改行・空行挿入によるメールテキストの整形

村田 匡輝† 大野 誠寛‡ 松原 茂樹†
†名古屋大学大学院情報科学研究科　‡名古屋大学情報基盤センター
murata@el.itc.nagoya-u.ac.jp,{ohno,matubara}@nagoya-u.jp

1　はじめに

電子メールは日常的なコミュニケーションツールであり，利用者はメールを読むために毎日多くの時間を割いている．効率的なメール閲覧のために，メールは受信者にとって読みやすく書かれていることが望ましい．読みやすいメールを書くためのテクニックの一つとして，改行・空行を適切に挿入して余白を有効に活用することが挙げられるが [1, 2, 3]，受信メールにおいて，必ずしも受信者が読みやすいと感じる位置に改行・空行が挿入されているわけではない．

本論文では，受信したメールに対し，受信者にとって読みやすい位置に改行・空行を挿入し整形する手法を提案する．本手法では，人がメールを書く際，改行や空行は送信者にとって読みやすい位置に挿入されていると想定し，受信者が書いた送信メールを学習データとする統計的アプローチにより，改行・空行挿入を実現する．

送信メールにおいて挿入されている改行・空行の挿入位置は，人によって傾向が異なる．改行・空行をどのような位置に挿入するかについて，関連性が見込まれる要因を整理し，それに基づき素性セットを定めた．ある個人の受信メールに改行・空行挿入を行うために有効な素性を，素性セットの中から選択して使用することにより，個人のメールの書き方に合ったメールテキストの整形を実現する．

2　改行・空行挿入の要因

人がメールテキストに対して改行や空行を挿入する際，いくつかの要因に基づき挿入位置を決定していると考えられる．それらの要因のうち，どれに着目しているかによって，その人の改行・空行挿入傾向，つまり，読みやすいと感じる改行・空行挿入位置の傾向が異なってくる．以下では，改行・空行挿入の際に一般的に関連性が見込まれる要因について整理する．

2.1　意味的な切れ目

メールテキスト中の各行が意味的なまとまりを構成することを重視する場合，改行は意味的な切れ目に挿入される．これにより，単語や文節の途中で行が区切られることが回避でき，効率的なメールテキストの閲覧に繋がる．

著者らはこれまでに読みやすい改行位置について分析している [4]．意味的な切れ目を示す情報として節境界や係り受け関係などが挙げられる．

2.2　行長

行長が何文字以内になるように改行を挿入するかは，個人のメーラ画面の横幅の設定と関連する．横幅の設定よりも長い行長で書かれたメールテキストを読むときには，スクロールが必要となるため，読みづらさを感じる場合がある．そのため，送信者は，自分のメーラの横幅の設定を考慮し，メールテキストの行長がその横幅よりも短かくなるようにメールテキストを記述すると考えられる．メーラ画面の横幅の設定は個人によって異なるため，メールテキストの1行が何文字以内になるように書くかは個人ごとに傾向の違いが存在する．

2.3　行長のバランス

行長について，1行が何文字以内で書かれるかという観点の他に，行長のバランスを揃えるという書き方が考えられる．一段落中の，各行の右端が不揃いになると読みにくいと感じる人もおり，そのような送信者は，他の行の行長を考慮し，行長のバランスを揃えるように改行を挿入する場合がある．改行位置をある程度そろえると読みやすくなることが指摘されており [1]，このように改行挿入をすることによって，読みやすく感じるテキストになると考えられる．

2.4　話題の切れ目

空行の働きの一つとして，メールテキスト中の話題の切れ目に挿入することによって，メールテキストを話題ごとに分割する役割がある．話題ごとに分割されていれば，効率的なメールテキストの閲覧が可能となる．話題の切れ目を示す情報として，例えば，文頭の接続詞の直前や，疑問文の直後などの手がかり語が挙げられる．

