語の分布と主題の展開に着目した文章の構造化

北條 孝† 田村 直良††

†横浜国立大学大学院 環境情報学府
††横浜国立大学大学院 環境情報研究院
{hojo,tam}@tamlab.eis.ynu.ac.jp

1 はじめに

本稿では、語の分布解析による大域的解析と主題関係による局所的構造解析を組み合わせた文章構造化手法について述べる。

大量に存在する電子化ドキュメントの処理を行なうにあたり、必要な情報を効率的に入手するための手段として自然言語処理に対する要求が高まっている。そのため、現在、情報検索などいくつかの分野においては、パターン駆動など表層的な処理により、効率的な処理システムが多く提案されている。一方、人間の文章理解に近いテキスト処理においては、なんらかの内容理解を基にしたより高度な処理が本質的である。テキストを意味的なセグメントによって構造化することは、テキストの意味理解の最初の一歩として特に重要な意味を持つ。セグメント間の関係の解析は、論旨の展開構造の理解や、筆者の主張の理解に通じるものである。

我々は、特に著者の論理構造解析手法を検討することを目的に、論旨の展開が明確で著者の主張が含まれている新聞の社説記事を対象としている。本稿では、大域的な論旨の展開を把握するための語の分布解析と個々の主張の展開を把握するための主題構造解析を組み合わせたセグメント境界検出手法について述べる。

文章の分割に関する研究には、固有表現を手がかりにする方法 [5] や単語の出現分布を利用する方法 [1] などがある。前者は、機械学習ベースであり、パラメータの選び方や訓練にコストがかかると思われる。後者は固定長のウィンドウを設けて語の出現をグループ化するが、一般に各語の出現が少数であるため、調整が微妙である。

我々の方法は、大域的な解析と隣接する文間の解析から成る。大域的な解析では、語が含まれているかどうかについての情報量（エントロピー）が最小になるようなセグメントの分割を求めることを基本原理とし、文間の解析では、主題の扱いに関する結束性に着目する。

2 語の分布解析に基づくテキスト・セグメンテーション

内容に関わる主要な語句は、テキスト全体で語彙連鎖を形成しているものの、テキスト中で一様に出現するのではなく、偏って現れる。本研究では、情報量（エントロピー）の観点で語の出現の極在性を扱い、エントロピーが最小になる分割を求めることにより大まかなテキスト・セグメンテーションを行う。

2.1 語の分布とテキスト・セグメンテーションのモデル

語が多く出現している個所は、その語に関わる内容が述べられていて一つのセグメントを成し、他の部分とは区別された役割を担っている。

まず、テキスト中の語の出現について、モデルを考える。各語は出現位置に偏りがあるとして、図1のようにモデル化する。

図において、横方向はテキストの先頭からの正規化された位置を表し、縦方向は出現頻度に相当する。モデルでは、語が出現しない部分と集中する部分があるとし、集中する部分を1～3個所としている。

語の出現は、直感的には、いくつかの正規分布を合成したもの（n次元正規分布）が考えられるが、提案のモデルではセグメント内の語の発生確率のみが問題となる。形式的には、語 w の出現によるセグメンテーションは、

