社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－NL－185 (7)
2008－SLP－ 71 (7)
2008／5／23

表層・語彙的特徴量に基づくブログの面白さ分析

萩行 正嗣 柴田 知秀 黒橋 禎夫
京都大学大学院情報学研究科
〒606-8501 京都府京都市左京区吉田本町
{hangyo,shibata,kuro}@nlp.kuee.kyoto-u.ac.jp

あらまし

近年、インターネット環境の普及とともに数多くの人がブログを通じて情報を発信するようになっている。それに伴い、大量に存在するブログから面白いものを探し出すことが困難になってきている。本研究では表層・語彙的特徴量に基づき、ブログの面白さを分析する手法を提案する。まず、ブログの記事から文字長などの表層的特徴量や評価表現などの語彙的特徴量といった様々な特徴量を抽出する。そして、これらを特徴量として与えて SVR を用いた機械学習を行なうことで、ブログの面白さを推定する。独自に設置したブログを用いて収集した 249 件のブログ記事とそれを採点したものを用いて実験を行なったところ、ベースラインを上回る精度を達成することができた。また、面白さの個人差の問題についてはドメインアダプテーションを用いることで対処した。最後に、学習されたモデルからブログの面白さの要因について考察を行なった。

キーワード ブログ, 面白さ, 評価表現, ドメインアダプテーション

Analyzing Interest in Blog Articles based on Surface and Lexical Features

Masatsugu Hangyo Tomohide Shibata Sadao Kurohashi
Graduate School of Information, Kyoto University
Yoshida-honmachi, Sakyo-ku,Kyoto, 606-8501, Japan
{hangyo,shibata,kuro}@nlp.kuee.kyoto-u.ac.jp

Abstract

Recently, with the prevalence of the Internet environment, a number of people transmit information through a blog. It is getting harder to search out the interesting one among a large amount of existing clog's. This paper describes a method for analyzing the interest of a blog based on automatically extracted surface and lexical features. From a blog text, surface features such as the number of characters and lexical features such as modalities and positive/negative expressions are extracted. Then, our system outputs an interest rate of the blog by a machine leaning method. We collected 249 blog articles, which are assigned to manual evaluations, and then experimented the proposed method on these articles. The experimental result showed that our method outperformed the baseline system. To deal with the difference among individuals, we adopt a domain adaptation technique, which regards an individual as a domain. Finally, we consider the elements of interest based on the learned model.

Key Words blog, interest, evaluative expression, domain adaptaion

1 はじめに

近年，インターネット環境の整備により、特別な知識なしに手軽に作成・投稿することができるブログが普及してきた。ブログは読者の購買意欲や、時には政治的、経済的な指向にすら影響を与える。またユーザが気軽に投稿でき、既存の文章表現に捉われないブログは、今後も表現手法の一つとして大きな位置を占めると考えられる。一方で、その手軽さからWeb上には多数のブログが存在するため[1]、自分にとって有用であったり、面白いと思うブログを見付け出すことが困難になってきている。

面白いブログを探し出すための一つの手段としてアクセス数やリンク数、あるいはブログの特徴の一つであるコメント数やトラックバック数を利用することが考えられる。例えばソーシャルブックマークはユーザが面白いと思うサイトを公開・共有する枠組みであり、これを利用することにより面白いブログを探し出すことができる。しかし、コメントやトラックバックなどが付くブログは一部のものであり、全てのブログから面白いものを探し出すことができるわけではない。

本研究では、ブログの面白さを自動判定する手法を提案する。まずブログの面白さについて考察するために、249件のブログ記事を投稿してもらい、各記事に対して12人が7段階で面白さを評価した。そして、作成したデータを分析することにより、面白さの要因について考察する。

次に、自動処理によりブログの面白さを推定する手法について述べる。ブログから表層・語彙的特徴量を抽出し、機械学習によって面白さを推定する。表層的特徴量としては、文字数や文字種、記号の割合などを用い、語彙的特徴量としては、評価表現、接続詞、TD·IDFなどを用いる。