2.5　段落の行数

上記で挙げた話題の切れ目以外のポイントとして，「空行」と「段落の行数」の関係がある．ディスプレイ上で読むメールテキストは，紙上の文章と比べ，ディスプレイの明るさやフォント等の関係から読みやすさが低下する．そのため，適宜空白を挿入して余白を作ることによって，メールテキストの読みやすさを向上させることが可能である．一段落の行数が何行以上になった場合に読みにくいと感じるかは，個人によって異なる．

3　メールテキストの整形手法

本手法の流れを図1に示す．まず，受信者の書いた送信メールを学習データとディベロップメントデータに分割し，**2**節で挙げた改行・空行挿入の要因に基づいて定めた素性セットのうちから，その受信者のメールの書き方の傾向を捉えるために有効と思われる素性

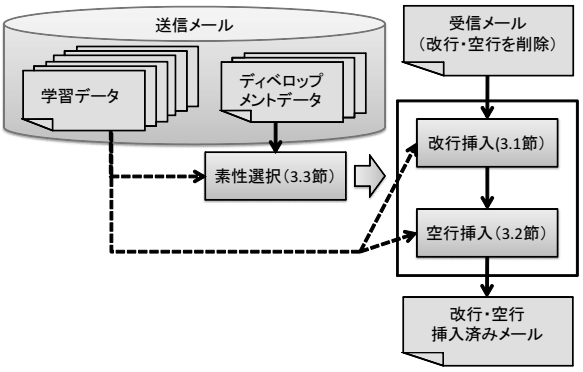

図 1: 本手法の流れ

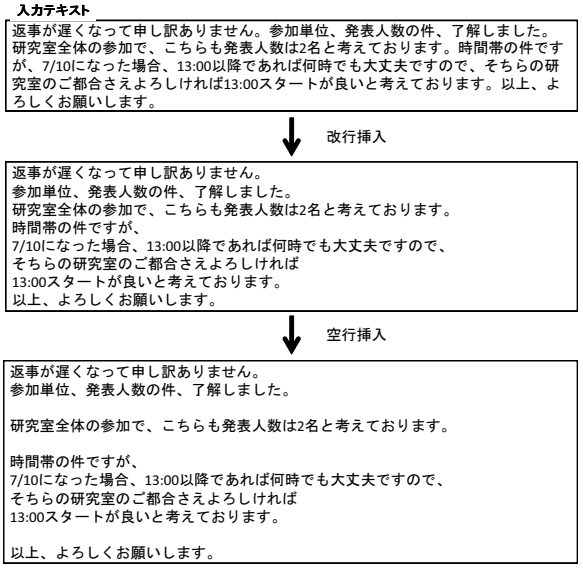

図 2: メールテキスト整形の例

を選択する．その後，改行・空行を除いた受信メールに対して，選択した素性を用いた改行挿入，空行挿入を順に実施する．その際には，受信者自身が読みやすいと感じる位置への改行・空行挿入を行うために，受信者が書いた送信メールを機械学習の学習データとして使用する．本手法によるメールテキスト整形の例を図 2 に示す．

3.1 改行挿入手法

改行挿入手法では，形態素解析，文節まとめ上げ，節境界解析，係り受け解析が与えられた文を入力とし，入力文中の各文節境界に対して，その位置に改行を挿入するか否かを同定する．入力文に対する適切な改行位置を同定するために，1 行あたりの文字数が最長文字数を超えないという条件の下，1 文中に挿入されうる改行位置の全ての組み合わせの中から，最適な組み合わせを確率モデルを用いて決定する．

以下では，n 個の文節からなる入力文を $B = b_1 \cdots b_n$ とするとき，改行結果を $O = o_1 \cdots o_n$ と記す．ここで，o_i は，文節 b_i の直後に改行が挿入されるか（$o_i = 1$）否か（$o_i = 0$）のいずれかの値をとる．なお，$o_n = 1$ である．

