ウェブ検索を利用したテキストセグメンテーション法

阿部　直人† 　田邊　勝義† 　奥田　英範†

† 日本電信電話株式会社 NTT サイバーソリューション研究所
〒 239-0847 神奈川県横須賀市光の丘 1-1
E-mail: †{abe.naoto,tanabe.katsuyoshi,okuda.hidenori}@lab.ntt.co.jp

あらまし　これまでにテキストセグメンテーションを行う手法として、語彙的結束性や統計的情報を用いた手法、或いは接続詞や副詞、文頭表現や文末表現などの単語の表層的情報を利用する手法など、様々な従来手法が提案されている。しかし、それらの統計的情報や表層的情報は事前に準備された学習用データから算出されるものが多い。一般的には、そのような学習用データを用意することなくテキストセグメンテーションを行えることが望ましい。そこで、本論文ではウェブ検索を利用して学習用データを必要としないテキストセグメンテーションを行う手法を提案する。goo ニュース記事を用いてテキストセグメンテーションの実験を行い提案手法の有効性を検証した。

キーワード　ウェブ検索, テキストセグメンテーション, 多重ジャンル抽出

Text Segmentation Method Using World Wide Web Search

Naoto ABE†, Katsuyoshi TANABE†, and Hidenori OKUDA†

† NTT Cyber Solutions Laboratories, NTT Corporation
Hikarinooka 1-1, Yokosuka-Shi, Kanagawa, 239-0847, JAPAN
E-mail: †{abe.naoto,tanabe.katsuyoshi,okuda.hidenori}@lab.ntt.co.jp

Abstract　Many methods have been proposed for text segmentation based on the cohesion scores of words, surface linguistic cues, and so on. These methods, however, demand the preparation of a training text database beforehand to calculate such statistical and semantic information. In general, it is desirable to be able to segment a text without recourse to a training database. This study, therefore, proposes a method for text segmentation based on world wide web searches instead of a training text database. An experiment is carried out to investigate the effectiveness of the proposed approach as text segmentation method using the Japanese texts known to goo news.

Key words　multi-genre extraction, world wide web search, text segmentation

1. はじめに

近年、膨大な量のテキストデータがウェブ上に溢れており、ユーザに必要な情報だけを取り出す技術が着目されている。テキストセグメンテーションは、テキストを内容的なまとまりの単位である意味段落に分割する処理である。収集したテキストから評判や意見を分析する際、従来はキーワードが含まれているテキストのページ全体、或いはキーワード周辺の限られた領域を取り出すことが多い。そのため、不要な意見や評判、或いは文が途切れたために意見や評判が抜ける等の問題があった。それに対し、テキストセグメンテーションによりキーワードに関係する意味段落を適切に抽出することで、意見や評判の解析精度の向上が期待できる。また、複数の内容が混在している未整形のテキスト文書に対して、テキストセグメンテーションにより意味的なまとまり単位に整形することで、計算機でテキストの自動分類を行うことや人間が閲覧し内容を理解するのに役立つと考えられる。

これまでにテキストセグメンテーションを行う手法として、語彙的結束性や文脈の変化を評価するパラメーター等の統計的情報を用いた手法 [1–4]、或いは接続詞や副詞、文頭表現や文末表現等の単語の表層的情報と統計的情報を組み合わせた手法 [5,6]、確率的統計モデルに基づく手法 [7,8] など、様々な手法が提案されている。それらの統計的情報や表層的情報は事前に準備された学習用データから算出されるものが多い。最近では、様々な言語資源を利用することが可能となり、その例としては新聞記事コーパスが挙げられる。しかし、テキストセグメンテーションを行うことができるテキストが、利用する学習用データに含まれている分野、或いは関連する分野に限定される問題がある。また、十分な精度でテキストセグメンテーションを行うためには、大規模な学習用データを作成する必要があり、

一般的に人手で作成することは困難である。可能ならば、そのような学習用データを必要とせずにテキストセグメンテーションを行うことが望ましい。

そこで、本論文では事前に学習用データを用意しない代わりに、ウェブ検索を用いてテキストセグメンテーションを行う手法を提案する。ウェブ上には膨大な情報が蓄積されていることから、ウェブを様々な情報をもつ辞書的な存在と見なすことができる。従って、学習用データが無くてもウェブ検索を利用してテキストの内容を追跡するために必要な語彙的、或いは統計的な情報を取得できると考えられる。また、ウェブには自動的に最新の情報が追加されるため、ウェブを利用することでテキストに含まれる内容の時期や分野を問わずテキストセグメンテーションを行うことできると考えられる。実際に、我々はウェブ上で検索を行うことで、過去や現在を問わず様々な分野における単語の意味や物事の内容を調べることが可能である。最近では、ウェブ検索を利用してある専門用語の分野を判定する研究が行われている [9]。

