで，文書要約や重要文抽出技術の精度向上が期待できる [2] [3].

本稿では，話し言葉に対するセグメンテーションとして，文書内の語の共起を考慮した語彙的結束性に基づくセグメンテーション法を提案し，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た場合の適合率・再現率により精度を評価し，提案手法の有効性を示す.

2　先行研究

2.1　セグメンテーション法

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の連鎖である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 [5].

- 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

講演の書き起こしなどの話し言葉文書のテキストセグメンテーションを行う際に考慮しなくてはならないのが最小単位である．話し言葉においては，音声認識結果の時点では，段落などの文書構造はもちろん，句読点も挿入されていない．従来の文を最小単位としている手法を音声認識結果に対して用いるためには，まず認識結果を文に分割する必要があり，その文に分割する際にセグメント境界となる文が明確に分割できていないと，セグメント境界と認識できない恐れがあるので，テキストセグメンテーションの段階では，文単位に分割する必要がない手法が望ましい.

2.2　TextTiling 法 [6]

TextTiling 法は分割型に属するテキストセグメンテーション手法であり，文書の意味的に関連の深い部分には，同一の語が繰り返し出現するという語彙的連鎖を利用する．まず，文書をある一定の長さの単語列に分割し，文書中のある単語列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，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話題の境界と推定する手法をとっている．窓間の類似度は，つぎに示す cosine measure で表される．

$$sim(wl, wr) = \frac{\sum_t f(t_{wl}) f(t_{wr})}{\sqrt{\sum_t f(t_{wl})^2 f(t_{wr})^2}} \qquad (1)$$

ここで，wl と wr は，それぞれ左窓と右窓であり，$f(t_{wl})$ と $f(t_{wr})$ 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話題の境界とする．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} を越えた場合に話題の境界とする．$\overline{S}$ は類似度の平均，σ は類似度の分散である.

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

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

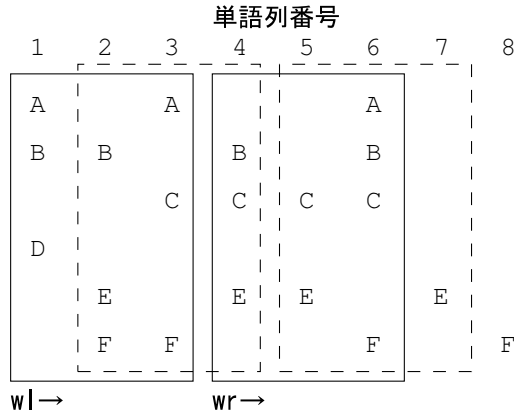

図 1: TextTiling 法

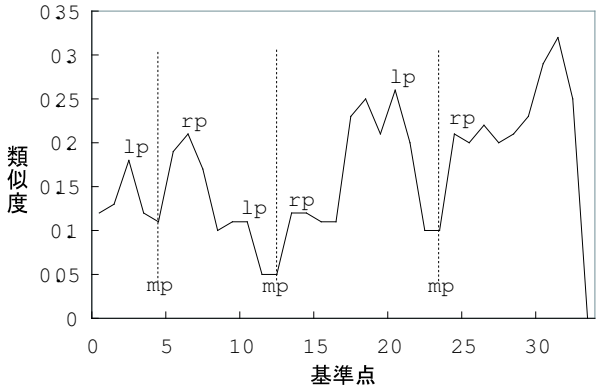

図 2: TextTiling 法によるセグメンテーションの例

この手法は，一定の語数にテキストを分割するので，テキストをあらかじめ句読点を挿入し，文に分割する必要がないという条件を満たしている．しかし，この手法は比較的長い書き言葉の文書を対象とした手法であり，話し言葉を対象とした場合には十分な精度が得られるとは限らない．