3.1.1 改行挿入のための確率モデル

本手法では，入力文の文節列を B とするとき，$P(O|B)$ を最大にする改行挿入結果 O を求める．入力文を m 行に分割した j 行目の文節列を $L_j = b^j_1 \cdots b^j_{n_j} (1 \leq j \leq m)$ とした場合，$1 \leq k < n_j$ のとき $o^j_k = 0$，$k = n_j$ のとき $o^j_k = 1$ となる．ここで，各文節境界に改行が挿入されるか否かは，直前の改行位置を除く，他の改行位置とは独立であると仮定すると，$P(O|B)$ は次のように計算できる．

$$
\begin{aligned}
& P(O|B) \qquad (1)\\
= \;& P(o^1_1 = 0, \cdots, o^1_{n_1-1} = 0, o^1_{n_1} = 1, \cdots,\\
& \quad o^m_1 = 0, \cdots, o^m_{n_m-1} = 0, o^m_{n_m} = 1 | B)\\
\cong \;& P(o^1_1 = 0|B) \times \cdots\\
& \times P(o^1_{n_1-1} = 0 | o^1_{n_1-2} = 0, \cdots, o^1_1 = 0, B)\\
& \times P(o^1_{n_1} = 1 | o^1_{n_1-1} = 0, \cdots, o^1_1 = 0, B) \times \cdots\\
& \times P(o^m_1 = 0 | o^{m-1}_{n_{m-1}} = 1, B) \times \cdots\\
& \times P(o^m_{n_m-1} = 0 | o^m_{n_m-2} = 0, \cdots, o^m_1 = 0, o^{m-1}_{n_{m-1}} = 1, B)\\
& \times P(o^m_{n_m} = 1 | o^m_{n_m-1} = 0, \cdots, o^m_1 = 0, o^{m-1}_{n_{m-1}} = 1, B)
\end{aligned}
$$

ここで，$P(o^j_k = 1 | o^j_{k-1} = 0, \cdots, o^j_1 = 0, o^{j-1}_{n_{j-1}} = 1, B)$ は，1 文の文節列 B が与えられ，$j-1$ 行目の行末位置が同定されているときに，文節 b^j_k の直後に改行が挿入される確率を表す．同様に，$P(o^j_k = 0 | o^j_{k-1} = 0, \cdots, o^j_1 = 0, o^{j-1}_{n_{j-1}} = 1, B)$ は，文節 b^j_k の直後に改行が挿入されない確率を表す．これらの確率を最大エントロピー法により推定した．

次に，$\arg\max P(O|B)$ の計算について述べる．2.3 節に挙げた行長のバランスを考慮して改行挿入を行うためには，各行の間で長さを比較することが考えられる．しかし単に各行の行長を比較するとなると，前後の改行位置の情報が必要となり，直前の改行位置以外の改行位置とは独立であるとする仮定が利用できなくなり，動的計画法を用いて効率的に $\arg\max P(O|B)$ を計算できなくなる．

そこで本研究では，できる限り効率的に計算を行うため，行長のバランスを考慮するための素性として，平均行長（1 文の文字数を，改行結果における 1 文全体の行数で割った値）に対して，各行の行長がどの程度近いのかという値を導入した．これは，行長のバランスが取れた改行結果では，各行の行長が入力文の文字数を行数で割った値に近づくためである．この素性を導入することにより，$\arg\max P(O|B)$ は以下のように計算できる．まず，入力文を l 行に改行する改行結果 $O \in \{O | \sum_{i=1}^{n} o_i = l\}$ の中で $P(O|B)$ を最大とする改行結果 O_l を求める．

$$
O_l = \mathop{\arg\max}_{O \in \{O | \sum_{i=1}^{n} o_i = l\}} P(O|B) (1 \leq l \leq n)
$$

