2008－NL－186（15）
2008 / 7 / 18

メールの文章における段落間の接続の強さの推定

西村　涼† 大田 康人†† 渡辺 靖彦† 村田 真樹††† 岡田 至弘†

† 龍谷大学大学院 理工学研究科 情報メディア学専攻
†† TIS 株式会社
††† 情報通信研究機構
E-mail: † r_nishimura@afc.ryukoku.ac.jp, †† is302403@tis.co.jp, † {watanabe,okada}@rins.ryukoku.ac.jp,
††† murata@nict.go.jp

あらまし メールの文章は他の文書なら改段落しない場合でも、「見やすさ」、「読みやすさ」を重視して改段落をする場合がある。こうした過剰で不要な段落わけは、メールの機械処理にとって問題である。そこで、メールの文章における段落間の接続の強さを機械学習によって推定する方法を提案し、過剰な段落わけを検出できることを示す。
キーワード メール、情報抽出、不要な段落わけ

Estimation of Connectivity between Paragraphs in Mails

Ryo NISHIMURA†, Yasuhito OOTA††, Yasuhiko WATANABE†, Masaki MURATA†††, and Yoshihiro OKADA†

† Ryukoku University Department of Media Informatics
†† TIS Inc.
††† National Institute of Information and Communications Technology
E-mail: † r_nishimura@afc.ryukoku.ac.jp, †† is302403@tis.co.jp, † {watanabe,okada}@rins.ryukoku.ac.jp,
††† murata@nict.go.jp

Abstract In order to improve the readability, we often segment mail text into smaller paragraphs than necessary. However, this oversegmentation is a problem of e-mail text processing. In this paper, we proposed an estimation method of connectivity between paragraphs in mails using machiene learning technieques, and showed that paragraphs which should be one paragraph can be found by detecting strong connectivity.
Key words mail, information extraction, oversegmentation

1. はじめに

メールの文章では、他の文章なら段落わけしない場合でも、「見やすい」「読みやすい」文章にしようとして段落わけをしている場合が多い。図 1 に Vine Users ML(注1)というメーリングリストに実際に投稿されたメールの文章を示す。この例ではメールの文章は 3 つの段落から構成されているが、その内容は、図 2 に示すように 1 つの段落で表現されていてもおかしくない。これらの例からメールでは意味的なつながりよりも「読みやすさ」や「見やすさ」を重視して段落わけをする傾向があると言える。

細かく段落わけすることは、メールを読むという観点から重要である。しかし、メールに対して機械処理を行う場合には、不利になることもある [1] [2] [3]。例えば、奥村ら [3] はメールの要点を抽出するために重要な段落を推定している。そこでは、メールで段落わけが行われている個所を、一般的な文章の場合と同様に、空行や記号などの形式的な手がかりを利用して決定している。しかし、メールの段落わけは過剰で不要なものも多く、他の文書に対しては有効な処理もメールの文章に対しては期待される精度では結果が得られないおそれがある。そこで本研究では、メールの文章における段落間の接続の強さを推定する方法について提案する。

従来から文章を意味的なまとまりに分割するテキストセグメンテーションはよく研究されており、文の構造解析や自動要約などの手がかりとして利用することや未整形のテキスト文書を人間が閲覧しやすいように整形することに有効であると考えら

(注1)：http://vinelinux.org/ml.html

今日はmuleを使っていておかしな所を見つけたので報告します。

Muleをウィンド上で立ちあげた状態で、cannaを辞書に登録しようと思い、M-x canna-tourokuをすると、Segmentation faultしてしまいました。

そこで、muleをkterm上からmule -nwで立ち上げM-x canna-tourokuを実行してみると「辞書を作りますか？」という感じのメッセージが出てきて、そのまま作業を続ける事が出来ました。

