メール文書からの感情等の特徴抽出について

平成 21 年 2 月 12 日

情報電子工学科 4 年
遠藤 建城

目 次

1 はじめに **1**

2 文章の評価について **1**

2.1 メールの文章を解析する方法 1

2.2 指標について 6

3 実験 **9**

3.1 実験の内容 9

3.2 実験に使用するソフト 10

3.3 実験に使用する文章 10

3.4 平均・標準偏差等について 14

3.5 実験結果 15

4 まとめ **18**

参考文献 **21**

概要

現在、メールというものは様々なところで使われており、そのメールの文章には相手の様々な感情が含まれていることが多い。人間はメールの文章を読めば含まれている感情を多少は理解できる。人間にできるのならばコンピュータにもメールの文章に含まれる相手の感情を多少は理解できるのではないかと考えた。本研究ではまずメールの文章から相手の感情等の情報を抽出する方法について幾つか挙げ、それぞれ考察する。次にその中から 1 つを選択し、その方法でメールの文章を解析する際に必要な指標を幾つか挙げ、それらの性質について具体的なブログの文章を用いて実験し考察した。

1　はじめに

現在、携帯電話やパソコン等でメールを受信することができる。その受信したメールには相手の感情が含まれていることが多く、判りやすい文章ならば読んだだけで相手の感情が多少は理解できる。人にできるのならば、コンピュータにも相手の感情を多少理解することができるのではないか、と考えた。コンピュータでメール受信時に自動的に感情を判断してくれれば、多数の人からのメールがたまってしまった際にどの人から優先的に返信すればいいか等を考える時の判断材料にすることが出来る。最終的に、メール着信時に自動的に文章を読み込み相手の情報を表示する方法を考察し、実際に相手の情報を表示するプログラムを作成することを目標とする。

2　文章の評価について

2.1　メールの文章を解析する方法

メールの文章を解析する方法には以下の方法を考えた。

- 全文を読み込み、構文解析を行いその結果を元に解析する方法

  全ての語句一つ一つに喜怒哀楽ごとに点数をあらかじめつけてデータベースに保存しておき、メールの文章を読み込んだ際にメールの文章に使われている語句毎に点数を呼び出し、その後一文毎に構文解析を行い、メールの文章全体の点数を出す。構文解析とは、ある一文の文法的な関係を説明すること。

- 語尾等の限られたところだけ着目し、解析する方法

  語尾等の一部の語句一つ一つに喜怒哀楽ごとに点数をあらかじめつけてデータベースに保存しておき、メールの文章を読み込んだ際に文章に使われている語尾の語句の点数を呼び出し、メールの文章全体の点数を出す。構文解析は行わない。

- 語句の意味を調べず、文の数等を調べ解析する方法

  語句の意味を全く調べないのであらかじめデータベースを作っておく必要もなく、構文解析の必要もない。前回までのデータとの比較で分析する方法。調査する項目は文の数、総字数、漢字の数等。

1 つ目の方法は似た研究を過去に他の研究室の松島明人氏 [1)] が研究しており、松島明人氏の研究は “チャットにおける感情の認識” である。松島明人氏の研究では語句毎の文法上の意味と感情の数値を入れたデータベースを作っておき、msn メッセンジャーからチャットの文章を横取りし、チャットの文章の構文解析を行い感情の数値の反転等をして、結果を出力する。ソフトは Perl と MySQL と kakasi を使用している。Perl は本研究でも使っ

ているプログラミング言語であり、MySQL とはデータベースを作るソフトで、kakasi とは文のわかち書き等ができるソフトであり、他にわかち書きができるソフトには ChaSen 等がある。Fig.1 に文章解析の流れを記載する。

