メールの文章における段落間の接続の強さの推定

大田康人　西村 涼　渡辺 靖彦　岡田 至弘
龍谷大学 理工学部 情報メディア学科
{y_oota,r_nishimura}@afc.ryukoku.ac.jp, {watanabe,okada}@rins.ryukoku.ac.jp

1　はじめに

メールの文章では、他の文書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よくある。図 1 に Vine Users ML [*1] というメーリングリスト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。細かく段落わけをすることは、見やすさ、読みやすさを向上させることもあるが、メールの文章を機械処理する場合には不利になることもある。過剰で不要な段落わけのため、他の文書に対しては有効な処理やアプリケーションをメールの文章に適用しても期待される精度では結果が得られないおそれがある [1]。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し、それにもとづいて過剰で不要な段落わけを取り除く方法について検討する。

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Xウィンド上で立ちあげた状態で、cannaで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文章の例

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Xウィンド上で立ちあげた状態で、cannaで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

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　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。メールの文章における段落間の接続の強さを推定し、段落わけの妥当性を判定する手がかりになると考えられるものを以下に示す。

1. 文頭の接続詞
   「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 段落は内容的なつながりが強く、図 3 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。
   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. 指示語
   「この」などの指示語が文頭にある段落は、直前の

[*1] http//vinelinux.org/ml.html (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)

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。

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 の例では、第 1 段落は挨拶の表現を含む 1 つの文で構成されていて、第 2 段落とは内容的なつながりは弱い。図 7 のように段落わけをするのがのぞましい例である。
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も含まれている場合、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 3 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 のように段落わけをせず、1 つにまとめてもよい例である。

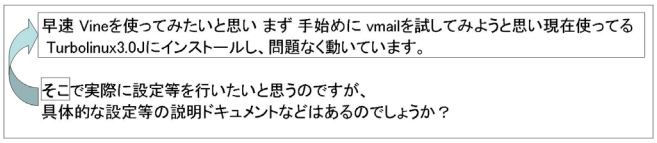

図 5　文頭に指示語をもつ文ではじまる段落の例

- 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

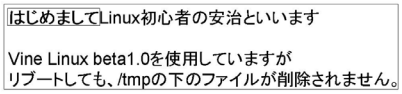

図 7　挨拶の表現を含む文でおわる段落の例

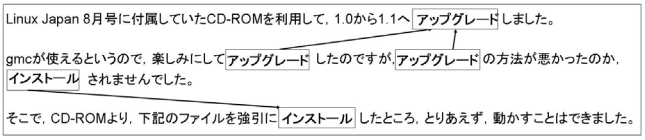

図 8　段落の最後の文に含まれる未定義語が直後の段落の最初の文にも含まれている文章の例

3　メールの段落間の接続の強さを推定する手法

2 章で述べた 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以下に述べる。

step 1　[挨拶の表現による推定]
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。

step 2　[文頭の接続詞による推定]
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。

step 3　[指示語による推定]
段落の最初の文が図 10 (a) に示す表現を文頭にもつ場合、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。

step 4　[未定義語による推定]
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。

step 5　[step 1 ～ 4 で推定できない場合]
step 1 ～ 4 で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。

- 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

4　実験結果と評価

3 章で提案した手法を評価するため、Vine Users ML に投稿された質問メール (返信のあるもの) を用いて以下の 2 種類の実験を行った。

- 質問メールから抽出された重要文 [2] の直前直後で行われた段落わけの妥当性の判定
- メールの文章で行われたすべての段落わけの妥当性の判定

段落わけの妥当性の判定には、段落間の接続の強さの推定結果を利用した。すなわち、段落間の接続が弱いと推定された段落わけは不要な段落わけと判定する。

- 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

4.1　メールの重要文の前後にある段落わけの妥当性の判定

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][3][4]。しかし、重要文の直前直後で行われた過剰で不要な段落わけが原因で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ためには、重要文の直前直後で行われた段落わけの妥当性を判定し、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 200 通取り出し、それらの重要文が段落の最初の文であるか、最後の文である場合の段落間の接続の強さを推定し、不要な段落わけをとりのぞく実験を行った。