本論文では、2 章で関連する研究について紹介する。そして、3 章において提案手法の概要を説明する。4 章では goo ニュース [11] を用いた提案手法の有効性を検証する実験内容を説明し結果を報告する。また、5 章にて実験結果の考察を行い、最後に 6 章において本論文のまとめと今後の課題について述べる。

2. 関 連 研 究

提案手法では事前に学習用データを用意することなく、ウェブ検索を利用してテキストセグメンテーションを行う。そのために、本論文では着目した関連研究が二つある。

一つは西脇らの関連記事を利用したテキストセグメンテーション手法である [2]。西脇らの手法は与えられたテキストに対して、大規模コーパスから関連記事を取得し、その情報を利用してテキストセグメンテーションを行う。大規模コーパスからの関連記事を利用することで、与えられたテキストの記述内容と関連した多くの単語をコーパスから取得できる。その結果、文脈を適切に考慮することができ、未知語が含まれている場合でも分割精度の劣化を低減することができたと報告されている。この観点から、関連記事を用いることで、与えられたテキストから得られた情報だけを利用する場合と比べて文脈の解析精度を改善できると考えられる。

もう一つの研究は、木田らのウェブを利用した専門用語の分野判定に関する研究である [9]。木田らの研究はテキストセグメンテーションに関するものではないものの、ウェブを利用してあるキーワードに対する検索結果に基づいて分野を特定する点に着目している。具体的には、ある専門用語が出現する文書を収集し、文書の内容の偏り度合を判定してその用語の分野を特定する点である。この観点から、適切な検索語を用いてウェブ検索を行うと、検索結果に現れるウェブページから検索語に関係の強い単語を得ることができると考えられる。

本論文ではウェブ検索を利用することで、あるキーワードに関連する複数の文書をウェブから取得できる点に着目した。ウェブ上から適切に関連記事を取得することができれば、西脇らの手法と同様にテキストの内容を追跡できる。その結果、事前に学習用データを用意しなくてもテキストセグメンテーションを行うことが可能となると考えられる。

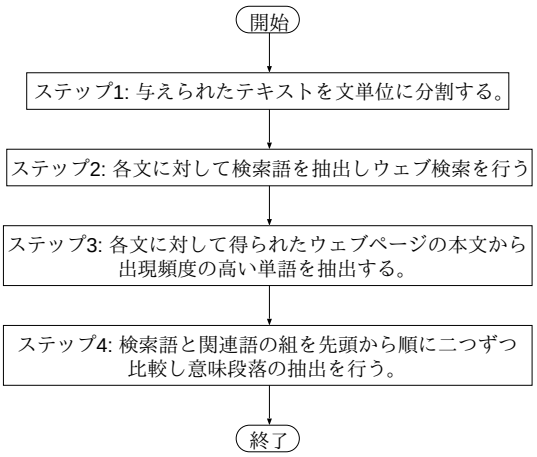


図 1　提案手法の処理の概要

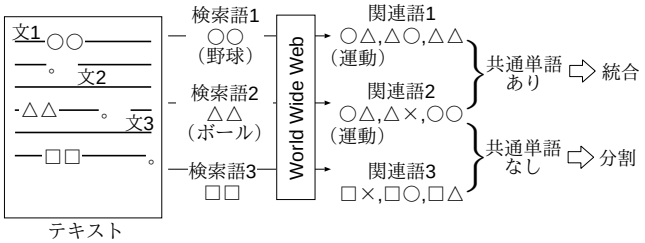


図 2　検索語と関連語を利用したテキストセグメンテーション

3. 提 案 手 法

3.1 概　　要

提案手法の概要を図 1 に示す。本論文では、主に名詞の単語を使用し、それ以外の品詞の単語は用いない。提案手法では、テキストの各文の内容をテキスト本文から抽出した名詞とウェブ検索により得られた名詞を用いて表現する。そして、異なる文の間で幾つかの共通する単語が存在する場合には、それらの文は同じ分野であると判定する。

名詞だけを利用する単純なテキストセグメンテーション方法として、与えられたテキストから抽出した名詞だけを利用し、先頭から順に文間の結束度を評価する方法が考えられる。しかし、各文から抽出した名詞はテキストの記述内容の分野のごく一部の範囲しか表していないことが多いため、テキストを過分割すると考えられる。例えば、図 2 の検索語 1 が「野球」で検索語 2 が「ボール」であるとする。この二つの単語は「スポーツ」の分野に属すると考えられる。しかし、文から抽出した名詞だけでは、この二つの単語は異なるため同じ分野と判定できず分割が発生する。

そこで、提案手法では文から抽出した名詞を検索語として利用しウェブ検索を行う。そして、検索されたウェブページは検索語に関連する記事であるという点から、得られたウェブページから検索語に関連する複数の単語を抽出する。例えば、図 2 の文 1 において「野球」を検索語にすると、スポーツ記事が検