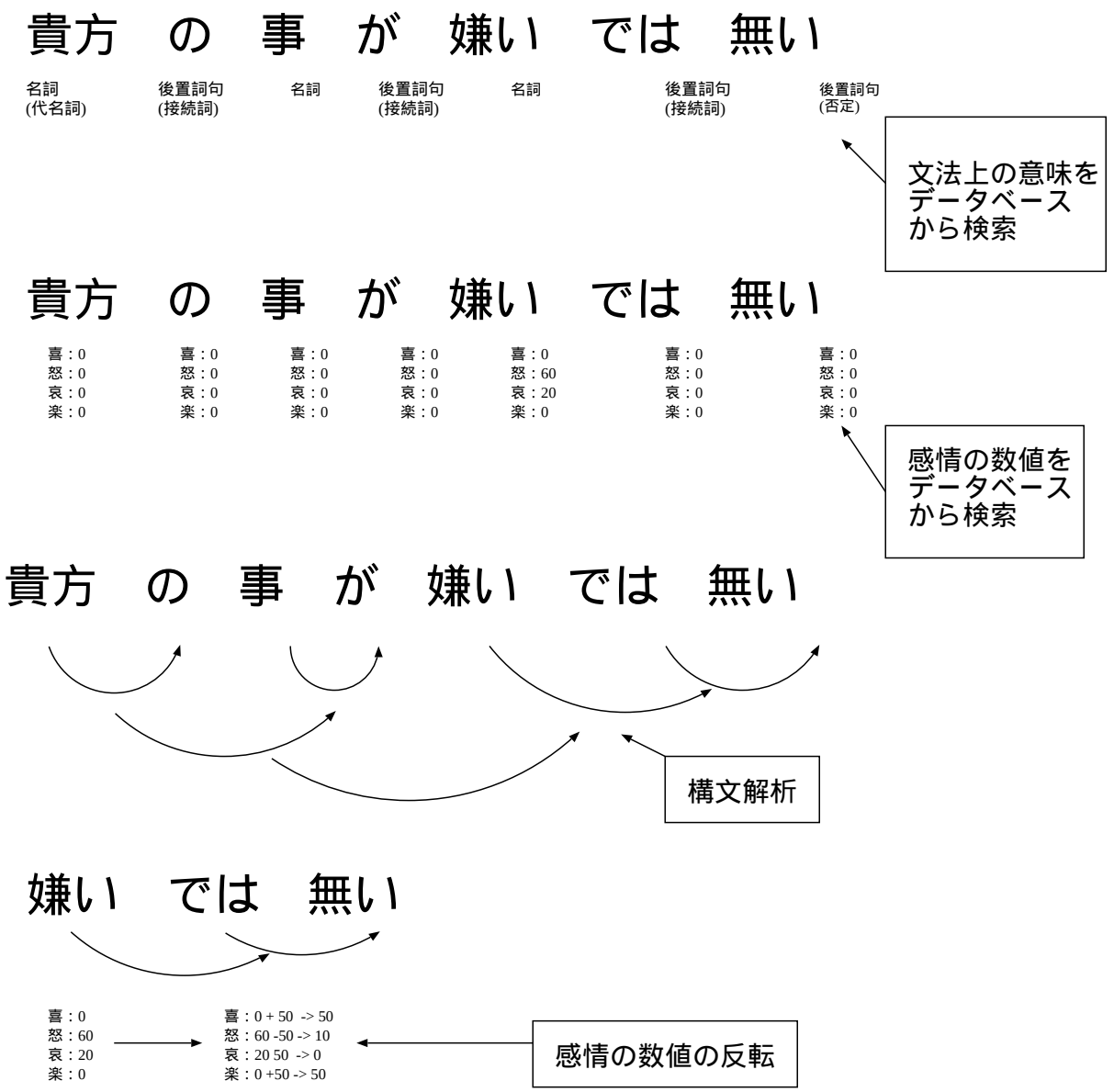

Fig. 1 “チャットにおける感情の認識” 文章解析の流れ

論文には Perl、MySQL、構文解析の内容が記載されている。以下にわかち書きができるソフトの kakasi と ChaSen について示す。

- kakasi

kakasi とは、漢字かなまじり文をひらがな文やローマ字文に変換することを目的として作成されたプログラムと辞書の総称である。更に、わかち書きも実装されている。わかち書きとは、日本語の文章を語句毎に分割し空白 (スペース) を入れた状態である。

UNIX で kakasi を使用するときは以下のように使用する

1. 標準入力から日本語文を入れる

   kterm 上で下記のように記述する。

   ```
   % echo "私は学校が好き。だから毎日学校へ行く。"|kakasi -w
   ```

2. 直接日本語文を入力する

   ```
   % kakasi -w
   ```
   と kterm 上で記述した後、

   私は学校が好き。だから毎日学校へ行く。

   のように入力する。終了は「Ctrl」キーと「D」を押す。

上のように実行した場合、出力結果は以下のようになる。

私 は 学校 が 好き 。 だから 毎日 学校 へ 行く 。

また、”kakasi -w” の”-w” の部分を変えることにより別の形で出力できる。例えば”-Ja” を入れた場合、文章中の全ての漢字をローマ字に変換してくれるし、”-KH” なら文章中の全ての片仮名を平仮名に変換する。このようにわかち書きと文字変換を簡単に出来るソフトである。

- ChaSen

  ChaSen は、kakasi と同じように単語の区切りを調べるのが難しい日本語を単語分割することができる。また、分割した単語の品詞を詳細に表示することができる。UNIX や Linux、MS-Windows で動作させることができる。

  UNIX で ChaSen を使用するときは以下のように使用する。

  1. ファイル (file) に日本語文を書いて、ファイルを ChaSen に読み込ませる

     kterm 上で下記のように記述する。

     ```
     % chasen file
     ```

2. 標準入力から日本語文を入れる。

kterm 上で下記のように記述する

```
% echo "私は学校が好き。だから毎日学校へ行く。"| chasen
```

3. 直接日本語文を入力する

% chasen と kterm 上で記述した後、

私は学校が好き。だから毎日学校へ行く。

のように入力する。終了は「Ctrl」キーと「D」を押す。

上のように実行した場合、出力結果は以下のようになる。

私	ワタシ	私	名詞-代名詞-一般		
は	ハ	は	助詞-係助詞		
学校	ガッコウ	学校	名詞-一般		
が	ガ	が	助詞-格助詞-一般		
好き	スキ	好き	名詞-形容動詞語幹		
。	。	。	記号-句点		
だから	ダカラ	だから	接続詞		
毎日	マイニチ	毎日	名詞-固有名詞-組織		
学校	ガッコウ	学校	名詞-一般		
へ	へ	へ	助詞-格助詞-一般		
行く	イク	行く	動詞-自立	五段・カ行促音便	基本形
。	。	。	記号-句点		
EOS					