図 1　見やすさ・読みやすさを意識して細かく段落わけされているメールの例

今日はmuleを使っていておかしな所を見つけたので報告します。
Muleをウィンド上で立ちあげた状態で、cannaを辞書に登録しようと思い、M-x canna-tourokuをすると、Segmentation faultしてしまいました。
そこで、muleをkterm上からmule -nwで立ち上げM-x canna-tourokuを実行してみると「辞書を作りますか？」という感じのメッセージが出てきて、そのまま作業を続ける事が出来ました。

図 2　図 1 の文章から過剰で不要な段落わけをとりのぞいた文章の例

れている [4] [5] [6] [7] [8]。これらの手法は、段落わけがなされていないテキストを意味的なまとまりに応じて分割することを目的としており、単語の結束性や接続表現などの表層的な情報を利用している。一方、人間によって過剰に段落わけが行われているテキストに対しては単語の結束性や接続表現などの表層的な情報だけでなく、段落わけの情報も利用した方が良いこともある。そこで、本研究では、人間が「読みやすさ」や「見やすさ」を重視して行った段落わけの情報を利用し、不要な段落わけを取り除くことで、適切な段落わけのみを残すというアプローチをとる。まず、2. でメールの段落わけの特徴を述べる。次に、3. でそれらの特徴を反映して段落間の接続の強さを推定する手法を説明し、問題点を考察する。最後に、4. で機械学習による手法を提案し、3. での精度が改善されたことを述べる。

2. メールの文章における段落間の接続の強さと段落わけの調査

メーリングリストには質問と回答のメールが繰り返し投稿されるものがある。たとえば、Vine Linux に関心のある人たちが情報を交換しているメーリングリスト（Vine Users ML）では質問と回答のメールがさかんに投稿されている。

メールの文章における段落わけの調査には、Vine Users ML に投稿された質問メール（返信のあるもの）を用いた。ここで、質問メールとは、Q&A 形式のやりとりを頻繁に行っているメーリングリストに最初に投稿されるメールのことである。人間がメールの文章における段落間の接続の強さを推定する際に利用する要因と考えられるものを以下に示す。

UNIX USER 4月号についている。Vine Linuxをいれてみました。
Xも無事動き、Netscapeもいれてppxpで問題なくつながりました。

しかし、IP Masquradeがうまく動作しません。
以前Turbo Linux 3.0を使っていたとき、同じように設定して動いていました。

図 3　「しかし」を文頭にもつ文ではじまる段落の例

この度、Slackware3.6から乗り換えました。
カッコ良いですねー、Vine Linux！
いたれり、つくせりのツールも最高です、大好きです。
もちろん、今、Vmailで送信しています。

ところで、私は/etc/inittabを編集してランレベル5で起動してxdmを利用しています。
Vineのウィンドウ画面はメチャカッコ良いのですが、xdmのログインマネージャでは基本が白黒で、イマイチです。

図 4　「ところで」を文頭にもつ文ではじまる段落の例

2.1　文頭の接続詞

接続詞は文頭において、先行する文とのつながりを示す役割を果たしている。このため、文頭に接続詞がある文とその前文はつながっていると考えることができる。例えば、「しかし」などの接続詞が文頭にある文ではじまる段落は、直前の段落とのつながりが強い。図 3 の例では、第 2 段落は「しかし」を文頭に持つ文からはじまっている。この第 2 段落と第 1 段落は内容的なつながりが強く、図 3 のように段落分けをせず、1 つにまとめてもよい例である。一方、話題を転換する時に使われる接続詞「ところで」「さて」を文頭にもつ文ではじまる段落は、直前の段落とのつながりが弱い。図 4 の例では、第 2 段落は「ところで」を文頭に持つ文から始まっている。この第 2 段落と第 1 段落は内容的なつながりが弱く、図 4 のように段落わけをするのがのぞましい例である。

