Vol.2011-MPS-82 No.15
2011/3/7

分割位置を教師値としたテキストの段落分割

但　馬　康　宏[†1]

テキストの段落分割は，変化点抽出法や HMM によりテキストの流れを理解する手法など，教師なし学習のモデルを適用する方法が知られている．本研究では，段落の区切りを明示された学習用テキストデータを用いて，教師あり学習のモデルを適用することにより段落分割を行う手法を提案する．その結果，HMM のみによる分割手法よりも一般の場合において高い精度で分割できることを確認した．

A text segmentation method from a boundary marked teaching set

Yasuhiro Tajima[†1]

Text segmentation problem is usually solved via unsupervised learning model for example HMM. We propose a new method for this problem via supervised learning model which is hybrid of HMM and a text clasifier. With this method, we evaluate a news text segmentation problem, then we confirm that our method has advantage over the simple HMM method.

1. はじめに

テキストをその意味による段落に分割する問題は，文脈理解における基本的な問題であり，テキストマイニングにおいては，処理対象を絞り込むために必要な技術である．この問題に対して，従来の研究では変化点抽出による方法や HMM の教師なし学習によるモデルを適用することが多く見られた．これに対し我々は以前に，テキストの各段落ごとにその段落が属するトピックのラベルが付けられている学習データを用いて，従来手法より精度の高い分割を行う手法を提案した[3]．これは，教師あり学習のモデルを本問題に適用したものである．本研究では，教師データに付加される情報をより減らし，分割位置のみが示された学習データを用いて段落分割を行う手法を示す．その結果，教師なし学習モデルである HMM を用いた方法よりも高性能な分割が可能であることを示す．

†1 岡山県立大学 情報システム工学科
Department of systems engineering, Okayama Prefectural University

2. 教師なし学習による段落分割

テキストの段落分割問題に対して，従来の研究には大きく分けて 2 通りの手法が存在する．第一はテキストの単語の出現傾向の変化点を抽出する手法であり，Hearst[1] の研究に端を発する手法である．この場合，何をもって変化点とするかはアルゴリズム側で決定する必要があり，また調整すべきパラメータも複数になる場合も多く，段落位置を望んだものにするためには難しさが伴うことがある．

第二は，テキストにおける段落はそこで述べられるトピックを表すものと仮定し，トピックの移り変わりをモデル化して段落の区切りを発見する手法である．これは主に HMM を用いて，各状態が一つのトピックを表すとし，出力記号を単語や文字とすることにより，テキストの文字列を出力記号に当てはめ，隠れ状態として文章の意味であるトピックとすることにより段落の区切り位置を推定する手法である[2]．HMM による時系列データの解析は，音声認識などの分野では一般的であり，テキストの意味理解においても効果を発揮している．隠れ状態の遷移に関する学習は教師なし学習により可能であるが，学習時間がかかりすぎる点や隠れ状態の遷移が文章の途中の単語で起きる可能性があるなど，第一の方法と同様にアルゴリズム側で調整すべき点が多い．

3. 提案手法

3.1 教師あり学習による段落分割

以前の研究において，教師あり学習を用いた段落分割手法を提案した[3]．この手法による分割では，分割対象となるテキストデータは，複数のトピックを含むテキストデータであるとし，1 つのトピックは 1 つの段落に対応するものと仮定している．したがって，段落分割はテキストデータをトピックごとに分割することにより達成される．ここで，分割対象のテキストにどのようなトピックが存在するかは，あらかじめ分割アルゴリズムに既知であるとする．すなわち，トピックの種類の数とそれらを区別するトピック番号が利用できるとする．

したがって，学習データは段落分割の位置と，その段落が既知のどのトピックであるかを示すトピック番号が付与されている．この学習データを用いて，以下のように 1 文のトピッ

ク判定器とトピックの流れを表す HMM を構成する．

(1) 学習データから各段落，各単語ごとにナイーブベイズで用いるパラメータを抽出する．
(2) 学習データのテキストを 1 文ごとにその文が属している段落番号に置き換え，テキスト一つに対して段落番号の列を一つ作成する．
(3) 上記で作成した段落番号の列を学習データとして，HMM を機械学習にて構成する．

次に，分割対象のテキストに対するモデルの適用方法を示す．

(1) 分割対象のテキストを 1 文もしくは 1 発話ごとに分解し，各文ごとにナイーブベイズにて段落番号を推定し，テキストで一つの段落番号の列を作成する．
(2) 推定された段落番号の列を最も高い確率で出力する状態遷移系列を求める．
(3) 得られた状態遷移系列から対象テキストの段落分割を行う．

