決定木獲得アルゴリズムを用いたテキストセグメンテーション

本田 岳夫, 望月 源, Ho Tu Bao, 奥村 学
北陸先端科学技術大学院大学 情報科学研究科
email:{honda,motizuki,bao,oku}@jaist.ac.jp

1 はじめに

テキストを構成する各文には何らかの意味的なつながりがある。このつながりを捉え、テキストを段落などの一貫性のある談話の単位に分割する研究は、談話構造解析の最初のステップに位置付けられる。

従来、手がかりとして、テキスト中に多数存在する表層的特徴が用いられてきた。最近になって、語彙的結束性 [1] の情報を使用した研究も行なわれている [3, 8]。本稿では、従来の表層的情報と語彙的結束性の情報を同時に使用して、テキストを段落への分割する手法について検討する。なお、本研究の段落とは意味段落をさす。

複数知識の統合手法として、我々は、段落境界を推定する時に、各境界前後の文の特徴をパラメタにとり、その各パラメタに対する重み付き総和をとり、その総和が高いものほど段落境界になり易いとして推定する“算術的” 分類モデルを用いてきた [6]。一方本稿では、複合知識の統合手法として、テキスト中の文と文との境目に対して、その前後の文の表層的な情報と語彙的結束性の特徴を C4.5 に与える属性とし、訓練テキストの正解となる段落境界／非境界をクラスとして決定木を獲得し、その決定木を用いて段落境界を推定する “論理的” 分類モデルによるアプローチをとる。重み付き総和を用いる手法では、離散的で排反な特徴でも別々のパラメタとして扱う必要があり、連続値とみなす必要があるものの、すべての段落境界候補で和が計算できるため、和の大きいものから N 個出力するということができる。一方、決定木を用いる本手法では、離散的で排反特徴を一つの属性の離散的な属性値として表現できる。本手法では、決定木獲得アルゴリズムとして C4.5[4] を使用する。

2 セグメンテーションに用いる知識

本研究で使用する知識は、重み付き総和による段落分割 [6] と同様であり、文と文との間を段落境界の候補として、その前後の文の特徴を属性にとる。これらは、大きく分けてテキストから直接抽出できる表層的特徴と、シソーラスを用いて計算される語彙的結束性の情報にわけられる。

表層的特徴は、着目する文の、文頭、文末、文中の 3 つに分けられる。文頭からは、接続詞の種類 (順接、逆接 etc) と前方照応の種類 (「あ」型、「こ」型、「そ」型) が得られる。文末の言い切りの形からは、文のタイプ (叙述文、判断文 etc) が得られる。文中からは、格助詞ガ、副助詞ハが語彙的連鎖を形成する語を伴って出現するかどうかを調べる。

語彙的結束性 [1] とは、テキスト中の語間の意味的つながりのことで、この語彙的結束性を持つ語の集まりを語彙的連鎖 [3] と呼ぶ。属性は、語彙的連鎖の開始、終了位置と、中断および再開 (これをギャップと呼ぶ) 位置に設定し、各段落境界候補でそれぞれ開始／終了している連鎖／ギャップの数を数え、連続値として与える。また、着目している段落境界候補を貫く主要な連鎖を形成する語について、その語を修飾する語が境界候補位置で変化する数も属性とする。なお語彙的連鎖は、角川類語新辞典 [5] を用い、本田 [7] のアルゴリズムにより計算する。

属性の取り方に付いて、次の例文[1] の 2 文と 3 文の間の境界候補を用いて説明する。

[1] 石田春夫,「学生のための自分学」より

1. だから科学の言葉は、科学者の自己を離れて 物のほうに張りついている。

2. 科学が主観を排除して、客観的事実だけを示すというのは、このように科学での言葉が (「ガ」の主語) 自己 を離れているということによるのである (断定文)。

3. これにたいして (逆接)、詩(修飾語の転換) での言葉は (「ハ」の主語) 物を表現するときでも、自己を離れるということはない (判断文)。

4. 物そのものが語っているように見えても、言葉は詩人の 自己 の手に握られているのである。

2 文目には「言葉が」、3 文目には「言葉は」が出現している。文のタイプはそれぞれ文末から、断定文、判断文となり、3 文目の文頭には逆接の接続詞がある。語彙的連鎖では、2 文目で [科学, 科学者, 科学, 科学] の連鎖が終っており、また 3 文目からは、[詩, 詩人] の連鎖が始まっている。また、この境界候補では、貫く主連鎖に [言葉, 言葉,...] があり、それぞれ「科学での言葉」「詩での言葉」と修飾語が転換している。