2.2　指　示　語

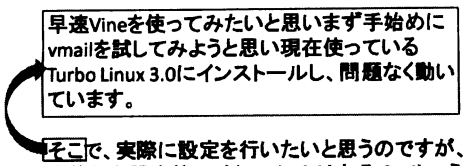
「この」などの指示語が文頭にある段落は、直前の段落とのつながりが強い。また、「以上のように」などの表現が文中にある段落も、直前の段落とのつながりが強い。一方、「以下のように」などの表現を含む文でおわる段落は、直後の段落とのつながりが強い。図 5 の例では、第 2 段落は指示語「そこ」を文頭にもつ文からはじまっている。この第 2 段落と第 1 段落は内容的に強いつながりをもっていて、図 5 のように段落わけをせず、1 つにまとめてもよい例である。

2.3　挨拶の表現

段落の最後の文が図 6 に示すような挨拶の表現を含む文である場合、直後の段落とのつながりは弱い。図 7 のように段落わけをするのがのぞましい例である。

2.4　未 定 義 語

形態素解析用の辞書で定義されていない未定義語は、専門用



そこで、実際に設定を行いたいと思うのですが、具体的な設定等のドキュメントはあるのでしょうか？

図 5　文頭に指示語をもつ文ではじまる段落の例

- はじめまして
- 初めまして
- こんにちは
- こんにちわ
- こんばんは
- こんばんわ
- といいます
- と言います
- もうします
- 申します
- このたび
- この度

図 6　挨拶の表現

はじめましてLinux初心者のXXXといいます。

Vine Linux 1.0を使用していますが、リブートしても、/tmp下のファイルが削除されません。

図 7　挨拶の表現を含む文でおわる段落の例（ただし、XXX の部分は人名）

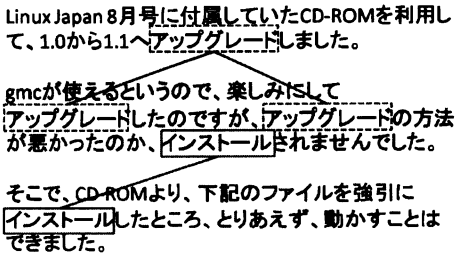


図 8　段落の最後の文に含まれる未定義語が直後の段落の最初の文にも含まれている文章の例

語であることが多い。そして専門用語は、文章で重要な役割をはたしていることが多い。段落の最後の文に含まれる未定義語が、直後の段落の最初の文に含まれている場合、それらの段落間のつながりは強い。図 8 の例では、第 1 段落の最後の文と第 2 段落の最初の文に「アップグレード」、第 2 段落の最後の文と第 1 段落の最初の文に「インストール」という未定義語がそれぞれ用いられている。これらの段落は内容的に強いつながりをもっていて、図 8 のように段落わけせず、1 つにまとめてもよい例である。

3. 手がかり表現によるメールの段落間の接続の強さの推定

われわれは以前に 2. で述べた 4 種類の手がかり表現を用いて、段落間の接続の強さを推定する方法を提案した [9]。以下にその概要を示す。

3.1 段落間の接続の強さを推定する手法の概要

入力には判定したい段落わけの直前直後の段落を用いた。また、文の形態素解析には Juman [10] を用いた。

Step1　[挨拶の表現による推定]

- また
- なおかつ
- で、
- まして
- したがって
- こうして
- すると
- じゃあ
- とすれば
- ついては
- しかし
- だけど
- だが
- でも
- 反対に
- もしくは
- 言わば
- つまり
- 要するに
- ただ
- そもそも
- なぜなら
- 続いて
- ゆえに
- そして
- しかも
- が
- だから
- よって
- かくして
- では
- そしたら
- ならば
- 次に
- けれど
- けど
- ところが
- 一方
- 逆に
- すなわち
- 言い換えれば
- 結局のところ
- 例えば
- ちなみに
- どうして
- だって
- 同様に
- あるいは

図 9　文頭にあって、段落間の接続が強いことを示す接続詞

段落わけの直前の段落の最後の文が図 6 に示す挨拶の表現を含む文である場合、その段落と直後の段落との接続は弱いと推定し、処理を終了する。

Step2　[文頭の接続詞による推定]

