原著論文

複数の印象辞書を相互利用する印象マイニング手法の提案†

熊本 忠彦 *1・河合 由起子 *2・張 建偉 *3

本論文では、「楽しい⇔悲しい」、「うれしい⇔怒り」、「のどか⇔緊迫」という3種類の印象を対象に、テキストを読んだ人々が感じる印象の強さを数値的に求めるための手法を提案する。印象の強さ(すなわち印象値)を算出するためには、テキストから抽出される特徴量がテキストの印象に及ぼす影響力を数値化し、印象辞書に登録しておく必要がある。このような印象辞書は、従来、印象の種類ごとに構築され、それぞれの印象値を算出するためにだけ用いられるが、本論文では、それぞれの印象が独立ではない点に着目し、各印象辞書を用いて算出されるテキストの印象値から、それぞれの印象値を算出し直すという新たなアプローチを試みる。具体的には、テキストから抽出する特徴量として、単語 unigram と単語 bigram の2種類を定義し、既存手法を用いて新聞記事データベースから6つの印象辞書(3種類の印象×2種類の特徴量)を構築する。各印象辞書を用いて算出されるテキストの印象値を説明変数(6個)、アンケート調査に基づいて数値化されるテキスト本来の印象値を目的変数とする重回帰分析を印象の種類ごとに行い、それぞれの対応関係を重回帰式という形で定式化する。未知データに対する提案手法の精度を5分割交差検定により調べてみたところ、それぞれの印象における平均誤差は、1~7の7段階評価尺度に対して0.68, 0.57, 0.62であり、その有効性が確認された。

キーワード: 印象マイニング, 評判分析, 感情分類, 重回帰分析

1. はじめに

近年、テキストから感情や印象、評価といった主観的な情報を抽出しようという研究[1, 2]が増加傾向にある。例えば、映画レビューをその内容に基づいて positive, negative の2クラス、あるいは neutral を加えた3クラスに分類するという研究[3, 4]や、ニュース記事を8個の感情クラス(Awesome, Heartwarming, Surprising, Sad, Useful, Happy, Boring, Angry)に分類するという研究[5]、入力されたテキストに任意の印象語(印象を表す形容詞など)を付与するという研究[6, 7]、特定の印象の強さを数値的に求めるという研究[8, 9]などがある。抽出された感情や印象、評価を有効利用するアプリケーションの開発も、情報可視化[10, 11]や情報推薦[8, 12, 13]、異メディアコンテンツ生成[14, 15]、情報の信頼性評価[16]といった様々な分野で行われている。しかしながら、読み手がどう

受け取るかという観点から読み手が感じる印象(特に「悲しい」や「のどか」などの感情的印象)の強さを数値的に求めるという研究はまだ少ない。

そこで本論文では、新聞記事を例に、記事を読んだ人々が感じる印象の強さを数値的に求めるための手法を提案する。本手法が対象とする印象は、「楽しい⇔悲しい」、「うれしい⇔怒り」、「のどか⇔緊迫」の3種類であり、それぞれの印象に対して「(左側の印象を感じる(1点), 割と感じる(2点), やや感じる(3点), どちらも感じない(4点), (右側の印象を)やや感じる(5点), 割と感じる(6点), 感じる(7点)」という7段階の評価スケール(以下、印象尺度と呼ぶ)を割り当てる。すなわち、本手法は、それぞれの印象尺度において、そのスケールに準じた1.0~7.0の実数値(以下、印象値と呼ぶ)を、印象の強さを表す指標として出力する。

このような手法を設計するに当たり、まず、出力された印象値が、読み手が感じる印象の強さを表していることをどうやって担保するのかという問題を考える。本論文では、新聞記事データベースにおける任意の単語と特定の印象語との(記事内)共起関係に基づいて印象辞書(記事から抽出される特徴量の記事印象への影響力を示すもの)[9]を印象尺度ごとに構築し、各印象辞書を用いて算出される記事の印象値とその記事を読んだ人々が感じる印象値との対応関係を定式化す

† Using Multiple Impression Lexicons to Quantify Readers' Impressions of News Articles

Tadahiko KUMAMOTO, Yukiko KAWAI and Jianwei ZHANG

*1 千葉工業大学 情報科学部 情報ネットワーク学科

Chiba Institute of Technology

*2 京都産業大学 コンピュータ理工学部 インテリジェントシステム学科

Kyoto Sangyo University

*3 筑波技術大学 産業技術学部 産業情報学科

Tsukuba University of Technology

ることで、単語間の共起関係という読み手が介在しない方法で算出される印象値を読み手が感じる印象の強さへと変換する。また、このとき、それぞれの印象が独立ではない点に着目し、それぞれの印象辞書を相互に利用することにより、各印象尺度における記事の印象値をより高精度に算出し直すという今までにない、新たなアプローチを試みる。さらに、記事から抽出する特徴量として単語unigramと単語bigramを採用する。単語unigramは、記事特徴量としての網羅性が高く、使い勝手がいいことから、数多くの研究[4, 6, 8, 9]で採用されている。特にPangらの研究[4]では他の特徴量(単語bigramのみ、単語unigramと単語bigramの組み合わせ、単語unigramと品詞情報との組み合わせなど)を用いるよりも単語unigramのみを用いる方が精度が良かったことが示されている。一方、単語bigramは、網羅性が低い反面、単語同士の組み合わせを考慮に入れることができることから、名詞の連接や述語構造などを考慮する研究[3, 7]で採用され、その有用性が示されている。そこで本論文では、単語unigramと単語bigramの両方を用いることにする。

結局、本論文では、新聞記事から抽出する特徴量として、単語unigramと単語bigramの2種類を定義し、既存の手法[9]を用いて6つの印象辞書(3種類の印象尺度×2種類の特徴量)を構築する。そして、各印象辞書を用いて算出される記事の印象値を説明変数(6個)、アンケート調査に基づいて数値化されるテキスト本来の印象値を目的変数とする重回帰分析[17]を印象尺度ごとに行い、それぞれの対応関係を重回帰式という形で定式化する。このとき、変数選択法として変数増加法[17]を採用することで、変数間の独立性が乏しいときに発生する多重共線性の問題を回避し、記事の印象値を求めるのに適した説明変数を取捨選択する。

以下、2章で関連研究を整理し、提案手法との違いを示す。3章で本研究で対象とする印象尺度を決定する。4章で本論文で用いる既存手法(文献[9]で提案された印象辞書構築手法)について説明する。5章で既存手法からの拡張点である単語bigramの生成法と、この拡張に伴い修正された単語bigramによる印象辞書構築手法及び印象辞書を用いた記事印象値の算出法を述べるとともに、6つの印象辞書を用いて各記事から算出される印象値(6個の説明変数)と人々がその記事を読んだときに感じる印象値(目的変数)との対応関係を重回帰分析により定式化する。6章で提案手法の精度を評価し、その有効性を検証する。最後に7章で本論文のまとめと今後の課題について述べる。

2. 関連研究

2.1 テキストからの印象マイニング

情報発信者が書いたテキスト(例えば映画レビューや書評など)をその内容に基づいてpositive, negativeの2クラス、あるいはneutralを加えた3クラスに分類するという研究が盛んに行われている[3, 4]。例えば、Turneyは、各種レビューを「recommended」か「not recommended」に分類する手法を提案している[3]。彼の手法は、入力テキストから特定パターン(例えば「形容詞+名詞」や「副詞+形容詞+名詞以外」など)のフレーズを抽出し、各フレーズと参照語「excellent」および「poor」との自己相互情報量[18]をそれぞれ求め、差を取ることで、各フレーズのSemantic Orientation(SO)を決定している。そして、全フレーズのSOを平均することにより、入力テキストのSOを求め、その値により「recommended」か「not recommended」かを決定している。しかしながら、この手法は書き手がテキストに込めた評価を対象としており、読み手を意識した設計にはなっていない。また、1本の印象尺度(recommended ⇔ not recommended)に特化しており、複数の印象を対象としていない点が大きく異なる。さらに、印象の強さを数値化するというよりは2クラスへの分類問題として扱っている点も異なっている。