$$S_w = \{p_1, p_2, \cdots, p_n\}$$

となる（$n+1$ 分割の場合）。ここで、S_i は、テキスト中でセグメントが開始される位置（例えば文番号）である。

テキストの（大まかな）セグメンテーションとは、語毎に出現する部分としない部分の境界を求め、類似した語を集めることによってテキスト全体を分割することである。

実現した解析システムでは、形態素解析システム茶筌 [4] によりテキスト全体を形態素に分割し、語

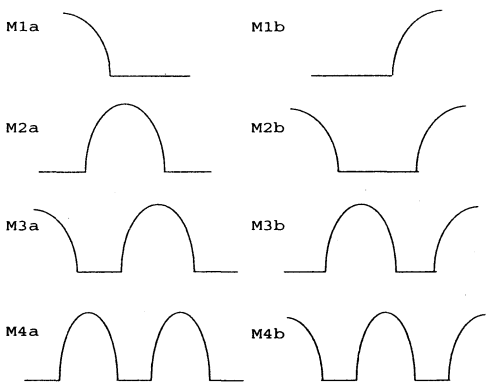

図 1: 語の出現のモデル

の分布を調べている。このとき、分類語彙表 [3] を用いて、最小項目まで一致する語は同一な語として扱っている。

2.2 エントロピーによるセグメントの分割

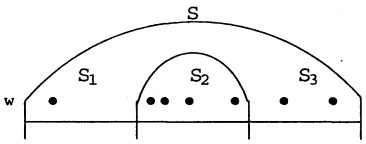

図 2: 語とセグメント

3 分割のモデルで検討する。図 2において、黒丸（・）は語 w の出現を表すものとする。セグメント全体（S）の文の数を $|S|$、語 w の出現数を $freq(w,S)$ とすると、語 w が S に含まれる生起確率は $\alpha = freq(w,S)/|S|$ であるので、w が S に含まれるかどうかを伝えるメッセージの情報量（エントロピー）は、

$$H(w,S) = \alpha \log \frac{1}{\alpha} + \bar{\alpha} \log \frac{1}{\bar{\alpha}}$$

となる。ただし、$\bar{\alpha} = 1 - \alpha$ である。同様に、w が S_i に含まれるかどうかを伝えるメッセージの情報量 $H(w,S_i)$ も定義できる。

セグメント S が S_1、S_2、S_3 に分割されたとき、各セグメントのエントロピーにセグメントの大きさで重み付け加算することにより平均のエントロピーを得られる。

$$H(w,\{S_1,S_2,S_3\}) = \sum_i \frac{|S_i|}{|S|} H(w,S_i)$$

3 分割のモデル（S_1、S_2、S_3）における語 w についての最良のセグメント分割は、

$$\underset{S_1,S_2,S_3}{\operatorname{argmax}} H(w,\{S_1,S_2,S_3\})$$

となる。

それぞれのモデルを仮定した場合についての最小のセグメント分割を求め、さらにそのなかで最小のエントロピーを持つモデルを採用する。

直感的な原理として、あるクラスの要素が特に多く含まれるなど偏った分割を行うことにより、セグメント内の要素があるクラスに属するかどうかを伝えるメッセージの情報量が、平均して下がることによる。

語がセグメント内に存在することと存在しないことはそれを伝えるメッセージとして共役の関係にある。つまり、モデルとして図 1を用いるが、M1a と M1b、〜、M4a と M4b は、エントロピーの算出としては等価な式となる。

2.3 セグメント分割、あるいは語の距離

前節までで各語毎に最良のセグメント分割を求めた。次に、類似した語を集めてクラスタを構成する。語義の観点からは必ずしも類似しているとは限らないが、筆者が意図する論旨の展開上、類似した出現をする語は構成するセグメントにおいて同様な重要性を持つと考えられる。

これに基づき、セグメントの分割数が等しくなる二つの語 w と w' の間の距離、すなわち語の出現により規定されたセグメント分割についての距離 $d(w,w')$ を次のように定義する (ユークリッド距離)。

$S_w = \{S_1, S_2, \cdots, S_n\}$
$S'_w = \{S'_1, S'_2, \cdots, S'_n\}$ のとき、

$$d(w,w') = \sqrt{\sum_i (S_i - S'_i)^2}$$

セグメントの分割数が一致しない場合は、差が 1 の場合のみ許し、それ以外は距離無限大とする。差が 1 の場合は、多い方のセグメントから任意に一致する数だけ分割点を選んで上式の値を求め、最小の値を採用することにする。

2.4 語のクラスタリング

前述の距離に基づいて語のクラスタリングを行う。用いたアルゴリズムとしては、各要素について（1）生成されたクラスタのどれかの要素と距離が近ければそのクラスタに要素を含める、（2）どれとも近くなければ新規にクラスタを生成する、といった単純なものである。