段落わけの直後の段落の最初の文が図 9 に示す接続詞を文頭にもつ場合、その段落と直前の段落との接続は強いと推定し、処理を終了する。逆に、段落わけの直後の段落の最初の文が、接続詞「ところで」「さて」を文頭にもつ場合、その段落と直前の段落との接続は弱いと推定し、処理を終了する。

Step3　[指示語による推定]

段落わけの直後の段落の最初の文が図 10(a) に示す表現を文頭に持つ場合あるいは図 10(b) に示す表現を文中に含む場合、その段落と直前の段落との接続は強いと推定し、処理を終了する。また、段落わけの直前の段落の最後の文が図 10(c) に示す表現を文中に含む場合、その段落と直後の段落との接続は強いと推定し、処理を終了する。

Step4　[未定義語による推定]

段落わけの直前の段落の最後の文に含まれる未定義語が、直後の段落の最初の文にも含まれている場合、それらの段落間の接続は強いと推定し、処理を終了する。

Step5　[Step1 ～ Step4 で推定できない場合]

Step1 ～ Step4 で段落わけの直前の段落の最後の文と直後の段落の最初の文との接続の強さが推定できない場合、それらの段落間の接続は弱いと推定し、処理を終了する。

3.2 実験結果と評価

3.1 での手法を実装し評価することで表層的な手がかり表現を利用した手法の有効性を確認する。Vine Users ML に投稿された質問メール（返信のあるもの）300 通を用いて以下の 2 つの実験を行った。

- 質問メールから抽出された重要文 [1] の直前直後にある段落わけにおける段落間の接続の強さの推定
- メールの文章で行われたすべての段落わけにおける段落間の接続の強さの推定

このとき、段落間の接続が強いと推定された段落わけは不要な

- これ
- それ
- あれ
- この
- その
- あの
- ここ
- そこ
- あそこ

(a) 段落の最初の文の文頭にあり、
直前の段落との接続が強いことを示す指示語

- 以上のように
- 以上の様に
- 以上のような
- 以上の様な
- 上記
- 前述
- 上述
- 前記

(b) 段落の最初の文に含まれていて、
直前の段落との接続が強いことを示す表現

- 以下のように
- 以下の様に
- 以下のような
- 以下の様な
- 下記
- 後述

(c) 段落の最後の文に含まれていて、
直後の段落との接続が強いことを示す表現

図 10 段落間の接続が強いことを示す指示語と表現

段落わけと判定する。

3.2.1 メールにおける段落間の接続の強さの推定

これまでにわれわれは、メーリングリストに投稿されたメールから重要文を抽出し、ユーザの質問に答えるための知識として利用できることを明らかにした [2]。しかし、重要文の直前直後で行われた段落わけを意味的なまとまりとして用いると、不要な段落わけのため、問い返しの知識の抽出に失敗することがあった。このため、メーリングリストに投稿されたメールを知識として利用するには、重要文の直前直後にある段落わけにおける段落間の接続の強さを推定し、不要な段落わけをとりのぞくことは重要である。そこで、Vine Users ML に投稿された質問メール（返信のあるもの）を利用して、それらの重要文の直前直後で段落わけが行われている場合の段落間の接続の強さを推定し、不要な段落わけを取り除く実験を行った。

実験に利用した 300 通のメールの重要文の直前直後には 176 個の不要な段落わけがあった。前節で述べた Step1 ～ Step5 の処理を適用した結果、表 1 に示すように適合率 91.7%、再現率 43.8%で重要文の直前直後で行われた不要な段落わけを判定することができた。不要な段落わけの判定にどの手がかり表現が役立ったかを表 2 に示す。このとき、Y は接続が弱い（段落が必要）ということを示し、N は接続が強い（段落が不必要）ということを示している。これらから、接続が弱い段落わけを判定する場合には、Step1 の挨拶表現による判定が適合率 91.7%と良く、接続が強い段落わけを判定するのには、Step4 の未定義語による判定が適合率 95.6%と良いことがわかる。また、未定義語が有効である場合が多いのは、技術的な情報をやり取りすることが目的のメールを対象としているからである。より一般的なメールを対象とする場合は、名詞などに置き換えて処理することが考えられる。Step3 の指示語による判定は適合率 80.0%と他のルールに比べて悪かった。