作成したデータを用いて実験を行なった。まず、各採点者の平均値を推定したところ、ベースラインよりも高い精度を達成することができた。しかし、ブログの面白さは人によって様々であるので、次に、採点者それぞれの評価を推定した。各採点者をドメインとみなすドメインアダプテーションの技術を用いることにより、精度向上がみられた。

また、学習されたモデルを考察することにより、人がブログのどのような要素に面白さを感じるかについての分析を行なった。

[1]英語版Wikipediaによれば、2007年12月時点で11.2億記事以上がWeb上に存在する。また米テクノラティ社による調査によれば2006年第4四半期に世界中で投稿されたブログ記事の37%が日本語によるものである。

2 ブログデータの作成と面白さの考察

まず、ブログの面白さを考察するために、ブログ記事を投稿してもらい、複数の採点者がそれらの面白さを評価した。本節ではデータ作成の方法と、採点者間にどれくらい相関があるのかを示す。また、作成したブログを分析することにより得られた面白さの要因について述べる。

2.1 データ作成と採点結果の傾向

2.1.1 データ

実験用のブログを作成するために、ウェブ上にブログを設置し、大学院生に記事を投稿してもらった。記事のテーマは[京都観光]・[携帯電話]・[グルメ]・[スポーツ]の4つとし、その中から重複を許して一人あたり2〜4記事を投稿してもらった。その結果集まった249記事を本研究での分析対象とする。

次に12人の採点者が249記事に対し7段階(7点:面白い、1点:面白くない)により、採点を行なった。記事の順序による点数の偏りを避けるため、採点者間では記事の採点の順番はランダムとなるようにした。投稿されたブログの例を図1、図2に示す。

2.1.2 採点結果の傾向

まず、すべての採点者、すべての記事の採点の平均点は4.05、各記事における採点者間の点数の分散の平均は1.76であった。分散の分布を図3に示す。

次に、採点者間のデータの比較を行なう。評価指標として、同一記事に対する評価間の平均二乗誤差(PMSE)とピアソンの相関係数を用いる。平均二乗誤差は小さいほどよく、また、相関係数は-1から1までの値をとり、1に近いほど2つのデータに相関があることを示す。任意の採点者間の相関係数の平均は0.21、二乗誤差平均の平均は1.85であった。採点者間では緩やかな正の相関が見られ、大半の採点者の評価が面白い、または、面白くないと一致する記事があったものの、評価が割れるものも多く見られたことから、採点結果にゆれがあることが分かる。

私の生きがい [スポーツ]

5 分前 …
3 分前 …
2 分前 …

「京都大学、東京大学 ‥‥
アテンション、ゴゥッ！」

さぁ!!スタートだ!!

ザババァ …

やばい … 誰かのオールがもぐって艇が揺れ、
東大に出られた …
でもまだまだ 2000 メートルもある !!
私たちはコンスタントが強い、水中も強い。
諦めずに 120 パーセントの力、だしていこう !!

500 メートル地点。
ちょっと東大が近づいてきたんじゃ？
いいぞ !! ちゃんと詰めてる !! いいぞ !!

1000 メートル地点。
同回生の子から「水中もっとだせー !! 」
という声が聞こえる。
もっと出さなきゃ … 。
しんどい … 。
でも確実に東大と並んでる。
いや、ちょっとでているかも？
いける !! これはもしかしたら …!!

図 1: ブログの例

—必要か否か— [携帯電話]

海外ではケータイを持っていない。友達と「～時に～で待ち合わせね」と約束して、一度別れてしまえば連絡手段はなし。普段は遅刻ばかりの私が、絶対にそこの場所に、定刻に向かいました。ケータイはないと不便、不安だとばかり思っていたけれど、持っていないことで良いこともあるのだなと実感しました。約束を守ることは当たり前のことだけれど、そんな当たり前のことを疎かにしていたのかと考えると、本当にダメですね。自分のダメな部分に気づくいい機会になりました。今の時代、ケータイを持たずに生活することは難しいでしょうが、にはケータイがない日があってもいいようがしました。(とはいえ、関空に到着してにケータイの電源を入れ、メールチェックかさずしてしまいましたが …)

