社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－NL－162　(22)
2004／7／16

単語の結束度と文の表層情報を組み合わせたテキストセグメンテーション

松井祥峰　　乾伸雄　　小谷善行

東京農工大学　工学教育部

概要

文章を意味的なまとまり(セグメント)に分割することは、段落や章立てなどの整形がされていないテキストを見やすく整形する場合や、文書要約をする際の一つの手がかりとなる。従来のテキストセグメンテーションの手法の一つに、単語の結束度を用いた手法がある。この手法は、テキスト中の単語の出現頻度に基づいたロバストな手法ではあるが、それだけでは正しい境界の近くを推定することはできても、正確な境界の位置を推定することは困難であった。

そこで本研究では、文の表層情報から分類学習を行う手法と単語の結束度による手法と組み合わせることによって、正確なセグメントの境界を推定することを試みた。その結果、単語の結束度の結果に比べて 20～30%程度精度が向上した。

Text Segmentation using Cohesion Scores of Words and Surface Linguistic Cues

Yoshitaka MATSUI　　Nobuo INUI　　Yoshiyuki KOTANI

Tokyo University of Agriculture and Technology

Abstract

Text Segmentation is a key-technology for text indentation and text summarization. The well-known method for text segmentation is cohesion scores of words. This method judges borders of texts by frequencies of words robustly. The borders extracted by this method are usually placed near the correct border, but rarely becomes the correct border.

We propose to combine two methods, cohesion scores of words and surface linguistic cues to improve the performance of text segmentation. Experimental results showed about 20-30% improvement in recall, compared with a pure cohesion scores method.

1．はじめに

近年では、膨大な量のテキストデータが世界中に溢れており、これらを処理する技術が注目されている。テキストセグメンテーションは、テキストを意味的なまとまりに分割するタスクであり、文の構造解析や自動要約などの手がかりとして利用することや、未整形のテキスト文書を人間が閲覧しやすいように整形することに有効であると考えられる。

従来のテキストセグメンテーションの手法として、意味的に類似した単語間の表層的関係である語彙的結束性を用いた研究[6]や、単

語の出現頻度を基にした単語の結束度による研究[1][2]、単語の概念ベクトルを用いた研究[4][5]などがある。また、語彙的結束性を用いた研究では、自立語以外の表層情報も境界推定のためのパラメータとして利用している。

単語の結束度や概念ベクトルを用いた研究では、セグメントの境界から前後ある一定の区間(窓とよぶ)をとり、窓内に出現する単語の頻度や概念ベクトルなどをもとに、その境界の区切れ易さを判定している。しかし、このような単語単位での推定では正しい境界の近くを推定することはできても、正確な境界の位置を推定することは困難である。

そこで本研究では、単語の結束度によって推定された境界付近を、自立語以外の表層情報を分類学習した結果を用いて補正することで、テキストセグメンテーションの精度を向上させることを試みた。

2．単語の結束度を用いた手法

Hearst による単語の結束度を用いたテキストセグメンテーションの手法[2]では、ある境界から前後一定範囲の窓内で単語の出現頻度によるベクトルを生成し、その余弦測度を境界前後の文章の関連度としている。

ある境界 i における単語の結束度 C(i)は、境界の前の窓を b1、後ろの窓を b2、$w_{t,b}$ を窓 b における単語 t の頻度とすると次の式で定義される。

$$C(i) = \frac{\sum_t w_{t,b1} w_{t,b2}}{\sqrt{\sum_t w_{t,b1}^2 \sum_t w_{t,b2}^2}}$$

