共起関係に基づくテキストの話題境界推定の試み

豊浦 潤* 木山 次郎† 伊藤 慶明‡ 岡 隆一*
*RWCP(Real World Computing Partnership) つくば研究センタ
†RWCP 新機能シャープ分散研究室
‡川崎製鉄 システム部 システム技術室

1 はじめに

最近、処理の対象となるテキストデータ集合の拡大に伴い、テキスト品質は多様化が問題になっている。この傾向は、私信である e-mail や、読者層の興味が予め想定できる Net-News などの場合に顕著である。また、広い読者層を想定している新聞記事の場合でも、1日分の記事を概観すると、実は全くタイプの異なるテキストの寄せ集めであることに気付く。このように、実世界に加工されないで存在しているテキストデータは膨大で、時間とともに増大していき、必ずしもデータ間での整合性がとれていないという問題がある。

数年前から、新聞記事の CD-ROM が製品化して、1年分の記事が1枚の CD-ROM に収録されるようになっている。各記事にはキーワードが付与されており、ユーザは、書誌情報やキーワードなどを手がかりに記事検索を行なうことになるが、多様な記事を、このような画一的な検索方法で扱うことに筆者は疑問を持っていた。例えば、記事にキーワードを付与する場合、大まかには、記事長 $\propto$ キーワード数 となる。そして、キーワード・ベースで記事をベクトルにすると、

長い記事 ： 密度の濃いベクトル $\vec{KL}$

短い記事 ： スパースなベクトル $\vec{KS}$

となる。今、$\vec{KL}, \vec{KS}$ は正規化されており、共にキーワード：A を含んでいるとする。このとき検索ベクトル: $\vec{A}$ に対する、一致度を通常行なわれるように cosine で評価すると、

$$\vec{KS} \cdot \vec{A} > \vec{KL}\vec{A} > 0$$

となり、短い記事の方が一致度が高いことになる。たとえ、長い記事が A について言葉多く説明した記事であってもである。これを避けるために $\vec{KL}, \vec{KS}$ を正規化しないとすると、$\vec{KL}$ は、$\vec{KS}$ よりずっと大きいノルムとなり、検索性能を低下させる noisy な記事になってしまう。

この問題に対する、1つの回答として、本稿では、テキストを話題境界でセグメンテーションする方法を提案する。長い新聞記事*は、しばしば、複数の事例について、列挙して述べていることが多い。こうした記事を、セグメンテーションし、検索のコーディングをセグメント毎に行なうことで、記事は複数のベクトルで構造的に表現されることになり、上記の問題も解消される。

2 話題境界推定

2.1 語彙的結束性

テキストセグメンテーションを行なうためには、何らかの尺度が必要であるが、これまでの研究では [2],[3],[1] が、語彙的結束性と共起関係の考え方が用いられている。ここで、語彙的結束性とは、与えられた単語集合の語間の類縁度の強さを表わすものである。

以下、テキストセグメンテーションの手順を、図1で簡単に説明する。

1. テキストからキーワード抽出を行なう
2. 出現部位が近傍のキーワード集合（共起性）に対して語彙的結束性を求める
3. 語彙的結束性が、相対的に小さい場所でセグメンテーションする

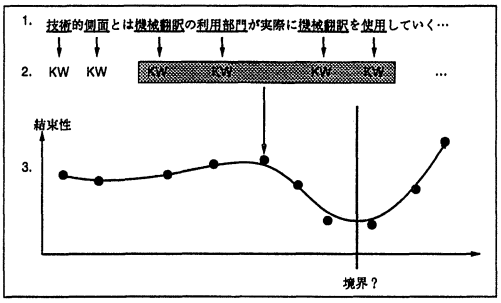

図 1: テキストセグメンテーションの手順

テキストセグメンテーションの精度は、語彙結束性の決め方にほぼ依存すると言って良い。これまでの研究例を概観すると、語彙結束性を

1. 単語の表層マッチング

2. 単語の意味の近さ

に注目して、シソーラスなどを利用して求めている。1 は、「同一の語が出現する区間は、同一話題である」を、2 は、「同一話題について述べる区間に出現する語は、意味的にも近い」ことを前提としている。確かに、同一話題区間では同語反復があるが、反復されるようなキーワードは少数に過ぎず、話題区間を推定するに足りる情報として充分であるかは疑問である。また、2 は百科全書の項目の分割などには有効かもしれないが、一般のテキストを対象とした場合、同一話題区間で意味集約されているかは疑問である。