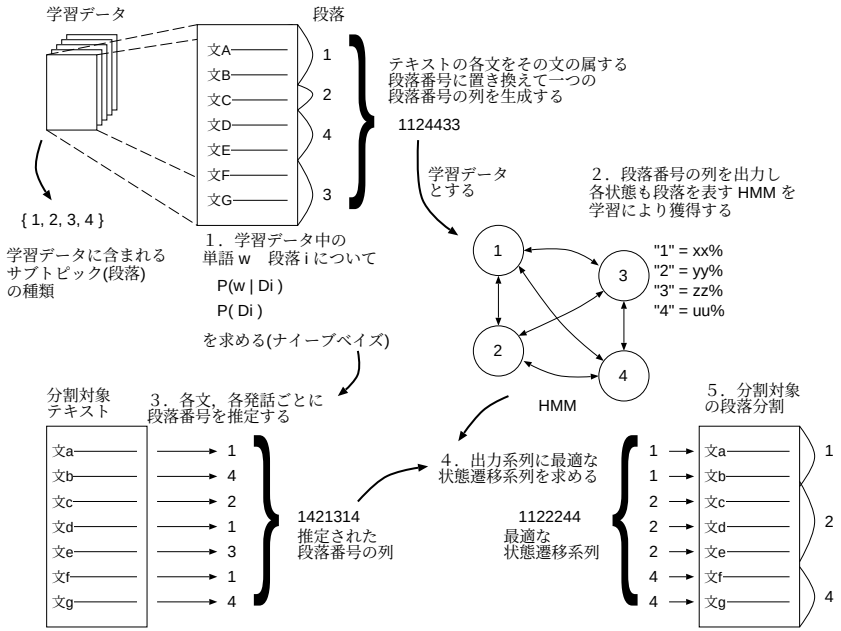

図 1 アルゴリズム 1 におけるデータと処理の流れ

この手法による段落分割方法を “アルゴリズム 1” と呼ぶ．図 1 にアルゴリズム 1 によるデータと処理の流れを示す．

3.2 分割位置を教師値とする段落分割

本研究では，学習データに付与される情報からトピック番号をなくし，段落の分割位置のみが与えられたテキストデータから 1 文の分類器とトピックの流れを表現する HMM を構成し，段落分割を行う手法を提案する．分割対象とするテキストデータは，一つの段落が一つのトピックを表す点は変わらないとし，以下の方法で学習データの各段落にトピック番号を分割アルゴリズムの側で付与することにより実現する．

(1) 学習データに含まれるすべての段落を対象にクラスタリングを行う
(2) クラスタ一つに対し一つのクラスタ番号を付与し，その番号をトピック番号として学習データに付加する
(3) アルゴリズム 1 を用いて 1 文の分類器および HMM を構成する
(4) 分割対象データに対する分割を行う

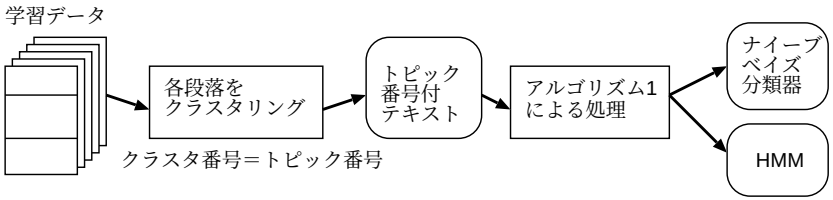

図 2 提案手法における処理の流れ

すなわち，アルゴリズム 1 では既知としていたトピックの内容をクラスタリングにより自動的に学習データに付与する方法である．図 2 に本手法による処理の流れを示す．

学習データにおける各段落は形態素解析を行い，段落内に出現する単語に対して 1，出現しない単語について 0 を割り当て，一つの段落について一つの $\{0,1\}^n$ ベクトル (段落ベクトル) を割り当てる．ここで n は学習データ全体に出現する単語の種類の数である．このベクトルを k-means を用いてクラスタリングし，本手法を適用した．段落ベクトル $v_i = (x_1, x_2, \cdots, x_n)$ とクラスタ重心 $c_j = (y_1, y_2, \cdots, y_n)$ との距離は

$$d(v_i, c_j) = |\{k \mid x_k = y_k\}|$$

Vol.2011-MPS-82 No.15
2011/3/7

表 1 データセット 1, 2 の仕様

データセット	テキスト数	1 記事の平均文数	1 記事の平均単語数
1	200	91.54	1868.32
2	200	96.69	2000.92

と定めた．すなわち，一致する要素の数である．ベクトルの定め方およびクラスタリングアルゴリズムについては，分割対象のテキストに応じた変更が分割精度の向上につながるが，本研究では上記の定義を用いた．