テキストをより多くのクラスに分類するという研究もある。例えば、Linらは、ニュース記事を8つの感情クラス(Awesome, Heartwarming, Surprising, Sad, Useful, Happy, Boring, Angry)に分類するSVMベースの手法を提案している[5]。具体的には、提示されたニュース記事に対して、指定された8つの感情のうちの1つを投票できるWeb上のニュースサイトを利用することで、それぞれのニュース記事本来の感情を決定し、各ニュース記事から抽出される特徴量(単語unigram, 文字bigram, affx similarityなど)と関連付けたものをSVMへの訓練データとして用いている。つまり、読み手を意識した研究であり、複数の感情を対象としている点で、著者らの研究と同じ方向性を持っているが、SVMによるクラス分類問題として扱っている点が印象の強さの数値化を目指している著者らの研究とは異なっている。

入力されたテキストに任意の印象語を付与するという研究がある。例えば、宮川らは、意味の数学モデル[19]を用いてテキストが有する任意の印象を抽出するための手法を提案している[6]。この意味の数学モデルは、文脈に応じた意味的連想を可能とする情報検索方式であり、検索に用いられるキーワード群を配置し

たメタデータ空間と呼ばれる正規直交空間から文脈を表す部分空間を選択し、その部分空間上での相関量に基づいて意味的に近いキーワード(印象語)の選択を可能にしている。この方式では、テキストの印象を表すキーワード(印象語)を文脈に応じて選択することが可能と考えられるが、特定の印象尺度に沿って印象の強さを数値的に求めることはできない。また、メタデータ空間を構築する際に、専門用語辞典や国語辞書といった言語資源における単語間の共起関係を用いているため、読み手を意識した手法とは言えない。

一方、清水らは、特定のフレーズパターンの出現頻度に基づいて形容詞同士の意味的關係や形容詞と名詞、形容詞と動詞との意味的關係を抽出し、さらに名詞と動詞の組み合わせに対する印象(特に「嬉しい」、「明るい」、「寒い」、「冷たい」、「重い」のような情景を表す形容詞)を推定する手法を提案している[7]。この手法の特徴として、印象推定の信頼性を印象適合値という数値で表している点が挙げられるが、印象の強さを数値的に求めることはできない。また、言語資源としてWebページを用いているため、読み手を意識した設計とは言えない。

著者らの研究と同様、特定の印象の強さを数値的に求めるという研究も行われている。例えば、阿部らは、「喜」、「怒」、「哀」、「楽」に「愛」を加えた5種類の感情を対象に、受信メールを読んだユーザが抱く感情の度合いを推定し、それぞれの感情の度合いに応じて楽曲を推薦する手法を提案している[8]。この手法の感情推定部分を設計するに当たり、阿部らは、訓練用のメールから抽出された特徴語(名詞、動詞、形容詞)のうち、tf・idf値が閾値以上のものを説明変数、各メールに対し、被験者が付与した各感情の評価点(5段階評価尺度)を目的変数とする重回帰分析を感情の種類ごとに行い、それぞれの対応関係を重回帰式で表している。しかしながら、メールから抽出される特徴語をそのまま説明変数としているため、汎用性(新規受信メールへの対処能力)という点で問題が生じる。すなわち、重回帰分析では、説明変数の数より多い数の訓練事例が必要となるが、この手法に汎用性を持たせるためには相当数の訓練事例が必要とされ、実際的ではない。もし訓練事例数が十分でないと、その分、説明変数の数を絞り込む必要があり、その結果、新規の受信メールに説明変数として選ばれた特徴語があまり含まれず、正確に感情推定できないということも考えられる。これに対し、著者らの研究では、記事から抽出された特徴量を用いて、仮の記事印象値を算出し、この印象値を説明変数として用いているので、記事から抽出する特徴量の数には制限がないという利点

がある。

一方、著者らは、先行研究[9]で、「楽しい⇔悲しい」、「うれしい⇔怒り」、「のどか⇔緊迫」という今回と同じ3種類の印象を対象に、印象辞書を用いて算出される記事の印象値を説明変数、その記事に対し被験者100人が与えた7段階評価値の平均値を目的変数とする重回帰分析を印象尺度ごとに行い、その結果得られた回帰式(3次関数あるいは5次関数)を用いて記事の印象値を補正するという手法を提案している。この研究も、読み手を意識しており、また印象の強さを数値化しようとしていることから、今回の研究と極めて近い研究といえるが、アプローチの仕方が異なっている。先行研究では、記事の印象値を補正するために重回帰分析を行っているが、今回の研究では重回帰分析を行っている。重回帰分析では説明変数が1つに制限されるが、重回帰分析では複数の説明変数を採用することができる。その結果、複数の印象辞書を用いて算出された印象値を線形結合することで記事の印象値を算出し直すことや記事特徴量として単語unigramだけでなく、単語bigramも利用することが可能となる。

2.2 印象を有効利用するアプリケーション

抽出された感情や印象、評価を有効利用するアプリケーションの開発も様々な分野で行われている。例えば、情報可視化の分野では、検索結果として得られるWebページ集合やWebニュースサイトから発信されているニュース記事集合の印象空間における分布を可視化するシステム[10, 11]が提案されており、情報推薦の分野では、話題と印象の両面においてユーザの興味を考慮したニュース記事推薦システム[12]や、ユーザの感性(「鮮やか」や「丸っこい」のような商品の物理的特徴を表す印象語と「クールな」や「かっこいい」のような商品全体のイメージを表す印象語との対応関係)をユーザとの対話により獲得し、商品の推薦に用いるシステム[13]、受信したメールを読んだユーザが抱く感情に合った楽曲を推薦するシステム[8]などが提案されている。また、異メディアコンテンツ生成の分野では、ニュース記事の印象に基づいてCGキャラクター(アナウンサ役)の声色やBGMを選択するニュース番組風アニメーション生成システム[14]や、物語のシーンの印象に合った声楽曲を生成するシステム[15]が提案されており、情報の信頼性に関する分野では、閲覧中のWebページの評価(ポジティブ、ネガティブ、中立)がWeb全体の中で多数派なのか少数派なのかを示すことにより、Web閲覧時の信頼性を示すシステム[16]が提案されている。

3. 印象尺度の設計

本研究では、著者の一人が先行研究[20]で設計した6本の印象尺度の中の3本を採用している。そこで本章では、その6本の印象尺度の設計手順と問題点を示し、「楽しい⇔悲しい」、「うれしい⇔怒り」、「のどか⇔緊迫」の3本を選択した根拠を明らかにする。

先行研究[20]では、新聞記事の印象特性(印象の現れ方に関する性質)を分析し、その結果に基づいて新聞記事を読んだ人々がその記事から受ける印象を定量的に捉えるための印象尺度(「楽しい⇔悲しい」、「うれしい⇔怒り」、「面白い⇔つまらない」、「楽観的⇔悲観的」、「のどか⇔緊迫」、「驚き⇔ありふれた」)を提案している。この6本の印象尺度の設計手順は以下のようになっている。

まずはじめに、新聞記事の印象特性を分析するために必要な印象評価データをアンケート調査に基づいて収集している。この調査では、被験者900人を9つのグループ(男女50人ずつ、計100人)に分け、グループごとに異なる記事を10記事用意している。各グループの被験者には、ランダムな順序で提示される10記事を順に読んでもらい、それぞれの記事の印象を印象尺度の候補となる42個の印象語を用いて「強い(1)、割と強い(2)、割と弱い(3)、弱い(4)、なし(5)」の5段階で評価してもらっている。次に、このようにして収集された印象評価データ(新聞記事10記事×印象語42語×被験者100人×9グループ)に対し、基本統計量(平均値と標準偏差)の分析や回帰分析、因子分析、クラスタ分析を行い、各印象語の特性や印象語間の関係を明らかにしている。その結果に基づいて、上述した6本の印象尺度が設計されている。

