Web検索を用いたテキストセグメンテーション

内山 俊郎（うちやま としお） /阿部 直人（あべ なおと）※
別所 克人（べっしょ かつじ） /内山 匡（うちやま ただす）
奥 雅博（おく まさひろ）
NTTサイバーソリューション研究所

テキストセグメンテーションは，与えられた文書を内容に応じて「意味段落」に分割する手法です．Web文書の解析技術は大きく進歩しましたが，処理する単位としては「文書」が使われています．しかし，処理する単位を意味段落にすれば，もっと「切れ味の鋭い」サービスが実現できます．

検索におけるテキストセグメンテーションの重要性

最初に検索を例として，テキストセグメンテーションの役割と重要性について説明します．

検索の目的は，大量の情報からユーザが取り出したい情報を見つけることです．検索方法として，検索クエリ（質問）を入力して，それに合った情報を検索対象の集合から探し出して出力する方法があります（**図1**）．このときに，「引き出す情報の単位」として，普通は文書全体を考えます．しかし，文書の中には「映画を見て感動した話を書き，続いて新しく買った携帯電話の機能のことを批評し，最後に友だちと行ったレストランの話で締めくくる」（**図2**）というように，さまざまな話題が含まれていることがあります．映画に関する話題を探していて，携帯電話の話まで出力されるのでは少々不便です．映画というキーワード周辺だけを取り出すこともできますが，過不足なく取り出すのは困難です．

このような問題に対しては，引き出す情報の単位を文書全体から「意味段落」（1つの意味を持った塊）に変更することが有効です．先ほどの文書を意味段落に分割すると，「映画」「携帯電話」「レストラン」となります．このように意味段落単位で処理すれば，映画に関する話題だけを取り出すのは簡単です．

ところで，文書中に意味段落の切れ目が明示的に示されているとは限りません．何らかの方法で文書の中身を解析し，意味段落を発見することが必要となります．ここで登場するのが，テキストセグメンテーション技術です．テキストセグメンテーションは，取り出したい（解析したい）情報だけにアクセスすることを可能にします．言うなれば，「切れ味の鋭い」検索サービスを実現してくれる重要な技術なのです．

テキストセグメンテーションの難しさ

これまでにテキストセグメンテーションに関する多くの手法が提案されています．例えば，与えられた学習データから求めた統計的情報や言語的情報に基づく方法[(1),(2)]，また事前に学習データを使用しない方法もあります[(3)]．前者の方法を用いる場合，セグメンテーションの精度が学習データに依存する問題があります．学習データと異なる分野の話には弱くなるわけです．一方，後者の方法として，共通して出現する

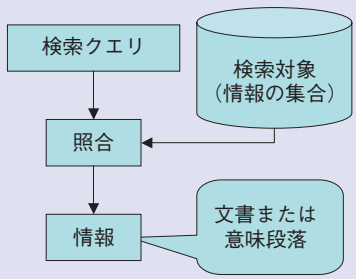

図1　検索の仕組み

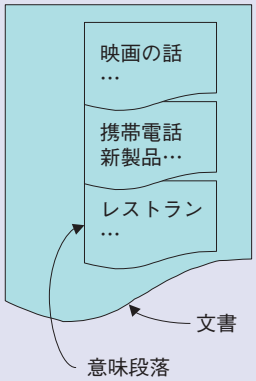

図2　文書と意味段落

※　現，NTTコミュニケーションズ

単語に基づいて文どうしのつながりを算出する方法（Hearst法[3]）があります．しかし，**図3**が示すように，話題が共通（文1と文2）していても同じ単語が出現するとは限りません．必要以上に分割され，細切れの意味段落が出力される問題があります．

Web検索の利用

これらの問題を解決するために，筆者らは事前に学習データを使用しない手法として，「Web検索を用いたテキストセグメンテーション方法」を提案しました[4],[5]．その基本アイデアを示したのが**図4**です．各文に含まれている単語を「検索語」として抽出し，その検索語を用いてWeb検索を行い，検索結果上位に多く含まれている単語を「関連語」として取り出します．そして検索語と関連語の両方を考慮したうえで，共通して出現する単語に基づいて文どうしのつながりを算出します．図4の例では，文1と文2は「選手」という共通の関連語を取り出しています．これを手掛かりとし，同じ話題の文として統合（同一の意味段落として）しています．このアイデアの背景にあるのは，「検索された記事内で出現頻度が高い単語は検索語に関連した単語であることが多い」という仮定です．このように関連語を考慮することにより，意味段落の境界を的確にとらえることが可能になります．さらに，Web検索を用いているので，学習データに依存する問題もなく，最新の話題に追従することも容易です．

実際の処理例

実際にWeb検索を用いたテキストセグメンテーション方法の適用例を紹介します．以下，処理の過程に沿って説明します．

① **図5**に示すテキスト（文書）を入力とします．この文書には9つの文が含まれています．

② 各文から検索語を取り出し，さらに検索語を用いてWeb検索を行って関連語を取得します．その結果を**表1**に示します（実際には検索語と関連語を合わせ，30単語を抽出しました．表1はその一部を掲載しています．また，形態素解析は，NTTサイバースペース研究所が開発したJTAGを用いています）．

③ 検索語と関連語を合わせた「キーワード集合」を用い，連接する文間の連結度を算出します．最終的に求めた連結度を**図6**に示します．文番号3と4の間と文番

文1：——野球ー ——.
文2：ー ー ー ー ー甲子園ー .
文3：ー ー□□ー ー.

