電子メールの重要文抽出と分類

角谷 由貴， 森田 和宏， 泓田 正雄， 青江 順一
徳島大学工学部知能情報工学科
E-mail: {kadoya, kam, fuketa, aoe}@is.tokushima-u.ac.jp

1. はじめに

インターネットの爆発的な普及に伴い，電子メールを利用する機会が多くなっている．しかし大量のメールを受信すると，それらすべてを確認して必要な情報を把握するのは困難である．このため，受信した大量のメールの中から自動的に必要な用件だけを抽出する自動要約技術が求められている．

自動要約の基本かつ主要な手法としては，テキスト中から重要な文を抽出して要約文とする重要文抽出法がある[1][2]．重要文抽出法では，何が重要であるかをどのように決定するかが問題となる[3]．このため，従来のメール要約では，スケジュール情報に特化した要約など[4]，通知内容が限定されたものしか対象とされていない．

本稿では，緊迫性や返信の必要性を持つ用件を重視し，時間表現や感情表現を抽出するとともに，依頼や質問などの表現パターンに着目してメールを分類することにより，一般的なメール文書の重要文抽出をおこなう手法を提案する．

2. 電子メールの重要表現

2.1 電子メールの重要文

2.1.1 通知内容を示す表現

一般にメール文書には，送信者から受信者に伝えたい用件が書かれており，この用件の内容によって，重要なメールかどうかが決まる．このため，メールの主題となる用件が把握できる文は重要文である．本手法では，8 つの分類を定義し，それぞれの分類を示す表現を抽出して重要度を判定する．以下にメール文の分類と表現例を示す．

- **通知**：スケジュールの通知や変更など
  例）次回のゼミは金曜日に変更します．
- **依頼**：仕事の依頼や提出物の催促など
  例）アンケートを提出してください．
- **質問**：問い合わせや相談，都合を伺う文など
  例）いつが都合がよろしいでしょうか？
- **苦情**：苦情や不満を含む問い合わせなど
  例）度々の納品遅延，甚だ困惑しております．
- **誘い**：イベントへの招待や勧誘など
  例）金曜日に一緒に飲みに行きましょう．
- **返事**：了解や断りなどの返事
  例）残念ながら欠席させていただきます．
- **励まし**：応援や慰めなど
  例）頑張ってください．
- **挨拶**：挨拶やお礼，祝いの表現など
  例）お久しぶりです，お元気ですか？

2.1.2 時間表現

イベント日程や作業の締め切り期限など，メール文書にはさまざまな時間情報が含まれる．急な仕事や変更事項など，早急に対処しなければならない用件が書かれている文は重要文である．以下に時間表現の例を示す．

- 日時を指定する表現
  例）3 月 15 日，10 時 30 分～
- 曜日の表現
  例）月曜日，火曜，（水）
- 現在からの相対的な時間を指定する表現
  例）今日，明日，来週，今月末
- 急ぎの副詞表現
  例）至急，急いで，早めに

2.1.3 感情表現

メール文書では，相手の感情を理解した上で適切に返事をする必要がある．例えば，「哀しみ」の感情が含まれている場合は慰めや励ましを，「怒り」が含まれている場合は謝罪やなだめの言葉を送る必要がある．特に，送信者が「怒り」の感情を抱いている場合は，早急に対応しなければ人間関係に支障をきたす恐れがある．本手法では，「喜・怒・哀・楽・驚き」の 5 つの感情を示す表現を抽出する．以下に，感情表現の例を示す．

- **喜び**：満足，感謝，祝いなどの表現
  例）おかげさまで無事合格しました．
- **怒り**：苦情，不満，迷惑などの表現
  例）料金が高すぎます．
- **哀しみ**：残念，心配，謝罪などの表現
  例）精一杯頑張ったけど，不合格でした．
- **楽しみ**：楽しいことや期待などの表現
  例）楽しみに待ってます♪
- **驚き**：驚いたことを表す表現
  例）広くてびっくりしました．

2.2 電子メールの不要文

2.1 節では，メール文書における重要表現について示した．これとは逆に，重要性の低い文であると判断できる表現も存在する．本稿では，用件とは関係のない文を不要文と定義し，不要文を判定することにより，重要文抽出の精度を向上させる．不要文と判定する内容を以下に示す．

