電子メールにおける重要文抽出と携帯電話向け要約システムへの適用

長谷川 隆明[†] 林 良彦[††] 山崎 毅文[†††]

インターネットに接続可能な携帯電話の普及により，携帯電話に電子メールを転送するユーザが増えている．外出先からでも携帯電話に電子メールの着信が通知されるので，日々大量に電子メールを受信するユーザにとって，携帯電話を通して即時に電子メールを読むことは有効である．しかしながら，画面の大きさや送信可能文字数の制約により，転送される電子メールの本文すべてを読むことは難しい．本論文では，この問題を解決するために，電子メールの特徴を利用した重要文抽出手法と携帯電話向け要約システムへの適用を提案する．本手法の特徴は，電子メールに特有な表現とスタイルの特徴を利用したルールにより重要文を抽出することである．本手法により，人手により作成された重要文の正解との比較において，特に複数の評価者間で正解に対するコンセンサスの得られる電子メールに対して，高い精度で正解と一致する重要文を抽出できた．また，携帯電話向け要約システムへの適用を考慮した評価においても，電子メールの取捨選択に十分利用できる高い精度が得られた．

Sentence Extraction from Emails and Its Use in Mobile Phones

Takaaki Hasegawa,[†] Yoshihiko Hayashi[††] and Takefumi Yamazaki[†††]

With the spread of mobile phones that can access the Internet, most people are forwarding their incoming email to their mobile phones. Mobile phones enable heavy users of emails to access them anywhere. The restricted capacities (display size and message limits) of mobile phones, however, prohibit whole messages from being read as is. We propose a new rule-based method of sentence extraction which takes the features of emails into consideration and describe its application to an email summarizer for mobile phones. Our method uses the expression and style specific to emails. The sentences extracted by the proposed method are virtually the same as those selected manually by majority voting. A task-oriented evaluation of the proposed method reveals its high precision and recall in screening for emails through mobile phones given a practical message size limit.

1. はじめに

インターネットに接続可能な携帯電話の普及により，パソコンで受信している電子メールを携帯電話に転送するユーザが増えている．携帯電話のキャリアにより提供されているプッシュ型の電子メールサービスにより，ユーザの携帯電話のメールアドレス宛てに電子メールが転送されればただちにユーザの携帯電話へその電子メールが通知される．これにより，ユーザは携帯電話を，パソコンのメールアドレス宛てに受信した電子メールの着信通知として有効に用いることができる．しかしながら，携帯電話のネットワークによる送信可能な文字数の制限や小さな携帯電話端末の画面の大きさにより，パソコンで受信することを前提とした電子メールは，携帯電話において読みやすいものではない．このため，携帯電話を用いてパソコンから転送されてくる電子メールを読むためには，事前に不要な部分を取り除いたり重要な部分だけを取り出したりすることによって，適切に電子メールを要約しておくことが必要となる．また，電子メールの要約を携帯電話に転送することは，要約による着信通知という新しいコンセプトをも生み出す．そこで本論文では，携帯電話を用いて電子メールを読むうえで有効な電子メールの要約方法とそれを用いたサービスシステムについて述べる．

本論文の構成は以下のとおりである．はじめに，2章で電子メールの特性と要約手法について述べる．次

† 日本電信電話株式会社サイバースペース研究所
Cyberspace Laboratories, Nippon Telegraph and Telephone Corporation
†† 大阪大学大学院言語文化研究科
Graduate School of Language and Culture, Osaka University
††† 株式会社 NTT レゾナント
NTT Resonant Inc.

に，3 章で本論文で提案する重要文抽出について述べ，4 章で重要文抽出を用いた携帯電話向け要約システムについて説明する．5 章で本手法による重要文に対して，人手による正解と比較すると同時に，携帯電話向け要約システムというタスクの中で評価する．6 章で携帯電話向け要約システムにとって考慮しなければならない制限文字数と本文の中にある重要箇所の抽出の単位に関する議論を行い，最後に 7 章で結論を述べる．

2. 電子メールの特性と要約手法

電子メールの要約は，これまで大きく分けて 2 つの方向で議論されてきた．1 つは，電子メールの交換により展開される議論を整理して要約する研究[8),11),14)]である．これらの応用として，ソフトウェアの開発を支援するためのツール[8)]や教育における学習者間のコミュニケーションの支援[14)]がある．もう 1 つは，携帯電話などへの配信を目的とした圧縮率の高い要約の研究であり，モバイル端末を意識したコンテンツ要約技術[16)]として近年注目されている．本論文の対象も後者であり，以下では電子メールの特徴をふまえながら従来技術を振り返る．

電子メールは，コミュニケーションの手段であるので，新聞記事のように一方的に何かを伝えるだけでなく，相手への質問や依頼・要求などを含むためインタラクションが起こることが特徴である．また，電子メールのテキストの質はそれほど高くはないことと，改行や空白などを用いてレイアウトが施されていることも特徴的である．モバイル端末を前提とした制限文字数がかなり限られた要約を生成するうえにおいて，これらの特徴こそ重要な差異化ポイントとなり，これらの特徴を考慮した要約手法が必要となる．いい換えれば，単語の頻度を利用した伝統的な重要文抽出[12)]や構文情報を用いた手法[7),13)]を適用しても，モバイル端末にとって適切な要約を生成することは難しい．さらに，電子メールには私的な個人情報が含まれるため，電子メールを大量に収集し要約としての正解を用意することが困難である．このため，ホームページのヘッドラインを生成するために用いられる統計的な手法[2),3)]も簡単には利用できない．また，機械学習により特徴的な単語を抽出する方法[15)]では，要約というよりは単語の羅列を生成してしまう．