2.2 統語的結束性

前節に述べた語彙的結束性は、テキストをキーワード・ベースでコーディングしており、統語情報は用いられていない。従来のテキスト処理では、処理時間と解析精度の問題から、統語情報をテキストコーディングに活用する研究は多くなかった。しかし、最近の計算機の高速化で、テキストを実時間で統語解析することも十分可能になっている。そこで、以下では統語情報を利用したテキストセグメンテーションを提案する。

*もちろん、新聞記事に限った話しではないが

話題とは何か、正確な定義を与えることは難しいが、ここでは「同じコトに対して述べている、連続する文」は同一話題であると考えることにする。これを以下、図を用いて説明する。図 2の一点破線で囲まれた領域は一つの文で、掛かり受けのある語を実線（以下、統語アークと呼ぶ）で結んでいる。このとき統語アークで結ばれた一連の語は、明らかに「首相の施政演説」という話題の内容を示している。換言すれば、統語アーク結ばれた語群は、話題に関して結束している。また、前後に出現する、「国会」、「首相」は、この文に出現する「国会」、「首相」は、前節に述べた共起性の仮説を採れば、文脈的に同じ話題に関連すると考えられる。そこで、これらを破線（以下、共起アークと呼ぶ）で結ぶ。但し、同一文章中でも極端に離れた位置に出現する場合は、もはや、同一話題に言及する語であるとは言い難いので、共起アークを結ぶ場合は語間の距離の制限が必要である。

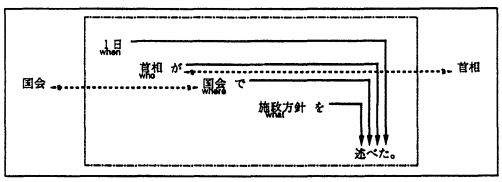

図 2: 話題を構成する語と関係

文章全体に関して、統語アーク、共起アークを張ると、同一話題に関する語群は閉グラフを構成することになる。生成されたグラフを $G_i(i = 1, 2, \cdots, n)$ とする。ここで、文章中の任意の地点: pos での話題の結束度が、$F(pos, G_i)$ で定義されれば、pos での話題は $F(pos, G_k) = \max_i^n(F(pos, G_i))$ を満たす G_k と決定できる。$F(pos, G_i)$ は、例えば

$F(pos, G_i)$ = 直線 $x = pos$ と G_i の交点数

とすれば良い。

3 評価実験

3.1 教師データの作成

これまでのテキストセグメンテーションの評価実験は比較的品位の良い文章に対して小

規模にしか行なわれていない。しかし最初に述べたように、今回筆者の目的は、品位の高くないテキストを大雑把にセグメンテーションできるかの可否にあるため、今回は、新聞記事†を対象にすることにした。

問題なのは、話題境界位置の決定である。多くのテキストに対して実験を行ないたいが、手作業で境界を探すのには大変な手間が掛かる。幸いにして、新聞記事では、先頭に "■" などの記号が来る場合は、その行が、小見出しになっている場合が殆んどであるので、これを手掛かりに、自動的に境界タグを付与することにした。この基準で付与したタグは、記事の全ての話題境界ではないかも知れないが、少なくとも付与したタグは話題境界になっていると考えられる。また、テキストセグメンテーションを行なうためには、記事はある程度長い必要がある。

具体的には、教師データを、以下の手順で決定した。

1. 一年分の新聞記事から 3000 文字以上の記事数を選ぶ
2. 文頭に、記号 ¦☆★○●◎◇◆□■△▲▽▼¦ いずれかが出現する行にタグを振る
3. タグ数が、文数の 1 割を越える記事を除く
4. タグ間の文数が、少ない記事を除く

3.2 実験方法

前節の手順で、作成された 531 記事を教師データとして、以下の 4 つの方法でのセグメンテーションを比較した。

1. 単語の表層マッチングによる方法（＝共起アークを用いる方法）
2. 単語の代わりに意味番号を用いる方法（小分類）
3. 単語の代わりに意味番号を用いる方法（中分類）
4. 統語的結束性を用いる方法

†具体的には、1989 年度の朝日新聞の記事を用いた

情報量を増やすために 1,4 では、部分列マッチを許した（比較する文字列の片方が 1 文字の場合は除外）。また、2,3 の意味番号は、角川類語辞典のものを用いた。