現在、私は横須賀の久里浜という所に住んでいます。意外だったのは、久里浜は富士山に結構近いということでした。職場から天気が良いときには富士山がよく見えます。しかもハッキリと。そんな訳で今度富士山の近辺をドライブして箱根にでも立ち寄ってみようと考えてます。箱根近辺には芦ノ湖など景色の良い場所がたくさんあるようなので、ドライブも楽しいのではないかと思います。研究の進捗はどうですか。今、研究室でパターン認識(俺のような特選選択とか)を取り組んでいる学生っているのかな。研究室を離れるときにウェブマイニングとか音声認識関係の研究を希望する学生が多かった気がするので、研究室の様子も変わったのではないでしょうか。

図 3　テキストの例

索結果の上位に占めることから、検索されたウェブページから関連語として「ボール」や「運動」といった単語が得られると考えられる。つまり、ある文のジャンルを表す単語、或いは他の文に出現するような単語が関連語として得られる。その結果、図 2 における先頭の二つの文が「野球」・「ボール」・「運動」という共通単語の存在から同じ内容であると判定できる。このように検索語と関連語を使用することで、幅広く文の内容の追跡できると考えられる。一方、共通する単語が存在しない場合 (図 2 では二つ目の三つ目の文) には文の内容が変化したと判断し、その二つの文の間でテキストを分割する。

以下のそれぞれの節において、与えられたテキストに対して、文単位に分割する方法、検索語を抽出する方法、関連語を取得する方法、そしてテキスト分割方法についてそれぞれ詳細を述べる。

3.2　文単位への分解

提案手法では、最初に与えられたテキストを文単位に分解し N 個の文を得る。ここで、文単位とは句点で区切られる一文のことを意味する。但し、会話文 (’「’ と’」’ で囲まれた文) 中に現れる句点では、会話内容を途中で切ることになるため分解を行わない。提案手法では、テキストの先頭から一文字ずつ調べ、句点が現れると一文として出力する。例えば、図 3 の例では、与えられたテキストは 9 個の文に分解される。

3.3　検索語の抽出

次に、各文に対して、ウェブで検索を行うための検索語を抽出する。本論文では、各文に対して形態素解析を行い、名詞と判定された単語を検索語として使用する。また、本論文では形態素解析に JTAG [10] を使用する。ここで、抽出された単語の中には、「年」・「ところ」・「それ」など、様々な文書の中に普遍的に存在する不要語 (ストップワード) が含まれる。また、検索語が少ない場合、文の内容とは関連性の弱い記事が得られることがある。そこで、提案手法では不要語を除いた後、抽出された検索語の個数 S が閾値 S_T(予備実験の結果から $S_T = 2$) 未満の場合には検索語の無い文として扱う。図 3 の例に対して、検索語を抽出した結果を表 1 に示す。

3.4　関連語の抽出

検索語の抽出が終了すると、各文に対して抽出した検索語を用いてウェブ検索を行い関連記事を取得する。本論文において関連記事とは検索結果の上位 P 件で参照されているウェブページのことを指す。また、本論文におけるウェブページとは、

日 月 年 人 約 日本 発表 東京 時 分 もの 午後 これ 午前
昨年 今年 予定 明らか 説明 可能性 事件 歳 今回 調査 現在
ところ 女性 調べ 回 私 方針 関係 近辺 中 円 問題 指摘
今後 場所 時間 意外 必要 この日 今度 訳 確認 位 様子
自分 こと 所 声 学生 それ 現在 後 疑い 様子

図 4　不要語辞書の例

表 1　図 3 の例に対する提案手法の検索語と関連語の例

文番号	検索語	関連語
1	横須賀, 久里浜	京急, 横浜, 駅, 野比, 地図, 施設, 三浦, 大船
2	久里浜, 富士山	横須賀, 富士, 逗子, 横浜, 天気, 京急, 鎌倉, 海岸
3	職場, 天気, 富士山	雨, 富士, 雪, 予報, 登山, 気温, 雲
4	(検索語無し)	(関連語無し)
5	ドライブ, 箱根, 富士山	温泉, 芦ノ湖, 御殿場, 観光, 旅行, 満喫, 紅葉
6	ドライブ, 芦ノ湖, 景色, 箱根	富士山, 温泉, 仙石原, 絶景, 観光, 日帰り
7	研究, 進捗	技術, 計画, 目的, 大学院, 開発, プロジェクト, 貢献
8	パターン認識, 研究室, 次元選択	研究, 学習, 線形, 設計, 識別, 技術, モデル
9	マイニング, 音声認識, 希望, 研究, 研究室	技術, 開発, 工学, 設計, モデル

HTML や XML などの構造化言語で記述されたテキストのことを指す。構造化言語で記述されたテキストは’<’ と’>’ で囲まれたタグを解析することによって、ウェブページに含まれる本文をテキスト形式で収集することができる。