なお、印象尺度の候補となった印象語42語は、以下の手順で類語辞典[21]から選ばれている。まず、類語辞典から新聞記事の印象を表しうる印象語として532語を抽出している。次に、文脈依存性が高い印象語(「近い」、「大きい」、「低い」など)や文脈理解を必要とする印象語(「いい加減な」、「面倒な」、「無関係な」など)、文章構成による影響が大きい印象語(「上手な」、「下手な」、「シンプルな」など)を候補から除外している。さらに、同義語・類義語を整理し、結果として42個の単語グループを形成している。最後に、それぞれの印象語を「記事」という単語と組み合わせてGoogle上でフレーズ検索を行い、ヒット件数を参考に各グループの代表語1語(計42語)を決定している。

しかしながら、文献[20]では、次のような問題点も示されている。

- 「驚き」に関する印象は、人によって大きく異なっており、かつ、その異なり具合はいずれの記事においても同程度である。したがって、「驚き」に関する印象を扱うためには、何らかの個人適応が必要と考えられる。
- 印象尺度の使い勝手に関するアンケート調査の結果から、「面白い⇔つまらない」と「楽観的⇔悲観的」という2つの印象尺度は、ある種の記事(例えば殺人事件や自殺、事故死等に関する記事)に対しては不適切であることがわかった。

そこで本研究でも、文献[9]と同様、この3本の印象尺度を取り扱わないことにした。すなわち、本論文において対象となる印象は、「楽しい⇔悲しい」、「うれしい⇔怒り」、「のどか⇔緊迫」の3種類であり、それぞれの印象に対し「左側の印象を感じる(1)、割と感じる(2)、やや感じる(3)、感じない(4)、右側の印象をやや感じる(5)、割と感じる(6)、感じる(7)」という7段階の評価スケールを設定している。

4. 印象辞書の自動構築

新聞記事の印象の強さを数値的に求めるためには、記事から抽出される特徴量が記事の印象に与える影響力をあらかじめ数値化し、印象辞書に登録しておく必要がある。本論文では、既存の手法[9]を用いて各印象尺度における印象辞書を構築する。

4.1 単語unigramの生成

今回用いた手法[9]では、新聞記事を単語に分解するための形態素解析システムとしてJuman[22]を用いているが、一部の単語に対してはJumanの出力結果をそのままの形では用いていない。すなわち、表1に示したルールを再帰的に適用することにより、Jumanの出力結果に対して後処理を行っている。例えば、「削除しない」のような語は、Jumanによりサ変名詞「削除」、動詞「する」、形容詞性述語接尾辞「ない」の3語に分けられるが、ルール8と9を順に適用することにより、「削除しない」という動詞1語に変換される。

以上の処理の結果から助詞、連体詞、指示詞を取り除いたものが単語unigramとして利用される。

4.2 印象辞書の自動構築

次に、前節で生成された単語unigramを用いて印象辞書を自動構築する。

まず、5年分の読売新聞記事データ(2002年版～2006年版)を用いて印象尺度ごとに任意の単語unigram U と対比的な印象を有する2つの印象語群との共起の仕

表1 Juman の出力結果を変換するためのルール [9]

1	形容詞／動詞＋名詞性述語接尾辞のとき、この2語を普通名詞1語に変換する
2	名詞／未定義語／形容詞（語幹）／動詞（基本連用形）＋名詞性名詞接尾辞（「化」を除く）のとき、この2語を普通名詞1語に変換する
3	名詞／未定義語／形容詞（語幹）／動詞（基本連用形）＋名詞性名詞接尾辞「化」のとき、この2語をサ変名詞1語に変換する
4	接頭辞「御／ご／お」＋動詞（基本連用形）のとき、この2語をサ変名詞1語に変換する
5	動詞（基本連用形）＋格助詞のとき、動詞（基本連用形）をサ変名詞に変換する
6	名詞／未定義語＋名詞性特殊接尾辞（「都、道、府、県、郡、市、町、村、区、州、省」を除く）のとき、この2語を未定義語1語に変換する
7	形容詞／動詞＋動詞性接尾辞のとき、この2語を動詞1語に変換する
8	サ変名詞／カタカナ／アルファベット／副詞／形容詞（基本連用形／ダ列基本連用形）＋動詞「する／できる」のとき、この2語を動詞1語に変換する
9	動詞＋助動詞「ぬ」／形容詞性述語接尾辞「ない」のとき、この2語を動詞1語に変換する
10	名詞（形式名詞と副詞的名詞を除く）／未定義語／動詞（基本連用形）／副詞＋判定詞のとき、この2語を形容詞1語に変換する
11	形容詞／動詞／判定詞＋形容詞性述語接尾辞（「ない」を除く）のとき、この2語を形容詞1語に変換する
12	名詞／未定義語／動詞／形容詞＋形容詞性名詞接尾辞のとき、この2語を形容詞1語に変換する
13	形容詞＋形容詞性述語接尾辞「ない」のとき、この2語を形容詞1語に変換する
14	形式名詞／副詞的名詞／助詞＋判定詞のとき、この2語を判定詞1語に変換する
15	判定詞＋形容詞性述語接尾辞「ない」のとき、この2語を判定詞1語に変換する
16	形容詞（ダ列タ系連用テ形／基本連用形）／動詞（タ系連用テ形）／判定詞（ダ列タ系連用テ形）＋副助詞「は／も」のとき、副助詞を削除する
17	接頭辞（「御、ご、お」を除く）＋任意の形態素のとき、この2語を1語にする

方を調べ、どちらの印象語群とより共起しやすいかを数値化したものを、 U の当該印象尺度における印象値として印象辞書に登録する。具体的な手順を以下に示す。

まずはじめに、各印象尺度の左側の印象を表す印象語群 IW_L と右側の印象を表す印象語群 IW_R を定義し、解析対象となる新聞記事データから印象語群 IW_L あるいは IW_R に含まれる印象語を1語以上含む記事を抽出するとともに、各記事に含まれる印象語の数を印象語群ごとに数える。以上の結果、印象語群 IW_L に属する印象語の数が印象語群 IW_R に属する印象語の数よりも多かった記事の集合を S_L (記事数を N_L) とし、逆に少なかった記事の集合を S_R (記事数を N_R) とする。このとき、同数だった記事の集合は利用せず、破棄する。次に、それぞれの記事集合 (S_L もしくは S_R) からすべての単語を抽出し、前節で述べた手法を用いて単語unigramを生成するとともに、その出現記事数

を数える。このとき、ある単語unigram U の記事集合 S_L における出現記事数を $N_L(U)$ 、記事集合 S_R における出現記事数を $N_R(U)$ とすると、それぞれの条件付き出現確率は次のように表される。

$$P_L(U) = \frac{N_L(U)}{N_L}, \quad P_R(U) = \frac{N_R(U)}{N_R}$$

この $P_L(U)$ と $P_R(U)$ を用いて、単語unigram U の印象値 $v(U)$ を次のような式で表している。

$$v(U) = \frac{P_L(U) * weight_L}{P_L(U) * weight_L + P_R(U) * weight_R}$$

$$weight_L = \log_{10} N_L, \quad weight_R = \log_{10} N_R$$

すなわち、単語unigram U の印象語群 IW_L に対する条件付き出現確率と印象語群 IW_R に対する条件付き出現確率の重み付き内分比を求め、単語unigram U の

印象尺度「 $IW_L \Leftrightarrow IW_R$ 」における印象値として印象辞書に登録している。なお、 $weight_L$ や $weight_R$ は重みであり、条件を満たす記事数 N_L や N_R が多いほど大きくなるように設計されている。

ここで、それぞれの印象尺度において印象辞書を構築するために用いた印象語群を表2に示す。これらの印象語群は、i)それぞれの印象尺度の印象を表す単語

(動詞もしくは形容詞)であること、ii)語義の多様性により他の印象を(なるべく)持たない単語であること、という基準に基づいて決められており、文献[9]と同じ構成になっている。

5. 重回帰分析に基づく記事印象値算出法の設計

本章では、今回用いた印象辞書構築手法[9]からの拡張点である単語bigramの生成法と、この拡張に伴い修正された単語bigramによる印象辞書構築手法及び印象辞書を用いた記事印象値の算出法を述べるとともに、重回帰分析に基づいて記事の印象値を算出し直すための重回帰式を生成する。