このように取り出した特徴を属性を定義する場合、いくつかのパターンが考えられる。まず、望月 [6] と同じようにすべての特徴が現れるか現れないかという形で与える方法がある。また、例えば、文頭に逆接の接続詞が出現したら、順接にはならないし、文タイプが判断文であれば、断定文にならない。このような特徴は、属性「文頭」の値を離散値で 順接, 逆説, ... と与えることもできる。また文タイプも境界候補の前後の文の文タイプが同じか異なるか、という属性、前の文と後の文の文タイプを別々の属性にするかの 2 通りが考えられる。本稿では、排反な属性をまとめる場合とまとめない場合の組合せも行なう。属性をまとめない場合には最大 23 属性、まとめる場合には、最少 13 属性になる。

3 C4.5

C4.5[4] は分類クラスといくつかの属性を持つ事例の集合からなる "フラットなファイル" から決定木を生成するシステムである。事例は、あらかじめ決められた属性とその属性値によって記述され、前もって定義された分類クラスに割り当てられている。属性値は、離散値でも連続値でも良い。

決定木の生成には分割統治法により、根ノードから後戻りなしに決定木を生成する。弁別ノードで調べる属性の選択には利得比基準により、事例の集合のエントロピーの減少がもっとも大きくなるような属性を選択する。また、決定木の単純化と過適合への対処のため、訓練事例から生成された決定木の枝刈りを行なう。枝刈りは、誤判別の確率を統計学的に考慮して、ある弁別ノードによって分けられる時の予測誤り率と、そのノードを葉ノードに置き換えた時の予測誤り率を比較して、葉に対する予測誤り率の方が低い時に行なわれる。

4 実験

実験条件は、いくつかの独立な条件が考えられる。本稿では、属性の与え方、属性値のグループ化、訓練データの選択について考慮する。

C4.5 に訓練集合を与える時の属性の与え方は 2節で述べたように、いくつか考えられる。ここでは、次に示すように属性を与える。

- 文頭からの情報 (接続詞、照応)
  - a1 「順接」の接続詞が存在するかしないか、のような 2 値属性
  - a2 文頭に出現するものの種類を属性値にとる離散的な多値属性
- 文タイプ
  - b1 境界候補の前後の文が同一の文タイプ
  - b2 前の文の文タイプ、後の文の文タイプという 2 つの多値属性

C4.5 はオプションで属性値をグループ化して決定木を生成させることができる。離散的な多値属性を与える場合には、グループ化するかしないかの 2 通りの選択がある (条件 c1,c2)。

実験用のテキストは、対象が中学校1,2年生から、大学受験程度までの範囲の国語現代文の問題集から、随筆および論説文を、24テキスト (平均境界候補数24.1) 選び出して使用する。正解の段落境界位置は、問題集の解答 (平均正解境界数 2.83, 全体の 11.8%) と心理実験の結果による解答 (平均正解境界数2.87, 全体の 12.0%) の2種類を用意した。心理実験は、5名の被験者に同じ問題を解くことを求めた。過半数の3名以上が、境界位置と認めた場所を正解とする[2]。

24(23) テキストを3つのグループに分け、3回の交差検定 (cross-validation) を行ない評価する。評価には、情報検索の分野で用いられる再現率 (recall) と適合率 (precision) を用い、各テキスト毎に適合率と再現率を計算し、訓練集合、評価集合それぞれの平均を用いる。適合率と再現率は次式で与えられる。

$$\text{再現率} = \frac{\text{出力結果に含まれる正解境界数}}{\text{正解境界数}}$$

$$\text{適合率} = \frac{\text{出力結果に含まれる正解境界数}}{\text{出力境界数}}$$

実験に用いる訓練集合の内訳を表1に示す。交差検定

	総数	境界数	非境界数
問題集1	390	41	349
問題集2	391	48	343
問題集3	375	47	328
問題集全体	578	68	510
心理1	360	37	323
心理2	361	46	315
心理3	375	49	326
心理全体	548	66	482

表 1: 実験データ

のそれぞれの訓練集合はいずれも非境界数に比べて境界数が非常に少ないので、すべてのデータを用いて学習すると、すべての事例が非境界と決定する根ノードのみの決定木が得られてしまう。信頼度を甘くして枝刈りを押えても、訓練集合に含まれるテキスト毎に見ると、境界に分類する事例がないものがある。この理由は、あるノードの枝刈りが行なわれると、そのノードからの部分木に含まれていた非境界に分類される事例の数が境界に分類される事例の数より多いため、置き換えられた葉は非境界クラスになってしまうためである。

この問題に対処するため、本稿では、非境界の数を制限して決定木を生成することにする。ここでは、境界数の 1/2、同数、1.5 倍、2.0 倍の非境界を訓練集合からランダムに選んだものと、境界事例すべてを実際の訓練集合として決定木を生成する。これをそれぞれ10回行ない、その平均をとる。

最終的に実験の条件は、属性の与え方 (2 * 2 通り)* グループ化 (2 通り)*非境界の数 (4 通り) の 32 通りを、問題集、心理実験を正解にした場合に適用する。評価集合に対する再現率–適合率のグラフを図 1,2に示す。