その後，$O_1, \cdots, O_n$ の中から $P(O|B)$ を最大とする改行結果を求める．

$$
\arg\max P(O|B) = \mathop{\arg\max}_{O \in \{O_1, \cdots, O_n\}} P(O|B)
$$

3.2 空行挿入手法

空行挿入手法では，形態素解析，改行位置が与えられた文から構成されるテキストを入力とし，入力テキスト中の各文境界に対して，その位置に空行を挿入するか否かを同定する．入力テキストに対する適切な空行位置を同定するために，1 テキスト中に挿入されう

る空行位置の全ての組み合わせの中から，最適な組み合わせを確率モデルを用いて決定する．

以下では，q 個の文からなる入力テキストを $S = s_1 \cdots s_q$ とするとき，空行結果を $T = t_1 \cdots t_q$ と記す．ここで，t_i は，文 s_i の直後に空行が挿入されるか（$t_i = 1$）否か（$t_i = 0$）のいずれかの値をとる．なお，$t_q = 1$ である．入力テキストを空行挿入によって r 個に分割した g 個目の文の列を $S_g = s_1^g \cdots s_{q_g}^g (1 \leq g \leq r)$ とした場合，$1 \leq h < q_g$ のとき $t_h^g = 0$，$h = q_g$ のとき $t_h^q = 1$ となる．

入力テキスト S が与えられたとき，$P(T|S)$ を最大にする空行挿入結果 T を求める．各文境界に空行が挿入されるか否かは，直前の空行位置を除く，他の改行位置とは独立であると仮定して，3.1.1 と同様の計算方法で $P(T|S)$ を求める．

3.3 素性の選択方法

文献 [4]，及び，**2** 節で挙げた改行・空行挿入の要因に基づき，改行・空行挿入に一般的に有効と考えられる素性セットを定めた．素性セットを表 1 に示す．

改行・空行挿入を行う際に考慮する要因は受信者によって異なると考えられるため，ある受信者の受信メールを改行・空行挿入によって自動整形する場合，必ずしも表 1 に示した全ての素性が有効であるわけではない．本手法では，受信者の送信メールの一部をディベロップメントデータとして，表 1 に示した素性のうち，全ての素性を使用する場合と一つずつ削除して使用する場合の改行・空行挿入の精度（4.1 節で示す再現率と適合率の調和平均である F 値）を比較する．削除することによって精度が増加する素性は，その受信者にとっての読みやすさの向上に貢献しないと考え，その受信者の受信メールに対して改行・空行挿入を行う際には使用しないこととする．すなわち，本研究では，一般的に有効と考えられる表 1 の素性セットの中から，対象の受信者にとって有効な素性を選択して使用することにより，その受信者のメールの書き方の傾向に合わせた改行・空行挿入を実現する．

4 予備実験

本手法の有効性を評価するためには，多数のメールユーザのデータを用いて実験する必要がある．そのための予備実験として，著者のうちの 1 名のメールデータを用いて整形実験を試みた．

4.1 実験の概要

まず，著者の送信メールを学習データとディベロップメントデータに分割し，3.3 節に示した方法を適用し，改行・空行挿入用の素性を決定した．決定した素性セットを使用した改行・空行挿入手法によってメールテキストの整形を実施した．学習データに著者の送信メール 517 通，素性選択のためのディベロップメントデータに送信メール 65 通を使用した．また，テストデータには，送信メール 65 通，受信メール 100 通を使用した．本研究の目的は，受信メールを改行・空行挿入によって受信者自身が読みやすいように整形することであるが，定量的な評価を行うために，著者の送信メールもテストデータとして用いた．