電子メールに特化した要約として，単語の文字列を同義のより短い文字列に置換する方法[4),10)]や電子メールの文面から意図の分類を行った後で，その意図を含む表現をより多く含むパラグラフを抽出する方法[6)]が提案されている．前者は圧縮率を優先するため文字列を記号などにも置換するため可読性が低下するという課題がある．後者は抽出の単位がパラグラフなので，可読性は良いが重要な部分が抽出されるパラグラフから漏れるという問題を含んでいる．

電子メールには，本文以外にもヘッダに情報が存在する．見出し情報を用いる方法[17)]を用いてヘッダのサブジェクトを利用して要約を生成することも考えられる．しかしながら，サブジェクトがそのまま変わらずにリプライが繰り返されると本文の内容が当初のサブジェクトに合致していた内容から離れていくため，サブジェクトが必ずしも適切でなくなるという問題が発生する．

3. 電子メールにおける重要文抽出

本論文では，電子メールに特化した重要文抽出の手法を提案する．本論文と従来手法の差異化ポイントは次の 2 点である．1 つは，電子メールの重要部分の選択の単位である．従来手法では，電子メールの重要部分の選択の単位はパラグラフまでで，さらに細かい単位に分けることは行っていない．しかしながら，電子メールには引用や署名あるいは箇条書きの混在などの独特のスタイルがあるため，重要文を抽出する前に抽出すべき単位となる文を切り出しておく必要がある．本論文では，意味のとれる範囲で選択の単位が最も細かい文を選択の単位とする．特に 1 つのパラグラフのサイズが大きい場合にはできるだけ原文の内容をカバーしながら，要約率を下げることは難しいので，選択の単位を文単位とすることは有効だと考えられる．

もう 1 つは，電子メールに特有な表現とスタイルの特徴を利用した重要文抽出である．電子メールの多くは，特定の個人や集団の相手に対する連絡である．連絡には，依頼や質問あるいは通知や報告といったはっきりした目的が存在していることが多く，本論文ではその範囲の電子メールを対象とする．また，電子メールの多くには，送信者を特定するための名乗りの表現や時候の挨拶が存在するなど独特のスタイルを持つ．本論文では，目的を表している表現と電子メールの持つスタイルに着目し，これをルールとして記述したうえでルールにマッチした文を抽出することにより，重要文抽出を実現する．反対に，時候の挨拶や決まり文句など不要な表現もルールとして記述すれば，これらも不要な箇所としての手がかりとして利用でき，重要文抽出の精度を高めることができる．

本論文で提案する重要文抽出のステップを**図 1** に示し，以下にまとめる．

ステップ 1 ヘッダや添付ファイルおよび署名や引用

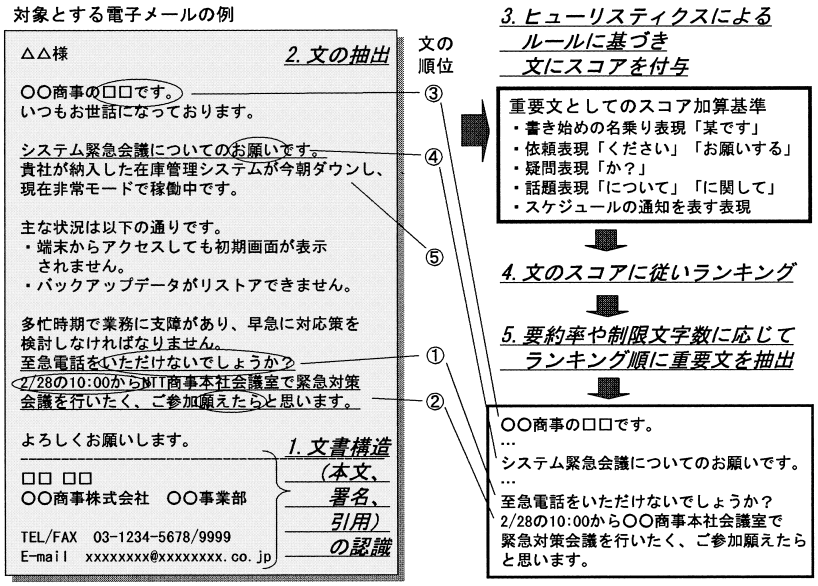

図 1 重要文抽出のステップ
Fig. 1 Summarization steps.

を除いて送信者が記述した本文のみを重要文抽出の対象とするため，電子メールの文書構造を認識し本文と署名や引用を行ごとに区別する．

ステップ 2 行ごとに区別された文書構造に加え，本文に混在する箇条書の認定や行間の接続を行い，重要文抽出の単位となる文の抽出を行う．

ステップ 3 電子メールの目的に着目した重要文抽出を行うために，これらの表現を含む文に対してルールに基づきスコアを付与する．具体的には，抽出された各文に対する形態素解析 JTAG[5]の後，形態素単位に分割された文に対して，形態素パターンとそれに対応するスコアからなるヒューリスティクスによるルールを適用し，形態素パターンにマッチする文に対して設定されたスコアを付与する．

ステップ 4 制限文字数の中でスコアの高い重要文を選択するために文のスコアに従って文をランキングする．

ステップ 5 要約率や制限文字数に応じてランキング順に文を選択し，選択されない脱落文を記号「...」で置換して出力する．

以下ではそれぞれの処理について詳説する．3.1 節では，重要文抽出の前処理となるステップ 1 とステップ 2 について述べる．ヘッダや添付ファイルを除いた本文の持つスタイルの特徴を整理し，これらの特徴に基づいて抽出の単位である文をどのように切り出すのかについて述べる．3.2 節では，ステップ 3 について述べる．特に電子メールの目的に着目した重要文抽出を実現するためのルールについて説明する．3.3 節では，後処理となるステップ 4 とステップ 5 について述べる．ステップ 3 によるルールの適用後に制限文字数の中でどのように重要文を選択するかについて述べる．