5.1 単語bigramの生成

本節では、4.1節に示した後処理の結果に、表3に示したルールを適用することにより、単語bigramを生成する。

表2 各印象尺度を構成する印象語群 [9]

印象尺度	印象語
楽しい $\Leftrightarrow$ 悲しい	楽しい, 楽しむ, 楽しみだ, 楽しげだ 悲しい, 悲しむ, 悲しみだ, 悲しげだ
うれしい $\Leftrightarrow$ 怒り	うれしい, 喜ばしい, 喜ぶ 怒る, 憤る, 激怒する
のどか $\Leftrightarrow$ 緊迫	のどかだ, 和やかだ, 素朴だ, 安心だ 緊迫する, 不気味だ, 不安だ, 恐れる

表3 単語 bigram 生成ルール

核となる単語が普通名詞／カタカナのとき

- 直前にある動詞／形容詞／判定詞。但し、その活用形が基本形／タ形／文語基本形／デアル列基本形／基本連体形／文語連体形／ダ列基本連体形／ダ列文語連体形／ダ列特殊連体形／その他の基本形／その他の連体形の場合のみ
- 直前にある名詞（形式名詞と副詞的名詞を除く）／未定義語／形容詞（語幹）。但し、核となる単語との間に1個以上の格助詞／接続助詞があってもよい

核となる単語がサ変名詞のとき

- 直前にある動詞／形容詞／判定詞。但し、その活用形が基本形／タ形／文語基本形／デアル列基本形／基本連体形／文語連体形／ダ列基本連体形／ダ列文語連体形／ダ列特殊連体形／その他の基本形／その他の連体形の場合のみ
- 前出する動詞／形容詞／判定詞。但し、その活用形が基本連用形／文語連用形／ダ列基本連用形／ダ列特殊連用形／ダ列文語連用形／その他の連用形の場合のみ
- 前出する名詞（形式名詞と副詞的名詞を除く）／未定義語／形容詞（語幹）／副詞

核となる単語が動詞／形容詞のとき

- 前出する動詞／形容詞／判定詞。但し、その活用形が基本連用形／文語連用形／ダ列基本連用形／ダ列特殊連用形／ダ列文語連用形／その他の連用形の場合のみ
- 前出する名詞（形式名詞と副詞的名詞を除く）／未定義語／形容詞（語幹）／副詞

核となる単語が判定詞のとき

- 直前にある動詞／形容詞。但し、その活用形が基本形／タ形／文語基本形／デアル列基本形／基本連体形／文語連体形／ダ列基本連体形／ダ列文語連体形／ダ列特殊連体形／その他の基本形／その他の連体形の場合のみ

単語bigramの核となる単語は、表3に示したように、普通名詞、カタカナ、サ変名詞、動詞、形容詞、判定詞であり、それぞれの単語に対してルールに整合する単語が同一文内にあるかどうかをチェックし、あれば、ペアとして単語bigramを生成する。このチェックは、核となる単語を基準に文頭方向に対して行われるが、動詞／形容詞／判定詞／サ変名詞が現れた時点で打ち切られる。但し、他の品詞(主に普通名詞や未定義語)が現れても打ち切られないので、その結果、1つの核となる単語から複数の単語bigramが生成されることもある。

5.2 単語bigramによる印象辞書の自動構築

単語bigramを記事特微量とする場合の印象辞書の自動構築は、4.2節で述べた、単語unigramの記事特微量とする場合の手法と同じアルゴリズムで行われる。すなわち、それぞれの式において単語unigram U の代わりに、単語bigram B を用いればよい。

5.3 印象辞書を用いた記事印象値の算出

印象辞書を用いて新聞記事の印象値を算出する手法を以下に示す。

まず、記事をJumanを用いて単語に分解し、表1に示したルールを用いて後処理を行った後、4.1節に示した方法で単語unigramを生成し、5.1節に示した方法で単語bigramを生成する。次に、生成された単語unigramの印象値をあらかじめ構築されている印象辞書から取り出し、印象尺度ごとに平均値を計算する。生成された単語bigramに対しても同様に、印象尺度ごとの平均値を計算する。以上の結果得られた単語unigramの平均値と単語bigramの平均値をそれぞれ、その記事の当該印象尺度における印象値として出力する。つまり、1つの記事に対して全部で6個(3種類の印象尺度×2種類の記事特微量)の印象値が算出されることになる。なお、単語unigramの処理と単語bigramの処理は独立に行われるので、1つの記事から共通部分を有する単語unigramと単語bigramが生成されても、それぞれ別個に計算される。

5.4 重回帰分析に資するデータの収集

まずはじめに、新聞記事とその記事から受ける印象との対応関係を示すデータを得るために、900人(男女450人ずつ)が参加するアンケート調査を行った。具体的には、回答者900人を9つのグループ(男女50人ずつ、計100人)に分け、各グループに毎日新聞の2002年版社会面[23]に掲載された10記事を提示した。この10記事はグループによって異なっており、全部で90

記事が重複しないように選ばれている。各回答者は、ランダムに提示される10記事の印象をランダムな順番で提示される3種類の印象尺度を用いて7段階評価した。すなわち、「楽しい⇔悲しい」、「うれしい⇔怒り

(a) 「楽しい⇔悲しい」の場合

(b) 「うれしい⇔怒り」の場合

(c) 「のどか⇔緊迫」の場合

図1 重回帰分析に資するデータの散布図

表4 印象辞書を用いて算出された換算値と回答者が付けた印象値の平均値との相関係数

換算値（記事特徴量別）	平均値（印象尺度別）		
	楽しい ⇄ 悲しい	うれしい ⇄ 怒り	のどか ⇄ 緊迫
Unigram 楽しい⇄悲しい	0.76	0.73	0.72
Bigram 楽しい⇄悲しい	0.72	0.63	0.66
Unigram うれしい⇄怒り	0.80	0.84	0.79
Bigram うれしい⇄怒り	0.76	0.81	0.75
Unigram のどか⇄緊迫	0.71	0.73	0.78
Bigram のどか⇄緊迫	0.63	0.58	0.71

り」,「のどか⇄緊迫」のそれぞれに対し, 対応する印象をどの程度感じるかを「(左側の印象を感じる(1点), 割と感じる(2点), やや感じる(3点), どちらも感じない(4点), (右側の印象を)やや感じる(5点), 割と感じる(6点), 感じる(7点)」の7段階で評価した。以上の結果得られたデータから各記事の各印象尺度における平均値を求めた。この平均値が重回帰分析における目的変数となる。なお, 各回答者に提示した記事は, 元の記事の第1段落のみであり, 第2段落以降は提示していない。これは, 記事の構成上, 第1段落を読めば記事の概要がわかるように書かれている点や段落ごとに記事の印象が変わる可能性がある点, 記事が長いと回答者にかかる負担が増大する点を考慮した結果である。但し, 将来的には1つの記事の中での印象の推移を追跡できる印象マイニングを実現したいと考えている。

次に, 4.2節と5.2節で構築した印象辞書と5.3節で示した記事印象値の算出法を用いて, 回答者らに提示した各記事(第1段落のみ)から6個の印象値を算出した。なお, 算出される印象値(算出値)は, 印象尺度の左側の印象が強いと1に近づき, 右側の印象が強いと0に近づくように設計されているが, アンケート調査では印象尺度の左側の印象が強いときは1に近づき, 右側の印象が強いときは7に近づくという設計になっていたので,

$$\text{換算値} = (1 - \text{算出値}) * 6 + 1$$

という式を用いて同じスケールになるよう算出値を換算した。

ここで, 全90記事に対する, 換算された印象値(6個の説明変数)と回答者らが付けた印象値の平均値(目的変数)との対応関係を印象尺度ごとに整理し, 図1に示す。図中の点は, 印象値の算出の際に用いた印象辞書(どの印象尺度のために用意された印象辞書か)と特徴量の種類(単語unigramによるものか単語bigram