一方、収集された本文のテキストから関連語を抽出する際、検索語によって取得できる関連語の個数は変化する。また、関連記事が少ない場合、関連語が特定の分野に偏りテキストを分割する際に悪影響を与える。そこで、提案手法では、取得した関連記事数 P が閾値 P_T(予備実験の結果から $P_T = 5$) 未満の場合には関連語を抽出しない。$P \geqq P_T$ の場合には、予め設定した検索語と関連語の合計個数 K を用いて、$K - S$ 個の関連語を関連記事から抽出する。

一つの文に対して提案手法における関連語の抽出手順を図 5 に示す。

提案手法では、図 5 のアルゴリズムを N 個の文にそれぞれ適用し、各文に対する関連語を得る。図 3 の例に対して提案手法により得た関連語を表 1 に示す。表 1 の例は $P = 20, K = 10, P_T = 5$ として提案手法を実行した例である。

3.5　テキスト分割

最後に、提案手法におけるテキスト分割の方法について説明する。提案手法では検索語と関連語を一つの組としたキーワード集合を作成し、キーワード集合を用いて内容的なまとまり (本論文では、これをブロックと呼ぶ)B_k(但し、$k = 1, 2, \ldots$)

(1) $K > S \geqq S_T$ の場合は (2) へ進む。それ以外の場合は関連語を抽出せずにアルゴリズムを終了する。
(2) 抽出した S 個の検索語を用いてウェブ検索を行う。
(3) 検索結果で参照されている上位 P 件のウェブページに対して、本文の内容を抽出し P 個の本文テキストを作成する。但し、$P < P_T$ の場合は本文テキストを作成せずアルゴリズムを終了する。
(4) P 個の本文テキストに対して、以下の (4.1) と (4.2) の二つの処理を行う。終了した場合には (5) に進む。
(4.1) 本文テキストにある文章に対して形態素解析を行い名詞を抽出する。
(4.2) 抽出された名詞で不要語辞書に登録されている単語を削除する。
(5) 残った名詞に対して文書頻度 (document freduency, DF) を計算する。
(6) 文書頻度の高い順に $K - S$ 個 の名詞を抽出し関連語とする。

図 5　一つの文に対する関連語抽出処理の手順

(1) $S \geqq S_T$ ならば (2) へ。それ以外の場合は空のキーワード集合を作成する。
(2) $K > S$ ならば (3) へ。それ以外の場合は、S 個の検索語からランダムに K 個を選択しキーワード集合の要素とする。
(3) $P \geqq P_T$ ならば (4) へ。それ以外の場合は、S 個の検索語だけを選択しキーワード集合の要素とする。
(4) S 個の検索語と $K - S$ 個の関連語を選択しキーワード集合の要素とする。

図 6　一つの文に対するキーワード集合の作成手順

の抽出を行う。一つの文に対するキーワード集合作成手順を図 6 に示す。図 6 で示すアルゴリズムを N 個の文に適用し、i 番目の文に対するキーワード集合 T_i(但し、$i = 1, 2, \ldots, N$) を作成する。

次に、作成されたキーワード集合を用いてブロックの抽出を行う。図 7 にテキスト分割の処理手順を示す。ここで、G はどのブロックにも割り当てられなかった文の集合である。一般的に、文章は先頭から順に書かれるため、本論文では先頭から順に二つの文に対応するキーワード集合 T_i と T_j(但し、$j \geqq i+1$) の比較を行う。具体的には、二つのキーワード集合に対して共通して存在する単語が閾値 C_T(予備実験から $C_T = 1$) 個以上ある場合には、前後の二つの文は内容が関連していると判断し、ブロック B_k に文の番号を要素として追加する。一方、共通単語の個数が C_T 個未満の場合には、その二つの文の間でテキストを分割しブロック B_k の要素を出力する。空のキーワード集合は無視し、代わりにその前後にある二つのキーワード集合で比較を行い、共通単語の有無に応じてテキストを分割する (図 8)。図 3 の例に対して、提案手法を適用してテキストセグメンテーションを行った結果を表 2 に示す。表 2 において、使用したパラメータは $S_T = 2, P_T = 5, C_T = 1, K = 10, P = 20$ である。