各文の境界ごとに単語の結束度を計算すると、その値は、セグメント内では大きく、セグメントの境界付近では小さくなると予想される。そこで、単語の結束度の極小点をセグメントの境界候補とし、前後の極大点との谷の深さを基準にその区切れやすさを判断する。ある極小点 i の谷の深さ D(i)は、前後の極大点を j、k とすると次の式で定義される。

$$D(i) = (C(j) - C(i)) + (C(k) - C(i))$$

また、局所的な極小点を除去するために、あらかじめ単語の結束度を前後一定範囲の結束度との平均値とするスムージングを行う。

単語の結束度が推定するセグメント境界は、D(i)の大きい方から順に、出力するセグメントの数や、あらかじめ設定された D(i)の閾値を基に決定される。

単語の結束度を用いた手法は、単語の出現頻度だけでテキストの境界を推定するため、テキストの種類を限定しないという特徴を持つ。しかし、単語ベクトルの窓幅に対して小さいセグメントでは、セグメントの検出が困難であったり、窓幅に対して大きいセグメントでは、そのセグメントの中でいくつもの境界を検出するといった欠点がある。一方で、セグメントの境界付近を推定できるものの、正確なセグメント境界を推定するには不十分であるといえる。

図１は、単語の結束度を用いてテキストセグメンテーションを行った結果の一部である。対象文書は、毎日新聞の社説 365 記事を連結したものであり、１つのセグメントが１つの記事に対応する。図中の縦線は正解の境界位置、単語の結束度によって推定された境界位置をそれぞれ示している。出力する境界数は記事数と同じ 365 とした。図１では４つのセグメント境界が存在する。単語の結束度が推定した境界のうち、126 文目と 157 文目の境界については正しい境界と一致している。しかし、197 文目、228 文目と推定した境界については、いずれも正しい境界との一文のずれがある。この文書における単語の結束度によって推定された境界の再現率は 31.2%であったが、前後一文のずれまでを正解とみなすと再現率は 87.7%となり、前後二文のずれまでを正解とみなすと 90.1%となった。このよ

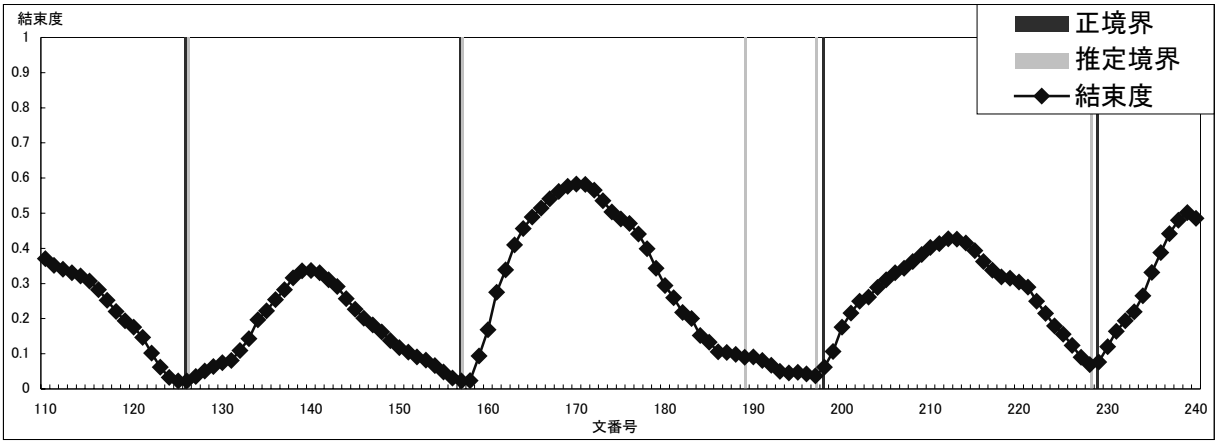


図１．単語の結束度とセグメントの境界

うに、単語の結束度による手法では、前後数文の誤差が生じ、セグメント境界の正確な位置を特定することは難しい。

３．文の表層情報を用いた手法

3.1. 学習パラメータ

単語の結束度によって推定された境界に生じる誤差を、セグメント境界付近の文の表層情報による分類学習によって修正することを考える。文の表層情報には、従来の研究[6][7]をもとに、一文に対して次のような手がかりを抽出した。

１）文頭表現

文の先頭に出現する接続詞や特定表現をさす。特定表現は、分類学習用のコーパスから抽出した文の先頭に出現する表現の中から 79 種類の表現を人手で選定した。また、接続詞については、EDR 電子辞書[8]を用いてその概念識別子によって 19 種類に分類した。学習パラメータとして、これら文頭表現の出現の有無とその種類を用いる。