- 署名
- 引用や転送内容
- メール文書に付属する広告

署名は，送信者の情報を得るには有用であるが，メールの用件とは無関係である．引用や転送は，以前にやり取りされた内容であるため，送信者が書いた本文に比べて重要性が低い．また，無料のメールアドレスやメーリングリストでは，メール本文の末尾に広告が挿入されることがあり，この付属広告についても，用件とは無関係である．図 1 に，メール文書中に含まれる不要な内容の例を示す．

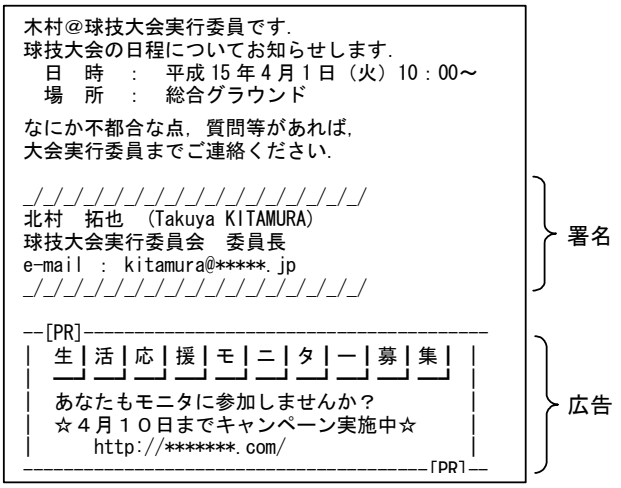

図 1. 電子メールの不要文の例

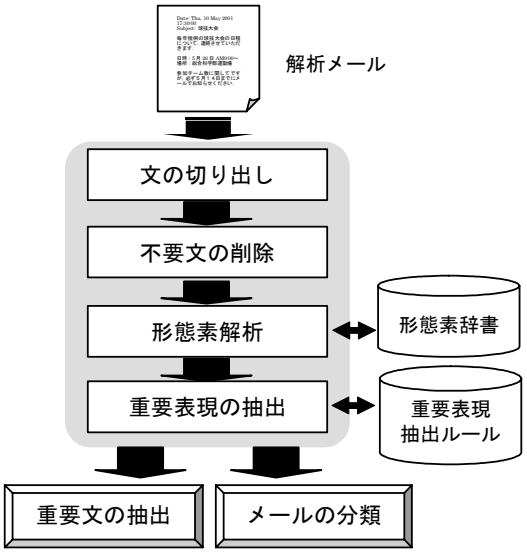

図 2. システム概要図

3. システム概要

本手法の重要文抽出手順を図 2 に示す．

(1) **文の切り出し**
1 文の区切り位置を判定し，文を切り出す．

(2) **不要文の削除**
署名，広告，引用などの不要文を判定する．

(3) **形態素解析**
不要文以外に対して，形態素解析を行う．

(4) **重要表現の抽出**
ルール照合によって時間表現，感情表現，通知内容の表現などの重要表現を抽出し，文の重要度と分類を判定する．

(5) **重要文の抽出とメールの分類判定**
抽出した重要表現の重要度から，重要文を決定して抽出する．また，抽出された重要文の分類によってメールの分類を判定する．

4. 重要文抽出法

4.1 通知内容を示す表現の抽出

重要表現のパターンは，形態素の表記や品詞情報を用いて記述する．さらに，抽出された表現の意味情報を用いて，意味と意味との組み合わせパターンを照合し，文の重要度と分類を求める．本稿では，1 段階目のルール照合を形態素照合，2 段階目のルール照合を意味照合と呼び，それぞれの照合で用いるルールを，形態素照合ルール，意味照合ルールと呼ぶ．

4.1.1 形態素照合

形態素照合では，形態素を組み合わせた意味の固まりを抽出する．例えば，「くださ＋い」「いただ＋け＋ませ＋ん＋か」「お＋願＋い＋し＋ます」などの形態素列を，依頼の文末表現と判定する．また，ルールによって，抽出した表現の重要度付けを行う．形態素照合ルールは，以下のような情報を持つ．

- 形態素情報を用いた重要表現のパターン
- 抽出する表現の重要度
- 抽出する表現の意味分類
- ルールの優先度

抽出する表現の意味分類は，依頼や質問などの通知内容と，目的語や用言の語幹，文末表現などの文法的な位置付けによっておこなう．例えば，「筆記用具を持参してください.」という文では，「(名詞) ＋を」から＜目的語＞，「持参」から＜依頼語幹＞，「く