図 2: ブログの例

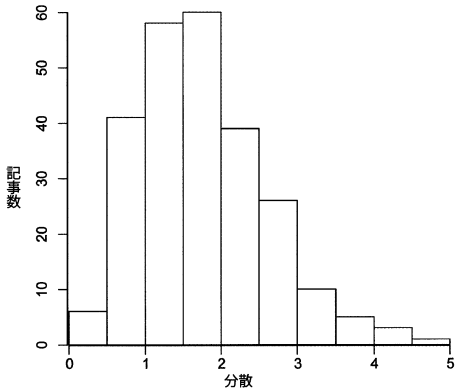

図 3: 各記事における採点の分散の分布

2.2 面白さの要因

一口にブログの面白さといっても様々な軸が考えられる。上記で構築したブログ記事と採点結果を分析すると以下のような面白さの要因があることがわかった。

修辞 読み手が文章表現からストレスを感じずに記事を読むことができる度合い

論理構成 記事の論理構成が正しい度合い

説得力 読み手が記事から感じる説得力の度合い

対話性 投稿者が読み手に対して語り掛ける度合い

キャラクタ性 記事から投稿者の人間性が伝わってくる度合い

共感 読み手が記事の内容に共感したり同意したりする度合い

情報量 読み手が記事の内容により、未知の事実を知る度合い

これらのいずれか、または組み合わさって、ブログが面白いか、または、面白くないを判断していると考えられる。

また、これらの要因は個人によって捉え方が異なると考えられる。例えば、記事の内容に共感するかどうかは個人差があり、また、情報量が高いと思うかどうかは、個人の知識量に依存すると思われる。この問題に対して、本研究では個人をドメインとみなしたドメインアダプテーションの技術で対処する。詳しくは 4 節で述べる。

3 ブログの面白さ判定の特徴量

本節ではブログの面白さを推定する手法について述べる。以下で述べる様々な特徴量をブログから自動抽出し、教師あり学習によって面白さを推定する。ここで、各記事に対して複数の採点結果がある

ので、どの点数を推定するかという問題がある。本研究では採点者の平均点を推定するタスクと、個人の採点を推定するタスクの二つのタスクを考える。詳しくは4節で述べる。

ブログ記事から自動獲得できる特徴量を以下にあげる。先にあげた面白さの要因と自動獲得できる特徴量の対応関係は3.4節で議論する。

3.1 表層から獲得できる特徴量

文字数、文数、行数、段落数 各記事の文字数、文数、行数、段落数、一文あたりの文字数の平均および分散、一段落あたりの文字数の平均および分散を求め、これらを特徴量として与えた。記事の文字数は記事の内容量を示している。文数、段落は記事の構成を表わしており、構造的な読み易さの指標となる。一文あたりの文字数の平均および分散は文のリズムを示す指標となる。

また $(\text{文字数} - L)^2$ および $(\text{一文あたりの文字数} - l)^2$ を特徴量として与えた。$(\text{文字数} - L)^2$ および $(\text{一文あたりの文字数} - l)^2$ は長すぎる記事や短かすぎる記事の指標となる。ここで、L は今回の実験で用いたブログ記事の平均の長さである483字、l は今回使用したブログの投稿フォームにおける一行の文字数である40字とした。

文字種 文字種は読者の視覚に直接作用する働きがある。例えば漢字が多ければ内容が難しそうだ、などである。また数字・アルファベットは内容の具体性を示す指標となる。そこで、各記事の漢字・ひらがな・カタカナ・数字・アルファベットの割合を特徴量として与えた。

記号 記号は直接的、間接的に文中で様々な働きをする。文頭の三点リードは前の文から間を取る働き、文中の三点リードは間を取り、文末の三点リードは言い淀む働きが考えられる。なお、Web上においては二点リードや中黒を三つ並べたものが三点リードの代用とされるので[10]、それらも三点リードとして数えた。疑問符、感嘆符は投稿者の感情を示す働きがある。読点は一文の中での区切りを表し、文の読み易さに影響すると考えられる。