3 話し言葉のセグメンテーション

3.1 フィラーによる影響

「えー」などフィラーは，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．左右の窓の類似度が 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，テキストセグメンテーションの精度が下がる．表 1 にフィラーがある場合とそれを取り除いた場合のテキストの窓枠間の類似度の計算において 0 が出現しないテキスト（以下判定可能テキスト）の割合の表を示す（表 1）．フィラーがある場合に比べ，ない場合では小さな窓でも判定可能テキストが多くなる．これはフィラーの削除によりテキストを単語の列として考えたときに意味のある単語の密度が上がり，類似度の計算がしやすくなったためである．

表 1: フィラーの有無による判定可能テキストの割合

(窓，移動)	フィラーあり	フィラーなし
(50,5)	0.0	0.0
(60,5)	2.5	5.0
(70,5)	7.5	10.0
(80,5)	10.0	15.0
(90,5)	17.5	20.0
(100,5)	20.0	27.5
(110,5)	30.0	42.5
(120,5)	52.5	60.0
(130,5)	57.5	75.0
(140,5)	70.0	80.0
(150,5)	77.5	85.5

3.2 テキストの長さと窓枠による影響

TextTiling 法では，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点がある [5]．左右の窓に含まれる単語数が数百以上であれば，両窓に境界を判定するための十分な同一の語が出現することが期待できる．しかし，小さな範囲内では，同一の語が繰り返し出現する期待値は低くなる．よって，短い文書を対象とする場合には，設定できる窓幅が小さくなることもあり，左右の窓の類似度において 0 が出現する箇所が増える．

本稿では実験の対象として，国立国語研究所の『話し言葉コーパス』（以下 CSJ）のうち，被験者 3 名が談話境界を与えているもの 40 テキストを用いた．これらのテキストの単語数は 1195〜5571 語の平均約 2500 語のテキスト群である．これらのテキスト群を単語数の

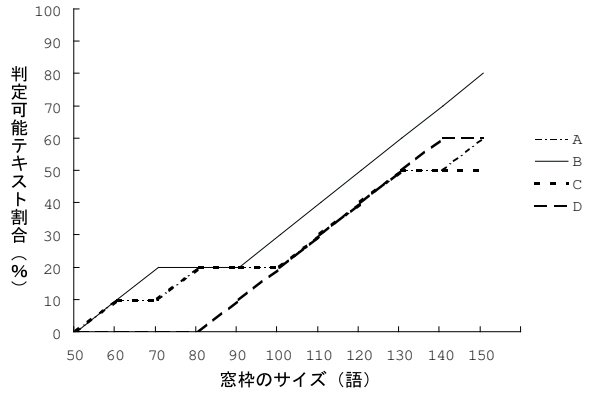

図 3: テキストの長さと窓枠の関係

少ない順に 10 テキストずつA～D（A：1195～1901 B：1943～2238 C：2240～3019 D：3127～5571）の 4 つに分けて移動幅は固定し，窓枠を変更しながら類似度を計算した．その判定可能テキストの割合を図 3 に示す．その結果，テキストの単語数毎では類似度の計算に大きな影響は見られず，判定可能テキストはテキストの単語数に影響しなかった．類似度において 0 が出現するのは，テキストのある箇所に同一の語が出現していないということであり，それはテキストの単語数に関わらず存在するという結果であった．

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傾向がある [7]．対象としたテキスト群にはセグメントとして 100 語以下のセグメントも見られた．フィラーを除いた結果においてもまだ幾つかの類似度に 0 が出現するテキストが見られ，それ以外のテキストの境界認識においても，テキストによっては境界認識の候補が殆ど検出されていないテキストもあった．これはセグメントに対して窓幅が大きすぎるということであり，より細かな測定が必要である．

4 TextTiling 法の話し言葉への適用

4.1 使用する語の選択

従来の名詞や記号などによる TextTiling 法では，フィラーなどもあり，類似度の計算に悪影響が出る．本手法では，これらを考慮して，まずどの語がセグメンテーションに効果的なのかを調べた．表 2 に用いた語の種類ごとの判定可能テキストの割合を示す．