表１．文頭表現の種類(一部)とその一例

種類	例	種類	例
0e84ad	しかし,あるいは,…	101a4d	なぜならば
3cfad7	従って,すると,…	3cfad7	それで,
3cf398	けれども,でも,…	特定表現	確かに,当然,…

２）文末表現

述語から文末までに出現する助詞や助動詞などによる表現をさす。抽出した文末表現を、単語とその品詞の種類によって 10 種類に分類した。尚、表現のあいまい性の問題は扱わず、１つの表現は１つの種類にしか属さないものとした。学習パラメータとして、これら文末表現の出現の有無とその種類を用いる。

表２．文末表現の種類

過去	た(av)	願望	たい(av)
現在	て(p)+いる(v)	限定	だけ(av)
推量	だ(av)+う(av)	可能	れる(v),できる(v)…
断定	だ(av),ない(a)	疑問	か(p)
義務	べし(av)	否定	ない(p)

a:形容詞　av:助動詞 p:助詞　v:動詞

(注：品詞は形態素解析ツール「茶筅」の結果による)

３）代名詞・連体詞

文に出現する「それ」「この」などの指示代名詞や連体詞をさす。学習パラメータとして、それぞれの品詞の出現の有無を用いる。

４）主語

副助詞「は」、または格助詞「が」に接続する名詞をさす。これを主語の候補とし、その単語が前後の窓内に出現しているかを調べる。

学習パラメータとしては、主語の候補となる単語の有無と、前後の窓内でのその単語の出現頻度をそれぞれ用いる。

3.2. 学習方法

表層情報を用いた推定は、ある境界がセグメント境界に成り得るかどうかという２値分類として扱う。本研究では、学習アルゴリズムとして、決定木学習（C4.5）と Support Vector Machine(以下 SVM)の二つを使い、比較した。

決定木学習法では、学習結果が木構造で表現できるため、人間にとっても見やすい学習結果が生成される。また、決定木を生成する過程で、重要度の高いと考えられるパラメータが木のルートに近い位置にくるため、各パラメータの重要度を人間が理解しやすい。

SVM は、近年よく用いられている２値分類学習の手法であり、テキスト分類のタスクなどで従来の学習法よりも優れていることが知られている。SVM の特徴として、サポートベクトルからのマージンが最大になるように境界を設定するため、一般的に、少ない事例やスパースネスなベクトルに対してもある程度の成果が得られることが知られている。また、高次元の入力に対して過学習しにくいという性質も持っている。さらに、カーネル関数を適用することによって非線形問題の学習も可能である。本研究では、多項式カーネルを用いて、表層情報を学習した。

学習に用いたパラメータ集合には、3.1 で述べた文の表層情報を前後一文の範囲で抽出する手法と、前後二文の範囲で抽出する手法の二種類を用いた。

3.3. 表層情報の学習

学習コーパスから文の表層情報を抽出し、決定木および SVM で学習させた。訓練データとして、正しい境界とその前後２文の範囲の境界をそれぞれランダムで選択した。

表３．決定木の学習結果

パラメータ	再現率	適合率
前後一文	77.4%	83.2%
前後二文	80.8%	87.8%

表４．SVM の学習結果

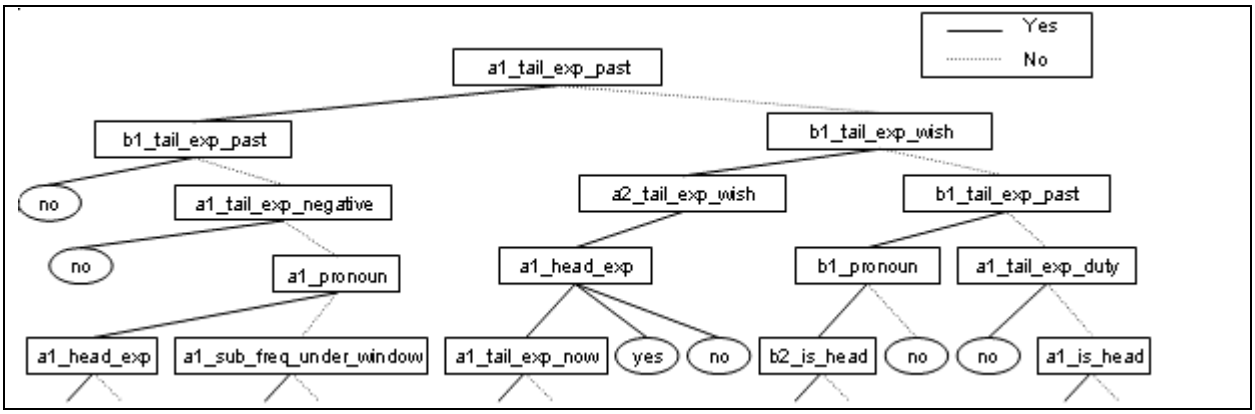
パラメータ	再現率	適合率
前後一文	71.4%	78.0%
前後二文	75.5%	81.2%