3.1 スタイルの特徴と文分割

電子メールテキストのスタイルの特徴として以下があげられる．

- 引用や署名などの構造がある．
- 文の途中であっても改行が挿入されることがある．
- 文章と箇条書きが混在することがある．

以上の特徴を持つ電子メールを対象とする場合には，重要な箇所を抽出する，すなわち重要文抽出を行う前に，文の単位で本文を切り出しておかなければならない．これを実現するために，次のようなステップを実行する．

(1) 文書構造の解析による署名の削除
署名は通常テキストの最後部に付いているので，末尾の行から上の方へ一定の行数だけ順に空行や記号のみの行あるいは行末が文末表現である行を検出する．検出された行の次の行から最後部までを署名と見なし[1]，これを削除する．

(2) 引用記号に基づく引用行の削除
各行に対して，行頭に英数字を含む引用記号が存在するかを調べることによって引用行を特定し，引用行を後述の処理対象から除外する．英数字のみの行やあらかじめ用意されたメールソフトにより付与さ

れる行も同様に除外する．ただし，コメントを表す記号を先頭に持つコメント行は除外しない．

(3) 英数字や記号または空白とその位置に基づく箇条書き行の特定

各行に対して，行頭に英数字やコロンなどの記号が存在する箇条書きラベルを持つ箇条書き行を特定する．あるいは箇条書きラベルがない場合には，空白がある一定の行頭からの位置以内に存在し，空白より後の文字列がある一定の長さ以内である行を箇条書きラベルなしの箇条書き行として特定する．

(4) 同じタイプの行の接続と文の切り出し

箇条書き行と連続する箇条書き行の行頭位置を比較し後続行の方が大きければ箇条書き行を接続する．また，箇条書き以外の行と連続する箇条書き以外の行を接続する．文末表現および文末表現がなくても行末に一定の行頭からの位置以内の特定の助詞や助動詞相当の文字列があればそこで文を分割する．コメント行については，コメント行が連続する場合のみ，後続する行の先頭にあるコメントを表す記号を除いて接続する．後述する処理のため，1行1文の形に整形しておく．

3.2 表現の特徴とルールに基づく重要文抽出

不要な表現を削除しながら電子メールの連絡の目的を表している重要な表現を抽出するためのルールについて詳説する．重要な表現や不要な表現は，位置情報にも密接に関係している．たとえば，時候の挨拶は本文の最初に存在する．前処理により抽出された各文は形態素解析により，形態素単位に分割され，以下で述べるルールが適用される．重要文抽出において位置情報を考慮するため，引用や署名を除いて1行1文の形式で抽出された本文は，ルールが適用されたときに動的に先頭文と冒頭文，通常文の位置情報に区分される．**図2**にルールの適用時に区分される電子メールの位置情報を示す．ルールは区分される位置情報に基づき以下の3種類に分類され，重要あるいは不要な表現を表す形態素パターンとそれが持つスコアから構成される．

先頭文ルール 本文の先頭から適用され，あらかじめ用意された時候の挨拶（「いつもお世話になっております.」など）や宛名表現（「△△様」など）にマッチする場合に適用され，その文のスコアを下げる．名乗り表現（「○○商事の□□です.」など）にマッチする場合はその文のスコアを上げる．

冒頭文ルール 先頭文ルールがマッチした次の文のみに対して，格助詞「を」格または「が」格を含む文があれば適用され，話題を含んでいるとしてその文のスコアを上げる．

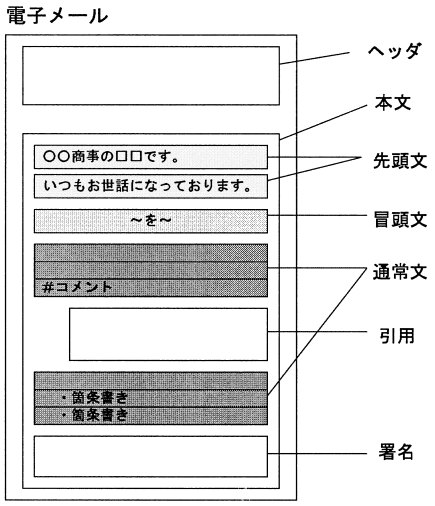

図2 重要文抽出ルールの適用時に割り当てられる文の位置情報
Fig. 2 Sentence position in email text dynamically assigned by using sentence extraction rules.

通常文ルール 先頭文ルールと冒頭文ルールのどちらも適用されないすべての文に対して適用され，あらかじめ指定した重要表現の形態素パターンにマッチした場合にその文のスコアを上げる．

以上のルールは入れ子構造が許されている．1つのルールの中には，入れ子として記述されるすべてのルールを満たさなければならないものと，どれかを満たせばよいものの指定が可能である．また，入れ子として例外を記述するルールも用意されている．たとえば，逆接の接続助詞「が」を含む文を重要だとしてスコアを付与するルールを定義する際に，たとえば「申し訳ありませんが」にマッチした文を重要としたくない場合には，これを例外として入れ子に記述することができる．