表 2 から見て解るように，単純に類似度を計算するための語を増やすほど類似度が 0 になる部分が少なくなり，TextTiling 法を行うための窓幅をセグメントを求めるのに十分な程，小さく設定することができる．しかし，この結果は助詞や助動詞はおろか，全ての語を用いて類似度を計算すれば，より小さな窓幅で類似度が 0 になる箇所は少なくなると指摘するものではない．

表 2: 類似度計算に用いた語による判定可能テキストの割合

(窓，移動)	名記	名記動	名記動形
(50,5)	0.0	20.0	22.5
(60,5)	5.0	50.0	60.0
(70,5)	10.0	85.0	87.5
(80,5)	15.0	85.0	87.5
(90,5)	20.0	95.0	97.5
(100,5)	27.5	100.0	100.0
(110,5)	30.0	100.0	100.0
(120,5)	52.5	100.0	100.0
(130,5)	57.5	100.0	100.0
(140,5)	70.0	100.0	100.0
(150,5)	77.5	100.0	100.0

そのように，単純に語を増やせば全体的に類似度が増加するが,「が」や「は」などの助詞,「です」や「ます」などの助動詞はテキスト全般において頻出する語であるので，フィラーと同じように，このような語でブロックの類似度が上がってもそのブロックの示す意味的な比較には役に立たず，類似度が平らになり逆にセグメントを求めるだけの十分な差が検出できなくなる．

4.2 セグメント境界に適した窓枠の大きさ

対象のテキストに対して窓枠と移動幅を変えて実験を行ったが，あるテキストにおいて十分な窓の大きさが別のテキストに対して上手く行くとは限らず，固定した大きさの窓と移動幅では十分な程のセグメント境界を得ることはできなかった．そこで，フィラーを除いたことによる結果の向上に基づいて，余計な語を取り除いて TextTiling 法を行った．余計な語とはフィラーや「が」や「は」などの助詞,「です」や「ます」などの助動詞が該当する．これらのテキスト全般において頻出するブロック比較には役に立たない語を取り除くことにより，窓の大きさをそのブロックに適した大きさに変更する．その意味的表現について図 4 で説明する．

本手法は，図 4 の [ ] で示されるブロックの類似度計算において必要のある語だけを取り出し，その語のみの窓幅で TextTiling 法を行う．これにより，類似度計算に用いる語のテキスト中の出現分布に応じて窓サイズを変更することが可能になる．