送信メールに対する評価は，元の送信メールテキスト中の改行・空行位置に対する再現率及び適合率により行った．再現率，適合率はそれぞれ，

表 2: 実験結果

	再現率	適合率	F 値
改行挿入	58.65% (61/104)	55.96% (61/109)	57.28
空行挿入	100.00% (123/123)	62.12% (123/198)	76.74

$$\text{再現率} = \frac{\text{正しく挿入された改行/空行数}}{\text{正解の改行/空行数}}$$

$$\text{適合率} = \frac{\text{正しく挿入された改行/空行数}}{\text{挿入された改行/空行数}}$$

を測定した．受信メールに対する評価は，著者によるメールテキストの主観的評価を行った．

メールテキストの内容部分のみを実験対象とするため，挨拶や署名，引用部分，転送部分を簡単なスクリプトを作成し除去した．言語情報は自動解析によって付与した．形態素解析には MeCab [5] を，文節まとめ上げ，係り受け解析には CaboCha [6] を，節境界解析には CBAP [7] をそれぞれ用いた．なお，実験のための最大エントロピー法のツールとしては，文献 [8] のものを利用した．オプションに関しては，学習アルゴリズムにおける繰り返し回数を 2,000 に設定し，それ以外はデフォルトのまま使用した．また，実験では，一行の最長文字数を 37 文字とした．

4.2 素性選択の結果

ディベロップメントデータに対して 3.3 節に示した方法を適用した結果，改行挿入においては，表 1 の 6. と 12. の素性を削除した場合に F 値がそれぞれ 1.16, 0.75 増加した．また，空行挿入においては，表 1 の 15., 17., 18. の素性を削除した場合に F 値がそれぞれ 19.56, 4.06 増加した．

以上より，実験では表 1 の素性から 6., 12., 15., 17., 18. を除いて使用した．

4.3 実験結果

送信メール 65 通に対する，提案手法による改行挿入，空行挿入の再現率と適合率を表 2 に示す．提案手法は，改行挿入において，再現率 58.65%，適合率 55.96%を達成した．また，空行挿入においては，再現率 100%，適合率 62.12%を示したが，この空行挿入結果は，テストデータ中の全ての文境界の直後に空行を挿入した結果と同一である．

提案手法による改行・空行挿入結果の例を図 3 に示す．図 3 に示した結果は，著者が書いた送信メールにおける改行・空行挿入結果と完全に一致している．本手法による改行・空行挿入の F 値は必ずしも高い値となっていないが，図 3 のように，著者の送信メールと完全一致する結果も出力できた．

4.4 主観的評価

著者の受信メール 100 通に本手法を適用し，テキストの主観的評価を実施した．元の受信メールと整形されたメールテキストを比較して「読みやすくなっている」,「読みにくくなっている」,「同程度」の 3 分類の判定を著者自身が行った．

100 通のメールテキストのうち，提案手法によって読みやすくなっていると判定したテキストは 17% (17/100) であった．また，同程度と判定したテキスト

表 1: 最大エントロピー法で用いた素性

改行挿入		
形態素情報	1.	b_k^j の主辞（品詞，活用形）と語形（品詞）
節境界情報	2.	b_k^j の直後に節境界があるか否か
	3.	b_k^j の直後の節境界の種類（節境界がある場合）
係り受け情報	4.	b_k^j が直後の文節に係るか否か
	5.	b_k^j が節末文節に係るか否か
	6.	b_k^j が行頭からの文字数が最長文字数以内の位置にある文節に係るか否か
	7.	b_k^j が連体節の節末文節から係られるか否か
	8.	b_k^j が直前の文節から係られるか否か
	9.	行頭文節 b_1^j から b_k^j までの間で係り受けが閉じているか否か
	10.	b_k^j の右側で，かつ，行頭からの文字数が最長文字数以内の位置にある文節の中で，b_k^j と同じ係り先をもつ文節があるか否か
行長	11.	最長文字数に対する行頭から b_k^j までの文字数の割合
行長のバランス	12.	平均行長に対する，行頭から b_k^j までの文字数と平均行長との差の割合
文節の第一形態素	13.	b_k^j の直後の文節の第一形態素の表層文字が「する，なる，思う，問題，必要」のいずれか，もしくはその品詞が「名詞-非自立-一般，名詞-非自立-副詞可能，名詞-ナイ形容詞語幹」のいずれかであるか否か
読点情報	14.	b_k^j の最終形態素が読点であるか否か
空行挿入		
行数	15.	直前の空行位置からの行数
	16.	s_{h+1}^g の行数
手がかり語	17.	s_{h+1}^g の第一形態素が接続詞であるか否か
	18.	s_{h+1}^g の第一形態素の表層形（接続詞である場合）
	19.	s_h^g が疑問文であるか否か