次に、重要文の前後に限定せず、メールに含まれる段落わけすべてを対象にして、そこでの段落わけが不要かどうか判定を行った。実験には、重要文の前後の段落わけの判定の実験に利用した Vine Users ML に投稿された質問メール（返信のあるもの）300 通を利用した。この 300 通のメールには 1723 個の段落わけがあり、そのうち 1166 個は不要な段落わけであった。この 1723 個の段落わけに対して前節で述べた Step1 ～ Step5 の処理を適用した結果、表 1 に示すように、適合率 92.7%、再現率 33.9%で不要な段落わけを判定することができた。不要な段落わけの判定にどの手がかり表現が役立ったかを表 3 に示す。挨拶表現が重要文の前後の段落わけを判定する時よりも有効ではなかった。これは、挨拶表現の直後が重要文であれば明確に段落間の接続は弱いと判定できるが、重要文でない場合は接続が弱いとは判定できない場合があるからである。

表 1 手がかり表現を利用した不要な段落わけの判定精度

	重要文の前後	メール全体
不要な段落わけ	176	1166
不要な段落わけを正しく判定	77	395
不要な段落わけと誤って判定	7	31
適合率	91.7%	92.7%
再現率	43.8%	33.9%
F 値	0.593	0.496

表 2 重要文の前後の段落わけの判定に利用した表層情報のうちわけ

(Answer/System)	Y/Y	Y/N	N/Y	N/N
Step1:[挨拶の表現]	33	0	3	0
Step2:[文頭の接続詞]	2	0	1	14
Step3:[指示語]	0	5	0	20
Step4:[未定義語]	0	2	0	43

表 3 すべての段落わけの判定に利用した表層情報のうちわけ

(Answer/System)	Y/Y	Y/N	N/Y	N/N
Step1:[挨拶の表現]	80	0	11	0
Step2:[文頭の接続詞]	5	7	1	72
Step3:[指示語]	0	15	0	83
Step4:[未定義語]	0	9	0	240

4. 機械学習法によりメールの段落間の接続の強さを推定する手法

メールの文章を調査した結果、表層的な情報で段落間の結びつきを推定できることがわかった。このような調査に基づき、3. でわれわれは表層的な情報を利用して段落間の接続の強さを推定する方法を提案した。しかし、表層的な情報に基づいたルールを段階的に適用していく方法では、それぞれのメールごとにチューニングする必要があり、汎用性は低い。また、表層的な情報に基づく抽出のルールをどのような順序で適用すべきかという問題もある。そこで、メールの種類によらず様々な表層的な情報を利用して推定することができる方法として機械学習による推定方法を提案する。

4.1 問題設定

段落間の接続の強さを推定するというタスクは、メールでなされている段落わけが必要か不必要かという 2 値分類を行えば達成できる。すなわち、段落わけが必要な場合は段落間の接続が弱いと推定し、段落わけが不必要な場合は段落間の接続が強

いと推定する。そこで、段落わけが必要か不必要かを、サポートベクトルマシン（SVM）と最大エントロピー法（MEM）を利用して求めることとした。

実験データには、3. で実験に利用した 300 通のメールを用いた。この 300 通のメールには 3605 文が含まれ、1723 個の段落わけがあり、そのうち不要なものは 1166 個であった。また、重要文の直前直後には 282 個の段落わけがあり、そのうち不要なものは 176 個であった。段落わけの直後の文を対象に直前の段落わけが必要か不必要かをタグ付けした。