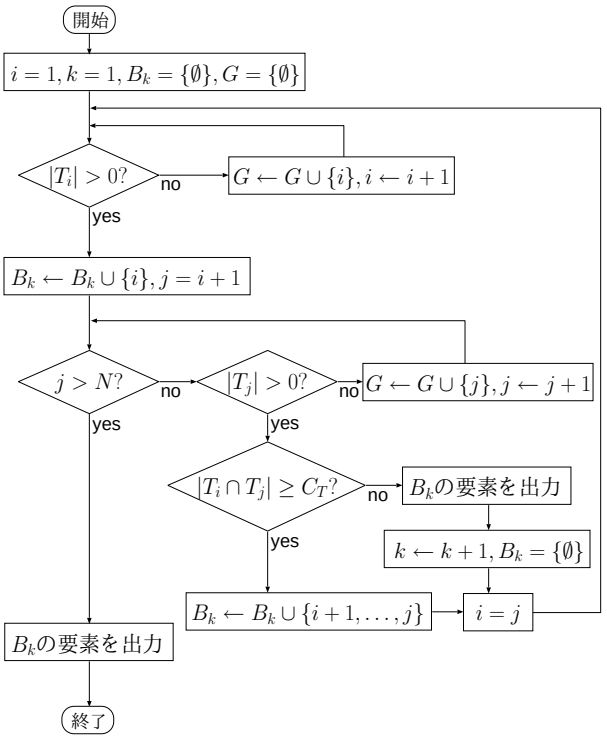


図 7　キーワード集合を利用したテキスト分割処理の処理手順

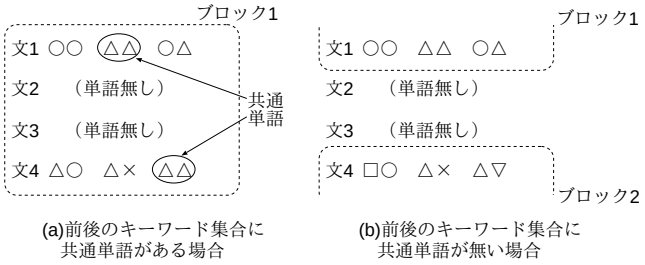


図 8　テキスト分割における検索語が無い文に対する処理

表 2　図 3 に例に対する提案手法のテキストセグメンテーションの例

ブロック	文番号	共通単語
1	1-6	芦ノ湖, 箱根, 富士, 京急, 久里浜, 逗子, 横須賀, 温泉, 横浜, 天気, 観光, ドライブ, 富士山
2	7-9	研究, 開発, 技術

4. 実　　験

4.1 使用したデータの概要

goo ニュース記事 [11] を用いて、ウェブ検索を利用したテキストセグメンテーションを行うことができるか検証する実験を行った。具体的には、goo ニュースの記事を用いて複数の分野の内容を一つにまとめた記事を作成し、提案手法を用いてテキストセグメンテーションを行った。goo ニュース記事の詳細を表 3 に示す。実験では表 3 の goo ニュース記事から三つの異なるカテゴリに属する記事をランダムに一つずつ収集し、一つの記事に連結する。連結した記事のデータを文の番号とカテゴリでまとめたものを表 4 に示す。表 4 の例では、「エンタメ」・「ライフ」・「政治」の三つのブロックで記事が構成されていること

表 3　使用した goo ニュース記事

カテゴリ (内容)	記事数	平均文章数
社会 (事件、災害、裁判 ...)	125	6.8
政治 (行政、政治、選挙 ...)	125	8.6
ビジネス (企業、金融、市況 ...)	125	7.9
ライフ (トレンド、健康、教育 ...)	94	6.6
エンタメ (映画、音楽、TV ...)	125	5.7
スポーツ (プロ野球、MLB、J リーグ ...)	125	7.3

表 4　実験で使用するテスト記事の正解ブロックの例

ブロック番号	文番号	カテゴリ
1	1-4	エンタメ
2	5-15	ライフ
3	16-18	政治

を表す。また、1 文目から 4 文目までが一つ目のブロック、5 文目から 15 文目までが二つ目のブロック、そして 16 文目から 18 文目までが三つ目のブロックであることを表す。

4.2　提案手法の性能評価

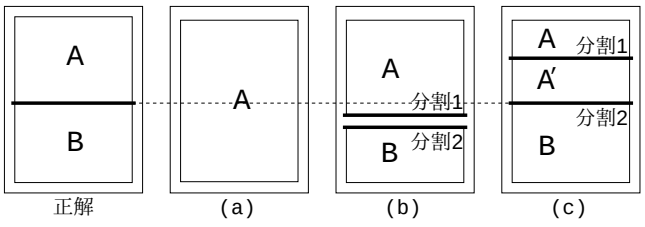
表 4 のようなテスト記事を作成する際、異なるカテゴリの記事を連結している。従って、記事の連結箇所で内容が変化するため、記事の連結部分で一つのブロックが検出できれば良いと考えられる。そこで、提案手法を用いてテキストセグメンテーションを行い、求まったブロックと表 4 の例のような正解ブロックとを比較する。そして、正しく検出できたブロック数とテスト記事に含まれるブロック数から適合率と再現率で評価する。適合率と再現率は以下の式で計算する。

$$\text{適合率} = \frac{\text{正しく検出されたブロックの個数}}{\text{提案手法で検出したブロックの総数}}$$

$$\text{再現率} = \frac{\text{正しく検出されたブロックの個数}}{\text{テスト記事内でのブロックの総数}}$$

従来研究では、検出した分割箇所と正解のテキスト結合部分との比較によって手法の性能を評価することが多い。一方で、抽出された分割箇所と正解の連結箇所とのずれをどの程度許容するかによって、手法の性能評価が大きく異なる。また、提案手法ではどのブロックにも属さない文が生じるために、その文に対して分割箇所の設定が容易ではなく、分割箇所に基づいて性能を評価するのが難しい。そこで、本論文では抽出されたブロックと正解のブロックを用いたブロック単位で手法の性能を評価する。本論文における抽出されたブロックと正解のブロックとの比較方法、及び適合率と再現率の算出方法の例を図 9 に示す。