提案方式によるセグメントの分割では、境界のみ得られるので、セグメントの分割が同一でも、それが語を含むセグメントなのか含まないセグメントなのかが判断できない。しかし、高々3分割のセグメンテーションでは、大きなセグメントが必ず含まれる。そのときその大きなセグメント内で特定の語の発生確率が十分に高くなることは稀であると思われることから、セグメントの分割が類似している場合は、語の出現も類似していると思われる[1]。

クラスタリングの実験結果の一部を図4に示す。

罫線で区切られた区間がクラスタリングで同一のクラスタに含まれると判断された名詞である。各名詞において、上段のM1,M2,…はモデル、1,2,3,…はセグメント区間を表し、数字の変化がセグメント境界を表す。下段は、名詞のテキスト内での出現位置を表す。

結果の例でも分かるように、セグメントを構成する語が意味的に近いという保証はない。これらの語の出現が、たまたま一致していたのか、あるいは何らかの関連があってのことなのかについては今後の課題とする。

2.5 テキストの（大まかな）セグメンテーション

前節のクラスタリングによりある程度大きなクラスタを構成する語のグループは、他の語と独立してセグメントを構成する。このようなセグメントは、セグメントどおしあまり重ならないものと予想され、これによりテキスト全体のセグメンテーションを行う。

以上のエントロピーを基にして得られるセグメントの境界は大域的な分割としては合理的であるものの絶対的なものではなく、文と文の隣接関係などの微妙な関係については別な制約を導入すべきである。次節では、隣接する文間の関係として主題構造に着目して検討を進める。

3 主題構造に基づくセグメント調整

主題構造解析により隣接文間の関係からを解析し、抽出されたセグメント境界の微調整を行なう。

主題とは、一文において話題の中心となる語句である。題述とは、主題以外の名詞句である。本研究では、各文に必ず一つの主題が存在するとして解析を行なう。

3.1 主題の抽出

主題構造解析をするために、トピックと主題の抽出を行う。トピックと主題の定義を以下に示す。

[1]このことは今回実験においても確かめられている。

- トピック：社説記事における記事タイトル中に出現する語句のような、文章全体の話題を表すような**名詞句**をトピックと定義する。
- **主題**と**題述** [2]：各文は主題構造を持つと仮定し、各文は主題と題述とから構成されているとする。具体的には、**は格**もしくは初出現の**が格**を主題と定義し、文の主題以外の残りの**名詞句** を題述と定義する。

3.2 主題の連鎖関係の種類

テキスト中の隣接する文間が下記の条件の6種類の連鎖関係のうちのいずれかを満たすものとし、結束性を持っているとする。

A **主題維持**：直前の文の主題と同一か、基準以上の類似性を持つ主題を持つ場合。

B **主題変化**：直前の文の題述のいずれかと同一か、基準以上の類似性がある主題を持つ場合。

C **主題回復**：最も近い主題変化の直前の主題と同一か、基準以上の類似性がある主題を持つ場合。

D **トピックの導入**：テキストのトピックと同一か、基準以上の類似性がある主題を持つ場合。

E **主題派生**：上記のいずれにも該当しない場合。この場合、直前の文やトピックとは関連性の低い文となる。

F **新規主題**: 最初に主題が出現した場合

ただし、基準以上の類似性とは分類語彙表 [3] の最小項目の範囲での同一性を、部分文字列関係に拡大して用いる。

3.3 主題の連鎖関係の決定

主題の連鎖関係を決定するルールを示す。

- ルール1：原則として、結束関係の強さはA＞B＞C＞D＞E＞Fとし、可能な限り結束性の高い連鎖を採用する。ただし、主題を抽出する際、主題が省略されている文に関しては、省略 (ellipsis) により結束構造 (cohesion)[2] があるものとして、主題維持と見なす。
- ルール2：最初に主題が出現するまでの文は「トピックの導入」とし、主題を定めない。最初に主題が出現した文を「新規主題」とし以降主題の連鎖関係を決定する。