次に，表現や位置情報を用いた詳細なマッチングを実現するための形態素パターンについて述べる．形態素パターンの記述は，形態素の満たすべき位置の制約と形態素自身を指定するための情報という2つの軸がある．形態素の位置の制約には，形態素が以降の形態素と連続しなければならないもの，形態素が以降の形態素とは連続しないもの，形態素が文頭になければならないもの，形態素が文末になければならないものの4種類が記述できる．形態素自身は，表記，読み，品詞，一般名詞カテゴリ☆が一致するかどうかにより指定される．これらの指定情報は単独でも組み合わせて

☆ 日本語語彙体系[9]に基づく意味属性である．

```
<SENTOU>
RULEG=nanori VAL=35
</SENTOU>

<RULEG nanori>
RULE=daresoredesu
RULE=at_mark
</RULEG>

<RULE daresoredesu>
ICAT=[363|5] NEXT=NEAR
HYOU=[です|と] NEXT=NEAR
HYOU=[. |申|もう]
</RULE>

<RULE at_mark>
ICAT=[363|5] NEXT=NEAR
HYOU=@ NEXT=FAR
HYOU=[です|申|もう]
</RULE>
```

図 3 形態素パターンとスコアから構成されるルールの例
Fig. 3 Examples of rules consisting of Japanese morphological patterns and their scores.

も用いることができる．たとえば，表記が「が」でかつ品詞が格助詞であるという具合に指定できる．また，表記と読みは部分一致も指定できる．たとえば，読みを「*ガツ」と指定すれば「1 月」から「12 月」まですべてを効率良く指定することができる．

ルールは形態素パターンとそれがマッチした文に加算あるいは減算するスコアからなる．1 つの文に対して各々の形態素パターンがマッチするたびに指定されたスコアが加算あるいは減算される．どのような表現がどのくらい重要と見なすかは人によってそれぞれ異なるため，一意に決定することはきわめて難しい問題である．このため，簡潔で分かりやすいルールの記述は，チューニングを容易にしておくうえで不可欠である．以下では，ルールを作成するうえで，具体的にどのような表現をどのくらい重要と見なすかについて述べる．なお，本論文では，ルールに与えられるスコアの大きさは，経験的に決定している．

本論文では，コミュニケーションにおいて重要と考えられる表現，すなわち依頼や要求，質問，名乗り，スケジュールの通知表現に焦点を当て，これらを重要と見なすルールを用意しておく．ここでは，ルールの一例として，名乗り表現を表すパターンとそのスコアを**図 3** にあげる．最初のタグ<SENTOU>は先頭文パターンを表し，<RULEG nanori>という形態素パターンにマッチすればスコアが 35 だけ付与されることを記述している．<RULEG nanori>は<RULE daresoredesu>と<RULE at_mark>のどちらかがマッチすれば全体がマッチしたと見なされる．これらの中で用いられている HYOU や ICAT は表記と一般名詞カテゴリ番号を表し，NEXT=NEAR と NEXT=FAR はそれぞれ次の形態素が隣接か否かを表している．表記は，論理和の形で記述されている．ちなみに一般名詞カテゴリ番号 363 は「機関」を 5 は「人間」を表している．たとえば，「○○商事の□□です.」のような文に対して，「□□」の一般名詞カテゴリが「機関」や「人間」であれば，このルールはマッチし，名乗りを表現する文という位置付けでスコア 35 が与えられる．スコアは数値自身に意味はなく，相対的な重要度を表現する．我々の作成したルールの中では，35 というスコアは，ほとんどの場合，文のランキングにおいて上位に残る程度の重要性があることを意味する．

3.3 制限文字数内でのランキング上位文の選択

指定される要約率や制限文字数の範囲内で，電子メール本文から重要文を抽出する方法について述べる．基本的には，1 つまたは複数のルールによってスコアが付与された各文に対して，スコアの大きい順にソートする．スコアの正規化およびソートの方法には様々なバリエーションがあるが，本論文では経験的に，1) 文の文字数により正規化されたスコアの総和，2) スコアの総和，3) スコアの最大値，の順でソートを行う．スコアの総和は複数のルールが付与したスコアの総和であり，スコアの最大値はその中で最も大きいスコアである．

要約率が指定される場合には，メールテキストの大きさと要約率から制限文字数を計算する．スコアに基づいてソートされた各文の順序に従って，指定された制限文字数を超えるまで，文を選択していく．このとき，引用行なども含めて出現順に付けられた文番号に基づいて選択された文どうしが連続しない場合には，文が脱落していることを明示するために記号「...」を挿入する．これは特に箇条書きの項目がとびとびに選択される場合などにおいて，内容についての誤解を与えないようにするのに役立つ．これらの記号を含めて選択した文が制限文字数を超える場合は，記号とその文の選択を断念し，文の選択を終了する．なお，要約を 1 行 1 文の形式で出力するために，各文と記号は改行コードを付けて出力し，それぞれの文末の改行コードの文字数も制限文字数の考慮に入れている．

4. 携帯電話向け要約システムへの適用

本論文で提案した重要文抽出手法を携帯電話向け要約システムへ適用する．本手法を適用した重要文抽出エンジンを搭載した SummaryBIFF システムの構成を**図 4** に示す．インターネットサービスプロバイ

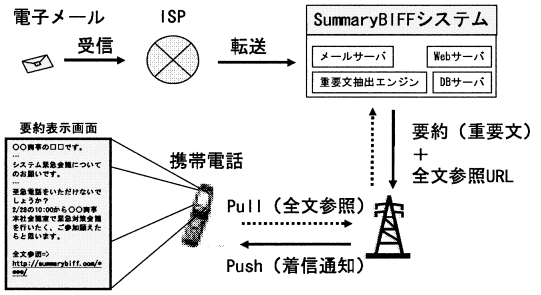

図 4 携帯電話向けメール要約システムの構成
Fig. 4 Architecture of email summarizer for mobile phones.