4.2 素　　性

機械学習の素性には、表 4 に示すものを利用した。これらの素性は 2. の調査結果を反映しているものもある。

「s1」は、段落間の接続の強さを推定するには段落わけの直後の文頭の表現が有効である、と仮定した素性である。「s2」と「s6」はメール独自の表層的な情報を素性としたものであり、新聞記事のようなテキストには有効でないと考えられる。「s3」「s4」「s7」は、新聞記事を対象にした処理でも利用される素性である [6]。「s5」は単語の結束性を反映した素性である。ただし、名詞ではなく未定義語を対象にしているのは、専門的な内容がやりとりされるメールにおいて専門用語として用いられている可能性が高い未定義語の結びつきを考慮したかったためである。「s8」「s9」「s10」「s11」「s12」の素性を取り出すための形態素解析には、Juman [10] を用いた。このとき「s9」「s10」「s11」「s12」の素性は、段落わけの直後の文とその前後の文から取り出した。これは、段落間の接続の強さを推定するためには、段落わけの直後の文の表層的な情報だけではなく、その前後の文の表層的な情報も利用するのがのぞましいと考えたからである。

4.3 段落間の接続の強さの判定手順

質問メールを入力とし、段落間の接続の強さを以下の手順で推定する。

Step 1 [文分割]
質問メールを 1 行 1 文の単位に分割する。

Step 2 [段落の判定]
質問メールで実際に行われている段落わけを、空行の情報をもとに判定する。

Step 3 [素性の付与]
機械学習で利用する素性を付与する。

Step 4 [段落の接続の強さの推定]
SVM と MEM を用いて段落の接続の強さを推定する。

4.4 実験と考察

SVM と MEM による段落間の接続の強さの推定の実験として以下の 2 つを行う。

実験 1 メールの重要文の直前直後にある段落わけにおける段落間の接続の強さの推定

実験 2 メールの文章に含まれるすべての段落わけにおける段落間の接続の強さの推定

実験はすべて 10 分割クロスバリデーションで行うが、テストデータに利用するメールから取り出した段落わけの情報は訓練データには用いない。本実験では、SVM には TinySVM [11] の多項式カーネルを利用し、オプションは d=1 c=1 とした。また、MEM には maxent [12] を用いた。

表 4　利用する素性の種類

s1	段落わけの直後の文の文頭の 1〜10 文字（例：「し」「しか」「しかし」）
s2	段落わけの直後の文の「挨拶表現」の有無とその種類（例：「はじめまして」、図 6）
s3	段落わけの直後の文に「接続表現」の有無とその種類（例：「また」「したがって」「こうして」、図 9）
s4	段落わけの直後の文に「指示語」の有無とその種類（例：「以上のように」「上記」、図 10）
s5	段落わけの直前の文と直後の文における同一の未定義語の有無とその種類
s6	段落わけの直前と直後の文の返信のメールでの引用の有無
s7	段落わけの直前の文の主語が段落わけの直後の文に出現している場合、あるいは段落わけの直後の文の主語が段落わけの直前の文に出現している場合における主語の出現の有無とその種類
s8	段落わけの直後の文のすべての形態素の unigram
s9	段落わけの直後の文とその文の前後 1 文のすべての形態素 unigram
s10	段落わけの直後の文とその文の前後 2 文のすべての形態素 unigram
s11	段落わけの直後の文とその文の前後 3 文のすべての形態素 unigram
s12	段落わけの直後の文とその文以外のメール全体の文のすべての形態素 unigram

なお、実験 1 は以下の 2 つの方法で実験する。

実験 1-1 重要文の直前直後にある段落わけにおける段落間の接続の強さを SVM と MEM を利用して推定する。

実験 1-2 重要文の直前直後にある段落わけにおける段落間の接続の強さを SVM と MEM を利用して推定する。ただし、学習データにはメールの文章に含まれるすべての段落わけを用い、テストデータには重要文の直前直後にある段落わけを用いる。