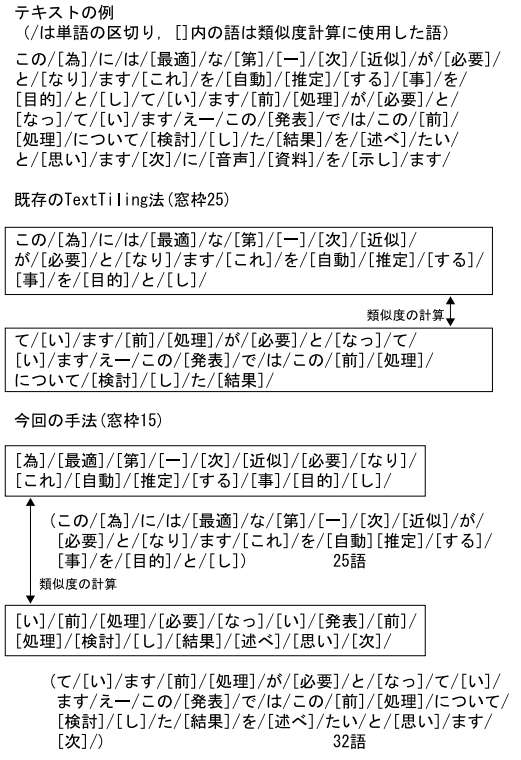

図 4: テキスト箇所に適した窓枠の変更

5　評価実験

4 章で説明した提案手法を実装し評価実験を行った．実験データと評価方法について説明する．今回実験の対象として，CSJ のうち談話境界を与えているもの 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．本稿では，フィラーや助詞助動詞を取り除く手法（4.1 節）と，窓に名詞未知語動詞および形容詞のみを用いた手法（4.2 節）の検討に，日本語形態素解析システム「茶筌」 [8] の形態素解析結果を用いた．テキストにより単語数やセグメントの大きさはそれぞれであるが，全ての文書に対して類似度に 0 が出ない最小単位の窓幅と移動幅において実験を行った．

評価指標としては，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-measure でもあわせて評価する．

$$F - measure = \frac{2PR}{P+R}$$

また，TextTiling 法では基準点の移動を文単位ではなく単語単位で行っている．そのため，その出力は文末にはならず，文の途中で現れる．そこで，評価の際には話題境界の候補とする出力の単語が決定されたら，対象テキストに対して自動的に文に分割し [1]，その単語が含まれる文が境界の文となっていれば正解としている．このように実際のセグメントに対してその近傍での出力も正解として評価している．実際の話題境界の正解を表す文が x 番目の文であった場合に，$x-1$ 番目の文を出力した場合も正解とする．適合率・再現率，F-measure で評価した結果を表 3 に示す．

表 3: TextTiling 法によるセグメンテーション結果

手法	再現率	適合率	F-measure
改良なし	17.3	42.1	25.8
フィラー等削除	22.8	43.4	27.1
窓サイズ可変	15.6	56.6	23.1
2 改良案併用	46.2	47.3	45.1

6　考察

TextTiling 法を話し言葉の文書に適用した場合，短い文書が多いため，類似度の問題が頻出し，人手で認定された境界に対してよりも大きな窓幅で計算しなければならなくなる．その結果，分割が少なくなり，従来の結果に比べ再現率が大幅に下がり，適合率も落ちた．そこで，4.1 節の手法により類似度の計算に用いる単語を増やすことで，類似度の計算において 0 の悪影響が減り，窓枠を小さくすることを可能にし，人手で分割したセグメントと近い数の境界の分割をすることができた．4.2 節の手法は，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語を省くということにより，適合率をあげることができた．しかし，意味的に必要な語のみ残すということは，全体的に類似度が平らになるということであるため，判定される境界はより少ないものとなり，再現率はさらに落ち込んだ．この両方の案を組み合わせることにより，テキストのセグメント境界を判定するのに必要な単語だけを用いて，それ以外の語を省き計算を行い，窓枠を小さく細かな判定をすることが可能となった．

この結果を用いて，さらに同一語の出現だけによるセグメンテーションではなく，新たに窓枠や閾値を変更する必要はあるが，書き言葉でも有効とされていた単語重要度やコーパスからの共起語を利用したセグメンテーションと関連させることにより，話し言葉においても精度の向上を期待できると考える．

7　まとめ

本稿では，話し言葉コーパスを対象としたセグメンテーション法を提案した．既存手法を適用した場合の問題を解決するため，対象語句の絞り込みに基づく TextTiling 法の改良手法を提案した．日本語話し言葉コーパスを対象として，人手で与えられたセグメント境界を正解とし適合率，再現率で評価した結果，それぞれ 28.9 %，5.2 %向上した．今後の課題としては，文書ごとに対する適切な窓枠と移動幅の関係，同一語以外の語彙的関係を使うためにコーパスを利用，または他の手法との融合などによる精度の向上が考えられる．

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,“ 講演のセグメンテーションを用いた音声要約手法の検討, ”日本音響学会 2004 年秋季講演論文集, 2-1-3, pp.41-42 2004.

[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] 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 2001.

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

[7] 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.3, pp. 83-112.

[8] 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』, 2003.11