ダ（ISP）で受信したユーザの電子メールを SummaryBIFF システムへ転送することにより，電子メールの要約（重要文）に原文へのリンク（全文参照 URL）が付与された新たな電子メールが作成される．ユーザ ID と携帯電話のメールアドレスが格納されている DB サーバとメールサーバを経由してユーザの携帯電話に即時に通知される（Push）．ユーザは重要文に目を通してそれが重要な電子メールであると判断すれば，暗号化されたユーザ ID を含む全文参照 URL から Web サーバとユーザ ID が格納されている DB サーバを経由することにより，その場で原文へアクセスできる（Pull）．このように，SummaryBIFF システムでは電子メールを受信するとすぐに要約による着信通知が届き，ユーザ自身が要約に基づいてすべての電子メールを取捨選択できるため，大量に電子メールを受信するユーザでも重要な電子メールだけに効率良くアクセスすることができる．

携帯電話の送信可能文字数が 500 byte と仮定すると，全文参照リンクの文字数を考慮すれば，重要文には 360 byte から 390 byte 程度の文字数しか割くことができない．このため，制限文字数を以上のように設定して重要文抽出エンジンを駆動することが妥当である．以下では，携帯電話向け要約システムを前提とした本手法の評価および考察を行う．

5. 評　　価

本論文で提案した重要文抽出手法の評価を行った．重要文抽出のためのルールは，先頭文ルールが 3 個，冒頭文ルールが 2 個，通常文ルールが 18 個である．通常文ルールの内訳は，設定されているスコアの高い順に，スケジュール通知表現（2 個），名乗り表現☆，依頼表現（3 個），意思を表す表現，疑問表現，期限を表す表現，話題表現，数値表現，主張を表す表現である．そのほかに，これらのルールにマッチしない場合でも順位が付けられるように，助詞「を」を含む表現，助詞「は」を含む表現，表記自身が重要さを表す文字列，否定表現，接続助詞を含む表現，断定表現を表すルールも用意した．ルールは図 3 全体の単位を 1 つと数えている．

評価者 3 名により評価者自身が受信した電子メール 139 通をテストセットとして準備した．テストセットは，メールニュースなどは含まず，特定の個人や集団を相手とした目的のはっきりした電子メールである．携帯電話向け要約システムを想定して，短すぎて要約する意味がないメッセージをテストセットから除外しておく必要がある．我々は評価者による重要文の選択を単純にすることも考慮して，メッセージから重要な 5 文を選択するというタスクを設定したことにより，テストセットは 6 文以上のメッセージとした．このため，本手法では対象外としている署名や引用行や英数字のみの行を除外して，重要文抽出の評価のために文の単位をあらかじめ揃えておく必要から，システムによる文分割を実行した．この結果，無視できない誤分割☆☆が存在した 22 通を除外し，117 通をあらためてテストセットとした．つまり，テストセットにおける文分割の精度は 84%であった．1 文あたりの平均バイト数は 52.8 byte☆☆☆（117 通）であったので，メッセージから 5 文の選択は平均で 264 byte となる．これは，携帯電話向け要約システムを想定したときの制限文字数を超えない適度な長さであると考えられる．

5.1 人手による正解との比較

はじめに，テストセットに対する重要文抽出の難易度を調べるために，テストセットがどのような長さのメッセージから構成されているかを，バイト数を基準にして調べた．また，人手による重要文抽出のコンセンサスの度合いを見るために，評価者 3 名がそれぞれ重要だと思われる 5 文を選択することにより重要文 5 文の正解を作成し，5 文のうち何文が一致するのかについて，その重なりの度合いでテストセットを分類した．3 名による選択文の重なりの度合いがメッセージの長さの違いによってどう変化するのかを見るため，

☆ 名乗り表現は通常先頭文ルールだけで十分であるが，電子メールが転送部分を含むことも考慮して転送元を抽出するために設定している．

☆☆ たとえば，文の中に箇条書きが割り込むような形で存在するような場合などが該当する．

☆☆☆ テストセットの中にはレイアウトのための空白文字が含まれるが，どの単位を文とするかにより不要な文間の空白文字となったり必要な文内の空白文字となったりするため，区別が難しい．そのため，単純に空白文字のバイト数は数えていない．

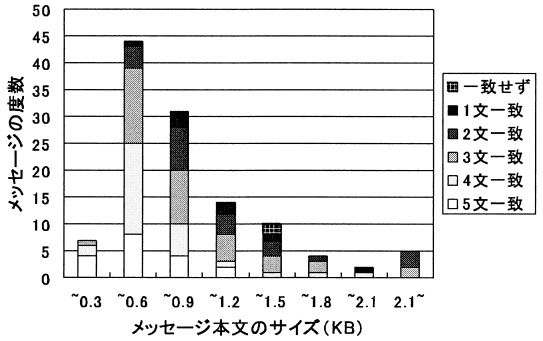

図 5 メッセージの長さの違いにおける複数の人手による選択文の重なり度合いの分布

Fig. 5 Distribution of email message sizes and various levels of consensus among the evaluators for manually selected sentences.

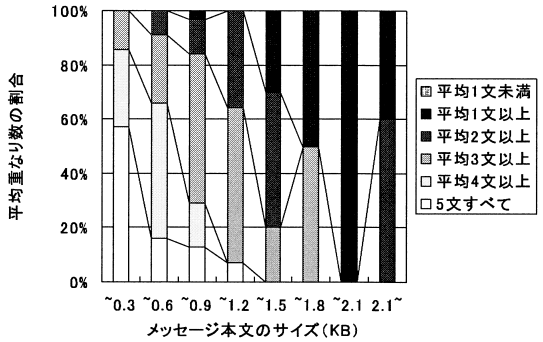

図 6 本手法による選択文 5 文と人手による選択文 5 文の平均重なり数の分布

Fig. 6 Distribution of percentage of correspondence among five sentences chosen by automatic extraction and manual selection from email messages.