ださ＋い」から＜依頼文末＞という 3 つの形態素パターンの意味分類が抽出される．

表現の重要度は，1 章で述べたように，緊迫性や返信の必要性に基づいて定める．例えば，「依頼」の文に対しては，受信者が依頼された内容を実行したり，報告や断りの返事をしたりする必要性があるため，重要度を高くする．逆に，「挨拶」の文に対しては緊急性も返信の必要性もないため，重要度を低くする．ルールの優先度は，抽出するパターンが重複している場合に，どちらのルールを適用するかを判断するための情報である．本手法では，約 500 パターンの形態素の組み合わせ表現を，約 100 種類の意味分類で抽出する．表 1 に形態素照合による意味分類の例を示す．

4.1.2 意味照合

意味照合では，形態素照合で抽出した意味分類を組み合わせて，文の分類を判定する．意味照合ルールは以下の情報を持つ．

- 重要表現の意味情報の組み合わせパターン
- 抽出する表現の重要度
- 抽出する表現（通知内容）の分類
- ルールの優先度

意味照合ルールの抽出方法やルールの役割などは形態素照合ルールと同様であるが，抽出するパターンは意味分類の組み合わせとなっている．形態素照合で得られる約 100 種類の意味分類を用いて，約 150 パターンの意味の組み合わせを判定し，通知内容の分類を決定する．意味の固まりを組み合わせることにより，部分的に目的語，動詞，文末などが同じ表現であっても，1 文としては異なる分類を判定できる．表 2 に，「ください」を含む文および「参加」を含む文の分類判定例を示す．

表 1. 形態素照合による意味分類の例

表現例	意味	重要度
開催，おこな（う），変更，中止	通知語幹	50
し+ます，予定+です	通知文末	10
くださ+い，お+願+い+し+ます	依頼文末	40
頑張（る），元気+を+出（す）	励まし	10

表 2. 意味照合による分類判定例

表現例	意味照合ルール	分類
開催してください	通知語幹＋依頼文末	依頼
頑張ってください	励まし＋依頼文末	励まし
いい加減にしてください	苦情＋依頼文末	苦情
参加してください	参加＋依頼文末	依頼
参加しますか	参加＋質問文末	質問
参加しませんか	参加＋誘い文末	誘い

表 3. 残り時間による時間表現の重要度

残り時間	重要度
1,440 分未満（1 日未満）	100
1,440 分～4,320 分未満（3 日未満）	70
4,320 分～1,0080 分未満（7 日未満）	50
10,080 分以上	20

表 4. 感情表現の重要度

感情の分類	喜び	怒り	哀しみ	楽しみ	驚き
重要度	10	35	10	10	10

4.2 時間表現の抽出

時間表現の抽出は，4.1 節で述べた通知内容の表現と同様に，形態素の表記や品詞情報を用いて記述したルールとの照合によっておこなう．2.1.2 節で述べた時間表現をメール本文から抽出し，現在時刻と比較して期限までの残り時間を算出する．この残り時間に応じて，時間表現を含む文に重要度を付与する．表 3 に，残り時間よる時間表現の重要度を示す．

4.3 感情表現の抽出

感情表現の抽出は，通知内容や時間表現と同様に，形態素の表記や品詞情報を用いて記述したルールとの照合によっておこなう．2.1.3 節で述べた感情表現をメール文から抽出し，感情の分類に応じて重要度を付与する．表 4 に，感情表現の重要度を示す．

4.4 不要文判定方法

4.4.1 署名の判定

電子メールの署名には，一般に，送信者の氏名や所属，アドレスなどが書かれている．このため，以下に示すような署名の特徴語を検出することで，署名を判定する．

- メールアドレスや電話番号
  「@」を含む半角英数字および記号の列
  「e-mail」「TEL」「FAX」などの文字列
- 住所
  「県」「市」「町」「村」「〒」などの文字
- 所属
  「大学」「会社」「部」「係」などの文字

また，署名は一般にメールの末尾に書かれている．このため，まず，メール末尾から記号列などで区切られた段落を抽出し，この段落中に上記の署名の特徴語が集中して含まれている場合，区切りの記号列も含めてその範囲にある文を全て署名とする．

4.4.2 引用や転送内容の判定

電子メールの引用部分や転送部分には一般に行頭に「>」の記号が付与される．また，引用内容の直前に「Original message」や「～wrote:」のような，メッセージがよく見られる．これらの記号や文字列を検出することで，引用部分を判定する．

