1E2-02

類似度情報を用いたグループ化による 電子メール返信文下書き自動生成手法について

Automatic Generation Method of Draft for Reply on E-mail Using Group Based on Similarity Degree

樋口 英司[*1] Eiji HIGUCHI　　荒木 健治[*1] Kenji ARAKI

[*1] 北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology Hokkaido University

In this paper, we have proposed a method of automatic generation of reply to new received e-mail. Especially, we use the similarity degrees between a new received e-mail and a previous received e-mail for judging the group for the new received e-mail. Moreover, we carry out the evaluation experiment using the proposal method, and describe the future problems.

1. はじめに

現在様々な場面で情報化が進み，それに伴い電子メールの利用者が急速に増大している．電子メールの普及率は，一般世帯において 2002 年の 7 月において 62.4%にまで増加している[1]．また，携帯電話を利用した電子メールの利用ということを考えると，さらに多くの人が電子メールを利用していると考えられる[2]．電子メールの利用には，ビジネスなどのフォーマルなものから，個人が行う日常的なコミュニケーションとしてまで幅広く利用されている．今後も，個人が行うコミュニケーションとしての電子メールの利用は増大していくと考えられる．このように電子メールのやりとりが増大するに伴って，その電子メールの処理には多大な労力や時間を要するようになった．

そのような背景の中，これまでさまざまな電子メールに関する研究が行われてきた．電子メールに関する研究としては，膨大なメールから有用なメールだけを選択しようとする情報フィルタリングの研究[3]や，蓄積された膨大な生データから価値ある情報を発掘するための，データベースからの知識獲得に関する研究[4]，本文内容から筆者を推定する研究[5]などが行われてきた．しかし，これまで受信メールに対して返信文を自動的に生成するといった研究はほとんど行われていない．実際には受信メールに対して返信文を作成するという作業は，労力や時間が非常にかかってしまう．そこで我々は実例から共通差異部抽出手法によって獲得したルールを用いて返信文を生成するシステムについて研究を行ってきた[6][7]．しかし，従来手法では共通部分として代名詞や当り障りのない形容詞（いい etc）のみを獲得するルールが多く，生成された返信文においても代名詞のみの生成文など実用に際して不十分な内容であることも多く，またルールの獲得が進むにつれこのような出力が増大した．

そこで本稿では受信メールに対し過去の履歴データとの類似度を計算し，メールデータをグループに分割することによりルールの獲得，適用対象を絞り込むことにより精度の向上を目指す．またルール獲得する対象文との類似度の値によってルールにランクを設定し，ルール適用の際にランク上位のルールから適用を行うことで代名詞等の適用を抑制する．本稿では実験を通じて提案手法の評価を行い，本手法の有効性について考察を行う．

連絡先：北海道大学大学院情報科学研究科
メディアネットワーク専攻言語メディア学研究室
札幌市北区北 14 条西 9 丁目，
Tel：0117067389
e-mail：higuchi@media.eng.hokudai.ac.jp

2. システム概要

今回提案するシステムの構成図を図1に示す．システムは次の六つから構成される．

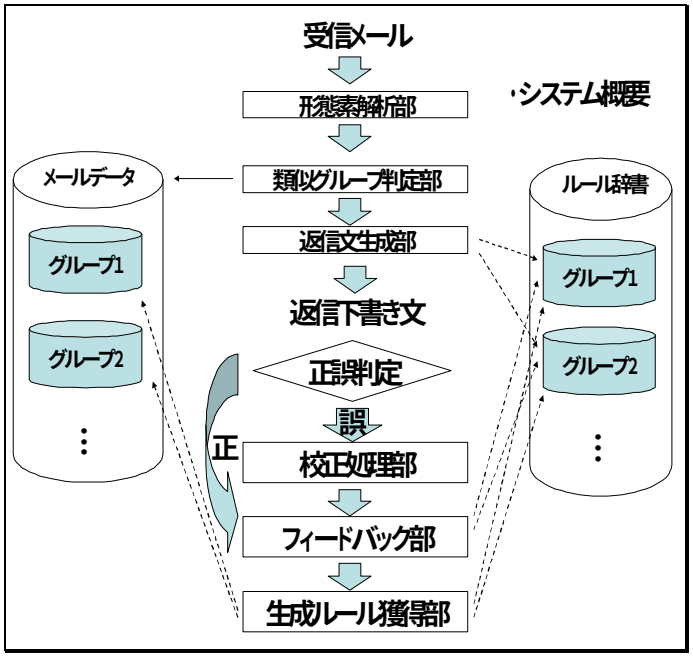

図1 システム構成図

形態素解析部　：入力された受信メール文の単語分割を行い，品詞情報を付与する．

類似グループ判定部：受信メールに対し，過去の履歴データとの類似度計算を行い，所属グループを決定する
返信文生成部　：解析を行った受信文と生成ルール辞書に保持している生成ルールとの比較を行い、使用する生成ルールを決定し返信文を生成する．
校正処理部　：出力された返信文候補のうち、不完全である文に対してはユーザーが人手で校正を行う．
フィードバック部　：返信文生成に使用した生成ルールが適切であるか否かにより，生成ルールの正適用回数，誤適用回数をそれぞれ増加させる．
生成ルール獲得部　：入力された受信文と校正済みの返信文の組を過去に入力されたメールデータの組と比較し、生成ルールを獲得する．その後，メールデータの格納を行う．