学習コーパスとして、毎日新聞の社説 1 年分（666 記事）を用意し、連結する順序を入れ替えることによって６つの文書を用意した。このうち、５つの文書の訓練データを学習させ、残りの文書の訓練データに対して、それぞれの学習アルゴリズム、パラメータ設定ごとに学習精度を求めた。表３は決定木、表４は SVM における学習結果である。値は６通りの平均値を示している。

表層情報を学習した結果では、決定木の方が SVM よりも学習精度が 5〜6%程度高くなった。また、パラメータ設定ではどちらの学習方法においても、前後二文のパラメータ設定の方が前後一文のものよりも 3〜4%程度高い結果となった。一方、再現率と適合率を比べると適合率の方が 6〜7%程度高く、これは訓練データにおける負例の個数が正例のものに比べて多かったためであると考えられる。

決定木で生成した木は、重要度の高いパラメータほどルートに近い位置で判定が行われる。図２は、前後二文をパラメータとしたときの決定木の一例である。ここでは、上位５階層までの木の状態を示している。また、表層情報のパラメータでは、境界の一文前に b1、二文前に b2、一文後に a1、二文後に a2 の接頭辞をそれぞれつけている。

表層情報の学習によって生成された分類木では、文末表現によるパラメータが上位階層に置かれる傾向がみられた。新聞の社説の場合、記事の先頭に題材とするニュースや事実などをとりあげ、記事の最後に筆者の主張をまとめるといった傾向が見られたため、事実と意見を区別する一つの手がかりである文末表現が重要度の高いパラメータとして選出されたのだと考えられる。



tail_exp_xx: 文末表現 xx の有無　sub_freq_under_window: 主語候補の境界より後ろの窓での頻度
is_head: 文頭表現の有無　head_exp : 文頭表現の種類　pronoun: 代名詞の有無

図２．表層情報を学習した決定木の一例（上位５階層まで）

一方で、境界前後の表層情報の方が上位階層に置かれるという傾向もみられた。

４．テキストセグメンテーション実験

4.1. 実験方法

毎日新聞の社説を対象に、各記事を連結したものを一つの文書として、正しい社説の境界を推定した。

まず、社説を学習用と実験用の２種類に分け、それぞれ連結する記事の順番を変えることによって複数の文書を作成した。学習用の文書は 3.2 で精度を調べたものである。また、実験用文書には 365 記事の社説を用意し、５通りの連結方法によって５つの文書を用意した(文書 A～E とする)。

