重回帰分析とクラスタ分析を用いたテキストセグメンテーション

望月 源, 本田 岳夫, 奥村 学
北陸先端科学技術大学院大学 情報科学研究科
e-mail:{motizuki,honda,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, 2] の情報を使用した研究も行われている．両者の組み合わせは，分割精度の向上に役立つと考えられるが，これまであまり行われていない．そこで本稿では，複数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提案する．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，我々は各パラメータに対する重み付き総和を使用する．このアプローチでは，精度向上のために，各パラメータの重要度に応じた重み付けと，各パラメータのデータに応じた組合せが必要である．

本稿では，統計的手法である重回帰分析を用いることにより，これらの問題を解決し，段落分割精度を向上する手法を提案する．また，さらに精度を向上させるために，クラスタ分析の手法の組み合わせについても議論する．

2 使用する知識

本研究で使用する知識は，[8] と同様であり，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．

語彙的結束性とは，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と呼ぶ．パラメータは，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) 位置に設定する．なお語彙的連鎖は，角川類語新辞典 [6] を用い，[9] のアルゴリズムにより計算する．

一方，表層的情報は，情報の種類によって助詞の情報，接続詞の情報，前方照応の情報，主語の有無の情報，同一の文タイプの連続の情報，修飾語の情報に分類をし，各情報ごとにパラメータを設定する．

テキストの n 文目と $n+1$ 文目の間を $point(n, n+1)$ とし，$point$ ごとにパラメータを抽出する．表1に，設定したパラメータと抽出のための規則を示す．

p	g	分類	抽出方法
1	1	助詞	「は」の付く文の前に 1 点
2			「は」の付く文の後に 1 点
3			「が」の付く文の前に 1 点
4			「が」の付く文の後に 1 点
5	2	接続詞	「添加」型の文頭出現，前に 1 点
6			「強調」型の文頭出現，前に 1 点
7			「説明」型の文頭出現，前に 1 点
8			「順接」型の文頭出現，前に 1 点
9			「逆接」型の文頭出現，前に 1 点
10			「転換」型の文頭出現，前に 1 点
11	3	前方照応	「あ」型の文頭出現，前に 1 点
12			「こ」型の文頭出現，前に 1 点
13			「そ」型の文頭出現，前に 1 点
14	4	主語	主語の無い文の前に 1 点
15	5	同一文タイプ	叙述文→叙述文の間に 1 点
16			判断文→判断文の間に 1 点
17			断定文→断定文の間に 1 点
18			その他→その他の間に 1 点
19	6	修飾語	主 chain の修飾語が変わったらその前に 1 点 (累積加算)
20	7	語彙的連鎖	連鎖の開始の前に+1 点
21			連鎖の終了の後に+1 点
22			ギャップの開始の後に+1 点
23			ギャップの終了の前に+1 点

表 1: 使用するパラメータ

3 パラメータの重み付け

重み付けの方法には，人手によるものと，自動的に計算されるものが考えられるが，労力の軽減，客観性のある値，再計算の容易さなどの観点から，人手によるよりも，自動化した方が大きなメリットがあると考

えられる．そこで本研究では，文間 $(point)$ の境界へのなり易さを，目的変数 y と置き，各パラメータを，説明変数 p_j と置くことで訓練セットを作成し，重回帰分析により自動的に重みを推定する [7]．

$y_i = w_0 + \sum_{j=1}^{k} w_j \times p_j$

境界位置の y には 10 を，非境界位置の y には-1 をそれぞれ与えた．

4　不要なパラメータの除去

本研究で設定したパラメータには，扱うデータに対して，すべてが有効かどうかがはっきりしないという問題点がある．そこで，データにあった最適なパラメータを選択する必要がある．対策として，不要なパラメータの除去が考えられるが，明確な除去基準が必要となる．重回帰分析では，最適な変数選択の手法が開発されている．そこで，本研究では，その変数選択手法を用いて，パラメータの最適化を行う．

具体的には，仮説 $H_0 : \omega_i = 0$ を検定することで，各変数が推定に必要かどうか判定する (ここで ω_i は，推定値 w_i の真の値)．仮説 H_0 が棄却されなければ，対応するパラメータ p_i は，不要なものと分析され，逆に棄却されれば，有用であると分析できる．検定は，各パラメータ (p_i) ごとに検定統計量 F_0 を求めて行う．$F_0 = (\frac{w_i}{SE(w_i)})^2$ (ここで $SE(w_i)$ は標準誤差と呼ばれ w_i の標準偏差を表す)

F_0 の分布は，自由度 $(p, n-p-1)$ の F 分布に従うので，$F_0 >= F_{(p,n-p-1)}(\alpha)$ なら，有意水準 α で仮説 H_0 は棄却され，p_i は含めても良いと検定できる．