結束性を表す指標として、1〜3 では単純に、$F(pos, \boldsymbol{G}_k)$ で述べたように、横断するアーク数を用い、図 1で説明したように谷になる地点を、話題境界の候補とした。ここで、一方、4 についても、単純な考察でわかるように、

$$y = \max_i^n(F(pos, \boldsymbol{G}_i))$$

の谷が話題境界の候補となる。共起アークの長さの制限は 200 語長とした、また、実際に谷を求める際は、平滑化のために ±10 語の区間で平均した値を用いた。

3.3 実験結果

前節に述べた 4 つの方法で、新聞記事の話題境界候補を求めた。これを、教師データの境界と比較して ±5 語の範囲にある場合を正解として、下記で定義する適合率・再現率を求めた。

$$\text{適合率} = \frac{\text{正解の境界数}}{\text{話題境界候補の数}}$$

$$\text{再現率} = \frac{\text{正解の境界数}}{\text{教師データの境界の数}}$$

教師データの全境界について、適合率・再現率を 1〜4 の方法で求めた結果を表 1に示す。

表 1: 方法 1〜4 の、適合率・再現率

	方法 1	方法 2	方法 3	方法 4
適合率	11.8	11.2	12.0	12.4
再現率	34.7	18.6	21.9	36.1

また、方法 4 で、記事単位に適合・再現率を求めた結果を、図 3に示す。

3.4 考察

図 3より明らかのように、適合率に比べて、再現率の方が高い値を示す傾向が見られた。これは、前に述べたように教師データの性質からある程度予想された傾向である。より公平に比較するなら、教師データの境界数を n として、実験で求めた境界候補のうち谷の度合

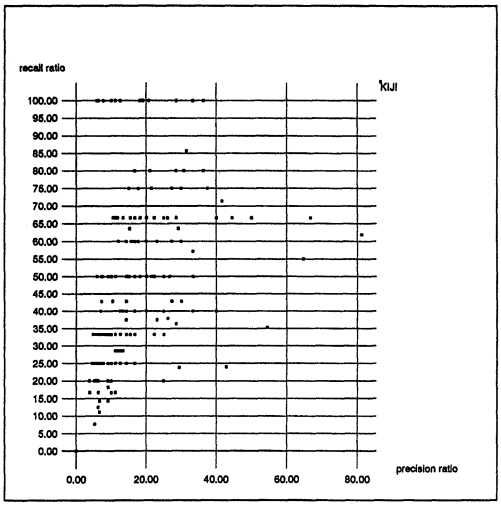

図 3: 方法 4 での記事毎の適合率・再現率

が大きいもの n を評価の対象とするべきだったかもしれない。しかし、現段階では、谷らしさの尺度が明確でなかったので[‡]、その評価は行なっていない。そもそも今回意図したセグメンテーションという作業自体、最初に分割の数を与えることは想定していないのである。また、適合率と再現率の間に、特に排他的な関係は見られなかった。

一方、表1の値は、方法 4 で、最善の値を示しているが、予期したほどの改善度は得られなかった。個々の記事に関して、目視で調査した結果、期待したように話題を構成する語間にアークが張られている事例も見られるが、以下のような傾向が見られた。

- 統語・共起アークのグラフ自体が、あまり適切に話題範囲を示していない
- 共起アークを張る語に "日本" など、話題を構成するにはあまりに一般的な語が多く、そうした一般語の出現の揺らぎの影響が大きい

こうした問題点を改善するためには、

- 共起アークを張る、部分列マッチングの条件を現在より厳しくする
- 統語解析を深くして、統語アークを張る場合を、who(subject),what(object) など、主格に対応する場合に限定し、when,where などの、自由格に対応する場合には統語アークを張らない

などの方式を改良していく必要がある。

また、今回実験に用いた教師データは急造のもので、若干信頼度に欠ける面もあったので、より精密な評価を行なうためには、精度の高い教師データの作成も必要である。

4 おわりに

統語的結束性を用いてテキストセグメンテーションを行なう方法を提案すると同時に、有効性を検証する実験を行なった。実験結果は必ずしも、期待通りのものではなかったが、方法を改善することにより精度を向上させる見通しはついた。

但し、最初に述べたように筆者の目的はテキストの構造化であり、質の良い $\boldsymbol{G}_i$ を求めることである。その意味でテキストセグメンテーションの精度を追究することはあまり考えていない。今後は、今回提案した方法などからテキストの構造化を進め、テキスト検索など具体事例への適用を検討していく予定である。

参考文献

[1] 木山 次郎, 他, Incremental Reference Interva-free 連続 DP を用いた任意話題音声の要約, 信学技報 *SP95-35*, pp.81-88, 1995.

[2] 小嶋 秀樹, 他, 単語の結束性にもとづいてテキストを場面に分割する試み, 情処研資 *NLP95-7*, pp.49-56,1995.

[3] 本田 岳夫, 他, 語彙的結束性に基づいたテキストセグメンテーション, 情処研資 *NLP95-7*, pp.25-32,1994.

[‡]谷底が低いのが谷らしいのか、谷底との隣接する山の頂上とのギャップが大きいのが谷らしいのか判然としない