2S-6

文体に着目したメール分類システムとその応用

辻野 友孝　白松 俊　大囿 忠親　新谷 虎松
名古屋工業大学大学院工学研究科情報工学専攻

1　はじめに

メールは便利なコミニュケーション手段として多くの人に利用されている．メールはコミニュケーション以外にもファイルの保管庫，バックアップ，メモなど様々な用途で利用されている．また，メールの中にはスパムメールなど必要の無いメールも多数存在する．そのため，ユーザは受信した大量のメールの中から必要な情報を取り出す必要がある．

本稿ではメールの文体に着目したメール分類を行うシステムを提案する．メールの分類を行う事でユーザのメールの閲覧，整理を支援し閲覧性の向上を図る．

2　メール分類手法

多くのメーラではキーワードを利用した分類を行うことができる．指定したキーワードにマッチするメールを分類する．この方法ではキーワードにマッチする物は必ず分類できるという利点がある，しかし，ユーザの求める分類ルールを作成する為に数多くのキーワードの登録が必要な場合や，分類する為のキーワードが分からない場合が考えられる．

本システムはメールの文体に着目した分類を行う．文体の特徴とは，文字や単語，文節，文，段落など文章の要素に関する書き手の独特の構成パターンである [1]．西原らは，文末表現に着目し，助詞・助動詞の組み合わせから発話分の役割を同定し，発話文の役割から人間関係の一つである上下関係を推定している [2]．金らは書き手が読点をどこに打つかによって書き手の特徴が表れると実証している [1]．また，形態素 N-gram は著者の推定に有効だと述べている．

本研究では文体の特徴の中でも隣接する語句と位置を利用して分類を行った．隣接する語句として形態素 N-gram を利用し，位置として句読点の直前の語句のみを利用した．また，本システムではユーザが受信メール一覧からメールを選択することにより分類ルールを作成する．そのため，分類ルールを作成する為に数多くのキーワードの登録をする必要は無い．また，受信メールからルールに適するメールを選択するので，キーワードが分からない場合でも分類が可能である．

3　メール分類システム

3.1　メール分類機構

メール分類機構ではメールの文から特徴を抽出し比較する事でメールの分類を行う．本システムはメールから特徴ベクトルを作成する為にメールの本文を MeCab [1] を利用し形態素解析を行った．解析結果を $tf-idf$ 値を用いて重み付けを行いカイ二乗値を利用し素性選択をした．本システムでは，データマイニングツール Weka [2] のライブラリを利用する事により素性選択を行った．Weka は、オープンソースのデータマイニングのフリーソフトである．本システムでは素性選択に weka.attributeSelection.ChiSquaredAttributeEval を利用し，

図 1: 分類ルール作成画面

カイ二乗値を利用した素性選択を行った．メールの分類では SVM(Support Vector Machine) を用いて分類を行った．また，本システムでは SVM に weka.classifiers.functions.SMO を用いてメールの分類を行った．

3.2　分類ルール作成

本システムではメール分類ルールの作成は web ページ上で行う．図 1 にメール分類ルール作成画面を示す．分類ルール作成画面はルール一覧，メール一覧，メール本文，ルール設定，分類結果からなる．メール一覧からメールを選択するとメール本文が表示される．チェックボックスにチェックを入れ，ルールに登録するメールを選択する．ルール設定では形態素 N-gram の N を変化させる，利用品詞，特徴ベクトルの次元数を選択することができる．精度選択ボタンを押すと，ルール設定で選択した条件を基に，選択したメールと選択していないメールを分類する為の特徴ベクトル作成される．10 - 交差検定を行い分類精度を求め，分類条件，実行時間，特徴語とともに表示する．選択終了後，ルールの名前を入力するとシステムがルールを作成する．作成されたルールはルールデータベースに保存され，分類ルール作成画面のルール一覧に追加される．ルールを選択することにより，そのルールに属するメールを確認することができる．

3.3　システム構成

本システムはローカルプロキシ上に構築した．システムをローカルプロキシ上に構築することにより既存のローカルメーラを継続して利用することができる．また，ローカルで管理するため，システム管理者にメールを閲覧される恐れも無い．

本システムの構成図を図 2 に示す．ユーザはシステムのルール作成機構にアクセスしルールを生成する．生成されたルールはルール登録機構によりルールデータベースに保存される．システムはユーザが作成したルールを基に受信メールの分類を行う．メール受信機構はメーラからのメール要求をトリガとして POP3 サーバーにメールの要求をする．POP3 サーバから受信したメッセージはメール分類機構によ

A Mail Classification System based on Different Stylistic and an Application

Tomotaka TSUJINO, Shun SHIRAMATSU, Tadachika OZONO, and Toramatsu SHINTANI
Dept. of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology Gokiso, Showa-ku, Nagoya, 466-8555 JAPAN