各パラメータについて検定を行うアルゴリズムは複数あるが，本研究では，最も一般的な stepwise 法 (変数増減法)[7] を用いた．これにより，不要なパラメータの除去を行う．

5　テキストのクラスタリング

重みの推定に使用する訓練セット内の各テキストには，パラメータの出現パターンにばらつきがあると考えられる．そのため，各テキストを，類似したテキストからなるいくつかのクラスタに分類した方が，より最適な重みの推定が行えると考えられる．そこで，本節では，訓練テキストのクラスタリングと，評価テキストの属するクラスタの決定について述べる．

テキスト i のパラメータの特徴ベクトル V_i は，$V_i = (v_{i1}, v_{i2}, \cdots, v_{im})$ と表され，V_i の要素 v_{ij} は，$v_{ij} = \sum_{k=1}^{n} p_{kj}/n$ (ここで n はテキスト i の文間数) と定義される．

テキスト間の類似尺度 $sim(V_a, V_b)$ には，
内積 $(\sum_{i=1}^{m} v_{a,i} v_{b,i})$，と
cosine 距離 $(\sum_{i=1}^{m} v_{a,i} v_{b,i} / \sqrt{\sum_{i=1}^{m} v_{a,i}^2 \times \sum_{i=1}^{m} v_{b,i}^2})$
を使用する [3]．

クラスタ間の類似度は，平均距離法で計算する．

5.1　訓練テキストのクラスタリング

訓練セットのクラスタリングアルゴリズムには，階層的クラスタリングを用いた．閾値には Sekine[4] の用いた基準を，基本的なアイディアとして使用する．

Sekine による基準は，次のようになる．まず，1 つのクラスタ内のテキストを 1 つのテキストと見立てて，情報理論からの概念である Uni-gram Entropy $H = -\sum_w p(w) \log_2(p(w))$ および Perplexity $PP = 2^H$ を計算する (確率 $p(w)$ はテキスト内の $token$ w の確率で，$p(w)$ = $token w$ の数/テキスト内の token の総出現数で求める)．

また，分類対象となるすべてのテキストから同数の $token$ を，ランダムに抜き出して，疑似的に同じサイズのクラスタを作り，同様に PP を計算する．

PP はクラスタ内の情報量を表す．同数の $token$ を持つ 2 つのクラスタでは，PP の小さい方が，同じ $token$ の出現割合が高いことを意味し，内容的にまとまっていると考えられる．類似度を計算して作られたクラスタの方が，ランダムに作られたクラスタより，まとまっていると考えられる．

そこで，実際のクラスタの $PP(PP_{act}$ と呼ぶ) と疑似的な $PP(PP_{rnd}$ と呼ぶ) の比率 (PP_{act}/PP_{rnd}) をクラスタリングの各段階ごとに計算し，比率の最小点での段階が，最もまとまった状態と考え，その時の分類を採用する．

本研究では，$token$ w に各パラメータをあてはめ，各段階のすべてのクラスタを考慮に入れた次式で比率を計算し，最小となる段階の分類を採用する．

$\frac{1}{c} \sum_c (\frac{1}{r} \sum_r \frac{pp_{act}(c)}{pp_{rnd}(cr)})$

(r は $token$ をランダムに選択する試行の回数を意味し，c はクラスタの数を意味する)

5.2 評価テキストのクラスタ決定

評価テキストのクラスタ決定は次のように行う．評価テキストと各訓練テキストとの類似度を計算する．平均距離によって訓練セットの各クラスタとの類似度を計算する．最も類似したクラスタとの類似度が，閾値以上ならば，そのクラスタに属するものとし，そうでなければどこにも属さないものとする．

閾値は，訓練テキストのクラスタリングで決定された段階のクラスタ間類似度を使用する．パラメータの重みは，決定されたクラスタで推定されたものを使用することになる．また，どこにも属さないテキストには，訓練セット全体で計算された重みを使用する．

6 実験

前節までに説明した各手法を組み合わせて，段落分割実験を行う．

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，問題集の解答と心理実験の結果による解答の 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[1]．

基本的な実験方法は，以下のようになる．

システムは，各文間 $(point)$ ごとに，段落境界へのなり易さ (scr) を計算する．各 $point(n, n+1)$ の scr_n はパラメータに重みをかけた値の総和で計算する．

$scr_n = w_0 + \sum_{i=1}^{m} w_i \times p_{i,n}$

ただし実験 1 の場合には w_0 は 0 とする

計算された scr の上位から順に境界候補数 $\times \frac{1}{3}$ 個を出力する．出力は，システムによって段落境界と見做された位置として扱い，2 種類の解答との比較を行う．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) という評価指標を用いる．