さらに、-F オプションというものがあり、chasen の後ろに-F ”format” を追加することで使用できる。”format”には出力フォーマットを指定するための変換文字を入れる。

```
% echo "私は学校が好き。だから毎日学校へ行く。"| chasen -F "%m\t%H\n"
```

kterm 上で上記のように記述したとき以下のように表示される。

私	名詞
は	助詞
学校	名詞
が	助詞
好き	名詞
。	記号
だから	接続詞
毎日	名詞
学校	名詞
へ	助詞
行く	動詞
。	記号
EOS	

%m は単語毎で分割 (わかち書き)、%H は品詞の表示を表す。\t は tab、\n は改行である。

kakasi は ChaSen に比べると軽いが、わかち書きの精度は ChaSen の方がいい。kakasi のわかち書きは文章を語句毎に分けてくれないことがある。例えば、この文

前から旧校舎は立ち入り禁止です。

を kakasi でわかち書きすると

前か ら 旧 校舎 は 立ち入り 禁止 です 。

と、”前”、”から”と分かれて欲しいのに、”前か”、”ら”と分かれてしまう。ChaSen では

前 から 旧 校舎 は 立ち入り 禁止 です 。

と正しく分かれてくれる。構文解析等をするのなら ChaSen の方がいいと考えられる。

2 つ目の方法ではあらかじめ語尾等の一部の語句に点数をつけてデータベースに格納しておかなければならないが、語句の点数を入れたデータベースを作成することが難しく断念した。

データベースを作成する時、語尾等の一部の語句のみでも語句は大量にあるので難しい。さらに、語句への点数の付け方にも問題がある。語句に点数を付ける基準がないので、プログラムの作成者の勝手な基準で点数を付けることになってしまう。そのような作

成者の勝手な基準で決めた点数ではよい結果が出力できるとは考えられないので断念した。同様に、1 つ目の方法でも同じ問題が発生する。

よって、本研究では 3 つ目の方法で文章の解析を行うことにした。

2.2 指標について

本研究に使用する指標を示す前に過去にメールの文章を調査したレポートがあったので、そちらを紹介する。早稲田大学の寺本美恵子氏の “文章の長さは何に関連しているか[3)]” というレポートがあった。このレポートは 70 通のメールについて、総文字数・一文の長さ・句読点等区切り相当部分文字数の長さに年齢・性別などで異なった傾向が見られるかどうかを調査したものである。このレポートでは以下の点を傾向を調査する際の指標としている。

- 本文相当行

  全体の行から、空行 (空けてある行)、引用行 (返信の場合そのまま受信したものを引用してあるもの、>の記号で表されることが多い)、署名と署名に相当する行 (名前、メールアドレス、電話番号など) を引いたもの。

- 総文字数

  本文の全ての文字の数。

- 一文の平均字数

  総文字数÷ 本文相当行、一文の平均字数。

- 区切文字数

  句読点等 (句読点と記号類！、？、「 」なども含む) で文章が区切れていると考えられる部分の文字数をそれぞれ数え、区切の数で割った、区切り内の文字数平均。

- 区切数

  区切総数÷ 本文、一文につき文章にいくつ区切れがあるかの平均値。

このレポートでは 5 つの指標を調査した際、男女共におおよそ 30 才を境として傾向が別れているという結果がでたようである。table 1 は全体の平均に比べ「多い」「少ない」を出したものである。table 1 の表だけでは分かりづらいので下に文章のイメージを載せる。

- 男性・30 才以下

  短い文章を多行書く。
  ーーーーーーーー、ーーーーーーーーー。

	30 才以下		30 才以上	
	男性	女性	男性	女性
総文字数	多い	多い	少ない	少ない
一文の平均字数	少ない	少ない	少ない	多い
区切文字数	少ない	多い	少ない	少ない
区切数	少ない	少ない	多い	多い

Table 1 文章の長さは何に関連しているか

ーーーーーーーーーーーーー。
ーーーーーーーーーーー。
ーーーーーーーーーー、ーーーーーーー。
ーーーーーーーーーーー。
ーーーーーーーーーーーー。
ーーーーーー、ーーーーーーーー。

- 女性・30 才以下

  基本的に 30 才以下男性と変わらないが、男性よりも一区切りの文字数が多い。
  ーーーーーーーーーーーーーーー。
  ーーーーーーーーーーーー、ーーーーーーーーーーーーーーーーーーー。
  ーーーーーーーーーーーーーーーーー。
  ーーーーーーーーーーーーーーーーーーーー。
  ーーーーーーーーーーーーーーーーーー、ーーーーーーーーーーーーーーーーーーー。
  ーーーーーーーーーーーーーーーーーーーーーーーーーーーーーーーーーーー。
  ーーーーーーーーーーーーーーーーーーーーーーーーーーーーー。