- 図 9(a) の例は、全てのブロックが一つのブロックとして抽出された場合を表している。このとき、本来検出するべき複数のブロックが発見できなかったことを表すので、正解ブロック A と B 共に検出に失敗しているとみなす。
- 図 9(b) の例では、一つのブロックが他のブロックの一部にはみ出している場合を表している。このとき、検出ブロック A は正しくないものとして扱う。一方、検出ブロック B は対応する正解ブロック B の一部分であるので正しく検出できたと扱う。尚、どのブロックにも割り当てられなかった文 (集合 G に含まれる文) は無視する。
- 一つのブロックを過分割した例を図 9(c) に示す。この例では、正解ブロック A を過分割し二つのブロックが検出されている。一つの正解ブロックに対し複数のブロックが得られた場合、一つのブロックを正解とし残りのブロックは全て不正解とする。具体的には、抽出したブロック A だけを正解とカウントし、残りの抽出ブロック A' は不正解とカウントする。



	(a)		(b)		(c)	
	適合率	再現率	適合率	再現率	適合率	再現率
ブロック評価	0/1	0/2	1/2	1/2	2/3	2/2
分割評価	0	0/1	0/2	0/1	1/2	1/1

図 9　分割結果の評価方法

以上の基準で図 9(a) から (c) までの例に対して算出した適合率と再現率を図 9 に示す。

4.3　実　験　1

実験 1 では第 4.1 節で説明したテストデータを 100 記事 (2180 文) 作成し、提案手法を用いてテキストセグメンテーションを行った。そして、抽出されたブロックと正解のブロックとを比較し適合率と再現率を求めた。提案手法におけるパラメータに関しては、検索語数の閾値、関連記事数の閾値、そして共通単語数の閾値はそれぞれ $S_T = 2, P_T = 5, C_T = 1$ とした。関連記事数 P と検索語と関連語の合計数 K のパラメータに関しては、P と K の値を適当な範囲で変化させた。具体的には、$P = \{5, 10, 20, 30, 40, 50\}$ であり、$K = \{5, 10, 20, 30, 40, 50\}$ とした。また、ウェブ検索を行う際に使用した検索エンジンは Google である。

実験 1 の結果を表 5 に示す。表 5 において、表中の数値は上段が適合率、下段が再現率を表す。また、表 5 において、F 尺度が最大となるパラメータの結果に対してボールド体で示している。F 尺度は適合率と再現率の調和平均であり、以下の式で算出される。

$$F\text{ 尺度} = \frac{2}{\frac{1}{\text{適合率}} + \frac{1}{\text{再現率}}}$$

適合率と再現率の値が共に大きいときに F 尺度は大きい値をとる。また、F 尺度の値が大きい程、正確にテキストセグメンテーションが行えたことを意味している。

表 5 から、パラメータ K と P の組み合わせの中で、$(K, P) = (20, 30)$ のときに F 尺度が最大となった。また、表 5 から検索語と関連語の合計数を決定するパラメータ K について見ると、K の値が小さい場合 (例えば、$K = 5$ や $K = 10$ の場合) において適合率が低く再現率が高いことが分かった。

表 5　各パラメータにおける適合率 (上段) と再現率 (下段)

関連記事数	合計単語数					
	K= 5	K=10	K=20	K=30	K=40	K=50
P= 5	26.6	46.3	61.8	66.1	65.4	59.5
	100.0	98.6	90.6	82.0	72.0	57.3
P=10	27.9	50.3	67.9	69.6	63.2	60.5
	100.0	98.3	91.3	81.0	63.0	54.6
P=20	28.1	53.2	71.3	71.6	66.3	62.6
	100.0	98.3	91.3	79.3	65.0	54.3
P=30	28.7	53.4	**73.8**	71.9	65.8	59.4
	100.0	98.0	**92.3**	79.3	64.3	51.3
P=40	28.9	54.6	72.0	71.5	65.3	58.4
	100.0	98.3	92.0	78.0	62.3	48.6
P=50	28.9	53.9	71.7	71.2	65.0	59.0
	100.0	96.6	89.0	78.6	63.3	50.0