実験 1-1 と実験 1-2 の実験結果を比較すれば、重要文の直前直後にある段落わけにおける段落間の接続の強さを推定するのに有効な手法を知ることができる。

段落間の接続の強さを推定する実験 1 および 2 で利用した素性の組み合わせを以下に示す。

s13 s1, s2, s3, s4, s5, s6, s7, s8
s14 s1, s2, s3, s4, s5, s6, s7, s9
s15 s1, s2, s3, s4, s5, s6, s7, s10
s16 s1, s2, s3, s4, s5, s6, s7, s11
s17 s1, s2, s3, s4, s5, s6, s7, s12

表 5 に実験結果を示す。この結果から、実験 1 の場合、メールの文章全体で学習した方 (実験 1-2) が性能が良く、最も性能が良い時で 89.7%で段落間の接続の強さを推定できることがわかった。一方、実験 2 では 89.0%の精度でメール全体の段落間の接続の強さを推定できることがわかった。利用した素性の性能を比較すると、「s13」の性能が低いことがわかる。これは、素性としての形態素 unigram を取り出す対象の文の違いが原因であると考えられる。すなわち、「s13」に含まれる「s8」では

段落わけの直後の文からしか素性を取り出していないが、「s14」「s15」「s16」「s17」では段落わけの直後の文とその前後の文からも素性を取り出している。これにより、段落間の接続の強さを推定するためには、段落わけの直後の文だけでなく、その前後の文からも形態素を取り出して素性として追加するのがのぞましいことがわかる。

次に、どの素性が段落間の接続の強さを推定するのに有効であったのかを調べるため、素性を取り除く実験を行った。素性を取り除く対象としてそれぞれの実験で最も性能が良かった素性の組合わせを選んだ。表 6 に実験結果を示す。この結果から、「s2」「s4」「s6」「s7」の素性は有効でない場合があるということがわかる。特に「s4」は取り除くことで推定精度が向上している場合が多い。これに対し、形態素の素性が段落間の接続の強さの推定に有効であるのがわかる。例えば、実験 1-2 で最も性能が良い「s17」について

- 「s17」から「s2」を取り除いた素性の組合わせ
- 「s17」から「s12」を取り除いた素性の組合わせ

をそれぞれ用いた SVM の実験結果を比較すると、前者に比べ後者の推定精度は 13.9%減少していた。これは形態素の素性が非常に有効であり、その他の素性では表現することができなかった情報を表現しているということを示唆している。実際、前者では正しく推定できたのに、後者では正しく推定できなかった例を調査すると、段落わけの直前の文に「いつもお世話になります」という表現を含むものがあった。これは挨拶表現であるが、今回の実験では「s2」の素性として登録されてはいなかった。このことから、形態素の素性を機械学習に利用すると、有効ではあるが人手では捉えられなかった情報を段落間の接続の強さの推定に利用できる可能性がある。

SVM と MEM によるメールの段落間の接続の強さの推定で、最も高い精度の実験結果が得られた実験と利用した素性の組合わせは、

重要文の前後の不要な段落の判定　実験 1-2 の実験条件で素性として「s1」「s3」「s4」「s5」「s6」「s7」「s12」を用い、SVM で推定する。

メールの文章全体の不要な段落の判定　実験 2 の実験条件で素性として「s1」「s2」「s3」「s4」「s5」「s7」「s11」を用い、SVM で推定する。

であった。これらの実験結果で段落間の接続が強いと推定された段落わけの精度、すなわち不要な段落わけの判定精度に注目した。その結果を表 7 に示す。この結果と 3. の手がかり表現による判定結果 (表 1) を比較すると、適合率は若干下がってしまったが、再現率には大幅な改善が見られた。また、F 値も重要文の前後の不要な段落わけの判定で 0.328 の改善が見られ、メール全体の不要な段落わけの判定で 0.426 の改善が見られた。この結果から、メールにおける段落間の接続の強さの推定は機械学習を利用することで改善できたといえる。

