2008－MPS－71 (23)
2008／9／19

HMM とテキスト分類器による対話の段落分割

但馬 康宏†，北出 大蔵‡，中野 未知子‡，中林 智††，藤本 浩司‡‡，小谷 善行†

†東京農工大学 工学部 情報工学科　‡トランス・コスモス株式会社
††株式会社金融エンジニアリング・グループ
‡‡テンソル・コンサルティング株式会社

あらまし テキストを段落に分割する問題に対して，本論文ではシナリオに基づいた分割を行う手法を提案する．本手法では，正確に段落分割された学習データから一つの文が属する段落を推定するナイーブベイズモデルおよび段落番号の列を出力とする HMM により学習データをモデル化する．分割対象テキストは，一文ごとにナイーブベイズにて属する段落番号を推定され，その段落番号の列の HMM における最適な状態遷移系列から段落分割を行う．本手法は特に，対話文などの間投詞や特徴的な単語の少ないテキストデータに対しても高い分割性能を得ることができる．評価実験として，実際の対話，およびウェブのニュース記事に対して段落分割を行い，本手法の有効性を確かめた．

A dialogue segmentation method via HMM and a text classifier

Yasuhiro Tajima†, Daizo Kitade‡, Michiko Nakano‡, Tomo Nakabayashi††,
Koji Fujimoto‡‡, Yoshiyuki Kotani†

†Tokyo University of Agriculture and Technology　‡Transcosmos, Inc.
††Financial Engineering Group, Inc.　‡‡Tnesor Consulting, Inc.

Abstract We present a new method for text segmentation via HMM and a text classifier. Our method has two stages to segment the target text. Every sentence or utterance is classified into a topic in the scenario at the firest stage. Then, in the second stage, the target text is segmented by an HMM which is a model of topic strings generated at the firest stage. Our HMM ourputs a string of topic numbers and it absorbs the miss classification in the first stage. For evaluation, we apply our method to a dialogue segmentation task and a news text segmentation task. We can confirm that our method performs better than the ordinal HMM segmentation method.

1 はじめに

談話などのテキストからその話題を抽出することは，談話理解における重要な問題のひとつである．特にテキストをその段落ごとに分割する問題は，テキスト分割（テキストセグメンテーション）と呼ばれる問題である．テキストを段落へ分割することにより，段落が扱うサブトピックに対するキーワード抽出や段落の要約などの処理に有益である．

テキスト分割の従来手法として，Hearst[3] による手法がよく知られている．これはある窓幅内のテキストに対する特徴ベクトルの変化点を抽出することにより分割位置を決定する手法であるが，変

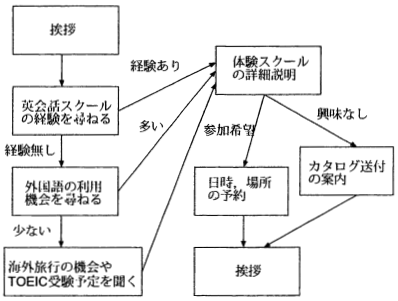

図 1: 対話のシナリオ

化点の閾値や窓幅などパラメータ調整が不可欠である．この手法の発展研究も多い [4] [5].

また，HMM によるテキスト分割も行われている [1][2]．すなわち，一つの状態はテキスト内の段落すなわちサブトピックを表すとし，テキストに出現する単語を出力記号とする HMM を機械学習により獲得して分割を行う手法である．この手法の改良研究も行われている [6].

本論文では，テキストに対するシナリオを仮定し，以下のような段落分割手法を提案する．まず，テキストを一文や一発話ごとに分割し，テキスト分類アルゴリズムによってそれらの文や発話があらかじめ所属する段落を推定する．その後，推定された段落番号を出力記号とする HMM によって分割位置を決定する二段階による方法である．

本手法の効果確認としてウェブニュース記事に対する段落分割と実際の対話記録に対する段落分割を行い，従来の HMM による手法と比較実験を行った．その結果，対話に対する段落分割において，従来手法よりも良い結果を得ることができた．また，ニュース記事に対する分割でも従来手法と劣らない性能であることを確認した．

2 提案手法

2.1 シナリオと分割対象テキスト

本論文で分割対象とするテキストは，段落が扱うサブトピックを図 1 に示すようなチャートに基づいているものとし，このようなチャートをシナリオと呼ぶ．

本論文では，シナリオは段落の種類を限定し，正解データを作成するために用いる．分割対象のテキストを以下のように定める．

- シナリオから，段落で扱うサブトピックが明らかである．
- 学習用の分割がなされたデータがあり，サブトピックに応じて段落番号が付与されている．すなわち，学習用データから同一の段落番号を持つ段落を抜き出すと，一つのサブトピックに対応する段落を集めることができる．