検索語や関連語の個数が少ない場合には前後の文で内容の関連性を十分に比較することができず過分割が発生した。その結果、正解のブロックは検出できているものの、無駄なブロックも多くなったと考えられる。一方、K の値が大きくなる (例えば、$K = 40$ と $K = 50$) と、適合率と再現率が共に低下する結果となった。これは、K の値を増やすことで、各文の内容と関連の弱い単語がキーワード集合に多く含まれるため、一つのブロックとしてまとまり易くなる。その結果、無駄なブロックは少なくなるものの、正解のブロックも検出できないと考えられる。一方、関連記事数のパラメータ P に関しては、値の変化に対し適合率と再現率の変化が大きく見られなかった。これは、検索エンジン Google の検索結果の上位には検索語に関連する記事が多く存在したため、関連記事数 P に依らず適切な関連語を抽出できたことが原因と考えられる。

また、実験結果からどのブロックにも割り当てられなかった検索語が無い文を調べたところ、全部で 326 文 (テスト記事全体の 15.0%) 存在した。検索語が無い文の代表的な例として、(1) ある物や人物に関して、予定や状況だけが書かれている短い文、(2) インタビュー時の人物の様子や気持ちなど、口語的な表現のみで名詞が無い文、が挙げれられる。しかし、内容や分野を判断できる文が検索語の無い文として扱われた例もあった。主な原因としては、人物名や商品名があるにも関わらず固有名詞の抽出に失敗し、検索語が少なくなったことが挙げられる。また、「労働時間」等の複合語が形態素解析により「労働」・「時間」のように複数の分解され、それぞれの単語が不要語辞書に登録されていたため削除されたことも検索語が少なくなった原因であった。

4.4 実　験　2

実験 2 では、提案手法が与えられたテキストの記述内容の分野と問わずテキストセグメンテーションを行うことができるか検証した。使用したテストデータは、実験 1 で作成した 100 記事のテストデータと同じである。評価方法は、表 3 に示したジャンル別にブロックの抽出結果を適合率と再現率で評価し、全体の結果における適合率と再現率と比較した。実験 2 で使用したパラメータ K, P は実験 1 の結果から適合率と再現率の合計が最大となった $(K, P) = (20, 30)$ を用いた。

表 6　ジャンル別で見た適合率と再現率

ジャンル	記事数	適合率	再現率
ライフ	42	**77.1**	88.1
ビジネス	57	**78.3**	**94.7**
スポーツ	58	66.7	**93.1**
社会	57	**80.6**	**94.7**
政治	39	65.5	**92.3**
エンタメ	47	68.9	89.4
全体	100	73.8	92.3

表 7　表 4 で示すテスト記事に提案手法を適用した結果

ブロック番号	文番号	出現頻度の高い単語
1	1-4	出演、撮影、映画、番組、ドラマ、···
2	5-15	洗面台、値上げ、小売、高騰、原材料、···
3	16-18	国連、外交、首相、国民、···

$(K, P) = (20, 30)$ を用いた際のジャンル別の適合率と再現率を表 6 に示す。表 6 の各ジャンルの適合率と再現率において、全体に対する結果以上の精度が得られた場合には、数値をボールド体で示している。表 6 から「エンタメ」の分野に関しては全体の結果と比べるとやや劣るものの、全体的に安定してテキストセグメンテーションを行えることが確認できた。この結果から、ウェブを利用することで様々な分野に対応したテキストセグメンテーションを行えることが確認できた。

5. 考　　察

表 6 の結果から複数の分野の記事に対して、全体的に安定してテキストセグメンテーションを行えることが確認できた。特に、「ビジネス」と「社会」の記事に関して高い精度でテキストセグメンテーションを行うことができた。この理由として、名前や場所、専門用語など、各文から複数個の検索語が安定して抽出できたことが考えられる。また、これらの分野の記事においては、文体が統一されていることや、文間での内容の変動が少ないことも、良い精度でテキストセグメンテーションを行えた理由として挙げられる。

また、表 4 で示すテスト記事に提案手法を適用し、抽出された各ブロックに対して出現頻度の高い単語を表 7 に示す。表 4 と表 7 から、提案手法が適切にテキストセグメンテーションを行うことができた場合には、表 7 で示す単語と文番号は与えられたテキストに付与するラベルとして利用できると考えられる。例えば、あるキーワードが与えられた場合に、付与されたラベルを参照し、テキストからキーワードに関連する文章だけを部分的に収集することが期待できる。この方法は、未整形のテキストに対して、与えられたキーワードに関係する記述だけを探す場合に有効であると考えられる。

一方、「エンタメ」の分野に関しては、全体の結果と比較して適合率と再現率が共にやや劣る結果となった。「スポーツ」と「政治」の分野では適合率が低く、「ライフ」の分野では再現率

が低くなった。適合率が低くなった原因として、挿入的な内容の文が多く混在し、その文で一旦分割が起こることが考えられる。失敗例を以下に挙げる。

- スポーツ記事の中で、選手が福祉関係の施設を訪れるという文において、検索語が「老人」・「施設」・「介護」・「訪問」となり、関連語が福祉関係や看護関係の単語が得られたため分割が発生した。
- スポーツ選手のインタビュー記事で、「本場」・「リング」の検索語から、「ダイヤモンド」・「ジュエリー」等のアクセサリーや装飾関係の関連語が抽出されたため分割が発生した。
- 政治に関する記事で、財政の内容から教育関係の話や外交関係の話題などに話が飛び、それぞれの内容毎に分割が発生した。