入力テキスト

毎日新聞データ集を確認したところ、データ集には朝刊と夕刊の記事が含まれていました。お渡ししたデータは朝刊、夕刊の区別なく文を抽出したデータを基に作成しているため、似たような文が見られたのだと思われます。よろしくお願い致します。

改行・空行挿入テキスト

毎日新聞データ集を確認したところ、
データ集には朝刊と夕刊の記事が含まれていました。

お渡ししたデータは
朝刊、夕刊の区別なく文を抽出したデータを基に作成しているため、
似たような文が見られたのだと思われます。

よろしくお願い致します。

図 3: 提案手法による改行・空行挿入テキストの例

は 23% (23/100) であり，60% (60/100) は読みにくくなっていると判定した．読みにくくなっていると判定したメールの割合が高くなった要因として，本文中に含まれるスペースや記号を考慮した改行が行えていないこと，箇条書きされたテキストへの改行挿入の性能の低さが挙げられる．これらの誤りは，著者のみでなく多くの人が読みにくいと思う要因であると考えられるため，メールテキストの構造を考慮した改行挿入を行うことが必要となる．

5 おわりに

本論文では，改行・空行挿入に基づくメールテキストの整形手法を提案した．本手法では，受信者の送信メールを学習データとし，素性セットのうちから，その受信者のメール整形に有効な素性を選択して機械学習に用いることにより，その受信者のメールの書き方の傾向に合った改行・空行挿入を実現する．著者のメールテキストを用いた改行・空行挿入実験では，改行挿入の F 値で 57.28，空行挿入の F 値で 76.74 を示した．また，メールテキストの主観的評価を行い，本手法の適用によって可読性が低下する要因を分析した．

今回の実験では，著者一名の送信メールデータを用いて実験を行った．今後，複数名のメールデータを用いて実験を実施し，それぞれの受信者にとって読みやすい改行・空行挿入が行えているかを評価する予定である．また，実験で使用した著者のメールデータにおいては，全ての文の直後に空行が挿入される結果となり，空行挿入手法が必ずしも有効に働いていないという結果となった．今後，空行挿入手法の改良を行う必要がある．

謝辞 本研究は一部，科研費挑戦的萌芽研究 (No.21650028) による．

参考文献

[1] 藤田，“メール文章力の基本” 日本実業出版社 (2010).
[2] 上田，細田，“超速マスター E メール・履歴書 エントリーシート 成功実例集” 高橋書店 (2009).
[3] 安藤，“E-メールハンドブック” 共立出版 (1998).
[4] 村田，大野，松原，“読みやすい字幕生成のための講演テキストへの改行挿入” 信学論，Vol.J92-D, No.9, pp.1621-1631 (2009).
[5] 工藤，山本，松本，“Conditional Random Fields を用いた日本語形態素解析” 情処研報．NL, Vol.2004, No.47, pp.89-96 (2004).
[6] 工藤，松本，“チャンキングの段階適用による日本語係り受け解析” 情処学論，Vol.43, No.6, pp.1834-1842 (2002).
[7] 丸山，柏岡，熊野，田中，“日本語節境界検出プログラム CBAP の開発と評価” 自然言語処理，Vol.11, No.3, pp.39-68 (2004).
[8] L. Zhang: Maximum entropy modeling toolkit for python and c++, `http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html` (2007).