図5に主題構造解析による抽出例を示す。

No. は文頭からの文番号を表し、 rel は主題関係を表す。 theme,rehme はそれぞれ各文の主題、題述を表す。

4 抽出実験

図3のような日経新聞1993年版の社説記事を対象に、提案手法により抽出実験を行った。

```
00 危うさを感じさせる急ピッチな株高（社説）
01　株価が上昇している。
02 売買高も大きく膨らんでおり、株式市場関係者の表情は久方ぶりに明るい。
03 ただ、景気の先行きはまだ安心できる状態ではなく、急ピッチな円高で企業業績はむしろ厳しさを増している。
04 危うさを感じさせる株高に幻惑されてはいけない。
05 不況克服の地道な努力を企業や政策当局に望みたい。
06　株価が上昇し始めてまだ一カ月だが、日経平均は二割近くも上昇し、ほぼ二万円を回復した。
07 理由はいくつかあげられる。
08 まず、公定歩合の引き下げで金利が低下し、債券市場などから株式市場に資金が移動したこと。
09　「金融相場」といわれるのはこのためで、景気の底の段階でしばしばみられる不況下の株高である。
10 多分、これが最大の要因だろう。
11　これに来週発表される追加景気対策を織り込むかたちで景気回復への期待感が高まったことや、天井圏で推移している欧米の株式市場から外国人投資家が日本市場に目を向けだしたことなどが株高をもたらした。
12　暗く長いトンネルが続いていただけに、今回の株高の心理的効果はまことに大きい。とりあえず銀行や事業会社の前三月期決算の負担を軽くしただけではない。
```

図 3: 処理対象記事（部分）

```
    0----+----10----+----20----+----30----+----40
    |----+----|----+----|----+----|----+----|
株式
M4 1112222333444555555555555555555555555
   0010000010010000000000000000000000000
市場
M4 1112222333444555555555555555555555555
   0010000002002000000000000000000000000
者
M3 1112222222233444444444444444444444444
   0010000000010000000000000000000000000
    |----+----|----+----|----+----|----+----|
```

図 4: クラスタリング結果の例（部分）

図4では、語の分布解析からセグメント位置は第2文と第3文（「株式」「市場」、「者」）、第6文と第7文—(「株式」、「市場」)、第8文と第9文(「株式」、「市場)、第11文と第12文(「株式」、「市場」、「者」)の付近に存在する。しかし、主題構造解析による関係では、第7文から第11文までは主題が一致していると判断されるため、第8文と第9文の間は分割点としては見なさない。全てのクラスタの境界について同様に分割点が調整される。

```
No.  6 rel: 新規主題
theme:
理由
rehme:
日経 平均
回復
二万 円
上昇
二 割 近く
一 カ月
上昇
株価

No.  7 rel: 主題継続
theme:

rehme:
いくつ

No.  8 rel: 主題継続
theme:

rehme:
移動
資金
株式 市場
債券 市場 など
低下
金利
引き下げ
公定歩合
```

図 5: 主題構造解析の例（部分）

5 おわりに

本稿では、語の分布に関する情報量（エントロピー）を最小にする分割を採用することを原理としたテキスト・セグメンテーション手法について述べた。語によって分割は異なるが、分割が類似している語を集めることによりセグメントが構成でき、それらの語が何らかの内容語となっていることを確認した。

また、隣接する文間の主題の連鎖関係に着目した主題構造解析について述べた。この関係は、テキストセグメンテーションが大域的な構造解析であるのに対して、局所的な解析をするものであり、セグメントの境界を微調整する。

両者を組み合わせることにより、より適切な構造化が行われることが確認された。

参考文献

[1] M. Hearst. Multi-paragraph segmentation of expository text. *Proceedings of the 32nd Annual Meeting of the Association for Computional Linguistics,pp.9-16*, 1994.

[2] M. A. K.Halliday. An introduction to functional grammar second edition. くろしお出版, 2001.

[3] 国立国語研究所. 分類語彙表. 秀英出版, 1994.

[4] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 浅原正幸. 日本語形態素解析システム「茶筌」version2.0 使用説明書 第二版. NAIST Technical Report, 奈良先端科学技術大学院大学 松本研究室, 1999.

[5] 望月源, 本田岳夫, 奥村学. 複数の表装手がかりを統合したテキストセグメンテーション. 自然言語処理学,Vol6,No.3,pp.43-58, 1999.