- 男性・30 才以上

  総量は少ないが、一文に長い区切り (文章) をいくつも入れて文を長くつなげる。
  ーーーーーーーーーーーーーーーーーーーーーーーーーーーーーーーーーーー、ーーーーーーーーーーーーー
  ーーーーーーーーーーーーー、ーーーーーーーーーーーーーーーーーーーーーーーーーーーー、 ーーーー
  ーーーーーーーーーーーーーーーーーーーーーー、ーーーーーーーーーーーーーーーーーーーーーー
  ーーーーーーーーーーーーーーーーーーーーーーーーーーーーーーーーーーー。

- 女性・30 才以上

  総量は少なく、一文に区切りをたくさん入れ、みじかい区切りを長くつなげる。
  ーーーーーーーーーーー、ーーーーーーーーー、ーーーーーーーー、ーーーーーーーー、ーーーーーー

ーーー、ーーーーーー、ーーーーーーーー、ーーーーー、ーーーーーーー、ーーーーーー、ーーーーーー、ーーーーーーー、ーーーーーーー。

この調査結果はあくまで平均のため、全ての人に当てはまるわけではないことに注意しなくてはならない。本研究が一人の人間の文章のみから調査しているのに対し “文章の長さは何に関連しているか” のレポートでは年齢性別がバラバラの多数の人間の文章から調査している。本研究とはメールの文章を解析するという点で同じなため参考にした。

メールの文章は全く同じ文章でも人によって込められた感情が違う。そのため相手の喜怒哀楽の絶対評価はできない。なので本研究では、送信者の過去の文章との相対的評価で過去のメールの文章とどのように違っているのかを考える。

まず、プログラムでメールの文章を読み込みその文章を数量化したものを保存する。次に、過去のメールの文章を数量化したものと比較する。全ての数量化した値から平均や標準偏差を求め、今回のメールの文章の偏差値を求める。偏差値の値でメールの文章の内容が今までに比べどうだったのかを調査する。偏差値の求め方は 3.3 節に示す。

“文章の長さは何に関連しているか” のレポートの指標を参考として本研究では以下のものをメールの文章を解析し数量化する際の指標とする。

- 文の数

  メールの文章の文の数を数え記録する。

- メール全体の文字数

  メールの文章全体の文字の数を数え記録する。

- 文中の漢字の数

  メールの文章全体の漢字の数を数え記録する。

- 文中の平仮名の数

  メールの文章全体の平仮名の数を数え記録する。

- 文中の片仮名の数

  メールの文章全体の片仮名の数を数え記録する。

- 文中の区切り文字の数

  メールの文章全体の区切り文字の数を数え記録する。区切り文字とは、文を途中で分ける “,” や “.”,“「”,“」”,“!”,“?” 等の記号である。

これらの指標から何が判るのかはメールの文章を書く人により変わる。普段よりメールの文章が短いと怒っている人や、普段よりメールの文章が長いと怒っている人など、人によりかなり差があるので偏差値が高いからこの人は怒っている、とは言い切れない。だが、メールの文章が短いと事務的な用件のみの文章が多く、メールの文章が長いと発信者の心情や近況を述べるものが多いという非常に大雑把な傾向はある。

他の指標として顔文字がある。顔文字を調べれば相手の感情を簡単に知ることが出来るが、本研究では顔文字を判断材料にはしない。人間が目で見て判るようなものをコンピュータに解析させても意味がない。顔文字を指標に入れてしまうと顔文字の有無で結果がかなり変わってしまうし、顔文字の点数をつけたデータベースを作っておかなければならなくなる。顔文字は 20000 種類以上あり、現在も増え続けているので完全なデータベースを作るのは、ほぼ無理である。

他の指標として敬体であるかないか、と言うものもあったが本稿ではこの指標は本稿では考えない。なお、敬体とは語尾がですます調になっている文のことである。

3 実験

3.1 実験の内容

本研究の最終的な目標は、文章から相手の感情を読みとるプログラムの作成であるが、本稿ではある文章から 2.2 節の各指標毎に平均や標準偏差等を計算する実験を行い、平均、標準偏差等からの指標の考察を行う。下に本実験の進め方を示す。

1. 特定の人間が書いたメールの文章を幾つもプログラムに入力する

   読み込む文章は 3.3 節に示す。プログラムは Perl というプログラミング言語を使用する。Perl については 3.2 節に示す。

2. 各メールの文章毎に各指標の値を出力させる

   指標ついては 2.2 節に示す。

3. その出力から平均や標準偏差等を計算する

   式は 3.3 節に示す。平均、標準偏差、各指標間の相関係数を求める。

4. 計算結果から指標毎に考察する

   平均、標準偏差からグラフを作成し考察する。各指標間の相関係数を見て指標間の相関があるのか考察する。

3.2 実験に使用するソフト

本実験では Perl というプログラミング言語を使用する。

本実験で使用するプログラミング言語。文字を扱う点で非常に優れている。Perl(パール、Practical Extraction and Report Language) はラリー・ウォール (Larry Wall) 氏によって作られたインタプリタ方式のプログラミング言語及びその処理系である。

コンパイルというコンピュータが理解できる機械語に翻訳処理する作業が必要がなく、プログラムがテキスト形式なので作成や実行や修正が簡単に出来る。C、awk、sed、シェルスクリプト等のほとんどすべての機能を取り込んでいる為、それらの言語でできることで Perl でできないことはほとんどないが、一つのことを実現するのに何通りでもできてしまうといった「副作用」がある。なお、Perl は MS-DOS、VMS、OS2、Plan 9、Macintosh、Windows の上でも動作出来る、移植性の高い言語である。