4.4.3 メール文書に付属する広告の判定

広告には，人目を引くように派手な文字飾りや勧誘の言葉がよく見られる．これらの特徴を検出する

ことにより，広告部分を判定する．以下に広告の特徴を示す．

・●，△，■，☆，┃，ー　などの特殊記号
・「当たる」「お得」「キャンペーン」などの特徴語
・URL

また付属広告は，一般に署名と同様にメール末尾に挿入される．このため，記号列などで区切られた段落を切り出し，この段落に対して上記の広告の特徴の検出をおこなう．

5. 要約結果の評価

5.1 分類判定の評価

本手法のメール文の分類判定の評価をおこなう．実験データは，電子メール 681 通の 5,859 文である．正解データは，人手によって分類を判定したものとする．表 5 に，メール文の分類判定結果を示す．正解データに含まれていない 2,559 文は，2.2 節で述べた不要文や，どの分類にも属さない文である．

5.2 重要文抽出の評価

本手法の重要文抽出の評価をおこなう．実験データは，業務用のメール 366 通と，私用のメール 315 通である．本文中から重要文が適切な数だけ抽出されているかどうかを ABC の 3 段階で評価する．評価基準は，以下のように定める．表 6 に重要文の抽出結果を示す．

A 判定：重要文が必要な数だけ抽出されている．
B 判定：重要文は抽出できているが，主要な用件ではない文も抽出されている．
C 判定：重要文が抽出できていない．

5.3 考察

分類判定の評価では，全ての分類で適合率，再現率とも 80%を超えており，本手法の分類の有効性が確認できた．誤判定例としては，以下のようなものがある．

表 5. メール文の分類判定結果

	正解データ	適合率	再現率
通知	1,696	88%	97%
依頼	738	96%	97%
質問	108	86%	85%
苦情	259	86%	83%
返事	42	85%	81%
誘い	44	89%	89%
励まし	82	90%	87%
挨拶	331	84%	97%
計	3,300	89%	95%

表 6. 重要文の抽出結果

	A 判定	B 判定	C 判定
業務用	59%	25%	16%
私用	45%	15%	40%

a）組み合わせによって意味が異なる表現
　例）ご了承いただけますか？
b）状況によって意味が異なる表現
　例）デモプログラムの作成を頼みます。

a）の例文は『質問』の文であるが，『依頼』と誤判定されていた．「いただけますか」は，「提出」や「準備」などの単語との組み合わせでは，『依頼』となる．しかし，「了承」などの承諾を求める単語との組み合わせでは『質問』と判定するのが適当である．また，b）の例文は『依頼』の文であるが，『通知』と誤判定されていた．「頼みます」という表現は，受信者に依頼している場合は『依頼』の文であるが，第 3 者に依頼することを伝えている場合は，『通知』の文となる．このように，状況によって意味が異なる表現は，判定が困難である．

重要文判定の評価は，業務用メールでは良好な結果が得られており，本手法の有効性が確認できた．しかし，私用メールでは，重要文が抽出できていない場合も多くも見られた．この原因として，私用メールでは挨拶や日記のようなメールが多いなど，業務用メールに比べて用件そのもの重要性が低いことがあげられる．

6. まとめと今後の課題

本稿では，メール文書における重要表現に着目して重要文を抽出する手法を述べた．重要表現としては，時間表現や感情表現，通知内容を示す表現をあげ，これらの表現パターンや意味の組み合わせを考慮して表現の重要度と分類を判定するルールを作成した．また，署名や広告などの不要文を判定して削除することで，重要文抽出の精度を向上させた．

今後の課題として，意味照合ルールの拡充により，さらに正確な分類判定をおこなうとともに，私用メールの重要文抽出の精度を向上させることがあげられる．

参考文献

[1] 奥村学，難波英嗣：“テキスト自動要約に関する研究動向”, 自然言語処理 6(6), pp.1-26, 1999.
[2] 森辰則：“検索結果表示向け文書要約における情報利得比に基づく後の重要度計算”, 自然言語処理，9(4)，pp.1-32，2002.
[3] 長谷川隆明，高木伸一郎：“電子メールコミュニケーションにおけるスケジュール情報の抽出”, 自然言語処理研究会, NL123-10, pp.73-80, 1998.
[4] 池田崇博，奥村明俊：“テキストからのスケジュール情報の抽出と自動通知”, 言語処理学会　第 6 回年次大会 ワークシップ論文集，pp.49-55，2000.