そこで、各記事の三点リード、疑問符、感嘆符、読点の数を数えた。なお、三点リードは文頭、文中、文末で異なる働きをすると考えられるので別に数えた。そして、これらを文数で割ったものを特徴量として与えた。

括弧対 鉤括弧で囲まれた部分は、専門用語あるいは特別な使われ方をしている語を示している場合が多く、専門性や特殊性を示す作用があると考えられる。丸括弧で囲まれた部分は、補足事項などが記されている場合が多く、内容の深さを示す指標となる。

各記事から鉤括弧対および丸括弧対の数を数えた。またこの時、顔文字を除くために、括弧対に挟まれた文字列中に記号のみしか存在しないものは除外した。

「笑」の割合 「笑」という表現は書き手のキャラクタ性を示す指標となる。

各記事の「笑」という文字の数を数えた。通常は(笑)のような使われ方が多いが、括弧を省略する場合も多く見られたので、記事中の「笑」の文字を数えることとし、これを文数で割ったものを特徴量として与えた。

3.2 形態素・構文情報から獲得できる特徴量

形態素解析器JUMAN[9]および構文解析器KNP[8]を用いてブログの形態素・構文解析を行ない、以下の特徴量を獲得する。

接続詞・感動詞の割合 JUMANの解析結果から得られた品詞情報を元に、文体に大きな影響を与えていると考えられる接続詞、感動詞の個数を数えた。そして接続詞・感動詞の個数それぞれを自立語数で割ったものを特徴量として与えた。

一人称の割合 一人称は書き手のキャラクタ性を示す指標となる。そこで各記事の一人称(私、僕、俺)の割合を特徴量として与えた。

疑問 KNPの解析結果から「疑問」という素性が付与された用言の数を数え、これを用言の数で割ったものを特徴量として与えた。

例 この小さな相棒の機嫌を損ねてしまったのだろうか。

敬語-丁寧表現 KNPの解析結果から「敬語-丁寧表現」という素性が付与された用言の数を数え、これを用言の数で割ったものを特徴量として与えた。

例 人に「やせたね」と言われるのがとても嬉しいです。

モダリティ　モダリティは書き手の判断や心的態度であり、受け手への様々な作用の原因となると考えられる。本研究では以下のモダリティを対象とする。

依頼 A、依頼 B、意志、認識-推量、認識-蓋然性、認識-証拠、勧誘、命令、評価-強、評価-弱、禁止

以下にモダリティの例を示す。

認識-推量: その正体が定かでなかったのだから不安が煽られたのでしょう。
命令: 京都で有名なお寺を挙げろと言われたら、
禁止: 「そもそも４人分食べるな」

KNP の解析結果より、各モダリティの数を数え、これらを用言の数で割ったものを特徴量として与えた。

3.3 語の知識より得られる特徴量

評価語　記事中での内容が肯定的か否定的かは、読み手の印象に大きな影響を与えると考えられる。そこで Web から共起パタンを利用して半自動収集された評価表現語辞書 [11] を用いて、各記事の肯定的な語および否定的な語の数を数えた。辞書の中で positive-、negative-とされる度合いの弱いものは別に数えた。これらを用言の数で割ったものを特徴量として与えた。

positive の例: お洒落、おいしい
positive-の例: しんみり、そわそわ
negative の例: 嫌い、悲しい
negative-の例: 何気ない、ファジー

接続詞の用法　接続詞は修辞に影響を与えると考えられる。接続詞は文頭では記事中での構成を示し、文中ではその文中での構成を示しているといった意味の違いがあると考えられるのでそれらを分けて数えた。また分類語彙表 [7] に基づいて、接続詞の用法を累加・展開・反対・選択・換言・補充・転換・理由に分けて数えた。これらを文数で割ったものを特徴量として与えた。