再現率と適合率を，各評価テキストごとに求め，その平均を評価セット全体の評価値とする．実験 2 から 4 では，訓練，評価セットの組合せを換えて，3 回の cross-validation[5] を行い平均で評価する．再現率，適合率は次式で与えられる．

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

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

実験は，以下の 4 通りの条件について行う．

1. 設定したすべてのパラメータを使用し，重みは人手により付ける．
2. 設定したすべてのパラメータを使用し，重みは自動的に推定する．
3. 最適なパラメータを選択し，選択されたパラメータを使用し，重みは自動的に推定する．
4. クラスタリングを行ってから，2 の手法を組み合わせる．

それぞれの実験結果を示す．解答欄の A は問題集の解答，B は心理実験による解答を示している．また，‘類’ 欄は類似度であり，I は内積，C は cosine 距離を用いたことを表す．

解答	再現率	適合率
A	0.492361	0.243690
B	0.374638	0.198673

表 2: 実験 1 の結果

解答	訓練セット		評価セット	
	再現率	適合率	再現率	適合率
A	0.653819	0.290501	0.506250	0.228428
B	0.650000	0.291410	0.463768	0.220415

表 3: 実験 2 の結果

解答	訓練		評価	
	再現率	適合率	再現率	適合率
A	0.612847	0.269031	0.575695	0.242811
B	0.614120	0.280612	0.479167	0.228870

表 4: 実験 3 の結果

7 考察

人手による重みを使用した実験 1 と，重回帰分析によって自動的に推定した重みを使用した実験 2 では，

問題集の解答で適合率が実験 2 によって下がるものの，全体としては実験 2 で精度が向上している．労力

[1] 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

類	解	訓練セット		評価セット	
		再現率	適合率	再現率	適合率
I	A	0.859028	0.376185	0.485417	0.218018
	B	0.705903	0.309709	0.442560	0.199551
C	A	0.876389	0.385157	0.490972	0.204483
	B	0.811111	0.360690	0.324638	0.158638

表 5: 実験 4 の結果

類	解	訓練セット		評価セット	
		再現率	適合率	再現率	適合率
I	A	0.982639	0.430251	0.554861	0.244052
	B	0.877222	0.388979	0.537004	0.251982
C	A	0.702083	0.318885	0.506250	0.225680
	B	0.760857	0.349720	0.541766	0.249452

表 6: クラスタリングの上限

の軽減，客観性，柔軟性の観点から，本来パラメータへの重み付けは，何らかの自動化手法を用いる方が望ましい．これらを考慮すると，今回用いた重回帰分析の手法は，自動化の 1 手法として有効であるといえる．

また，実験 2 と，不要なパラメータを除去した実験 3 では，問題集，心理実験のいずれの解答に対しても，実験 3 によって更に精度が向上している．この結果から設定したパラメータをそのまま使用せず，データに合わせて除去することは有効であることがわかった．その除去基準として今回用いた重回帰分析の変数選択法が，段落分割において有効なパラメータ除去手法の 1 つであることが言える．

一方，クラスタリングを実験 2 に組み合わせた実験 4 では，実験 2 を上回る精度を出すことができなかった．その理由として，閾値の設定に問題があったことがあげられる．クラスタリングの各段階での評価を行い一番良かった結果を集めた上限を表 6に示す．上限では，実験 2 の結果を上回っており，閾値の設定がうまくいけば，クラスタリングの組み合せは有効であると考えられる．

8　おわりに

本稿では，複数の知識を組み合わせて用い，文章を意味段落に分割する手法について述べた．本研究では，複数知識の統合手法として，パラメータの重み付き総和を用いるアプローチをとっている．このアプローチで精度向上のために問題となる，パラメータの重み付けには，重回帰分析の手法を用いた．また，設定したパラメータからデータに合わない不要なパラメータを除去する手法として，重回帰分析の変数選択法を使用することにより，最も良い精度で分割が行えた．訓練セットのクラスタ分析の組み合わせも試みたが，閾値の設定がうまくいかなかったため，良い結果が得られなかった．最適な閾値を設定することが，今後の課題としてあげられる．

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] 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.

[3] Gerard Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Addison-Wesley, 1988.

[4] Satoshi Sekine. A New Direction for Sublanguage N.L.P. *Journal of Natural Language Processing Vol.2 No.2*, pp. 75–87, 1995.

[5] Sholom M. Weiss and Casimir Kulikowski. *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine larning, and expert systems.* Morgan Kaufmann, 1991.

[6] 大野晋, 浜西正人. 角川類語新辞典. 角川書店, 1981.

[7] 田中豊, 垂水共之. Windows 版 統計解析ハンドブック 多変量解析. 共立出版, 1995.

[8] 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. 情報処理学会研究会資料 NL109-7, pp. 47–54, 1995.

[9] 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.