なお本研究では文章を解析する際、kakasi や ChaSen を使用していた。本実験では kakasi や ChaSen を使用しなかったが、この研究を更に進めるときにおそらく使用することになると考えられる。なお、2.2 節の文章を解析する方法の 1 つ目、及び 2 つ目の方法で研究するときにも kakasi や ChaSen を使用するだろう。

3.3 実験に使用する文章

本実験は文章を読み込み、そこから情報を読み取るものである。故に、読み込む文章が無ければならない。本研究の読み込むサンプルの文章には以下の候補があった。

- メールの文章

  私が友人から頂いたメールの文章、又は私が友人に送信したメールの文章を使用する。本研究では最終的にメールの文章から感情を読み込むプログラムの作成なので、メール文章をサンプルとして使うのが必要がある。下にメール文章を載せる。

  > 友人からのメール文章
  >
  > すいませんてか 5 時までかと思ったら 15 時までだったんだねさっきのメール見間違えてたわ

- web 上の電子掲示版の文章

  インターネットの電子掲示板に書き込まれた文章を使用する。候補として “電車男 [5]” や “ブラック会社に勤めてるんだが、もう俺は限界かも知れない [4]” の書籍化された 2 冊がある。 この 2 冊は、2 ちゃんねるへの書き込みを基にした本で、2 ちゃ

んねるを知らない人にもスレッドの雰囲気を伝えることを狙ったらしく、小説化はせず、アスキーアートなども含めてスレッドの書き込みをそのまま掲載している。そのため、メールから情報を読み取るという本研究には向いてると考えられる。スレッド (省略形で “スレ” ともいう) とは、おおもとは 糸 という意味であり、一連の話の流れという意味で使われている。たくさんある電子掲示板群の中の、特定の話題を扱ったひとつの話題ツリーのことを指す。以下に 2 冊の書籍の紹介と文章の一部を示す。

– 電車男

  インターネットの電子掲示板である 2 ちゃんねるへの書き込みを基にしたラブストーリー。漫画化・映画化もしている。2 ちゃんねるの独身男性板 (通称毒男板) の「男達が後ろから撃たれるスレ」というものがあり、このスレッドに「731」により投稿された一見何気ない書き込み (749) が、発端である。アキバ系ヲタクを自認する主人公がスレッドに相談を書き込み、応援するスレ住人が様々なアドバイスをしていく話である。以下に文章の一部を示す。

  731 名前 : Mr. 名無しさん 投稿日 : 04/03/14 21:25

  すまん。俺も裏ぐった。
  文才が無いから、過程は書けないけど。

  このレスまじで魔力ありすぎ …
  おまいらにも光あれ …

– ブラック会社に勤めてるんだが、もう俺は限界かも知れない

  電車男と同じく 2 ちゃんねるへの書き込みを基にしたビジネス論の書籍である。プログラマになった主人公が入社日からの苛酷な日々を書き込み、それに数人の人か書き込んでいく話である。以下に文章の一部を示す。

  1 : 名前 : 名無しの VIP 投稿日 : 07/11/24 21:38:07.44

  職業はプログラマ。この職業、マジでやばすぎる。

  入社日での出来事。
  パソコンを渡される→指示された通り、色々なものをインストール→設計書を渡される。

  「これでおっけーと。んじゃ作れ」
  「え？」

「いや作れって」
「あ、え？は、はい」
「みんな忙しいから、出来る限り自分で解決しろよ」

そう言って去って行くチームリーダー。
このまま悩んでもしょうがない。とりあえず設計書を見てみるか。
フレームワークがどうのこうの、うんたらかんたら … 。テストにはどうのこうの … 。

ワケわからんぞ

- ブログの文章

ブログとは人や数人のグループで運営し、日々更新される日記的な Web サイトの総称である。内容は、ニュースや専門的トピックスに関して自らの専門や立場に根ざした分析や意見を表明したり、他のサイトの著者と議論したりする形式が多く、従来からある日記サイトとは区別されることが多い。一部のブログは、書籍化されている。

本研究では MORI LOG ACADEMY[6] というブログをサンプルとして使用する。下に MORI LOG ACADEMY のブログの文章を載せる。

【ＨＲ】　ひさしぶりの日記

子供の頃にも日記はつけて続いたためしはなかったし、大人になってもそういった習慣はないのに、インターネットのサイトでかつて日記を始めたら、これが 5 年も続いて 5 冊の本になった。今まで書いてきたどの著作よりも、その日記が自分では一番の力作だと思っている。最初から、出版するつもりで書いたから、つまり仕事としてやったから続いたと思う。でも、マンネリを感じて 2001 年でいったんやめてしまって、そのときは「ああ、もう書かなくて良いのだ」と本当に嬉しかった。きっと、小説をやめたら、同じくらい嬉しいだろう。
さて、日記をもう一度書いてみることにした。今回も仕事として書くし、出版物になる予定だから、続くだろう。前回は、「ネットの日記を本にするなんて」という目で見られたけれど、今では普通になった。もの凄い大勢の人がネットで日記を書いている。いったい読み手はいるのか、と不思議。そういう話はまたそのうちゆっくりと……。