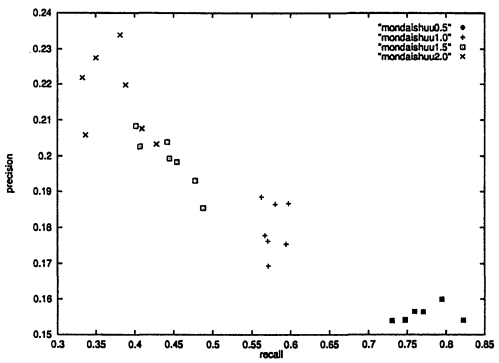

図 1: 結果 (問題集)

5 考察

問題集、心理実験ともに訓練に与える非境界の数が少ないほど再現率が高く、多くなるに従って再現率 が下がっている。問題集では、若干ではあるが、非境界数が多いほど適合率が高い。心理実験では、1.5 倍の時に適合率がもっとも高いものがある。これは、枝刈りをする時に非境界の数が少ないほど、境界になる葉が多くなり、境界と推定される評価事例の数が多くなるた

[2]3名以上の同意がなく、境界なしとなったテキストが1つあったので、心理実験では23テキストが対象となる

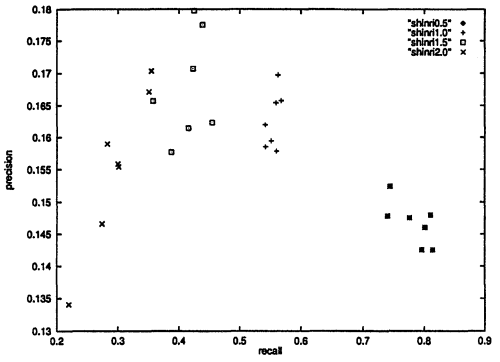

図 2: 結果 (心理実験)

めである。心理実験の適合率が 2.0 倍のところで下がっているのは、境界と推定するものが少なくなり過ぎて、適合率、再現率が 0 になるものが多かったためである。

属性の与え方に関しては、あまり明確な結果の差が出なかったが、文頭の情報を 2 値属性にとり (a1)、前後の文タイプを 2 つの属性に分けたもの (b2) が総じて良く、問題集では属性のグループ化をしない場合 (c1)、心理実験では、グループ化をしたもの (c2) がもっとも良かった。

適合率が総じて低いのは、境界と推定された非境界の数が多いためであり、これには次のような理由が考えられる。境界の、語彙的連鎖の属性を除いた属性の値の列と全く同じ属性値列を持つ非境界 (ニアミスと呼ぶことにする) の数を調べると、問題集の場合、境界 68 事例に対し、ニアミス 87 事例あった。このニアミスと境界を分割するためには語彙的連鎖の属性により分割しなければならないが、もし訓練集合に含まれていないニアミスがある時には境界と推定されてしまう。これにより境界と推定する数が多くなり、適合率が下がることになる。また、ニアミス数が境界数よりも多いため、訓練集合にニアミスを多く含んだ場合には、枝刈りが起こった時にニアミスと境界を含んだ葉は非境界の葉になる可能性が高く、この場合には再現率が下がる。

非境界の数の制限は、枝刈りにより境界となる葉がなくなることに対処するためであった。しかし、実際に評価で与える事例は非境界が 10 倍近く多く、精度が下がってしまう。これを解決するためには、境界に対する枝刈りと非境界に対する枝刈りの尺度を変えて対処する必要があるのではないかと考えられる。

謝辞

本研究で使用した、日本語シソーラス「角川類語新辞典」の使用許可を下さった (株) 角川書店に感謝します。

参考文献

[1] Michael Halliday and Ruqaiya Hasan. *Cohesion in English.* Longman, 1976.

[2] Diane J. Litman and Rebecca J. Passonneau. Combining Multiple Knowledge Sources for Discourse Segmentation. In *33rd Annual Meeting of the Association for Computational Linguistics*, 1995.

[3] Jame Morris and Graeme Hirst. Lexical Cohesion Computed by Thesaurual Relations as an Indicator of the Structure of Text. *Computational Linguistics*, Vol. 17, pp. 21–48, 1991.

[4] J. Ross Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, 1993. (邦訳:AI によるデータ解析, 古川康一 監訳, トッパン,1995).

[5] 大野晋, 浜西正人. 角川類語新辞典, 1981.

[6] 望月源, 本田岳夫, 奥村学. 複数の知識の組合せを用いたテキストセグメンテーション. 情報処理学会自然言語処理研究会資料 109-7, pp. 47–54, 1995.

[7] 本田岳夫, 奥村学. 語義曖昧性を考慮した有意な語彙連鎖の生成. 情報処理学会研究会資料 NL97-14, pp. 95–102, 1993.

[8] 本田岳夫, 奥村学. 語彙的結束性に基づいたテキストセグメンテーション. 情報処理学会研究会資料 NL102-4, pp. 25–32, 1994.