Webニュース記事を対象とする喜怒哀楽抽出システム

熊本 忠彦 †　　田中 克己 †,††

† 独立行政法人情報通信研究機構 メディアインタラクショングループ
†† 京都大学 大学院 情報学研究科 社会情報学専攻

1 まえがき

言葉はイメージであり，明示的な意味のほかにも，様々な感情を伝える．ある言葉がどのような感情をどの程度伝えるのかを定式化できれば，電子書籍リーダや Web ラジオ，視覚障碍者向けホームページリーダなど，応用範囲は広い．本稿では，様々な感情のうち喜怒哀楽に焦点を当て，Web ニュース記事に含まれている喜怒哀楽の程度を決定するためのシステムを提案する．

なお，提案システムにおいて，喜怒哀楽及びその程度は，2 つの感情尺度「悲しい—うれしい」，「怒る—喜ぶ」に対する評価値（0〜1 の実数値）という形で記述され，その値は，入力記事に現れる単語（普通名詞，サ変名詞，動詞，形容詞，カタカナ）の種類から求められる．

2 設計方針

提案システムは，コストと実用性の観点から，以下の 3 つの仕様を満たす必要がある．

【感情辞書を自動構築できる】単語と感情の対応関係を示すものを「感情辞書」と呼ぶ．このような辞書を人手で構築するという方法には，作業者の主観が入りやすいという欠点がある．

【任意の感情をその程度とともに抽出できる】推定すべき感情の種類は応用分野やその時々の状況によって異なる．また，抽出された複数の感情の中から特徴的な感情を決定するためには，それぞれの感情の程度を定量的に示す必要がある．

【正解（学習）データを必要としない】単語と感情の対応関係を明示的に示す正解データを人手で作るのは，容易でない上，任意の感情すべてに対し，あらかじめ作るというのは不可能に近い．

以上の 3 仕様を満たすために，ヒューリスティックな知識を導入する．すなわち，「感情語 e を含む記事はその感情語が表す感情を伝える」という仮定のもと，新聞記事データベースに現れる各単語が感情尺度を構成する 2 つの感情語のどちらと，より高い確率で共起するかという観点でシステムの設計を行う．なお，新聞記事データベースには，日経新聞全文記事データベース（1990〜2001 年版）[1] の 200 万記事を利用する．

3 喜怒哀楽抽出システム

3.1 感情辞書構築システム

感情辞書には，各単語の感情尺度値とその重みが登録される．まず，感情尺度値の計算手法を示す．

y 年版に掲載された記事のうち，感情語 e を含む記事の数を $N(y,e)$，感情語 e と対象語 w を同時に含む記事の数を $N(y,e\&w)$ とすると，対象語 w の感情語 e に対する出現確率 $P(y,e,w)$ は，

$$P(y,e,w) = N(y,e\&w)/N(y,e)$$

と表される．ここで，対象語 w の感情語 e_1 に対する出現確率と感情語 e_2 に対する出現確率の比 $R(y,e_1,e_2,w)$ を計算し，対象語 w が感情語 e_1 と e_2 のどちらと共起する確率が高いかを示す指標とする．

$$R(y,e_1,e_2,w) = \frac{P(y,e_1,w)}{P(y,e_1,w)+P(y,e_2,w)}$$

但し，分母が 0 となるときは，便宜的に $R=0$ として処理する．この R を各年版ごとに求め，平均することにより，対象語 w の感情尺度「e_1—e_2」における $S(e_1,e_2,w)$ を求める．

$$S(e_1,e_2,w) = \frac{\sum_{y=1990}^{2001} R(y,e_1,e_2,w)}{\sum_{y=1990}^{2001} T(y,e_1,e_2,w)}$$

但し，$N(y,e_1\&w)+N(y,e_2\&w) > 0$ のとき，$T=1$ であり，それ以外のときは，$T=0$ となる．

次に，感情尺度値 $S(e_1,e_2,w)$ に対する重み $M(e_1,e_2,w)$ を，対象語 w と感情語 e_1，e_2 とが共起した年数と頻度の総和（12 年間分）に応じて増減するよう定義する．

$$M(e_1,e_2,w) = log_{12} \sum_{y=1990}^{2001} T(y,e_1,e_2,w) \times log_{144} \sum_{y=1990}^{2001} (N(y,e_1\&w)+N(y,e_2\&w))$$

以上の方法で構築された感情辞書の一部を表 1 に示す．表 1 には，感情尺度値が 0.8 以上の単語及び 0.2 以下の単語のうち，重みの最も大きい単語が示される．

A System for Extracting Feelings from Articles on Web News Sites, Tadahiko Kumamoto† and Katsumi Tanaka†,††, †National Institute of Information and Communications Technology, ††Kyoto University