平均 IDF・TF·IDF　記事から得られる情報量の指標として平均 IDF と TF·IDF を用いる。各記事に含まれる名詞 (単名詞及び複合名詞) の IDF の平均および一記事中での合計である TF·IDF を特徴量として与えた。IDF の計算には検索エンジン TSUBAKI[2] [13] で検索対象となっている 1 億ページから獲得された複合名詞データベースから計算されたものを用いた。なお、一般的な語を排除するために Web で出現するページ数が 400,000 以下の名詞についてのみ IDF の値を計算した。

感情形容詞　形容詞のうちで、分類語彙表の「3.30-心」の項目に含まれるものの割合を特徴量として与えた。

例 しんどい、悲しい、かわいい、楽しい

話し言葉らしさ　記事中の機能語の話し言葉らしさ [6] の平均、および文平均、分散を特徴量として与えた。

話し言葉らしさが正の例 でもやっぱりスポーツをしていてしみじみ痛感するのは、男と女の体は根本的に違うなぁってこと。
話し言葉らしさが負の例 自分の身長よりも長い長刀を使う「しかけ応じ」は決まった型を演じるのだが、何度見てもあきない。

3.4 人の感じる面白さの要因と自動獲得できる特徴量の対応

人の感じる面白さの要因と自動獲得できる特徴量は表 1 のような対応があると考えられる。例えば、文長や行数などは修辞を近似する特徴量であり、IDF や数字表現などは情報量を近似する特徴量である。

4 ブログの面白さ判定

2 節で述べたブログデータを用いて面白さを推定する実験を行なった。タスクとしては採点者の平均点を推定するタスクと、個人の採点を推定するタスクの二つを行なった。

4.1 平均点の推定

各記事の平均点を学習データとして、未知の記事に対して、SVR (Support Vector Regression) によりその点数を推定した。SVR のパッケージとしては TinySVM[3]を用い、カーネルは二次多項式カーネルを用いた。249 件のブログの平均点を用いて、5 分割交差検定によって分類器を学習した。図 4 で説明すると、灰色の部分の素性と平均点から学習したモデルを用いて、1、2、3 番の記事の平均点を推定することになる。

[2] http://tsubaki.ixnlp.nii.ac.jp/index.cgi

[3] http://chasen.org/~ taku/software/TinySVM/

表 1: 面白さの要因と自動獲得できる要素の対応

面白さの要因	特徴量
修辞	文長 , 行数 , 段落数 , 一文長 , 文字種
論理構成	接続詞 , 段落数 , 段落長
共感	感動詞 , 感情形容詞 , モダリティ-禁止, 勧誘 , 命令
情報量	IDF , 鉤括弧対 , 数字表現 , 評価表現
説得力	数字表現 , IDF , モダリティ-認識
対話性	疑問符 , 疑問 , 話言葉らしさ, 評価表現
キャラクタ性	一人称 , 話し言葉らしさ, 敬語, モダリティ-禁止 , 命令

表 2: 平均点を推定した結果

	採点者間	システム	全て 4
二乗誤差平均	0.792	0.657	0.752
相関係数	0.401	0.485	-

結果を表 2 に示す。全て 4 点を出力したものをベースラインとした[4]。ここで「採点者間」は、採点者とそれ以外の採点者の平均点を比較したものであり、「システム」と「全て 4」は出力とすべての採点者の平均点を比較したものである。個人によって採点にばらつきがあるため、平均値を推定するシステムの方が、採点者間よりも数字がよくなる。また、結果を見ると二乗誤差平均、相関係数共にベースラインよりもシステムの方がよい結果であることがわかる。

4.2 個人の採点の推定

上のようにブログの平均的な面白さを求める問題について、ベースラインを上回る結果を得ることができた。しかし、採点者間の点数の分布からも分かるように個人間で面白さの基準は大きく異なる。そこで、各採点者ごとの各記事の点数を推定を行なうこととした。図 4 で説明すると、さきほどは平均点を推定していたが、今回は各マス目の点数 (例えば、1 番の記事の B さんの採点結果) を推定する。