図3　話題が共通（文1と文2）でも同じ単語が出現しない例

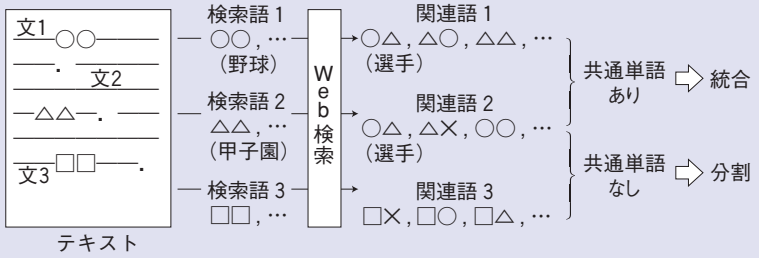

図4　Web検索を用いたテキストセグメンテーション法の概要

ドライブが好きな私は大阪まで高速道路で行くことが多い．札幌にいたときから日帰りで網走目指してドライブしたりしてたから，長距離運転はだいぶ慣れたほうだと思う．
最近はガソリン代が高いのでちょっとどうしようかなと考え気味．
ドライブしたついでに温泉にも入りたい．
温泉で絶景をながめながらゆっくりしたいし，いろいろ美味しいもの食べたいし．
日帰りでドライブと温泉が楽しめる場所をかいたくしてみるか….
最近，ゴルフを始めていろいろ本を買って打ちっぱなしで練習中．
近くに24時間営業のゴルフ打ちっぱなし場があるから，仕事帰りに練習してるんだけど…，低弾道のショットが多い気がする．
まあ，そんな簡単にうまく打てるようになるとは思っていないので，時間を見つけて打ちっぱなしで気長に練習をするか．

図5　文書例（文献(5)から引用）

表１　文書例に対する検索語と関連語

文番号	検索語	関連語
1	ドライブ，高速道路	料金，道路，ETC，渋滞，車，北海道，通行止め，工事
2	ドライブ，運転，慣れる，札幌，長距離，日帰り，網走	北海道，風，走る
3	ガソリン代	ガソリン，車，価格，節約，ガソリンスタンド，カード，高騰，給油，家計
4	ドライブ，温泉	露天風呂，楽しむ，ホテル，風呂，楽しめる，満喫，日帰り，食べる
5	ゆっくり，温泉，食べる，絶景，眺める，美味しい	料理，旅館，ホテル，観光
6	ドライブ，温泉，開拓，楽しめる，日帰り	観光，楽しむ，日帰り温泉，料金，満喫
7	ゴルフ，始める，打ちっぱなし，練習	打つ，ボール，ゴルフ練習，当たる，教える，スコア
8	ゴルフ，ショット，仕事，打ちっぱなし，弾道，練習	スコア，打つ，飛距離，スイング
9	うまい，気長，見つける，打てる，練習	無理，欲しい，通る，打つ，立つ

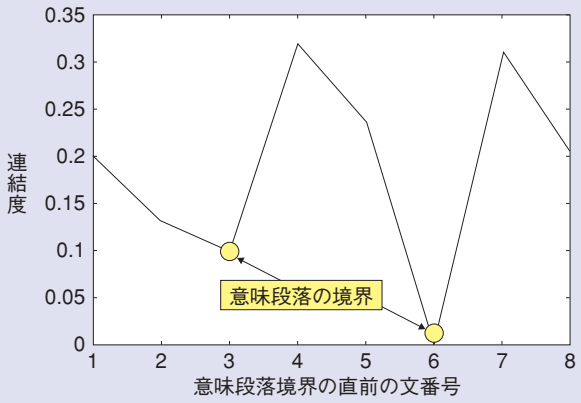

図６　文書例に対する連結度

号６と７の間に連結度の谷があることを示しています（この谷は実際に意味段落の境界と一致しています）．

④　適当なしきい値を用いて，意味段落を決定します．最終的に得られた意味段落と各意味段落で２回以上出現したキーワードを**表２**に示します．結果として，

・ドライブの話題
・旅行先での楽しみについて
・ゴルフの話題

という意味段落が抽出できました（文献(5)では，しきい値の決定法も提案しています）．

このように，与えられたテキスト（文書）から意味段落を自動的に抽出できることが分かると思います．

性能評価

ニュース記事やブログを用いた評価実験を行ったところ，従来手法[(3)]よりもかなり高い精度（分割位置の正確性）でテキストセグメンテーションできることが確認できました．従来手法は，与えられたテキストから抽出した単語だけを利用しているため，表記揺れや単語の出現の仕方に影響を受けやすくなっています．これに対し，提案手法では検索された記事から求めた関連語も利用することで，意味段落内において共通する単語の個数を増やすことができます．そして，意味段落の境界ではキーワード集合（検索語＋関連語）に含まれる単語の変化が明瞭になるため，適切な意味段落境界を検出できたためと考えられます．