3.2 記事を対象とする感情抽出システム

記事 $text$ を Web ニュースサイトから獲得し，形態素解析システム juman[2] を用いて，記事に含まれている単語を調べる．次に，感情辞書から各異なり単語の感情尺度値 S と重み M を取得し，記事の感情尺度値 $O(text)$ を算出する．

$$O = \sum^{text} S \times |2S-1| \times M \Big/ \sum^{text} |2S-1| \times M$$

但し，$|2S-1|$ は傾斜配分であり，感情尺度と関係の少ない一般的な単語（感情尺度値は 0.5 に近い値をとる）が O 式の平均操作に及ぼす悪影響を軽減するために導入されている．

4 性能評価

Yahoo ニュース[*1] から記事 100 件を収集し，各記事ごとに，被験者 50 人（20 代から 60 代の女性 30 名，男性 20 名）が与えた評価結果と提案システムが算出した感情尺度値を比較する．

まず，被験者に「もし自分がアナウンサーになって，かつ感情を込めて記事を読み上げるとしたら，どのような感情を込めるか？このとき，様々な感情を込めることが予想されるが，そのうち，喜怒哀楽という感情に関しては，どの程度の感情を込めるのか？」という問題を提示した．被験者は，記事を順に読み，2 つの評価尺度「悲しそうに（怒りを込めて）⇔どちらかといえば悲しそうに（怒った感じで）⇔中間／どちらともいえない／どちらでもない⇔どちらかといえばうれしそうに（喜びを込めて）⇔うれしそうに（喜びを込めて）」のそれぞれに対し，5 段階評価（5 点，4 点，$\cdots$，1 点）を行った．

次に，感情抽出システムを用いて各記事の感情尺度値を求め，0.570 以上のときを被験者の評価の 5 点／ 4 点に，0.343 以下のときを 2 点／ 1 点に，それ以外のときを 3 点に対応させた．被験者の評価結果と一致した記事の数（一致数）と割合（一致率），ならびに最多クラス（3 点）を常に出力する場合の一致率（チャンス率），各記事ごとに最多クラス／最少クラスを出力する場合の一致率（最高一致率／最低一致率）を表 2 にまとめる．なお，閾値は実験的に設定した．

表 2 から，感情尺度「怒る—喜ぶ」における一致率は比較的高いが，「悲しい—うれしい」に対しては十分と言えない．しかしながら，いずれにせよ，実用レベル（一致率 90% 以上）には程遠い．ただ，理論上の最高一致率が 74.6%，77.1% であることを考えると，単に感情推定手法を複雑にすればよいというものではなく，ユーザの知識や感性（し好や興味，性格），状態（気分や体調，忙しさ），そして購読環境（時間帯や場所，購読順序）に応じた処理が必要と考えられる．

[*1] http://dailynews.yahoo.co.jp/fc/

表 1 各感情尺度と共起の高い単語

対象語		感情尺度値	重み
感情尺度「悲しい—うれしい」			
死	普通名詞	0.839	1.170
離婚	サ変名詞	0.805	0.877
亡くす	動詞	0.801	0.997
悲しい	形容詞	0.986	1.535
レヴィ	カタカナ	0.857	0.346
笑み	普通名詞	0.162	1.083
製造	サ変名詞	0.197	1.086
勝てる	動詞	0.097	1.060
好調だ	形容詞	0.189	1.144
バラエティー	カタカナ	0.175	0.806
感情尺度「怒る—喜ぶ」			
怒り	普通名詞	0.926	1.135
抗議	サ変名詞	0.850	0.958
怒る	動詞	0.985	1.611
バカだ	形容詞	0.817	0.901
ヤジ	カタカナ	0.861	0.625
手放し	普通名詞	0.016	1.128
誘致	サ変名詞	0.173	1.072
喜ぶ	動詞	0.048	1.840
割安だ	形容詞	0.152	0.879
バイオ	カタカナ	0.154	0.697

表 2 被験者の評価結果とシステム出力値の比較

感情尺度	悲しい うれしい	怒る 喜ぶ
一致数	2,614	3,046
一致率	52.3%	60.9%
チャンス率	48.3%	48.9%
最高一致率（理論値）	74.6%	77.1%
最低一致率（理論値）	1.6%	0.7%

5 まとめ

本稿では，Web ニュースサイトから得られる記事の喜怒哀楽を 2 つの感情尺度「悲しい—うれしい」，「怒る—喜ぶ」に対する評価値（0〜1 の実数値）として出力するシステムを提案した．

今後は，今回得られた評価結果の解析をもとに，より高精度な感情推定手法の開発を進めるとともに，ユーザへの個人適応を視野にテキストと感情の対応関係に影響を及ぼす要因を調べていきたい．

参考文献

[1] 日経全文記事データベース DVD-ROM 版，日本経済新聞社．

[2] 黒橋禎夫，長尾真，日本語形態素解析システム JUMAN version 3.61，1999．