このように挿入的な内容の文が混在している場合、検索語に関連する単語が抽出できているものの、記事全体の内容とは無関係な記事が得られる。その結果、このような文が単独で一つのブロックを成し過分割の原因となることが分かった。このような文が多く存在する記事としてはブログ記事も該当する。ブログ記事では、一般的に文体が統一されていないことが多いため、各文から名詞を安定して抽出するのは難しいと予想され、提案手法をブログ記事にそのまま適用すると、精度の低いテキストセグメンテーション結果が得られると考えられる。提案手法を様々な種類の記事に適用するためには、検索語の抽出方法や意味段落の抽出方法を更に工夫する必要がある。具体的な解決策としては、複数文から検索語を抽出する方法や、名詞以外にも形容詞や動詞などの他の品詞を用いる方法が考えられる。

再現率が低くなる原因として、様々な分野に共通して使用される単語が存在し、分割が行えなかったことが考えられる。以下に失敗例を示す。

- 記事内容が政治のジャンルから芸能のジャンルに変化した際、ある人物の名前が共通単語として抽出され分割ができなかった。
- 学術的な貢献をした人の内容から教育問題に話題が変化した際、教育や学校という単語が抽出され分割ができなかった。

本論文では、このような様々な分野に存在するような単語は不要語として扱うことを想定している。しかし、今回の実験で使用した不要語辞書は人手で作成されたため、不要語の選定基準が曖昧であった。可能ならば、ウェブ全体に存在するような単語を自動的に収集し、不要語辞書を作成することが好ましい。そのような不要語辞書を用いることで、上記の問題が解決できるかどうかを検証する必要がある。

提案手法では、関連記事の取得にウェブ検索を利用した。これは同時に、提案手法は利用する検索エンジンに結果が依存することを意味している。例えば、事件に関する文から名詞を取り出し検索を行った際、事件に関連する単語だけでなく、最近話題になっている地名や人名も関連語に多く含まれることがあった。このことから、ウェブ検索を利用することで、利用する検索エンジンのランキング手法によって関連語が変化することや、最近の話題になっている内容に関連語が偏るという傾向が分かった。現段階では、何らかの基準を用いて検索語に関連する記事を十分取得できたと判定できるまで、複数の検索エンジンを利用する等、利用上の工夫が必要であると考えられる。

6. おわりに

本論文では、ウェブ検索を利用したテキストセグメンテーション法を提案した。goo ニュース記事を用いた実験から、事前に学習用データを用意することなくテキストセグメンテーションを行うことができ、提案手法の有効性を確認することができた。今後の方針として、従来手法との比較により提案手法の長所や短所を明らかにすることや、新聞記事データやブログ記事などを使用して提案手法の有効性を検証する必要がある。また、挿入的な内容の文に対して文脈の前後関係を利用して過分割を防ぐ方法について検証する。更に、単語の不適切な組み合わせにより関連語が得られない場合において、検索語内の単語を補正し適切な関連記事をウェブ上から取得できる方法についても検討する。

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, **42**(11), 2650–2662, 2001.

[2] 西脇 正通, 田中 英輝, 関連記事を利用したテキストセグメンテーション. 情報処理学会研究報告, **2002-NL-152**, 79–84, 2002.

[3] 西澤 信一郎, 中川 裕志, 名詞の文書内頻度を利用したテキストセグメンテーション. 情報処理学会研究報告, **97-NL-117**, 145–152, 1997.

[4] M. Hearst, TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. *Computational Linguistics*, **23**(1), 33–64, 1997.

[5] 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，**6**(3), 43–58, 1999.

[6] 松井 祥峰, 乾 伸雄, 小谷 善行, 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. 情報処理学会研究報告, **2004-NL-162**, 151–158, 2004.

[7] 越仲 孝文, 奥村 明俊, 磯谷 亮輔, HMM の変分ベイズ学習によるテキストセグメンテーション及びその映像インデキシングへの応用. 電子情報通信学会論文誌, **J89-D**(9), 2113-2122, 2006.

[8] D. Beeferman, A. Berger and J. D. Lafferty, Statistical Models for Text Segmentation, *Machine Learning*, **34**(1–3), 177–210, 1999.

[9] 木田 充洋, 外池 昌嗣, 宇津呂武仁, 佐藤 理史, ウェブを利用した専門用語の分野判定. 電子情報通信学会論文誌, **J89-D**(11), 2470–2482, 2006.

[10] T. Fuchi and S. Takagi, Japanese Morphological Analyzer using Word Co-occurence – JTAG. In *Proceedings of the 36th Annual Meeting of the Association for Conputational Linguistics*, 1998, pp.409–413.

[11] goo ニュース, `http://news.goo.ne.jp/`