2.1 類似グループ判定部

類似グループ判定部では，新規受信メール文と過去の履歴データ中の受信メール文との間で類似度計算を行い，その結果により新規受信メールの所属グループを決定する．

類似度の計算手法として，文書データ D の本文中に出現する単語に対し tf･idf[8]値を計算，各単語に対する tf･idf 値をベクトルの要素としてベクトルを生成し，これを文書 D の文書ベクトルとする．その後比較するニ文書の文書ベクトル間のコサイン距離計算より類似度を判定する．tf･idf 値計算では各単語の重み付けを行うため，代名詞よりもその他の自立語を重視して類似度計算が行われると思われる．

新規受信メール文書 M 中に出現した単語wに関するtfと idf 及び v(w)の値を以下の式(1)～(3)で求める．

$$\mathrm{tf}(w) = \text{M中でのwの出現数}/\text{総単語数} \quad (1)$$
$$\mathrm{idf}(w) = \log(\text{全文書数}/\text{wの出現文書数}) \quad (2)$$
$$v(w) = v(w)\cdot \mathrm{idf}(w) \quad (3)$$

メール文書を構成する全単語に対し v(w)を求め，文書ベクトル V(M)を生成する(4).

$$V(M) = (v(w_1), v(w_1), v(w_1), \cdots) \quad (4)$$

生成された V(M)と，過去の履歴におけるグループの代表ベクトル V (G) との間での類似度 S (M, G) をベクトルのコサイン距離を用いて式(5)により求める．

$$S(M,G) = V(M)\cdot V(G) / \|V(M)\| \cdot \|V(G)\| \quad (5)$$

各グループに対し類似度 S を計算し，予め設定した閾値以上で最大のグループを新規受信メール M のグループとする．閾値を超えるグループが存在しない場合には新規グループを生成する．

2.2 ルール構成

ルールの構成を以下に示す．

- 受信部：受信文の比較から獲得された共通部分
- 返信部：返信文の比較から獲得された共通部分
- 差異部：返信文のうち，獲得対象となった文から共通部分を変数化した差異部分
- ルール保持値：正適用回数，誤適用回数，文中の出現単語をもとに計算される文の対応関係の値，ルールの基となる文を構成する単語数，そして生成ルール中返信部の一致部分の単語数を持つ．
- ランク：ルール獲得の際に計算された類似度を予め設定した各ランクの閾値と比較し決定

2.3 返信文生成部

類似グループ判定部で得られたグループ G_M を基に返信文生成部では返信文の下書きを生成する．まず，適用ルール判定を行う．ここで G_M に所属するルールを適用可能なルールとする．G_M 内のルールの中で，ランクの高いルールより適用可能判定を行い，使用するルールを決定，返信文の生成に移る．

返信文の生成は返信部をベースとしルールの追加適用を行いながら生成する返信部ベースの生成手法と差異部分をベースとし返信部と差異部を組み合わせて生成を行う差異部ベースの手法という二通りの生成手法によって行う．これは，返信部ベールの生成手法では返信文において最適と思われる部分のみを出力とし，差異部ベースの生成手法では完全な文を出力とすることを目的としている．

2.4 生成ルール獲得部

生成ルール獲得部では新規受信メールとそれに対する返信メールの組と過去の履歴中の受信･返信メールの組との比較を行い，ルールの獲得を行う．このとき，比較対象となるのは新規受信メールの所属するグループと同一のグループの履歴データのみである．これはルールの獲得対象を類似度の高いグループ間にのみ限定することで返信文の生成に不適切なルールの獲得を抑制するために獲得対象の限定を行った．

続いて受信メール同士の比較から共通部分を，返信メール同士の比較から共通部分，差異部分を抽出しルール内容を獲得する．このときルールのランクの決定も行う．まず比較対象との間での類似度を式(5)に従い計算する．計算結果の値と各ランクの閾値を比較しランクを決定する．

3. 評価実験

第一著者の携帯電話での受信メールと返信メールの組 150 組を用いて評価実験を行った．形態素解析ツールとして JUMAN[8]を用いた．生成された出力に対し評価を行った．予備実験より最適な値として今回は類似度の閾値を 0.15 に設定し，各ランクは閾値の整数倍を境界とした．出力内容に応じ以下の 5 つに分類する．

5．出力内容が全て本来の受信文に含まれる
2. 出力内容の一部を削除することで本来の受信文に含まれる
3. 本来の受信文には含まれないが，追加しても問題のないもの
4. 半分以上の削除が必要なもの
5. 未出力

正適用を 1～3，誤適用を 4, 5 とし，式(6)～(8)に基づき適合率，再現率，F 値を求める．

再現率(R) = 出力生成回数/入力総数　(6)
適合率(P) = 正適用回数/出力生成回数　(7)
F値 = 2• R • P/R + P　(8)