249 件のブログの採点結果と特徴量を入力とし、SVR を用いて 5 分割交差検定により点数を推定した。以下の 4 つの手法を比較した。

1 つ目は、平均点により学習を行なった上記のモデルを用いて点数を推定した。システムの推定する点数は採点者によらず一定である。

2 つ目はその採点者の採点結果のみから学習したモデルを用いて点数を推定した (図 4 の太線内のみを利用する)。

3 つ目は、個々人の採点結果 (図 4 の太線内) をドメインだと考えてドメインアダプテーションを行なった。具体的には [1] の手法を用いた。$\mathbf{\Phi}$ を特徴

[4]この場合は一方が 4 で変動しないために相関係数は計算することができない

採点者

記事	A	B	C	…	J	K	L	平均点
1	5	5	5	…	7	4	7	5.6
2	7	4	4	…	3	5	6	5.3
3	6	4	6	…	4	4	5	4.8
4	3	5	4	…	3	3	2	3.5
5	1	3	5	…	7	4	6	3.8
⋮	…	…	…	…	…	…	…	…
⋮	…	…	…	…	…	…	…	…
246	3	4	2	…	3	6	2	3.3
247	4	3	3	…	2	6	1	2.8
248	3	5	1	…	1	1	3	2.3
249	1	2	1	…	2	2	7	2.1

図 4: 学習に用いる記事

量ベクトル (上記の実験で用いたもの)、$\mathbf{0}$ を $\mathbf{\Phi}$ と同次元の 0 ベクトルとすると、素性ベクトルは以下のようになる。

$$\boldsymbol{x} = (\mathbf{\Phi}, \mathbf{0}, ..., \mathbf{0}, \mathbf{\Phi}, \mathbf{0}, ..., \mathbf{0}) \quad (1)$$

上記のベクトルは、最初に $\mathbf{\Phi}$ をおき、その次に採点者のところを $\mathbf{\Phi}$、それ以外を $\mathbf{0}$ として採点者の人数だけ並べたものである。例えば、A さんの素性ベクトルは、

$$\boldsymbol{x} = (\mathbf{\Phi}, \mathbf{\Phi}, \mathbf{0}, \mathbf{0}, \cdots, \mathbf{0}) \quad (2)$$

となり、B さんの素性ベクトルは、

$$\boldsymbol{x} = (\mathbf{\Phi}, \mathbf{0}, \mathbf{\Phi}, \mathbf{0}, \cdots, \mathbf{0}) \quad (3)$$

となる。

このように素性ベクトルを与えることにより、$\mathbf{\Phi}$ のうち全採点者に共通で有効な素性は最初の $\mathbf{\Phi}$ で重みが学習され、各採点者の $\mathbf{\Phi}$ において、採点者固有の有効な素性の重みが学習される。

4 つ目は上記の実験でも利用した全て 4 を出力するベースラインである。

実験の結果を表 3 に示す。平均点しか用いない場合、各採点者の好みを反映させることができないためよい結果が得られないと考えられる。一方で個人の採点しか用いない場合には、個人の好みを反映することはできるが、データ数が少ないために一般的

表 3: 各採点者の点数の推定結果

	採点者間	平均値から学習	推定する採点者のみから学習	domain adaptation	全て 4
二乗誤差平均	1.85	1.38	1.38	1.36	1.43
相関係数	0.22	0.27	0.22	0.28	-

な面白さの要素の学習ができなかったためによい結果が得らなかったと考えられる。

ドメインアダプテーションを行なうことで、各人共通の一般的な面白さの評価と個人の好みの両方を反映させられることで、精度が向上したもの考えられる。

採点者間では 2.1.1 節で示したように、個々人で採点結果に開きがあるために低い値となっている。これは無作為に抽出された人間と好みの一致する割合は低く、個人の好みを反映させたシステムの方がより個人の好みに近い判断をする可能性を示せたと言える。