図 2 に本手法でのデータと処理の流れを示す．本手法では，シナリオおよび学習データを用いて以下の手順で分割モデルを構成し，そのモデルを用いて段落分割を行う．まず，モデルの構成手順を述べる．

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．T を一つのテキストとし，i 番目の文を u_i と表し，m_i 個の単語 $w_1^{(i)}, w_2^{(i)}, \cdots, w_{m_i}^{(i)}$ の連接であるとする．与えられた文が段落番号 k の段落に含まれる事象を D_k と表すと，文 u_i が段落番号 k に含まれる確率は，$P(D_k|u_i)$ であり，

$$P(D_k|u_i) = \frac{P(u_i|D_k)P(D_k)}{P(u_i)}$$

として求められる．ここでさらに，

$$P(u_i|D_k) = P(w_1^{(i)}|D_k) \cdots P(w_{m_i}^{(i)}|D_k)$$

と近似し，単語 w について，

$$P(w|D_k) = \frac{\sum_{u \in H_k} c(w,u)}{\sum_{u \in H_k} \sum_{v \in W_u} c(v,u)}$$

とする．ただし，H_k は学習データにおいて段落番号 k に属するすべての文の集合であり，W_u は文 u に現れるすべての単語の集合，$c(w,u)$ は文 u における単語 w の出現回数である．

次に HMM を構成する．学習データの文 u_i に付けられた段落番号を $o(u_i)$ とする．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．それら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．ここで HMM の出力記号も隠れ状態もそれぞれが段落番号に対応している．

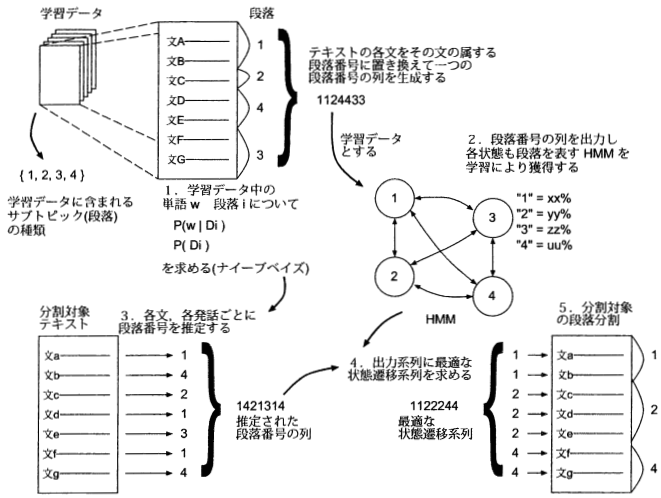

図 2: 本手法でのデータと処理の流れ

分割対象のテキスト $T = u_1 u_2 \cdots u_n$ から，前節のナイーブベイズにて 1 文ごとに段落番号の推定を行い，段落番号の列 $O = o(u_1)o(u_2)\cdots o(u_n)$ を得た後，O を出力する最適な状態遷移系列 $q_1 q_2 \cdots q_n$ を求める．ここで，各 q_i $(i = 1, 2, \cdots, n)$ は HMM の状態である．最終的に文 u_i は段落 q_i に属すると推定され，段落番号の変化点，すなわち，$q_j \neq q_{j+1}$ となる位置で段落分割を行う．

3　評価実験

3.1　対話に対する段落分割

ある架空の英会話教室への勧誘を目的とした対話を設定し，以下の条件で実際の対話を行い，その書き起こし記録に対して本手法を適用し評価を行った．被験者は，勧誘オペレータ 4 名，被勧誘者 62 名である．一つの対話あたりの平均発話数は 167.06 であり，一つの発話に含まれる平均単語数は 10.58 であった．図 1 に示すシナリオを元に正解データを作成して，全 62 対話を 5 分割交差検定にて行った．テキストの形態素解析には mecab 0.91 を使用した．

比較対象として，以下の 2 種類の分割方法も同時に行った．

HMM： 従来から知られている単語を出力記号とした HMM による分割結果．一つの発話の中で最も多くの単語が対応づけられた状態をその発話が属する段落として，段落境界が常に発話と発話の間となるようにした．さらに，計算精度不足を補うため，学習中にリスケーリングを行った．学習アルゴリズムは Baum-Welch である．

bayes： ナイーブベイズで 1 発話の属する段落を推定した結果をそのまま採用した場合の結果．

表 1 上段に結果を示す．正しい分割位置から前後 1 発話以内の分割を正解とした．最下段の “発話割合” は，テキスト中の全発話のうち正しい段落番号が付けられた発話の割合である．