によるものか)によって区分されており, 各区分ごとに90個の点がプロットされている。但し, プロットが重なる場合は, 「楽しい⇄悲しい」が上, 「のどか⇄緊迫」が下となり, 印象尺度が同じ場合は, 単語bigramの方が上になる。

図1に示した散布図を見てみると, それぞれの換算値と平均値の間には正の相関があることがわかる。実際にそれぞれの相関係数を計算してみると, 表4のとおりであり, 0.58~0.84という値であった。また, それぞれの特徴量の分布範囲に違いがあることもわかる。これらのことは, 6個の換算値と平均値との対応関係を重回帰分析により定式化できることを示唆している。なお, 表4を見てみると, それぞれの換算値に対し最も高い相関係数を持つ平均値は, 同じ印象尺度のものとなっており, 表2で設定した印象語群の妥当性を示している。

5.5 重回帰分析による重回帰式の生成

前節で収集したデータを用いて, 換算された印象値を説明変数, 回答者らが付けた印象値の平均値を目的変数とする重回帰分析を印象尺度ごとに行った。このとき, 多重共線性の問題を引き起こす可能性のある説明変数を除去するために変数増加法[17]を採用し, 説明変数の取捨選択を行った。以上の処理の結果, 表5に示す重回帰式が得られた。なお, 参考のために, 表5には各説明変数の重回帰式への寄与の大きさを表す標準偏回帰係数も示す。

結局, ある新聞記事の, ある印象尺度における印象値は, 対応する重回帰式に, その記事から算出される各説明変数の値を代入することにより, 得られる。例えば, ある記事の印象尺度「のどか⇄緊迫」における印象値 y は, その記事から算出される説明変数「Unigram うれしい⇄怒り」, 「Bigram うれしい⇄怒り」, 「Unigram のどか⇄緊迫」, 「Bigram のどか⇄緊迫」の値(換算値)をそれぞれ x_1, x_2, x_3, x_4 とすると,

表5 重回帰分析により設計された重回帰式

(a) 印象尺度「楽しい⇔悲しい」

説明変数	偏回帰 係数	標準偏回帰 係数
Bigram 楽しい⇔悲しい	0.249	0.262
Unigram うれしい⇔怒り	1.221	0.597
(定数項)	-1.259	

(b) 印象尺度「うれしい⇔怒り」

説明変数	偏回帰 係数	標準偏回帰 係数
Unigram うれしい⇔怒り	1.146	0.575
Bigram うれしい⇔怒り	0.267	0.304
(定数項)	-1.102	

(c) 印象尺度「のどか⇔緊迫」

説明変数	偏回帰 係数	標準偏回帰 係数
Unigram うれしい⇔怒り	0.399	0.203
Bigram うれしい⇔怒り	0.250	0.289
Unigram のどか⇔緊迫	0.540	0.202
Bigram のどか⇔緊迫	0.298	0.238
(定数項)	-2.221	

$y = 0.399x_1 + 0.250x_2 + 0.540x_3 + 0.298x_4 - 2.221$ という式で算出される。ここで提案手法が出力する印象値と回答者らが付けた印象値の平均値との相関係数を求めてみると、0.81～0.85という値であり、5.4節で示した換算値と平均値との相関係数(0.58～0.84)に比べ、高くなっているのがわかる。なお、各重回帰式の決定係数は、0.66～0.73であり、いずれの印象尺度においても0.5より高く、重回帰分析の結果が良好であったことを示している。

5.6 生成された重回帰式の解釈

表5に示した各重回帰式において、標準偏回帰係数を見てみると、印象尺度「楽しい⇔悲しい」では、「Unigram うれしい⇔怒り」が0.597、「Bigram 楽しい⇔悲しい」が0.262となっており、「Unigram うれしい⇔怒り」が目的変数をよく説明していることになる。言い換えれば、表2に示した印象尺度「うれしい⇔怒り」のための印象語群のうち、「うれしい、喜ばしい、喜ぶ」と共起しやすく、「怒る、憤る、激怒する」と共起しにくい単語unigramは、楽しいという印象を生じさ

せ、その逆の共起パターンを持つ単語unigramは、悲しいという印象を生じさせるということの意味している。印象尺度「うれしい⇔怒り」でも同様に、「Unigram うれしい⇔怒り」が0.575、「Bigram うれしい⇔怒り」が0.304と、「Unigram うれしい⇔怒り」の値が最も高いが、次点の説明変数も同じ印象尺度の特徴量なので、結局、「うれしい、喜ばしい、喜ぶ」と共起しやすく、「怒る、憤る、激怒する」と共起しにくい単語unigram・単語bigramは、うれしいという印象を生じさせ、その逆の共起パターンを持つ単語unigram・単語bigramは、怒りという印象を生じさせるということの意味している。一方、印象尺度「のどか⇔緊迫」では、「Bigram うれしい⇔怒り」の値がわずかに高いが、0.289と低い値であり、4つの説明変数の値はほぼ拮抗している。これは、表2の印象尺度「うれしい⇔怒り」と「のどか⇔緊迫」において、上段の印象語群と共起しやすく、下段の印象語群と共起しにくい単語unigram・単語bigramは、のどかという印象を生じさせ、その逆の共起パターンを持つ単語unigram・単語bigramは、緊迫という印象を生じさせるということの意味している。

ここで、それぞれの重回帰式において、目的変数と各説明変数の相関係数を表4により調べてみると、いずれの目的変数に対しても「Unigram うれしい⇔怒り」が最も高い値であることがわかる。複数の印象尺度を設計するとき、それぞれの印象尺度の違いを印象語群にどう反映させるかは、1本の評価尺度を対象とする評判分析研究にはない、難しい問題であり、現段階では印象尺度を構成する印象語(楽しい、悲しい、うれしい、怒り、のどか、緊迫)の同義語・類義語をベースに試すということを行っている。重回帰分析の結果、いずれの目的変数に対しても「Unigram うれしい⇔怒り」が最も高い相関係数を持つ説明変数となったわけだが、この説明変数だけでは十分ではなく、「楽しい⇔悲しい」に対しては「Bigram 楽しい⇔悲しい」が採用され、「のどか⇔緊迫」に対しては「Unigram のどか⇔緊迫」と「Bigram のどか⇔緊迫」が採用されている。これは、「Unigram うれしい⇔怒り」とそれぞれの印象尺度に対応した説明変数とが組み合わせることで、より高い精度が実現されていることを示している。一方、重回帰分析(変数増加法)を採用した今回のアプローチでは、精度の向上に貢献する印象語群を自由に設計し、追加することができる。すなわち、それぞれの印象尺度を表現する印象語群を一意に決定するのが困難な中、いろいろな印象語群を試してみることができるというのも本アプローチの利点の一つと考えられる。

表6 重回帰式の導入による誤差（Root-Mean-Square Error）の減少

印象尺度	導入前		導入後 (提案手法)	改善率	
	単語 unigram	単語 bigram		単語 unigram	単語 bigram
楽しい ⇔ 悲しい	0.94	1.34	0.64	32.0%	52.1%
うれしい ⇔ 怒り	0.83	1.05	0.56	32.4%	46.9%
のどか ⇔ 緊迫	0.82	1.22	0.57	30.1%	52.8%

6. 提案手法の有効性検証

本章では、学習データと未知データに対する提案手法の精度を評価し、その有効性を検証する。

6.1 学習データに対する精度評価

まず、重回帰分析の際に用いた全90記事を対象に、提案手法が出力する印象値と回答者らが付与した印象値の平均値との誤差が重回帰式を導入する前と後でどう変化するかを調べた。但し、重回帰式を導入する前は、複数の記事特徴量を同時に利用することができないので、それぞれの印象尺度において算出された単語 unigram (あるいは単語 bigram) の平均値 (すなわち算出値) を5.4節に示した換算式に代入し、スケール変換した結果 (すなわち換算値) を提案手法が出力する印象値としている。一方、重回帰式を導入した後の印象値は、表5の重回帰式に各印象尺度の換算値を代入することによって算出された理論値のことを指している。調べた結果を表6にまとめる。なお、誤差は、全90記事に対する平均値と換算値 (もしくは理論値) の差分平方和を記事数 (=90) で割り、平方根をとったものであり、いわゆるRoot-Mean-Square Error (RMSE) となっている。また、改善率は、

$$\text{改善率} = \frac{\text{導入前の誤差} - \text{導入後の誤差}}{\text{導入前の誤差}} \times 100$$

と定義する。

表6によれば、すべての印象尺度において、特徴量の種類に関係なく、高い改善率を示しており、重回帰式を導入することで、精度が向上することが確認された。一方、導入前の単語 unigram と単語 bigram の精度を比べてみると、単語 bigram の方が精度が低い。そこで、それぞれの印象尺度における換算値 (同じ印象尺度のみ) と平均値の相関係数を表4で調べてみると、単語 unigram の場合で0.76～0.84、単語 bigram の場合で0.71～0.81と単語 bigram の方が若干低くなっているが、それほど差ではない。ここで、図1の散布