3.1 実験結果

出力の内訳は表1のようになった．また，共通部をベースにした生成手法，差異部をベースにした生成手法それぞれの再現率，適合率，F 値の値を従来手法と今回提案した手法の間で比較したものを表 1，2 に示す．

	分類1	分類 2	分類 3	分類 4	分類 5
共通部ベース	12	4	8	42	84
差異部ベース	4	5	0	44	97

表1　出力内容の内訳

	従来手法		提案手法	
	全体	最終 50 組	全体	最終 50 組
再現率	57.3	80.0	44.0	62.0
適合率	25.6	35.5	36.4	38.7
F 値	35.3	40.9	39.8	47.7

表2　共通部ベースの生成手法による結果

	従来手法		提案手法	
	全体	最終 50 組	全体	最終 50 組
再現率	46.0	74.0	35.3	64.0
適合率	18.8	16.2	17.0	18.8
F 値	26.7	26.6	22.9	29.0

表3　差異部ベースの生成手法による結果

共通部ベースの生成手法では，従来手法では F 値が全体で 35.3%，最終 50 組で 40.9%となり，提案手法では全体で 39.8%，最終 50 組で 47.7%という結果が得られた．また，差異部ベースの生成手法では従来手法で F 値が全体で 26.7%，最終 50 組で 26.6%となり，提案手法では全体で 22.9%，最終 50 組で 29.0%であった．今回の実験で獲得されたルールの総数は 682 個で，そのうちランク1が 153 個，ランク2が 404 個，ランク3が 122 個，ランク4が 3 個という結果になった．またグループ数は 18 個であった．

4. 考察

共通部ベースの生成手法において主語（俺 etc）のみの出力であった割合は，類似度によるグループ化適用前では 23.2%であったのに対し適用後では 9.4%へと減少した．これはルール間にランクを設けることにより返信文の生成に関する部分が主語のみといった生成ルールの適用が抑えられたためだと考えられる．また，適合率，F 値の面でも若干の向上がみられた．しかし，差異部ベースの生成手法や共通部ベースの生成手法における追加適用を行った際の結果では改善は少なかった．これは，ルールをグループに限定，ランクを優先して適用を行うため複数ルールを組み合わせて生成を行った際に不適格な組み合わせが生まれたためである．特にランク最上位のルールに適用可能なルールが存在した際に，システムは最上位ランクのルールのみ組み合わせの対象としてしまうので，適した文を生成するには不十分であったことが考えられる．また，ルールの獲得，適用をグループ内に限定したため再現率は類似度によるグループ化適用前よりも共通部ベースで 13.3%，差異部ベースで 10.7%低い値となっている．

5. まとめ

本稿では，返信文生成におけるルールの獲得，生成の精度向上を目指し，新規受信メールと過去の履歴データとの類似度を計算し，類似度情報に基づくデータ，及びルール獲得，適用時の対象のグループ化を導入した．そして実験を通じて有効性の検討を行った．共通部ベースの生成手法では再現率，適合率，F 値の面で上昇をみせ，また代名詞のみの出力割合の面でも 13.8%ほど減少し，文中で省略されることも多い代名詞を提示するのではなくその他の名詞や同士を提示するようになり内容面での向上がみられた．一方差異部ベースの生成手法では大きな変化はみられなかった．今後の課題として，ルール適用の際にランクの高いものを優先することと同時にルール間の組み合わせを考慮し低ランクのルールからも条件によっては採用することが挙げられる．また，今後は返信文の生成だけではなくメール処理に関する機能の全体を踏まえた統合システムの構築も検討していきたい．

参考文献

[1] 日経 BP：日本のネット人口(2002 年 7 月), http://www.nikkeibp.co.jp/.

[2] 携帯電話/IP 接続サービス（携帯）/PHS/無線呼び出し契約数, http://www.tca.or.jp/japan/database/daisu/.

[3] 安藤一秋，青江順一，獅々堀正幹:多属性項目の履歴情報に基づく電子メイル文書のフィルタリング手法，自然言語処理 133，情報処理学会，1999.

[4] 山見太郎，村越広享，島津明，落水浩一郎:電子メールを利用したコミュニケーションにおける討議スレッド自動抽出手法の実装と評価，自然言語処理 137，情報処理学会，2000.

[5] Malcolm Corney, Olivier de Vel, Alison Anderson, George Mohay:Gender-Preferential Text Mining of E-mail Discourse, 18th Annual Computer Security Applications Conference(ACSAC), 2002.

[6] 樋口英司，荒木健治:電子メールにおける返信文自動生成手法の有効性について，北海道情報処理シンポジウム 2004 講演論文集，情報処理学会北海道支部，2004.

[7] Satoshi Uematsu , Kenji Araki , Koji Tochinai: Improvement of Automatic Generation Method of Reply Sentence by Inductive Learning Using Common Portions on E-mail , Proceedings of FIRST INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY & APPLICATONS (ICITA2002), 2002.

[8] 徳永健伸:情報検索と言語処理，東京大学出版会，1999.

[9] 黒橋禎夫，長尾真:日本語形態素解析システム JUMAN，1999.