本研究は、メールの文章を読み込み感情等の情報を抽出することを最終目標とするが、まず指標に関する考察を行う。

実験の再現性を確保するために、更新・削除が頻繁に行われるデータ情報ではなく、書籍化されているものが本研究のサンプルとして望ましい。また、特定の人物のメールの文章だけでは指標の検証で偏った結果が出てしまう恐れがある。よって、メールの文章をサンプルとして使用するのは難しい。

サンプルとして使用する文章はある程度長い方がいい。本研究では文章全体の字数を指標にもするため、短い文章をサンプルとして使用してしまった場合、例えば普段 10 字くらいの文章を書く人の文章をサンプルとしたら、8 字のときと 12 字のときとでは差が大きくなってしまう。それでは指標の検証ができないので短過ぎる文章は使えない。更に、本研究では顔文字を指標として使用できないので顔文字が多く含まれている文章はあまり使いたくない。

web 上の電子掲示版の文章の候補の二冊の本には顔文字や記号で作られた絵 (アスキーアート) が多く書かれている。アスキーアートとは table2 のようなものを指す。

Table 2 アスキーアートの例

```
                /
            ノ/
          /.::: /        /
        /:::  ::::  /    //
      /.:   :::  :::  /   /:::/
     /.:::   :::  :::  ::: ‘’ :::三
    / ◎   ::::  :::  :::: .:::. ::::三
    >  :::   :::   :::  :::  :::: ::::三
    \  :::   :::  :::  \’\  ::\
       \   ::::  ::: \     \\
          ヾ_,-ヽ\        \
            ヽ\  \
              ヽ’  \|
```

このようなアスキーアートを解析するのは非常に難しいため、本研究では考えない。なので、このようなアスキーアート多く使われている文章はあまり使いたくない。また、顔文字は先に書いたように指標として扱っていない。なので、“電車男” や “ブラック会社に勤めてるんだが、もう俺は限界かも知れない” のような顔文字、アスキーコードが多く含まれるスレッドの文章をサンプルとして使用するのは難しい。web 上の電子掲示版の文章は不特定多数の人間が書き込んでいるので、特定の人の文章を解析したい本研究ではサン

プルとして使用するのは難しい。

ブログの文章は書籍化されているのも多数あり、またアスキーアートがほとんど含まれていないものもあるので、その点で他の 2 つと比べ本研究には向いている為、ブログの文章をサンプルとして使用することにする。

3.4 平均・標準偏差等について

本実験ではブログの文章をプログラムに入力し、指標毎に値を出力する。その作業を何回も繰り返し平均と偏差、及び各指標間の相関係数を求め考察する。本稿では記載していないがメールの文章を入力し指標毎に値を出力する際に、指標毎の偏差値も計算する。この偏差値を見て今までのメールとどのくらい違うのかを知ることができる。

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{2}$$

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{3}$$

$$t = \frac{x_i-\bar{x}}{s}\times 10+50 \tag{4}$$

x_i、y_i は各指標の値で $\bar{x}$、$\bar{y}$ は各指標の値の平均、s は不偏標準偏差、r は指標間の相関係数、t は偏差値を表す。

不偏標準偏差の式 (2) は上の式のままだとプログラムに組み込んだ際ループが 1 回増えてしまうので下のように変形して組み込む。

$$\begin{aligned} s &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i^2-2x_i\bar{x}+\bar{x}^2)} \\ &= \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2-\sum_{i=1}^{n}2x_i\bar{x}+\sum_{i=1}^{n}\bar{x}^2\right)} = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2-2\bar{x}\sum_{i=1}^{n}x_i+n\bar{x}^2\right)} \\ &= \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2-2\bar{x}n\bar{x}+n\bar{x}^2\right)} = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2-2n\bar{x}^2+n\bar{x}^2\right)} \end{aligned}$$

$$= \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}$$

table 3 に前のセクションに記載した MORI LOG ACADEMY の文章を入力した場合の出力の例を示す。

文の数	全文中の字数	全文中の漢字の字数	全文中の平仮名の字数	全文中の片仮名の字数	全文中の区切り字数
13	419	91	258	19	34

Table 3 解析した指標毎の値

table 3 の表のような値を文章を読み込むたびに出力し、それを何回も繰り返す。平均と偏差、及び各指標間の相関係数は上の (1) ~ (3) の計算式で求める。

なお、標準偏差は標本分散の正の平方根の標準偏差ではなく、不偏分散の正の平方根の不偏標準偏差を使用する。標準偏差と不偏標準偏差の式の違いは分母に"n" か"$n-1$" のどちらが入っているかである。不偏標準偏差は母集団の標準偏差を推定するために用いる。$n-1$ で割ることで、不偏性を持った不偏推定量になる。