表7 記事特徴量の有効性検証

印象尺度	単語 unigram	単語 bigram	提案手法 (併用)
楽しい⇔悲しい	0.66	0.65	0.64
うれしい⇔怒り	0.58	0.61	0.56
のどか⇔緊迫	0.61	0.60	0.57

図を見てみると、単語 bigram における換算値の分布範囲が平均値の分布範囲と若干ずれているのがわかる。このずれが精度低下の主な原因と考えられる。

次に、記事特徴量として、単語 unigram のみを用いる場合、単語 bigram のみを用いる場合、単語 unigram と単語 bigram を併用する場合 (提案手法) の各々の場合において、全90記事を対象とする重回帰分析を印象尺度ごとに行い、重回帰式を生成した。そして、それぞれの重回帰式を用いて算出される記事の印象値と回答者らが付与した印象値の平均値との誤差を調べた。結果を表7にまとめるとともに、生成された重回帰式を表8と表9に示す。なお、提案手法の重回帰式はすでに表5に示されているので、ここでは示さない。

表7から、単語 unigram と単語 bigram を併用した場合に最も高い精度を得ているのがわかる。さらに、重回帰式を導入して複数の印象辞書の相互利用を可能にすることで、単語 unigram のみの場合や単語 bigram のみの場合でも、ほぼ同程度の高い精度が得られているのがわかる。特に、単語 bigram のみの場合の精度の向上が著しいのは、前述したずれを重回帰式が補正したためと考えられる。実際、重回帰式の当てはまり具合を表す決定係数は、単語 unigram のみの場合で0.64～0.71、単語 bigram のみの場合で0.66～0.68とほぼ同程度であった。ちなみに、単語 unigram と単語 bigram を併用した場合の決定係数は、0.66～0.73と若干高めであった。

ここで、提案手法 (表5) と単語 unigram のみを用いた手法 (表8) の重回帰式を比べてみると、以下のよう

表8 重回帰分析により設計された重回帰式（単語 unigram のみを用いた場合）

(a) 印象尺度「楽しい⇔悲しい」

説明変数	偏回帰 係数	標準偏回帰 係数
Unigram 楽しい⇔悲しい	0.570	0.237
Unigram うれしい⇔怒り	1.195	0.584
(定数項)	-2.570	

(b) 印象尺度「うれしい⇔怒り」

説明変数	偏回帰 係数	標準偏回帰 係数
Unigram うれしい⇔怒り	1.680	0.843
(定数項)	-2.247	

(c) 印象尺度「のどか⇔緊迫」

説明変数	偏回帰 係数	標準偏回帰 係数
Unigram うれしい⇔怒り	0.945	0.482
Unigram のどか⇔緊迫	0.996	0.372
(定数項)	-3.969	

まず、印象尺度「楽しい⇔悲しい」において、提案手法では「Unigram うれしい⇔怒り」に加え、「Bigram 楽しい⇔悲しい」が採用されているが、単語 unigram のみを用いた手法ではこの「Bigram 楽しい⇔悲しい」の代わりに「Unigram 楽しい⇔悲しい」が採用されている。この2つの説明変数「Bigram 楽しい⇔悲しい」、「Unigram 楽しい⇔悲しい」と目的変数の相関係数を表4で調べてみると、それぞれ0.72、0.76となっており、「Unigram 楽しい⇔悲しい」の方が若干高い。にもかかわらず、決定係数は、提案手法が0.66、単語 unigram のみを用いた手法が0.64と、提案手法の方が少し高い値となっている。このことは、「Unigram うれしい⇔怒り」に「Bigram 楽しい⇔悲しい」を組み合わせることで、相補的な効果が生まれ、決定係数が向上したことを示している。一方、印象尺度「うれしい⇔怒り」と「のどか⇔緊迫」に対しては、単語 unigram のみを用いた手法の説明変数が提案手法の説明変数のサブセットとなっており、また決定係数も、提案手法が0.73と0.71、単語 unigram のみを用いた手法が0.71と0.67と、どちらの印象尺度においても提案手法の方が高くなっていることから、提案手法の方が高い精度を得て

表9 重回帰分析により設計された重回帰式（単語 bigram のみを用いた場合）

(a) 印象尺度「楽しい⇔悲しい」

説明変数	偏回帰 係数	標準偏回帰 係数
Bigram 楽しい⇔悲しい	0.216	0.227
Bigram うれしい⇔怒り	0.414	0.459
Bigram のどか⇔緊迫	0.296	0.227
(定数項)	0.644	

(b) 印象尺度「うれしい⇔怒り」

説明変数	偏回帰 係数	標準偏回帰 係数
Bigram うれしい⇔怒り	0.625	0.712
Bigram のどか⇔緊迫	0.216	0.170
(定数項)	1.013	

(c) 印象尺度「のどか⇔緊迫」

説明変数	偏回帰 係数	標準偏回帰 係数
Bigram うれしい⇔怒り	0.446	0.516
Bigram のどか⇔緊迫	0.515	0.412
(定数項)	-0.058	

いるのは当然のことと言える。

次に、提案手法(表5)と単語 bigram のみを用いた手法(表9)の重回帰式を比べてみると、以下のようなことがわかる。

まず、印象尺度「楽しい⇔悲しい」において、提案手法では「Bigram 楽しい⇔悲しい」と「Unigram うれしい⇔怒り」が採用されているが、単語 bigram のみを用いた手法では「Unigram うれしい⇔怒り」の代わりに、「Bigram うれしい⇔怒り」と「Bigram のどか⇔緊迫」が採用されている。各説明変数と目的変数の相関係数を表4で調べてみると、「Unigram うれしい⇔怒り」が0.80なのに比べ、「Bigram うれしい⇔怒り」が0.76、「Bigram のどか⇔緊迫」が0.63と若干低めの値となっているのがわかる。すなわち、「Unigram うれしい⇔怒り」の重回帰式への寄与を「Bigram うれしい⇔怒り」と「Bigram のどか⇔緊迫」とで代替しようとしており、その結果、0.66という同じ値の決定係数が得られている。一方、印象尺度「うれしい⇔怒り」に対しては、「Unigram うれしい⇔怒り」(提案手法)と「Bigram

のどか⇄緊迫」(単語bigramのみを用いた手法)が置き換わっており、それぞれの目的変数との相関係数が0.84と0.58であることから、その差が精度の差となって現れているのがわかる。また、印象尺度「のどか⇄緊迫」においては、単語bigramのみを用いた手法の説明変数が提案手法の説明変数のサブセットとなっており、また決定係数も、提案手法が0.71、単語bigramのみを用いた手法が0.68となっていることから、精度にも若干の差が出ている。

6.2 未知データに対する精度評価

次に、5分割交差検定により未知データに対する精度評価を行った。具体的には、(1)90記事分の学習データを5分割し、18記事に対する換算値と平均値のデータセットを5つ作成する、(2)この5つのデータセットのうちの4つ(72記事分の換算値と平均値)を用いて重回帰分析を行い、それぞれの印象尺度に対して重回帰式を生成する、(3)残りのデータセット(18記事分の換算値と平均値)を未知データとし、換算値を重回帰式に代入することにより、理論値を得る、(4)この理論値と平均値の誤差を求める、という手順をすべての組合せ(5通り)に対して行った。その結果得られた誤差(5回分)の平均誤差を表10に示す。また、比較のため、2章で示した著者らの先行研究[9]の未知データに対する誤差も表10に示す。