実験では、単語の結束度における窓の大きさを 150 単語とし、結束度のスムージングの範囲を１とした。実験用文書における１セグメントの文数、単語数を表５に示す。

表５．１セグメントあたりのデータ

	平均	最小	最大
文数	29.6	19	56
単語数	351.5	261	740

それぞれの境界候補 i に対して次の処理を行う

Step1. i が境界として正しいと判断された場合
→ **i を境界とする**

Step2. n=1 から訂正範囲Nだけ次の処理を行う

Step2.1. i + n および i – n が境界として正しいかどうか調べる

Step2.2. i + n のみが境界として正しいと判断された場合
→ **i + n を境界とする**

Step2.3. i - n のみが境界として正しいと判断された場合
→ **i - n を境界とする**

Step2.4. i + n と i - n の両方が境界として正しいと判断された場合
→ **単語の結束度の小さい方を境界とする**

Step3. いずれの境界も正しいと判断されなかった場合
→ **i を境界とする**

図３．境界候補の修正アルゴリズム

4.2. アルゴリズム

単語の結束度と文の表層情報を組み合わせたテキストセグメンテーションのアルゴリズムについて説明する。

まず、単語の結束度を用いてセグメントの

表６．学習法とパラメータごとの再現率

	文書 A	文書 B	文書 C	文書 D	文書 E
決定木(前後一文)	61.1%	48.8%	51.8%	47.1%	48.8%
決定木(前後二文)	58.6%	47.4%	48.5%	47.4%	49.3%
SVM(前後一文)	66.6%	53.2%	54.5%	51.5%	54.5%
SVM(前後二文)	64.9%	53.2%	53.7%	51.5%	54.0%

表７．SVM を用いた前後１文パラメータにおける精度

	本手法	単語の結束度	前後一文許容	訂正率	誤り率
文書 A	66.6%	31.2%	87.7%	67.5%(139)	8.8%(10)
文書 B	53.2%	28.5%	69.6%	66.0%(99)	8.7%(9)
文書 C	54.5%	26.0%	71.2%	70.3%(116)	12.6%(12)
文書 D	51.5%	29.0%	69.3%	66.6%(98)	15.1%(16)
文書 E	54.5%	28.5%	70.7%	68.8%(106)	10.6%(11)

境界を推定する。その後、表層情報を分類学習させた結果を図３のように適用する。
図３では、単語の結束度が導き出した境界候補を表層情報を分類学習させた結果によって境界位置を前後 N 文の範囲で修正している。このアルゴリズムでは、はなるべく単語の結束度による境界候補に近い方を優先して修正している。また、表層情報からセグメントの境界が決定されなかった場合には、単語の結束度による境界候補を境界としている。

4.3. 実験結果

4.3.1 学習方法別の結果

3.2 で述べた、それぞれの学習方法ごとに実験を行った。単語の結束度による境界の推定では、実際の記事数と同じ 365 個の境界を決定するようにし、表層情報を用いた推定では、前後１文の範囲(N=1)で境界を訂正した。表６は文書 A における実験結果である。

学習方法別では、決定木よりも SVM の方が精度が高かった。学習パラメータ別では、前後一文の方が精度が高くなる傾向はあるが、それほど大きな差は見られなかった。

3.3 では決定木の方が学習精度が高かったにも関わらず、実験結果では SVM の方が精度が高くなっている。これは決定木が過学習をしてしまったためだと考えられる。3.3 で用いた文書は社説の連結順序を変えたものであり、抽出される文の表層情報には制限があった。従って、ある特定の表層情報に偏った決定木が生成されてしまったと考えられる。

表７は、SVM の前後一文の表層情報をパラメータとしたときの文書ごとの精度である。本手法と単語の結束度の精度は再現率を示している。また、前後一文許容とは、単語の結束度によって推定された境界が、正しい境界と前後一つの範囲でずれていた場合も正解とみなしたときの再現率である。これは、表層情報を用いた推定をした後の目標値となる。

訂正率とは、表層情報を用いた推定によって訂正が可能であった境界のうち、実際に正しく訂正が行われた割合である。

$$\text{訂正率} = \frac{\text{正しく訂正された数}}{(\text{前後一文許容の正解数}) - (\text{単語の結束度の正解数})}$$

誤り率は、単語の結束度によって正しく推定できた境界数のうち、表層情報を用いた推定で誤って訂正してしまった割合である。カッコ内は個数を表している。

表８．取捨選択アルゴリズムによる SVM 前後一文パラメータの精度

	再現率	適合率	単語の結束度	前後一文許容	訂正率	誤り率	削除率
文書 A	60.3%	78.3%	31.8%	90.7%	67.9%(140)	36.2%(42)	65.2%
文書 B	47.4%	68.1%	29.6%	72.3%	66.0%(103)	35.2%(38)	63.2%
文書 C	49.3%	68.4%	27.1%	73.7%	70.3%(118)	37.4%(37)	63.3%
文書 D	47.4%	67.6%	29.0%	71.0%	67.3%(103)	34.0%(36)	61.7%
文書 E	48.5%	66.8%	29.0%	72.6%	69.2%(110)	36.8%(39)	53.3%