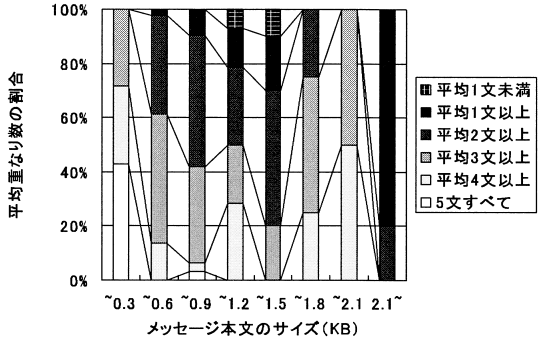

図 7 リード手法による 5 文と人手による選択文 5 文の平均重なり数の分布

Fig. 7 Distribution of percentage of correspondence among five sentences chosen by the lead method and manual selection from email messages.

図 5 にメッセージのバイト数と 3 名の選択文の重なりの度合いの関係を示す．たとえば，表中の「5 文一致」とは，評価者 3 名の選択した 5 文がすべて一致することを表し，「1 文一致」とは評価者 3 名の選択したそれぞれ 5 文の中で一致する文が 1 文であることを表す．テストセットは 1.5 KB 以内のメッセージが多数を占め，コンセンサスがとれるメッセージはバイト数が小さいメッセージに偏り，コンセンサスがとれないメッセージはバイト数が大きいメッセージに分布していることが分かる．これは要約率の大小と関係するためで，メッセージの長さが長くなれば要約率は小さくなり，メッセージの長さが短くなれば要約率は大きくなるからであると考えられる．なお，評価者 3 名の選択文が 1 文も一致せず，まったくコンセンサスがとれないデータは 2 通であった．

次に，本手法により出力される重要文の評価を行った．本手法による重要度順上位 5 文と各評価者の選択した 5 文の重なりを平均し，平均重なり数の割合の分布をメッセージの大きさごとに**図 6** に示す．表中の「5 文すべて」は評価者 3 名のいずれもが選択した 5 文とシステムが抽出した 5 文が完全に一致することを表し，「平均 4 文以上」は評価者 3 名の選択したそれぞれ 5 文とシステムが抽出した 5 文の重なりが平均で 4 文以上 5 文未満であることを示す．「平均 3 文以上」，「平均 2 文以上」，「平均 1 文以上」についても同様で，「平均 1 文未満」は評価者 3 名のそれぞれが選択した 5 文とシステムが抽出した 5 文の重なりが平均で 1 文未満であることを示す．ベースラインは文の数を揃えるために先頭から 5 文を選択したリード手法とする．ベースラインについても同様に平均重なり数の割合の分布を**図 7** に示す．これらより，テストセットの多数を占める，メッセージの大きさが 1.5 kb 以内までは，本手法の方がベースラインを大きく上回ることが分かる．特に，図 5 より，117 通のうち 5 文すべてが一致するのは 1.2 KB 以内に集中する 18 通（15%）であったが，本手法ではこのうちの 16 通（89%）が正解に完全一致するが，ベースライン手法では，4 通（22%）しか完全一致しない．これは，コンセンサスが得られたメッセージ，つまり，誰もがこのデータに対してはこのように要約されると良いと考えるメッセージに対しては，リード手法では不十分であり，本手法は非常に高い精度で重要文を抽出することができる．逆に，1.5 kb を超えるメッセージは母数が少ないものの，ベースラインの方が良い．これは，評価者は先頭から多く重要文を選択したことを意味しており，長いメッセージになると先頭に全体の内容を表す要約が存在するためと考えられる．このことから，極度に長いメッセージを除けば，本手法は人手による選択文をよ

表 1 着信通知タスクにおける重要文による電子メールの取捨選択の精度
Table 1 Precision and recall of important email screening in incoming email notification task.

	評価者 A	評価者 B	評価者 C	合計
正解データ数	83	81	71	235
抽出データ数	75	80	84	239
正解抽出数	74	72	71	217
再現率	89%	89%	100%	92%
適合率	99%	90%	85%	91%

り多く含む適切な重要文を抽出することができるといえる．

5.2 着信通知タスクにおける評価

本手法を携帯電話向け要約システムに適用した場合のタスクに基づく評価を行うために，要約による着信通知として評価を行うときの観点を整理する．要約されたテキストを見て，本当に重要なメールが着信したことがどれくらい分かるのかがポイントだと考えられる．そこで，テストセットの 117 通を対象に評価者 3 名が本手法により電子メールから抽出された重要文と原文を別々に読んで重要かそうでないかをそれぞれ判定し，評価者ごとに本当に重要だと判断される原文に対する再現率と適合率を求めた．ただし，実際のところは送信者の情報などは重要であるが，重要文自体を適切に評価するために，重要文抽出に用いていないヘッダおよび引用部と署名は原文には含まれていない．重要かどうかの判断は，着信通知タスクを考慮しすぐに電子メールを読んでおく必要があるかどうかを基準とした．原文を読んで重要と判断した電子メールを正解データとし，重要文を読んで重要と判断した電子メールを抽出データとした．抽出データと正解データの重なりを正解抽出数，正解データ数に対する正解抽出数の割合を再現率，抽出データ数に対する正解抽出数の割合を適合率と定義した．評価者ごとの再現率と適合率を**表 1** に示す．