[1] http://mecab.sourceforge.net
[2] http://www.cs.waikato.ac.nz/ml/weka/

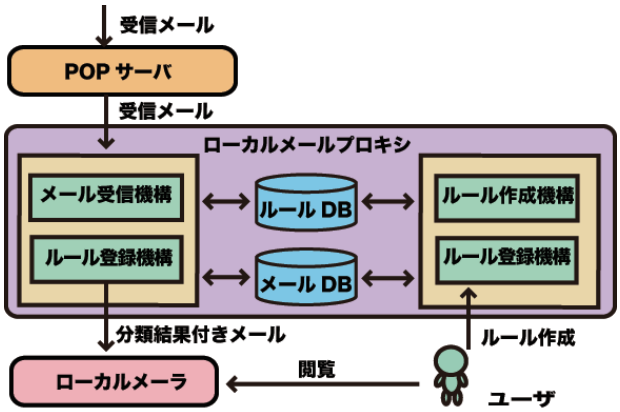

図 2: システム構成図

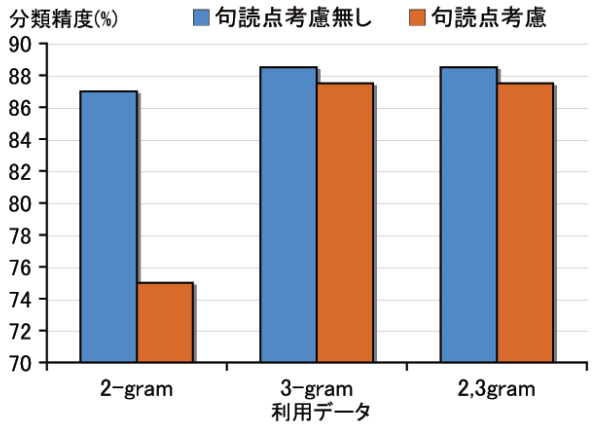

図 3: 受信メール分類精度

り，ルールデータベースに保存されているルールに適合するか調べる．システムは適合したルール名をメールのヘッダに付与してメーラに送信する．メーラではメールのヘッダに付与されたルール名をキーワードとして分類を行う事で本システムによる分類結果を利用する事ができる．また，システムは次回からのルール作成利用するため，受信したメールをメールデータベースに保存する．

4 実験と評価

4.1 実験データ

本研究では文体に着目し目上の人からメールとそれ以外の人からのメールを分類した．本実験では目上の人の人からのメールとは研究室の教授と先輩の 6 人からのメールとした．実験データは大学院生 1 名の研究室 PC で受信した最新の 200 通のメールを利用した．最新の 200 通のメールの中には，サービスへの登録メールや学会のメーリングリストなど研究室外からのメールも含まれている．そのうち目上の人からのメールは 63 通であった．

4.2 隣接語句と位置に着目

特徴ベクトルを作成する際に句読点の前の形態素 N-gram のみを利用して特徴ベクトルを作成したものと，位置を考慮しない形態素 N-gram を利用して特徴ベクトルを作成したものの精度の比較を行った．事前実験で最も精度の良かった，全ての品詞の形態素 N-gram を用い，素性選択を行い上位 250 個の素性を利用してベクトルを作成した．本実験では bi-gram，tri-gram, bi-gram + tri-gram 3 パターンそれぞれに対して，句読点と直前の形態素を利用した場合と位置を考慮しない場合の計 6 パターンで実験を行った．メールの分類精度を求める際に 10 - 交差検定を用いた．

図 3 に句読点の前の形態素 N-gram を利用して分類を行った結果を示す．図の縦軸は精度で横軸は利用データを示している．結果より bi-gram の場合は句読点と直前の形態素を利用した場合では特徴が抽出できず，本実験では分類精度が低くなってしまった．tri-gram, bi-gram + tri-gram では位置を考慮しない場合より若干分類精度が良くなるという結果になった．これらより，句読点の前の表現はメールの分類に有用であることがわかった．また，bi-gram を用いた場合よりも tri-gram を用いた場合の方が分類精度が高くなっていた．この事より，隣接語句を利用した方法はメールの分類に有用であることがわかった．

5 誤送信メール防止システム

本研究の応用として人間関係を利用した宛先間違いによる誤送信防止システムを試作した [3]．メールは受信者が決まっているので相手との人間関係によりメールの文体が決まる．文体に着目する事で相手との人間関係を利用した誤送信防止を行う．本研究では人間関係とは上司と部下の上下関係，社内と社外の所属関係，知り合いと友達の友人関係などの関係を人間関係と呼ぶ．本稿では宛先間違いによる誤送信とは，送信メールの件名や内容が宛先と送信者の関係に適切ではないメールと定義する．宛先間違による誤送信を防止する方法として，キーワードマッチを利用した防止システムがある．キーワードマッチを利用したシステムでは，特定のドメイン以外に送信するメールの文中に危険ワードが混入していた場合に誤送信と判断する．しかし，既存手法ではメール本文の内容を考慮していないため宛先間違いによる誤送信防止に有用な方法とは言えない．

本システムは送信メール履歴を利用して送信メールの本文の文体から送信相手との人間関係の推定を行う．システムはローカルプロキシサーバ上に構築し，メーラーから送信されたメールを受け取る．システムはメール本文から推定した人間関係と，事前に登録してある宛先との人間関係を比較することにより宛先誤りによる誤送信を防止する．誤送信と判断したメールは送信せず，ユーザに誤送信の可能性があることを通知する．目上の人に不適切な内容のメールを送ってしまう宛先間違い誤送信は 93 %で誤送信を判定することが可能であった．

6 まとめ

本稿では文体に着目したメール分類ルールについて述べた．本研究では文体の中でも，語の連なりと語の位置に着目した分類システムを作成した．語の連なりを考慮するために形態素 N-gram を用い，語の位置にとして句読点の前の語句を利用した．実験の結果，語の連なりと位置を考慮した，句読点の前の形態素 tri-gram を利用して分類を行った場合が最も分類精度が高かった．また，文体に着目したメール分類システムの応用として人間関係を利用した宛先間違いによる誤送信防止システムについて述べた．

参考文献

[1] 計量国語学会 “計量国語学事典”，朝倉書店，Nov. 2009.

[2] 西原陽子，砂山渡，谷内田正彦，“発話テキストからの人間の中の良さと上下関係の推定”，電子情報通信学会論文誌. Vol.J91D, No.1 pp. 77-88，Jan. 2008.

[3] 辻野友孝, 白松俊, 大囿忠親, 新谷虎松，“人間関係を利用した誤送信メール防止システムの試作”, FIT2010 第 9 回情報科学技術フォーラム, Sep. 2010.