$$\text{誤り率} = \frac{\text{訂正誤り数}}{\text{単語の結束度の正解数}}$$

$$\text{削除率} = \frac{\text{誤りを削除した数}}{(\text{境界候補数}) - (\text{前後一文許容の正解数})}$$

表層情報を用いることによって、決定木で 50～60%程度、SVM で 65～70%程度の割合で正しい境界に訂正が行えた。また、逆に表層情報を用いたことによって、誤って訂正してしまった割合は、決定木学習で 10～20%程度、SVM で 10%前後と訂正率に比べ低い値であった。

4.3.2 取捨選択アルゴリズムによる結果

図３のアルゴリズムでは、単語の結束度によって推定された境界候補の数だけ境界を決定していた。従って、単語の結束度によってその近傍が境界候補とならなかったセグメント境界については、表層情報を用いても抽出することはできなかった。そこで、単語の結束度によって推定される境界候補の数を増やし、表層情報を用いてこれらの境界の取捨選択を行うようにアルゴリズムを修正した。アルゴリズムの修正は図３における Step3 を取り除くだけである。

単語の結束度によって推定される境界数を、400 に増やして実験を行った。学習アルゴリズムは 4.3.1 の実験で最も精度の高かった、前後一文の表層情報をパラメータとした SVM を用いた。訂正範囲は前後１文とした。実験結果を表８に示す。削除率とは、単語の結束度が誤って推定した境界のうち、表層情報を用いて削除できた割合である。

取捨選択アルゴリズムを適用することによって、誤り率が 35%前後に増加し、結果的に再現率が 5～8%程度低下してしまった。しかし、一方では、単語の結束度によって間違って推定された境界候補を 60％程度の割合で削除したことで適合率が向上している。

５．考察

学習方法別の実験と取捨選択アルゴリズムによる実験のいずれの場合でも、表層情報を分類学習させる手法を用いることによって精度が向上した。これは、訂正率が誤り率より十分高く、単語の結束度が推定した境界に、訂正が必要な境界が多く含まれていたためだと考えられる。本研究で用いた社説のような文書では、単語単位でのテキストセグメンテーションでは正しいセグメントの境界を特定することは難しいと予想されるため、本研究で提案する手法は有効であるといえる。

しかし、境界候補を取捨選択した場合の誤り率が高くなったため、再現率が下がってしまった。今回用いた学習パラメータは、主に境界付近の文のみに着目しており、それより前後の文脈を考慮していない。今後の課題は、文を構文を解析するなどして前後の文脈を考慮した学習パラメータを設計することであると考える。

６．おわりに

本研究では、単語の結束度によるテキストセグメンテーションに、自立語以外の表層情報を分類学習させる手法を組み合わせて実験を行った。その結果、単語の結束度における手法に比べ、正しいセグメント境界を推定する精度が向上した。今後は、分類学習の精度を上げるため、前後の文脈を考慮した学習パラメータの設計が望まれる。

謝辞

本研究の一部は、日本学術振興会科学研究費(15300269)の援助を受けて行われた。

参考文献

1) Hearst , M.A. : Multi-Paragraph Segmentation of Expository Text, Computational Linguistics,
2) Hearst, M.A : TextTiling: Segmenting Text, Computational Linguistics Vol.23, No.1 pp34-64(1997)
3) Thorsten Joachims：Text Categorization with Support Vector Machines: Learning with Many Relevant Features, Proceedings of ECML-98, 10th European Conference on Machine Learning(1997)
4) 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,pp2650-2662(2001)
5）別所克人：クラスター内変動最小アルゴリズムに基づくトピックセグメンテーション,自然言語処理研究会, 154-25, pp177-183 (2003)
6) 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, pp43-58 (1999)
7) 福本淳一, 安原宏：文の連接関係解析に基づく文章構造解析, 自然言語処理研究会, 88-2,(1992)
8) 日本電子化辞書研究所：EDR 電子化辞書使用説明書, 1995