原文を読んで重要と判断した正解データが 3 名の合計で 235 通なので，全体（評価者 3 名，117 通）に占める割合は，67%である．3 名の合計で，正解データ 235 通のうち重要文を読んでそれが重要だと判断できた正解抽出数は 217 通だったので，92%の再現率で各ユーザにとって重要な電子メールを網羅できる．一方，3 名の合計で，重要文を読んで重要と判断した抽出データ 239 通のうち，原文を読んで重要だと判断できた正解抽出数は 217 通だったので，91%の適合率で重要文からその電子メールが各ユーザにとって重要だと推定できる．いずれの評価者についても再現率適合率ともに高い値を示しているので，重要文だけを見て電子メールの重要性が判断でき，要約による着信通知タスクにおいては電子メールの取捨選択に十分使えるということがいえる．これは，SummaryBIFF システムを利用すれば，重要な電子メールにだけ原文にアクセスすればよいので，効率的な電子メールアクセスが実現可能となることを意味する．

5.3 アンケートによる評価

ここまでに示した評価は，小規模な主観評価によるものであるため，提案手法の有効性や，実装した重要文抽出ルールの実フィールドでの網羅性を直接実証するものとはいいがたい．しかしながら，メールという個人性の高い情報を対象とすることから，大規模な実験を行うことも困難である．そこで，SummaryBIFF システムを約 20 名のユーザに 3 カ月利用してもらい，その使用感・有効性をアンケートにより分析した．その結果，要約サービスの品質に対して不満を持つユーザは 21%にとどまり，58%のユーザは，要約結果に満足していることが分かった．また，重要文抽出の精度についても 74%のユーザが「重要文はほぼ抽出されている」と回答した．これらより，電子メールに特有な表現とスタイルの特徴を利用したたかだか 23 個のルールでも，本システムはほぼ実用レベルに達しているものと考えられる．

6. 考　　察

本章では，着信通知というタスクにおける重要文抽出の単位について検討する．本論文では抽出の単位を文としたが，文への分割を行うことなく，パラグラフ（空行で区切られる形式段落）を抽出の単位とすることも可能である．仮に，抽出単位をパラグラフとした場合は，文分割の必要がないため原文のつながりがそのまま維持され可読性に優れる反面，制限文字数が限られている場合には他の重要な部分を落としてしまうという欠点が存在する．一方，抽出単位を文にした場合は，文分割時の誤りや脱落文のために文のつながりが失われ可読性が犠牲になるが，短い制限文字数の中においても重要な箇所は網羅できるという利点が存在する．我々は，以上を検証するために，テストセットについてパラグラフに関する調査を行った．

バイト数を基準にして，各評価者の選択した重要文を含むパラグラフの長さの平均値とその重要文のパラグラフあたりに占める割合を調査した結果を**表 2** に示す．比較のため，重要文を含まないものも含めてすべてのパラグラフの個数と長さの平均値についても示した．重要文を含むパラグラフあたりに占める重要文のバイト数の割合は 73.1%であった．しかし，この中

表 2 重要文を含むパラグラフの長さの平均値と重要文の占める割合
Table 2 Average number of bytes per paragraph including selected sentences and average percentage of sentence selection per paragraph.

	すべて	重要文を含む	複数文に限定
個数	947	443	258 (58%)
バイト数 (byte)	105.5	120.9	159.3
重要文の割合 (%)	–	73.1	64.9

には文を単位とした場合と共通となる，1 文だけで構成されるパラグラフが含まれている．1 文だけのパラグラフを比較対象から除外すると，重要文を含むパラグラフのうち 58%のパラグラフが複数の文から構成されることが分かり，重要文の占める割合は 64.9%であった．これは，抽出の単位をパラグラフとした場合には，重要でない部分が 35%程度混入することを意味する．さらに，実際のアプリケーションでは，単純に空行で分割可能なパラグラフを単位とするので，本論文では対象としなかった記号のみの行や英数字のみの行のバイト数も加算されるため，パラグラフ内の非重要な部分はさらに多くなるであろう．

また，表 2 に示すように，すべてのパラグラフの長さの平均値が 105.5 byte であるのに対して，重要文を含むパラグラフの長さの平均値は 120.9 byte であった．一方，すべての文の長さの平均値が 52.8 byte であったのに対して，重要文の長さの平均値は 63.9 byte であったことから，その分だけ重要文を含むパラグラフは長くなったと考えられる．特に，重要文を含む複数文からなるパラグラフだけに限れば，その長さは 159.3 byte となり，重要文の長さである 63.9 byte の約 2.5 倍になる．これは，パラグラフを抽出単位とした場合，少ない制限文字数の中では選択できるパラグラフの個数が少なくなることを意味する．以上より，パラグラフ内の非重要な部分の割合とパラグラフ全体の長さという 2 つの面から，携帯電話向け要約システムへの適用時には，抽出単位を文単位として重要箇所の網羅性を上げる方が有利であると考えている．

7. おわりに

本論文では，電子メールにおける重要文抽出手法を提案し，携帯電話向け要約システムに適用した場合の評価について報告した．本手法の特徴は，電子メールのメッセージに対して，比較的短い文の単位で分割を行い，電子メールに特有な表現とスタイルの特徴を利用したルールを適用することにより，電子メールから重要な個所を網羅的に抽出することである．実験結果とアンケートの分析から，重要文抽出は精度が高く，携帯電話向け要約システムは実用レベルに達していることが分かった．利用コミュニティなどにルールをチューニングすることにより，さらに実用性は高まるものと考えられる．