この 200 通のメールの重要文の前後には 110 個の不要な段落わけがあった。3 章で述べた step1〜5 の処理を適用した結果、表 1 に示すように、適合率 93%、再現率 38% 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。不要な段落わけを正しく判定した 42 例のう

表 1　段落わけの妥当性の判定結果

	重要文の前後	メール全体
不要な段落わけ	110	363
不要な段落わけを正しく判定	42	126
不要な段落わけと誤って判定	3	7

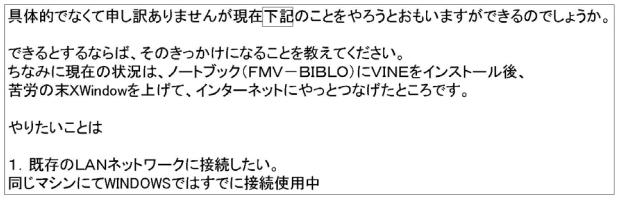

図 11　指示語による推定では不十分にしか不要な段落わけを判定できなかった例

ち、step 1〜3 (挨拶文・文頭の接続詞・指示語による推定) で判定したものが 21 例、step 4 (未定義語による判定) で判定したものが 21 例であった。未定義語を利用した判定が他の手がかり表現に比べて多いのは、Vine User ML に投稿されるメールが専門的な内容を扱うものが多く、その重要文には未定義語 (専門用語) が多く含まれているためであると考えられる。図 11 の例では、第 1 段落と第 2 段落の段落わけは不要であると正しく判定している。ただしこれは、第 1 段落の最後の文に「下記」が含まれているため step 3 (指示語による推定) によって不要な段落わけであると判定されたのであるが、その「下記」が指す内容は第 2 段落ではなく、その後の第 3 段落で述べられている。そして、第 2 段落と第 3 段落の段落わけは step 1 〜 5 の処理では不要なものであると判定はできなかった。

4.2　メールの文章における段落わけの妥当性の判定

次に、重要文の前後に限定せず、メールに含まれる段落すべてを対象にして、その段落わけが不要かどうか判定を行った。実験には、Vine Users ML に投稿されたメールから無作為に取りだした 100 通を利用した。

この 100 通のメールには 951 例の段落わけがあり、そのうち 363 例は不要な段落わけであった。この 951 個の段落わけに対して 3 章で述べた step1〜5 の処理を適用した結果、表 1 に示すように、適合率 95%、再現率 35% で不要な段落わけを判定することができた。不要な段落わけを正しく判定した 126 例のうち、step 1〜3(挨拶文・文頭の接続詞・指示語による推定) で判定したものが 110 例、step 4(未定義語による推定) で判定したものが 16 例であった。4.1 節の場合とは逆に、未定義語を利用した判定の例が少ない。これは、重要文ではない文には未定義語 (専門用語) が重要文ほど多く含まれていないためであると考えられる。また、メールの文章には文の区切りがあいまいなものが多く、文の区切りに失敗したため、不要な段落わけの判定に失敗したものが本実験では 24 例あった。

5　終わりに

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手法について述べた。この手法は、文頭の接続詞、指示語、未定義語などの手がかり表現のみを用いているが、比較的精度よく段落間の接続の強さを推定することができた。現在は段落の最後の文と直後の段落の最初の文だけを用いて段落間の接続の強さを推定している。今後はそれら以外の文も用いてより精度よく推定したい。また、提案手法を用いることによってメーリングリストに投稿されているメールからの知識の抽出がどの程度改善されるのかについても検討したい。

参考文献

[1] 田村 晃裕, 高村 大也, 奥村 学, “複数文質問のタイプ同定”, 言語処理学会第 11 回年次大会, D5-5, (2005).

[2] 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, (2005).

[3] 西村 涼, 渡辺 靖彦, 岡田 至弘, “同義語を用いた質問文の拡張による係り受け関係の柔軟な照合”, 情報処理学会研究報告, 2006–NL–176, (2006).

[4] 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).

[5] 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).