まず、表10に示された未知データに対する精度と表6に示された学習データに対する精度を比べてみると、未知データに対する精度は若干低くなるが、ほぼ同等であることがわかる。次に、先行研究の手法と精度を比較してみると、「うれしい⇄怒り」に対しては精度が低くなったが、「楽しい⇄悲しい」と「のどか⇄緊迫」に対しては改善しており、今回提案したアプローチでも先行研究の手法と同等以上の精度を実現できるということがわかった。著者らの手法の長所は、説明変数の数を増やすことができる点にあり、印象尺度(すなわち印象辞書)を追加したり、高次のN-gramを導入したり、あるいは先行研究の手法を取り込んだりすることで、更なる精度向上が期待できる。

表10 5分割交差検定による精度評価

印象尺度	提案手法	先行研究 [9]
楽しい ⇄ 悲しい	0.68	0.69
うれしい ⇄ 怒り	0.57	0.49
のどか ⇄ 緊迫	0.62	0.64

6.3 学習データに対する誤差解析

ここで、学習データ(90記事×3種類の印象尺度)を対象に誤差(平均値と理論値の差の絶対値)が大きかった記事の内容を調べてみた。

まず、学習データから平均値と理論値の差が大きかった5記事を印象尺度ごとに抽出し、15記事を得た。各印象尺度における誤差の分布は、「楽しい⇄悲しい」の場合で0.98～1.27、「うれしい⇄怒り」の場合で0.91～1.17、「のどか⇄緊迫」の場合で0.96～1.28であった。

この15記事の各々の平均値は、4.99～6.33の範囲に分布しており、総じて負の印象(悲しい、怒り、緊迫)が強めであることがわかる。つまり、提案手法は、この15記事の印象を実際よりも弱く評価していたことになる。各記事の内容を調べてみると、その主な原因として、以下の2点に気付く。一つは、負の印象の強い単語があっても、そうでない単語が多いと、記事の印象(理論値)が弱められてしまうという点であり、これは、記事から抽出される特徴量の印象値を単純に平均している点に問題があるといえる。もう一つは、個々の単語にはあまり負の印象の強いものはないが、記事全体としては強い印象を感じる場合があるという点であり、これは、提案手法が個々の単語の印象値のみを処理対象とし、話題などの大局的な情報を取り入れていない点に問題があるといえる。

次に、逆の場合、すなわち理論値と平均値の差が大きかった5記事を印象尺度ごとに抽出し、15記事を得た。各印象尺度における誤差の分布は、「楽しい⇄悲しい」の場合で1.06～1.66、「うれしい⇄怒り」の場合で0.91～1.32、「のどか⇄緊迫」の場合で0.95～1.30であった。

この15記事の各々の平均値は、2.75～5.18と広い範囲に分布しているが、詳しく見てみると、正の印象(楽しい、うれしい、のどか)がやや強いもの(平均値は2.75～3.21に分布)と負の印象がやや強いもの(平均値は3.95～5.18に分布)とに分けることができた。正の印象がやや強い記事は4つあり、その内容を調べてみると、平均値と理論値の差が大きかった場合と同様、正の印象の強い単語よりそうでない単語が多く、記事の印象(理論値)が弱められてしまうということ、並びに個々の単語にはあまり正の印象の強いものはないが、記事全体としては強い印象を感じる場合があるということがわかった。一方、負の印象がやや強い記事は11個あったが、これは提案手法がこの11個の印象を実際よりも強く評価したことを意味しており、今までとは逆のパターンになっている。そこで、各記事の内容を調べてみると、その主な原因として、一つの

ことに気付く。すなわち、負の印象の単語が比較的多く用いられている割に、記事の内容がさほど深刻ではないということである。例えば、警察官の汚職事件やダイヤの不法所持・密輸事件、名誉毀損や安眠妨害に対する損害賠償裁判などが相当している。

以上、本節で述べたような問題の解決に際し、いくつかの側面からのアプローチを考える。まず、新たな説明変数の導入を検討する。例えば、記事から抽出される特徴量の印象値がどのように分布しているかを表す指標として、現在用いている平均値に加え、最大値や最小値、あるいは第1四分位数や第3四分位数といった統計量とその組み合わせを試してみたい。次に、重要文抽出技術との組み合わせを考える。悲惨な状況下で起きた明るい出来事を伝える記事やその逆など正の印象の単語と負の印象の単語が混在する記事も見受けられることから、記事の印象を決定づけるような文(あるいは事象)を抽出した後、その文(あるいは事象)を対象に印象マイニングを行うという方法も有効かもしれない。また、話題のタイプによって人々の感じる印象が強くなったり、逆に弱くなったりすることもあるので、記事の話題タイプを決定する技術と組み合わせた上で、印象値を算出するための重回帰式を話題タイプごとに設計するという方法も考えられる。今後の課題として取り組んでいきたい。

7. むすび

本論文では、新聞記事を読んだ人々が感じる印象の強さを数値的に求めるための手法を提案した。本手法が対象とする印象は、「楽しい⇔悲しい」、「うれしい⇔怒り」、「のどか⇔緊迫」の3種類であり、それぞれの印象に対して「(左側の印象を感じる(1点)、割と感じる(2点)、やや感じる(3点)、どちらも感じない(4点)、(右側の印象を)やや感じる(5点)、割と感じる(6点)、感じる(7点)」という7段階の評価スケール(印象尺度)を設定している。提案手法は、それぞれの印象尺度において、このスケールに準じた1.0~7.0の実数値を出力する。

まず、記事から抽出する特徴量として単語unigramと単語bigramの2種類を定義し、既存手法[9]を用いて新聞記事データベースから6つの印象辞書(3種類の印象尺度×2種類の特徴量)を構築した。そして、各印象辞書を用いて算出される記事の印象値を説明変数(6個)、アンケート調査に基づいて数値化される記事本来の印象値を目的変数とする重回帰分析を印象尺度ごとに行い、それぞれの対応関係を示す重回帰式を設計した。この重回帰式に入力記事から算出される各説明変数の値(換算値)を代入すれば、その記事の当該

印象尺度における印象の強さを求めることができる。

提案手法の精度を学習データを用いて評価してみたところ、重回帰式を導入して複数の印象辞書の相互利用を可能にすることにより、精度が大きく改善することが確認された。また、5分割交差検定により未知データに対する精度を調べてみたところ、それぞれの印象尺度における平均誤差は、0.68, 0.57, 0.62であり、学習データに対する精度とほぼ同等であることが確認された。

今後の課題としては、5.4節で述べたような1つの記事の中での印象の推移を追跡できる印象マイニング手法の開発や、精度の向上に貢献する印象語群の設計、印象の感じ方の違いを補正する個人化手法の開発、あるいは6.3節で述べたような新たな説明変数の導入や他の技術(重要文抽出技術、話題タイプ決定技術)との統合などが挙げられる。

謝辞

本研究は、SCOPE若手ICT研究者育成型研究開発(課題番号:102107001)、筑波技術大学平成24年度競争的教育研究プロジェクト事業、福田将治奨学金寄付金(52303)、千葉工業大学附属総合研究所科学研究助成金、ならびにJSPS科研費24500134による助成の成果であり、ここに記して謝意を表すものとする。

参考文献

- [1] 山本和英, 集合知言語処理の研究動向, 第3回集合知シンポジウム, <http://goo.gl/7GOn6>, 2012.
- [2] 乾孝司, 奥村学, テキストを対象とした評価情報の分析に関する研究動向, 自然言語処理, Vol.13, No.3, pp.201-241, 2006.
- [3] Peter D. Turney, Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, Proc. of the 40th Annual Meeting on Association for Computational Linguistics, pp.417-424, Philadelphia, USA, 2002.
- [4] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, Thumbs up? Sentiment Classification using Machine Learning Techniques, Proc. of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Vol.10, pp.79-86, 2002.
- [5] Kevin Hsin-Yih Lin, Changhua Yang, Hsin-Hsi Chen, Emotion Classification of Online News Articles from the Reader's Perspective, Proc. of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Vol.1, pp. 220-226, 2008.
- [6] 宮川祥子, 清水康, 特定分野ドキュメントを対象とした意味的連想検索のためのメタデータ空間生成方式, 情報処理学会論文誌: データベース, Vol.40, No.SIG5 (TOD2), pp.15-28, 1999.
- [7] 清水浩平, 萩原将文, 名詞と動詞の組み合わせに対す