今後は，文分割の誤りによる可読性の犠牲を最小限に抑えるために文分割の精度をさらに上げ，限られた文字数の中でも重要文の網羅性を高めることに取り組んでいきたい．

謝辞 本論文を執筆するにあたり，本研究の機会を与えていただきました NTT サイバースペース研究所の小原永主席研究員，NTT レゾナントの松岡浩司氏ならびに NTT 東日本の堀井統之氏，SummaryBIFF システムの運用にご尽力いただきました NTT サイバーソリューション研究所の廣嶋伸章研究員，本研究を支援していただいた NTT サイバースペース研究所の皆様，および，貴重なコメントをいただきました査読者の方々に謹んで感謝の意を表します．

参考文献

1) 浅野久子，加藤恒明，高木伸一郎：Signature の局所的パターンマッチによる電子メールからの送信元住所録情報抽出とそれを用いた住所録管理システム，情報処理学会論文誌，Vol.39, No.7, pp.2196–2206 (1998).
2) Banko, M., Mittal, V. and Witbrock, M.: Headline Generation Based on Statistical Translation, *Proc. 38th Annual Meeting of the Association of the Computational Linguistics* (*ACL-2000*), pp.318–325 (2000).
3) Berger, A. and Mittal, V.: OCELOT: A System for Summarizing Web Pages, *Proc. 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR-2000*), pp.144–151 (2000).
4) Corston-Oliver, S.: Text Compaction for Display on Very Small Screens, *Proc. Workshop on Automatic Summarization at the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics* (*NAACL-2001*), pp.89–98 (2001).
5) Fuchi, T. and Takagi, S.: Japanese Morphological Analyzer Using Word Co-occurrence — JTAG, *Proc. 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics* (*COLING-ACL'98*), pp.409–413 (1998).
6) 福本淳一，桝井文人：メールロボ：インターネットメールからの情報抽出，沖電気研究開発第 181 号，Vol.66, No.2, pp.55–58 (1999).
7) 畑山満美子，松尾義博，白井 諭：重要語句抽出

による新聞記事自動要約，自然言語処理，Vol.9, No.4, pp.55–70 (2002).
8) 伊知地宏，倉部　淳：メイルを用いたソフトウェア開発を支援するツール，情報処理振興事業協会（ipa）平成13年度成果報告集，情報処理振興事業協会（IPA）(2001). http://www.ipa.go.jp/NBP/13nendo/re-ports/explorafft/mailide/mailide.pdf
9) 池原　悟，宮崎正弘，白井　諭，横尾昭男，中岩浩巳，小倉健太郎，大山芳史，林　良彦（編），NTTコミュニケーション科学研究所（監修）：日本語語彙体系，岩波書店 (1997).
10) 稲垣博人，早川和宏，井上孝史，田中一男：モバイル端末の表示特性に応じたメッセージ要約方式の提案，情報処理学会第56回全国大会講演論文集（分冊2），pp.255–256 (1998).
11) Lam, D., Rohall, S.L., Schmandt, C. and Stern, M.: Exploiting E-mail Structure to Improve Summarization, IBM Watson Research Center Technical Report 02-02, IBM Watson Research Center, Cambridge, MA, USA (2002).
12) Luhn, H.: The automatic creation of literature abstracts, *IBM Journal of Research and Development*, Vol.2, No.2, pp.159–165 (1958).
13) 望月　源，奥村　学：読みやすさの向上と冗長性の排除を考慮した重要箇所抽出型要約，情報処理学会自然言語処理研究会報告139-3，pp.17–24 (2000).
14) 村越広亨，山見太郎，島津　明，落水浩一郎：電子メールを利用した学習者間のコミュニケーション支援技術の開発，教育システム情報学会誌，Vol.18, No.3-4, pp.308–318 (2001).
15) Muresan, S., Tzoukermann, E. and Klavans, J.: Combining Linguistics and Machine Learning Techniques for Email Summarization, *Proc. 5th Workshop on Computational Language Learning* (*CoNLL-2001*), pp.152–159 (2001).
16) 中川裕志，渡部聡彦：携帯端末向けコンテンツ変換と自然言語処理，情報処理，Vol.43, No.12, pp.1300–1304 (2002).
17) 仲尾由雄：見出しを利用した新聞・レポートからのダイジェスト情報の抽出，情報処理学会自然言語処理研究会報告117-17，pp.121–128 (1997).

(平成15年6月18日受付)
(平成16年5月11日採録)

長谷川隆明（正会員）

1969年生．1994年慶應義塾大学大学院理工学研究科計算機科学専攻修士課程修了．同年日本電信電話（株）入社．2003年から2004年にかけて，ニューヨーク大学客員研究員．現在，NTTサイバースペース研究所研究主任．自然言語処理，情報抽出の研究開発に従事．日本ソフトウェア科学会会員．

林　　良彦（正会員）

1959年生．1983年早稲田大学大学院理工学研究科博士前期課程修了．同年日本電信電話公社横須賀電気通信研究所入社．2004年3月NTTサイバースペース研究所退社．この間，1994年〜1995年スタンフォード大学言語情報研究センター滞在研究員．2004年4月大阪大学大学院言語文化研究科言語情報科学講座教授．博士（工学）．自然言語処理・知的情報アクセスの研究に従事．電子情報通信学会，人工知能学会，言語処理学会の各会員．

山崎　毅文（正会員）

1963年生．1986年東京大学工学部物理工学科卒業．同年日本電信電話株式会社入社．以来，機械学習，自然言語処理の研究に従事．現在，（株）NTTレゾナントにおいて，映像コミュニケーションサービスの事業化推進担当．人工知能学会会員．