さきにあげた図 1 のような読み易い文体の記事などは面白さを正しく推定することができた。一方、図 2 のようなオチのある記事は正しく推定することができなかった。

5 考察

特徴量の考察を行なうために全記事を対象に線形カーネル SVR による学習を行ない得られた一次式

$$y = \boldsymbol{w} \cdot \boldsymbol{x} + b \tag{4}$$

の重みベクトル $\boldsymbol{w}$ のうち、大きく影響を与えているものを表 4 に示す。以下ではこれらからブログにおける面白さの要因を考察する。

修辞　文や記事の長さに関する特徴量が大きな影響力を与えていることが分かる。記事長は長い方が好まれるが、$(\text{記事長} - L)^2$ の重みが負であることから、長過ぎるものは評価が下がる傾向にあることも分かる。文数や一文長平均・分散の要素からは短かい文を連ねた形で、文の長さは大きな変化がない方がより好まれることが分かる。

また、一般に文章の上手な人間は接続詞を多用することなく読み易い文を書くために、接続詞全般が負の影響を与えていると考えられる。

情報量　TFIDF や鉤括弧対、数字の重みが正の数値である。これらは全て情報量を示す特徴量であり、これらが好意的に評価されているということは、読み手がブログに対して情報を得ることを求めている結果であると言える。

高圧的　評価語-negative やモダリティ禁止・命令などの重みから、記事の内容が高圧的なものは低い評価に繋がっていると思われる。

対話性　モダリティの勧誘が負の重みを持っている。これは今回のブログの場合では、投稿者と読み手が無関係な他人、コメントなどのコミュニケーションが行なわれていない、などによって勧誘に実効性が存在しないことが原因と考えられる。

話し言葉らしさは、語平均の話し言葉らしさについては負の重みが与えられる一方、文ごとの分散に対しては正の重みが与えられている。これらから分かることは、記事全体としては、話し言葉のような語り掛ける文体は嫌われる傾向にある。しかし記事中の一部においては投稿者の人間性が伝わるような要素が求められていると考えられる。

6 関連研究

我々の知る限りブログの内容の面白さを判定するシステムは現在存在しない。

小論文における自動評価は盛んに研究されており、英語では PEG [3]、e-rater [5]、IEA [2]、BETSY [4] などが、日本語では Jess [14] がある。

PEG ではまず、評価モデルを説明するための 2 つの概念を導入している。一つは、trins であり、流暢さ、語法、文法などといった、人が認識する要素である。これらは直接測ることができないため、その代用として、proxes を考える。proxes は trins の近似であり、自動獲得できるものである。例えば「流暢さ」という trins は「語数」という proxes と強い相関がある。そして、人手で採点された点数を目的変数、proxes を説明変数とする重回帰分析を行なうことで偏回帰係数を推定し、小論文のスコアを決定する。

e-rater でも同様に重回帰分析の手法を用いているが、手本となるエッセイとの tf-idf のコサイン類似度を用いることで内容がテーマに則しているかなども要素として取り入れている。

日本語の小論文自動採点システムである Jess はプロのライターの書いた文章を手本として小論文を評価するシステムである。採点基準としては、修辞、

表 4: SVR の重みベクトル

	正の重み	負の重み
修辞	文数 (0.36), 記事長 (0.27)	接続詞 (-0.27), 一文長平均 (-0.10), (記事長 $-\ L)^2(-0.08)$
情報量	鉤括弧対 (0.45), TF·IDF(0.24), 数字表現 (0.24)	
共感		感情形容詞 (-0.19), モダリティ-勧誘 (-0.17), 感動詞 (-0.07)
高圧的	評価語-positive(0.39), 敬語-丁寧表現 (0.09)	評価語-negative(-0.31), 命令 (-0.16), 評価強 (-0.07)
		モダリティ-禁止 (-0.05)
対話性	話し言葉らしさ分散 (0.06)	話し言葉らしさ単語平均 (-0.28), モダリティ-勧誘 (-0.17),
		疑問符 (-0.08)