- る印象推定法, 日本感性工学会論文誌, Vol.10, No.4, pp.505-514, 2011.
- [8] 阿部健一, 藤本悠, 大原剛三, ユーザーが受信メールから抱く感情に合わせた音楽推薦システム, 第4回データ工学と情報マネジメントに関するフォーラム, No.A9-1, 2012.
- [9] 熊本忠彦, 河合由起子, 田中克己, 新聞記事を対象とするテキスト印象マイニング手法の設計と評価, 信学論 (D), Vol.J94-D, No.3, pp.540-548, 2011.
- [10] Tadahiko Kumamoto, and Katsumi Tanaka, Web Opinion Poll: Extensive Collection and Impression-based Visualization of People's Opinions, *Advances in Communication Systems and Electrical Engineering*, Chapter17, Vol.4, pp. 229-243, Springer US, 2008.
- [11] 張建偉, 河合由起子, 熊本忠彦, 田中克己, 地域性に基づく発信者の観点差異を可視化するセンチメントマップシステムの提案, 情報処理学会論文誌データベース, Vol.3, No.1 (TOD45), pp.38-48, 2010.
- [12] 河合由起子, 熊本忠彦, 田中克己, 印象と興味に基づくユーザ選好のモデル化手法の提案とニュースサイトへの応用, 知能と情報, Vol.18, No.2, pp.173-183, 2006.
- [13] 矢野絵美, 北野有亮, 末吉恵美, 篠原勲, ピンヤボンシニーナット, 加藤俊一, 消費者の感性モデルを利用したレコメンデーションシステムの構築, 情報処理学会論文誌: データベース, Vol.44, No.SIG8 (TOD18), pp.46-54, 2003.
- [14] 熊本忠彦, 灘本明代, 田中克己, 記事の印象を伝達するニュース番組生成システム wEE の設計と評価, 信学論 (D), Vol.J90-D, No.2, pp.185-195, 2007.
- [15] 石塚賢吉, 鬼沢武久, 加藤茂, 物語のシーンの印象に基づいた声楽曲の生成, 日本感性工学会論文誌, Vol.10, No.4, pp.523-534, 2011.
- [16] 山本祐輔, 手塚太郎, アダムヤフト, 田中克己, WebAlert: Web 情報の印象集約を利用した閲覧ページ内容に対する反対意見提示, 日本データベース学会論文誌, Vol.7, No.1, pp.251-256, 2008.
- [17] 菅民郎, 多変量統計分析, 現代数学社, 京都, 2000.
- [18] Kenneth W. Church, and Patrick Hanks, *Word Association Norms, Mutual Information, and Lexicography*, *Computational Linguistics*, Vol.16, Issue1, pp.22-29, MIT Press, Cambridge, MA, USA, 1990.
- [19] 清水康, 金子昌史, 北川高嗣, 意味の数学モデルによる画像データベース探索方式とその学習機構, 信学論 (D-II), Vol.J79-D-II, No.4, pp.509-519, 1996.
- [20] Tadahiko Kumamoto, Design of Impression Scales for Assessing Impressions of News Articles, *Lecture Notes in Computer Science*, Vol. 6193, pp.285-295, Springer-Verlag, In International Workshop on Social Networks and Social Media Mining on the Web, Tsukuba, Japan, 2010.
- [21] 大野晋, 浜西正人 (編), 類語国語辞典, 角川書店, 東京, 1986.
- [22] 黒橋禎夫, 河原大輔, 日本語形態素解析システム JUMAN version5.1, 2005.
- [23] CD-毎日新聞データ集 2002 年版, 毎日新聞社.
(2012年3月14日 受付)
(2012年8月23日 採録)

[問い合わせ先]

〒275-0016 千葉県習志野市津田沼2-17-1

千葉工業大学 情報科学部 情報ネットワーク学科

熊本 忠彦

TEL: 047-478-0443

FAX: 047-478-0443

E-mail: kumamoto@net.it-chiba.ac.jp

著者紹介

くまもと ただひこ
熊本 忠彦 [正会員]

千葉工業大学情報科学部教授。1988年筑波大学第三学群情報学類卒業。1990年同大学大学院理工学研究科修士課程修了。同年郵政省通信総合研究所(現、独立行政法人情報通信研究機構)。2007年千葉工業大学情報科学部情報ネットワーク学科准教授を経て、2010年より現職。テキスト印象マイニングとその応用(検索、推薦、可視化、他)に関する研究に従事。1996年筑波大学博士(工学)取得。FIT2004論文賞、電子情報通信学会、情報処理学会、人工知能学会、言語処理学会、日本データベース学会、日本感性工学会各会員。

かわい ゆきこ
河合 由起子 [非会員]

京都産業大学コンピュータ理工学部准教授。1997年九州工業大学情報工学部電子情報工学科卒業。2001年奈良先端科学技術大学院大学情報科学研究科情報システム学博士後期課程修了。同年独立行政法人情報通信研究機構。2006年京都産業大学理学部コンピュータ科学科講師を経て、2008年より現職。博士(工学)。情報推薦、Webマイニング、信憑性分析の研究に従事。情報処理学会、電子情報通信学会、日本データベース学会各会員。

張 建偉 [非会員]

筑波技術大学産業技術学部助教。2005年筑波大学大学院システム情報工学研究科博士前期課程修了。2008年筑波大学大学院システム情報工学研究科博士後期課程修了。博士(工学)。埼玉大学情報メディア基盤センター産学官連携研究員、京都産業大学コンピュータ理工学部特定研究員を経て、2012年より現職。Webマイニング、Web情報システム、情報保障の研究に従事。情報処理学会、日本データベース学会各会員。

Using Multiple Impression Lexicons to Quantify Readers' Impressions of News Articles
by
Tadahiko KUMAMOTO, Yukiko KAWAI and Jianwei ZHANG

Abstract :

The authors examine the impressions people gain from reading newspaper articles and propose a method for quantifying the strength of these impressions. Our target impressions are limited to those represented by three bipolar scales, "Happy-Sad," "Glad-Angry," and "Peaceful-Strained." In order to compute the strength of each impression as an "impression value," that is, a real number between 1 and 7, it is generally required to quantify the power of features extracted from articles to influence on their impressions and record it in an impression lexicon. An impression lexicon is usually constructed for each kind of impression and is used only to compute an impression value for an impression. In this study, we focus on the fact that the impressions are not independent of one another and adopt a new approach that recalculates each value with the values that were computed using each impression lexicon. First, we define "word unigrams" and "word bigrams" as features extracted from articles and construct a total of six impression lexicons (i.e., one lexicon for each of the three kinds of impressions multiplied by two kinds of features) from a newspaper database. Next, we apply multiple regression analysis for each impression, where the values computed from articles using each lexicon are used as one of the explanatory variables, and an average of the values that respondents used to rate each article using the corresponding impression scale in questionnaire surveys is used as the objective variable. Consequently, a multiple regression equation is obtained for each impression, which represents a correspondence relationship between the variables. We also perform five-fold cross-validation using the data obtained in the surveys to verify the effectiveness of the proposed method. The results show that the average root-mean-square errors for unlearned data are 0.68, 0.57, and 0.62 for respective impressions.

Keywords : impression mining, sentiment analysis, emotion classification, multiple regression analysis

Contact Address : **Tadahiko KUMAMOTO**

Chiba Institute of Technology

2-17-1, Tsudanuma, Narashino, Chiba, 275-0016, JAPAN

TEL : 047-478-0443

FAX : 047-478-0443

E-mail : kumamoto@net.it-chiba.ac.jp