相関係数は 2 つの指標の値の類似性の度合いを示す統計学的指標である。単位はなく、-1 から 1 の間の実数値をとり、1 に近いときは 2 つの確率変数には正の相関があることを意味し、-1 に近ければ負の相関があるということを意味する。0 に近いときはもとの確率変数の相関は弱いことを意味する。指標間の相関係数の値が 1 に近い場合、2 つの指標は連動して一定の関係で動くことになり独立した指標とは言えないことになる。

3.5 実験結果

本研究では 398 日分 (13ヵ月分) のブログの文章をサンプルとして使用した。table 4 に実験により得られた、MORI LOG ACADEMY のブログの文章を解析した指標毎の値のデータの一部を示す。

table 4 のデータより求めた各指標毎の平均、不偏標準偏差、度数分布を table 5 に示す。度数分布の 2.5 ~ -2.5 の値は $\bar{x}+\sigma s$ の σ の値であり、σ が 0 の時、$\bar{x}-0.25s$ ~ $\bar{x}+0.25s$ の度数が入る。度数分布は 100 分率 (%) で表している。

Fig 2 は、table 5 の表の各指標毎の値を使用したものである。y 軸は 100 分率 (%) で表し、x 軸の値は table 5 の表と同様に $\bar{x}+\sigma s$ の σ である。σ が 0 の時、$\bar{x}-0.25s$ ~ $\bar{x}+0.25s$

文の数	全文中の字数	全文中の漢字の字数	全文中の平仮名の字数	全文中の片仮名の字数	全文中の区切り字数
62	1601	354	910	83	147
55	1461	350	791	79	140
37	999	216	560	51	119
40	1149	274	625	102	107
49	1199	311	642	81	118
55	1562	376	847	108	149
45	1231	318	592	124	111
42	1106	246	654	33	137
56	1667	419	910	99	152
50	1196	320	593	114	98
44	1135	320	595	75	98
55	1516	315	877	106	169
57	1484	385	829	81	136
39	1078	271	600	27	99
59	1633	345	916	124	152
59	1824	450	1095	82	143
63	1716	448	965	105	128
61	1636	409	844	109	158
55	1303	358	705	45	114
39	1135	237	669	61	121
56	1495	407	846	60	132
50	1432	414	801	50	121
55	1479	360	811	69	144
45	1505	361	899	50	150
59	1439	354	728	170	129
37	1038	247	604	47	103
53	1557	399	919	37	140
53	1718	440	1013	89	128
32	1039	234	617	68	82
40	975	219	534	93	90
50	1424	320	769	84	179

Table 4 解析した指標毎の値のデータ

	文の数	総字数	漢字の総数	平仮名の総数	片仮名の総数	区切りの総数
$\bar{x}$ (平均)	66.0	1900	456	1057	120	174
s (標準偏差)	12.1	340	89.3	202	49.0	35.4
σ						
2.5	0.251	1.51	1.01	1.26	0.25	1.51
2	2.51	2.26	3.27	2.51	2.76	2.01
1.5	5.53	5.03	4.27	5.03	6.28	5.53
1	12.8	11.8	9.55	13.3	11.1	11.1
0.5	18.8	20.1	19.6	17.8	14.8	18.1
0	20.6	22.6	22.9	21.1	19.1	20.1
−0.5	16.6	13.3	16.8	16.3	19.8	19.6
−1	12.8	13.6	13.6	11.1	16.1	10.1
−1.5	5.53	4.77	4.77	6.53	7.54	7.54
−2	2.76	2.51	1.26	3.01	1.01	2.51
−2.5	0.75	2.26	2.26	1.76	0.00	1.26

Table 5 各指標毎の平均、不偏標準偏差、度数分布

	文の数	総字数	漢字の総数	平仮名の総数	片仮名の総数	区切りの総数
文の数		0.850	0.764	0.805	0.364	0.723
総字数			0.888	0.955	0.432	0.867
漢字の総数				0.815	0.217	0.713
平仮名の総数					0.264	0.810
片仮名の総数						0.348
区切りの総数						

Table 6 各指標毎の相関係数

の度数が入る。

このグラフは、全ての指標が正規分布に似た形にはなっているが、正規分布のグラフに比べて、ばらつきが多いことが判る。σ が 1 ~ -1 の間に全体の 68%以上なくてはならないのに本研究の全ての指標は 60%にすら届いていない。

次に各指標毎の相関係数を table 6 に示す。本研究では相関係数は table 6 の表のような値となった。文の数や総字数と他の指標との相関係数は片仮名の総数以外は 1 に近く、正の相関があると言える。ただ、文の数や総字数が増えれば漢字や平仮名等の数が増えるのは当然の事なので、どちらかの指標を不要だと考えていいものなのか判断しづらい。また、漢字や平仮名や区切りの 3 つの指標も正の相関があると考えられるが、漢字の数が多ければその漢字を含んでいる総字数が多く、総字数が多ければ平仮名や区切りも必然的に多くなると間接的に正の相関になっていると考えられる。故に、先程と同じようにどの指標を不要だと考えていいものなのか判断しづらい。