技術の適用先

冒頭では，検索を例としてテキストセグメンテーションの役割と重要性について説明しましたが，適用先は検索に限りません．ここでは，いくつかの適用先について紹介します．**図７**が適用先のイメージ図です．文書を入力とし，テキストセグメンテーションを施して意味段落を出力します．この意味段落を入力として，さまざまな文書解析処理を行うことで，処理単位が「文書」であったときに比べて切れ味の鋭い解析が実現できます．文書解析の例

表２　文書例に対して提案手法を適用した結果

段落番号	文番号	出現単語
1	1，2，3	車，高速道路，価格，ドライブ，北海道
2	4，5，6	楽しめる，温泉，食べる，走る，ホテル，楽しむ，日帰り，観光，露天風呂，ドライブ
3	7，8，9	打つ，練習，スイング，上達，探す，打てる，ゴルフ，アイアン，打ちっぱなし，仕事，スコア，ドライバー，ボール，教える

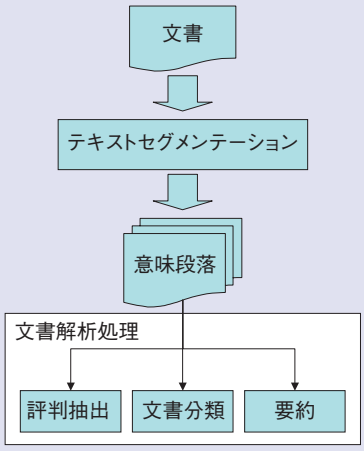

図７　テキストセグメンテーションの適用先

として，「評判抽出」「文書分類」「要約」を取り上げて説明します．

⑴　評判抽出

ブログや掲示板などに書かれている商品の評判情報やクチコミ情報は，商品を探しているユーザや商品のメーカにとってますます重要な情報となり，自動抽出の研究が盛んです．評判情報を自動的に抽出する際，入力が意味段落単位であれば，精度向上が期待できます．例えば，携帯電話の機能についての批評を集めているときに，映画の話やレストランの話が混入することがなく，間違った情報を取り出す危険性が減らせます．

⑵　文書分類

文書の内容に応じた広告配信の実現や，検索における絞込みを行う際は，文書分類技術が欠かせません．しかし，さまざまな話題が混在している文書についてその分野を適切に推定することは困難です．意味段落ごとに分野を推定することで，精度向上が期待できます．

⑶　要　約

文書要約においても，別々の話が連結された文書を正しく要約するのは大変です．意味段落単位で要約すれば，関係ない事柄を結びつけるなどの間違いを未然に防ぐことができます．

今後の取り組み

ここまで解説したように，テキストセグメンテーション技術は，処理単位を従来の文書から意味段落にすることができます．今後，本技術とさまざまな文書解析技術とを組み合わせることにより，切れ味の鋭いサービス実現を目指したいと思います．

■参考文献
⑴ 別所：“クラスタ内変動最小基準に基づくテキストセグメンテーション，”情処学論，Vol.47, No.3, pp.957-967, 2006.
⑵ 西脇・田中：“関連記事を利用したテキストセグメンテーション，”情処学研報，Vol.2002, No.104, pp.79-84, 2002.
⑶ M. Hearst：“TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages,” Computational Linguistics, Vol.23, No.1, pp.33-64, 1997.
⑷ 阿部・田邊・奥田：“ウェブ検索に基づくテキストセグメンテーション，”信学論 ⑴，Vol.J91-D, No.3, pp.723-732, 2008.
⑸ 阿部・内山・内山・奥：“ウェブ検索を利用したしきい値選択型テキストセグメンテーション，”情処学論，Vol.49, No.12, 2008.

（後列左から）内山　匡/ 奥　雅博/ 阿部　直人（右上）
（前列左から）内山　俊郎/ 別所　克人

次の時代の検索サービスの中に，世の中の人が「これはすごい！」と感じるようなアイデアを盛り込めたら良いなと思っています．

◆問い合わせ先
NTTサイバーソリューション研究所
TEL　046-859-2672
FAX　046-855-1730
E-mail　uchiyama.toshio@lab.ntt.co.jp