表 5　段落間の接続の強さの推定結果

素性の種類	実験 1-1(%) SVM/MEM	実験 1-2(%) SVM/MEM	実験 2(%) SVM/MEM
s13	77.7/76.2	80.5/80.5	83.1/84.0
s14	86.5/85.8	84.0/**85.8**	87.5/**86.8**
s15	**86.9**/**88.3**	85.1/84.4	88.7/86.7
s16	85.5/86.9	86.5/83.0	**89.0**/86.2
s17	84.8/84.4	**89.7**/83.7	87.1/85.3

表 6　素性を取り除いた場合の段落間の接続の強さの推定結果

取り除いた素性の種類	実験 1-1(%) SVM/MEM	実験 1-2(%) SVM/MEM	実験 2(%) SVM/MEM
s1	86.5/87.2	89.7/83.0	88.7/85.8
s2	85.8/87.2	**90.1**/84.0	88.7/85.8
s3	86.5/88.3	89.7/85.5	89.0/87.7
s4	86.5/**88.7**	89.7/85.8	**89.1**/**86.9**
s5	86.9/86.9	88.7/82.6	89.0/86.1
s6	86.5/87.9	89.4/83.7	**89.2**/86.8
s7	86.5/**88.7**	89.7/85.8	89.0/87.1
s9/s10/s11/s12	**76.2**/**77.7**	**76.2**/**77.0**	**79.7**/**81.6**

表 7　不要な段落わけの判定精度

	重要文の前後	メール全体
不要な段落わけ	176	1166
不要な段落わけを正しく判定	166	1097
不要な段落わけと誤って判定	18	117
適合率	90.2%	90.4%
再現率	94.3%	94.1%
F 値	0.921	0.922

文　献

[1] 渡辺靖彦, 横溝一哉, 西村涼, 岡田至弘: “メーリングリストを利用した質問応答システムのための知識獲得.”, 自然言語処理, vol.12, No.6, pp.25-44, (2005).

[2] 西村涼, 渡辺靖彦, 岡田至弘: “メーリングリストに投稿されたメールを利用してあいまいな質問に問い返す質問応答システムの作成”, 言語処理学会第 13 回年次大会, E5-2, (2007).

[3] 奥村晃弘, 野中雅人, 濱口佳孝, 野崎正典, 奥村幸治, 清水泰志: “メール要点抽出&転送システム/早解メール”, 沖テクニカルレビュー, 第 192 号 vol.69 No.4, (2002).

[4] Masao Uchiyama, Hitoshi Isahara: “A Statistical Model for Domain-Independent Text Segmentation”, ACL/EACL-2001, pp.491–498, (2001).

[5] 阿部直人, 内山俊郎, 内山匡, 奥雅博: “ウェブ検索を利用したブログテキストセグメンテーション法”, DEWS2008 B4-5, (2008).

[6] 松井祥峰, 乾伸雄, 小谷善行: “単語の結束度と文の表層情報を組み合わせたテキストセグメンテーション”, 情報処理学会研究報告, NL-162, (2004).

[7] 中野滋徳, 足立顕, 牧野武則: “語の近接性に基づいた意味段落境界の判定手法”, 情報処理学会研究報告, NL-166, (2005).

[8] Hearst, M.A: “TextTiling: Segmenting Text”, Computational Linguistics, Vol.23 No.1 pp.34-64, (1997).

[9] 大田康人, 西村涼, 渡辺靖彦, 岡田至弘: “メールの文章における段落間の接続の強さの推定”, 言語処理学会第 14 回年次大会, PC3-5, (2008).

[10] 黒橋禎夫, 河原大輔: “日本語形態素解析システム JUMAN version 5.1 使用説明書.”, 京都大学, (2005).

[11] Taku Kudoh: TinySVM: Support Vector Machines, (http://chasen.org/taku/software/TinySVM/index.html, 2002).

[12] Masao Uchiyama: Maximum Entropy Modeling Packages, (http://www2.nict.go.jp/x/x161/members/mutiyama/software.html#maxent, 2007).