4. 評価実験

ウェブニュース記事に対するトピックごとの分割を行った．国内，海外，経済，エンターテイメント，スポーツ，テクノロジーの 6 つのトピックの記事を集め，以下の 2 種類のデータを作成した．

(1) Left-to-Right モデルに沿ったシナリオ．国内，海外，経済，エンターテイメント，スポーツ，テクノロジーの 6 つのトピックのうち 2 つのトピックを削除し，残りの 4 つのトピックがこの順番で出現するデータ．以後，データセット 1 と呼ぶ．

(2) すべてのトピックがランダムに出現するシナリオ．6 つのトピックの記事からランダムに 4 つの記事を選択して，一つのテキストデータとしたもの．以後，データセット 2 と呼ぶ．

データセット 1 およびデータセット 2 の内容を表 1 に示す． データセット 1,2 それぞれについて，5 分割交差実験にて分割性能を調べた．比較対象は，単語を出力記号とする HMM による分割 (HMM) とアルゴリズム 1 による分割である．アルゴリズム 1 による分割は，学習データの各段落に正しいトピック番号が付与されているものを利用して実験を行った．

提案手法は，単語を出力記号とする HMM による分割に比べ，left-to-right モデルでの性能は劣るが，トピックの移り変わりに規則性が少ないモデルでは性能向上が見られた．しかし，いずれの場合でも段落分類をあらかじめ与えるアルゴリズム 1 に比べると劣っている．

5. まとめ

テキストをトピックごとの段落に分割する問題について，従来の手法に比べ自然な設定における教師データを用いて分割を行う手法を示し，評価実験を行った．従来手法である，単語を出力記号とする HMM に比べトピック出現の規則性の少ない状況では良い性能が得られたが，トピックごとにあらかじめ分類された教師データを用いて分割を行う手法に比べると性能が劣ることが確認された．今後の課題として，クラスタリングアルゴリズムの改善，1 文の分類器の構成方法の改善などが挙げられる．

表 2 データセット 1,2 での分割性能（前後 1 文許容）

データセット 1	data1	data2	data3	data4	data5	平均
本手法（精度）	0.2925	0.3172	0.2889	0.3117	0.2414	0.2903
本手法（再現率）	0.2889	0.3528	0.3278	0.3944	0.3111	0.3350
本手法（F 値）	0.2907	0.3340	0.3071	0.3482	0.2719	0.3111
HMM（精度）	0.6118	0.4047	0.7124	0.6117	0.6398	0.5960
HMM（再現率）	0.6708	0.3069	0.7556	0.6652	0.6917	0.6181
HMM（F 値）	0.6400	0.3491	0.7334	0.6373	0.6647	0.6069
アルゴリズム 1（精度）	0.6394	0.5911	0.5575	0.6153	0.5694	0.5946
アルゴリズム 1（再現率）	0.6333	0.5389	0.5417	0.6194	0.6056	0.5878
アルゴリズム 1（F 値）	0.6364	0.5638	0.5495	0.6174	0.5869	0.5911
データセット 2	data1	data2	data3	data4	data5	平均
本手法（精度）	0.2834	0.3551	0.2939	0.3759	0.3847	0.3386
本手法（再現率）	0.4278	0.4194	0.3528	0.4417	0.4389	0.4161
本手法（F 値）	0.3410	0.3846	0.3207	0.4061	0.4100	0.3734
HMM（精度）	0.1674	0.1306	0.1128	0.1189	0.1203	0.1300
HMM（再現率）	0.9569	0.9514	0.9611	0.9431	0.9333	0.9491
HMM（F 値）	0.2849	0.2296	0.2019	0.2111	0.2131	0.2287
アルゴリズム 1（精度）	0.3661	0.3422	0.3069	0.4872	0.4094	0.3824
アルゴリズム 1（再現率）	0.7306	0.6861	0.5861	0.7417	0.6667	0.6822
アルゴリズム 1（F 値）	0.4878	0.4567	0.4029	0.5881	0.5073	0.4900

謝辞 本研究の一部は科学研究費補助金 (No.21700007) の補助を受けている．

参考文献

1) Hearst, M. A.: Texttiling: segmentaing text into multi-paragraph subtopic passages, Computaional Linguistics, Vol. 23, pp.33-64 (1997)
2) Ostendorf, M., Digalakis, V. V. and Kimball, O. A.: From HMM's to segment models: a unified view of stochastic modeling for speech recognition, IEEE Transactions on speech and audio processing, Vol. 4, No.5, pp.360–378 (1996)
3) 但馬康宏，北出大蔵，中林智，藤本浩司，小谷善行：HMM とテキスト分類器による対話の段落分割，情報処理学会論文誌 数理モデル化と応用，vol.2, no.2, pp.70–79 (2009)