いずれも本手法が最も F 値が高い結果となった．精度は本手法が最も高く，再現率はナイーブベイズが最も高い値となる．これは，発話ごとに段落番号が変化するため，段落分割点を過剰に生成しているためである．

3.2　ニュースのトピック分割

本手法と従来型の HMM による段落分割およびナイーブベイズでの分割の性能を従来法で評価対

表 1: 分割性能（前後 1 発話，1 文許容）

対話の分割	data1	data2	data3	data4	data5	平均
本手法（精度）	0.3510	0.4597	0.4781	0.4723	0.4325	0.4391
本手法（再現率）	0.5944	0.5016	0.6746	0.6863	0.5799	0.6081
本手法（F 値）	0.4414	0.4797	0.5596	0.5596	0.4954	0.5100
HMM（精度）	0.1504	0.2075	0.1948	0.2045	0.1394	0.1791
HMM（再現率）	0.9032	0.8666	0.9519	0.8709	0.8420	0.8859
HMM（F 値）	0.2579	0.3348	0.3234	0.3313	0.2392	0.2980
bayes（精度）	0.1194	0.1228	0.1702	0.1599	0.1183	0.1381
bayes（再現率）	0.9667	0.9429	0.9603	0.9872	0.9633	0.9645
bayes（F 値）	0.2125	0.2173	0.2889	0.2753	0.2108	0.2416
本手法（発話割合）	0.5660	0.6960	0.6218	0.6365	0.7013	0.6451
HMM（発話割合）	0.5496	0.6403	0.5818	0.5433	0.5472	0.5716
bayes（発話割合）	0.5113	0.5226	0.5244	0.5175	0.5532	0.5261
ニュースの分割	data1	data2	data3	data4	data5	平均
本手法（精度）	0.6750	0.6139	0.5431	0.6167	0.6333	0.6164
本手法（再現率）	0.6444	0.5722	0.5556	0.6111	0.6389	0.6044
本手法（F 値）	0.6594	0.5923	0.5492	0.6139	0.6361	0.6104
HMM（精度）	0.6294	0.4033	0.7524	0.5978	0.6189	0.6004
HMM（再現率）	0.6792	0.3069	0.7889	0.6458	0.6667	0.6175
HMM（F 値）	0.6534	0.3486	0.7702	0.6209	0.6419	0.6088
bayes（精度）	0.0952	0.0824	0.0805	0.0881	0.0907	0.0874
bayes（再現率）	0.9944	0.9889	0.9444	0.9722	0.9722	0.9744
bayes（F 値）	0.1737	0.1522	0.1483	0.1616	0.1659	0.1604
本手法（発話割合）	0.7354	0.7374	0.7110	0.7435	0.7678	0.7390
HMM（発話割合）	0.8307	0.5392	0.8290	0.7722	0.8006	0.7543
bayes（発話割合）	0.5409	0.5311	0.5182	0.5512	0.5543	0.5391

象として多く取り上げられているニュース記事のトピック分割において比較した．

ウェブのニュース記事より，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，上記のトピックの順にランダムに 4 つの記事をつなげてテキストとした．一つの記事の平均文数は 91.54，平均単語数は 1868.32 であった．

表 1 下段に結果を示す．本手法と従来の HMM による手法との差が対話の場合よりも少なくなっている．これは，ニュースのテキストは，各トピックごとに特徴的な単語が明確であり，従来手法での分割が効果的であることを示している．

参考文献

[1] 越仲 孝文，奥村 明俊，磯谷 亮輔，HMM の変分ベイズ学習によるテキストセグメンテーション及びその映像インデキシングへの応用，信学論 J89-D-9, pp.2113-2122, 2006.

[2] 今井 亨，R. Schwartz, 小林 彰夫，安藤 彰男，話題混合モデルによる放送ニュースからの話題抽出，信学論 J81-D-II-9, pp.1955-1964, 1998.

[3] M. A. Hearst, Texttiling: segmentaing text into multi-paragraph subtopic passages, Computaional Linguistics, 23, pp.33-64, 1997.

[4] 別所 克人，クラスタ内変動最小基準に基づくテキストセグメンテーション，情報処理学会論文誌，47(3), pp.957-967, 2006.

[5] D. Beeferman, A. Berger, and J. Lafferty, Statistical models for text segmentation, Machine Learning, 34(1-3), pp.177-210, 1999.

[6] M. Ostendorf, V. V. Digalakis, and O. A. Kimball, From HMM's to segment models: a unified view of stochastic modeling for speech recognition, IEEE Transactions on speech and audio processing, 4(5), pp.360–378, 1996.