平仮名の字数と総字数の相関係数は 0.955 で、この 2 つの指標はほぼ連動していると考えられ、独立した指標とは言えないことが判る。

4 まとめ

本研究の最終的な目標はメールの文章を読み込んで相手の喜怒哀楽等の感情を出力するプログラムを作成することであるが、本稿では文章から感情等の情報を読み取る方法につ

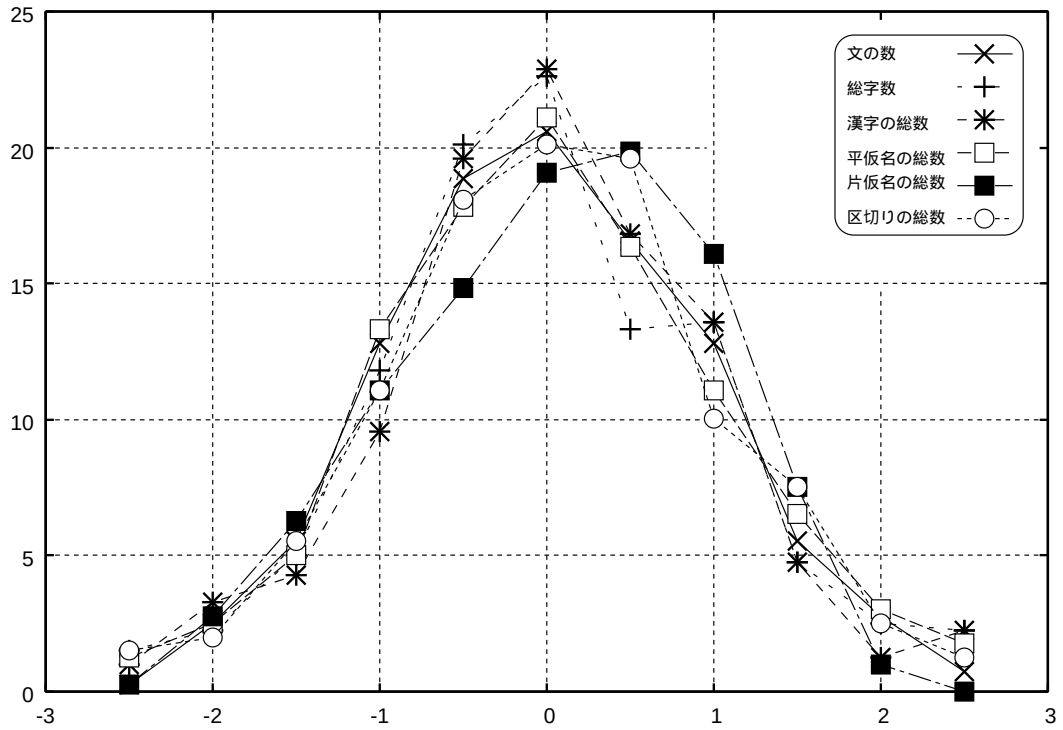

Fig. 2 各指標毎の度数分布

いて考察し、文章から情報を読み取る際の指標について実験、考察した。

文章から感情等の情報を読み取る方法についてだが、本稿では 3 つの方法を挙げ考察した。その内の 1 つは過去に似た研究をしていた方がいたのでその論文を読んだが、語句 1 つ毎に喜怒哀楽の数値を入力したデータベースを作るというとてつもない作業が必要だったので、本研究では相手の文章の語句一つ一つの意味を調べずに感情等の情報を読み取る方法を考察した。

だが、この方法では喜怒哀楽という感情をコンピュータに絶対的な評価をさせるのは無理なので、過去の文章との比較による相対的な評価を行った。この方法で評価するときにメールの文章を数量化するために指標というものが必要になる。本稿では “文の数”,“総字数”,“漢字の総数”,“平仮名の総数”,“片仮名の総数”,“区切りの総数” の 6 つの指標を挙げ、その指標の有用性について実験し考察した。今回のブログの文章を人間が感情を判断して分類しておき、それと指標をつき合わせるという作業の実験をこの次にやりたかったが時間がなかった。今回の指標で文章から感情が抜きだせるのかの検証をする必要がある。これは今後の課題であろう。

指標は実験の結果 “総字数” と “平仮名の総数” の相関係数が非常に高い値を出しており、指標はほぼ連動していると考えられ独立した指標とは言えないことが判り、 “平仮名の総数” は、無くてもいいのではないかとも思われるが、本研究では MORI LOG ACADEMY のブログの文章のみで実験をしたので、他の文章を使用したとき “総字数” と “平仮名の総数” 相関係数がどうなるのかはまた、何回も別の人の文章を読み込んで実験しなければ判らない。その点も今後の課題である。

参考文献

[1] 松島 明人: “チャットにおける感情の認識”, 新潟工科大学工学部情報電子工学科卒業論文, (2004)

[2] 三間 優: “方言学習と方言変換ソフトの改善方法の考察”, 新潟工科大学工学部情報電子工学科卒業論文, (2007)

[3] 寺本 美恵子: “文章の長さは何に関連しているか”, 早稲田大学文学学術院 上野和昭研究室 国語学研究班レポート 電子メールの言葉, (2001)

[4] 黒井 勇人 : “ブラック会社に勤めてるんだが、もう俺は限界かも知れない”, (新潮社,2008)

[5] 中野 独人 : “電車男”, (新潮社,2004)

[6] 森 博嗣: “MORI LOG ACADEMY 1 ~ 3”, (メディアファクトリー,2006)