論理構成、内容の 3 つを採用している。評価は 10 点満点とし、修辞 5 点・論理構成 2 点・内容 3 点としている。

本研究でも小論文の自動評価システムと同じように、面白さ判定を回帰問題として扱っている。しかし、テーマがあらかじめ決まっており、また正解を用意することのできる小論文に対して、ブログの内容は多様である、面白さの感じ方が人によって異なる、正解を用意することができない、などの点において大きく異なる。また、小論文の自動評価でも用いられている修辞・論理構成などの特徴量を本研究でも取り入れたが、評価表現・話し言葉らしさなどの特徴量は小論文の評価では用いられていない。

一方 Web 上でのコミュニケーションの研究としては、松村らによる「２ちゃんねる」の研究がある [12]。「２ちゃんねる」の盛り上がりを、議論深化傾向と議論発散傾向に分類し、共分散構造分析を用いることで各指標を用いた因果モデルを構築している。

7 おわりに

本研究では、自然言語処理に基づき、ブログの面白さの評価を行なうシステムを構築した。249 記事のブログを 12 人の採点者とシステムに評価させ、システムの出力と採点者の値の比較を行なった。その結果、ドメインアダプテーションの手法を利用することにより精度向上がみられた。したがって、個々の読み手の好みを反映することで、より読み手が面白いと思う記事を見つけ出すことができる可能性を示した。

今後の課題としては、特徴量を再度検討するとともに、話のオチといった要素を持つ記事を正しく評価できるような特徴量を取り入れることが考えられる。

また本研究ではブログの評価を行なうことに重点を置いたが、SVR の重みベクトルを研究することで人間がブログを採点する際に重視する点や個々の人間の好みを調べることも可能である。今後このような心理学的な立場に基づいた研究に利用することも検討している。

参考文献

[1] Hal Daume III. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 256–263, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[2] Landauer, T.K., Laham, D., Foltz, P.W. The intelligent essay assessor. *Debate on Automated Essay Grading*, Vol. 15, No. 5, pp. 27–31, 2000.

[3] E.B. Page. The imminence of grading essays by computer. *Phi Delta Kappa*, pp. 238–243, 1966.

[4] Runder, L.M., L Liang. Automated essay scoring using bayes' theorem. *National Council on Measurement in Education*, 2002.

[5] Jill Bursein Yigal Attali. Automated essay scoring with e-rater v.2. *The Jounal of Technology, Learning, and Assessment*, Vol. 4, No. 3, 2006.

[6] 玉城伸仁, 黒橋禎夫. 機能語句の話し言葉らしさ指標. 言語処理学会第 14 回年次大会, pp. 436–439, 2008.

[7] 国立国語研究所. 分類語彙表, 増補改訂版. 大日本図書, 2004.

[8] 黒橋禎夫, 河原大輔. 日本語構文解析システム KNP verson 2.0 使用説明書. 東京大学大学院情報理工学研究科, 2005.

[9] 黒橋禎夫, 河原大輔. 日本語形態素解析システム JUMAN verson 6.0 使用説明書. 京都大学大学院情報学研究科, 2007.

[10] 小学館辞典編集部編. 句読点、記号・符号活用辞典。小学館, 2007.

[11] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一. 意見抽出のための評価表現の収集. 自然言語処理, Vol. 12, No. 3, pp. 203–222, 2005.

[12] 松村真宏, 三浦麻子, 柴内康文, 大澤幸生, 石塚満. ２ちゃんねるが盛り上がるダイナミズム. 情報処理学会, Vol. 45, No. 3, pp. 1053–1061, 2004.

[13] 新里圭司, 柴田知秀, 河原大輔, 黒橋禎夫. 大規模日本語ウェブ文書を対象とした開放型検索エンジン基盤の構築. 言語処理学会第 13 回年次大会, pp. 1117–1120, 2007.

[14] 石岡恒憲. 小論文およびエッセイの自動評価採点における研究動向. 人工知能学会誌, Vol. 23, No. 1, pp. 17–24, 2008.