2G-4

Webニュース記事からの印象の自動抽出

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

新聞記事を読めば，そこに書かれてある事実関係がわかるだけでなく，様々な印象を受ける．しかしながら，記事の中に「明るい記事」,「悲しい記事」といったことが明示的に示されているわけではない．

そこで，本稿では，Web ニュース記事を対象とする印象抽出手法を提案する．具体的には，日経新聞全文記事データベース [1] に現れる単語（普通名詞，サ変名詞，動詞，形容詞，カタカナ）と任意の感情尺度（印象語の対からなる評価尺度であり，0〜1 の実数値をとる）との対応関係を定量化し，感情辞書に登録するための手法，ならびに感情辞書を用いて，Web ニュース記事の，ユーザ指定の感情尺度における評価値を算出するための手法を提案する．また，提案手法の有効性を検証するために，Web 上の Yahoo ニュースから記事 100 件を収集し，各記事に対して被験者 50 人が評価した結果と提案手法が算出した感情尺度値を比較する．但し，感情尺度には，評価のしやすさという観点から，「悲しい—うれしい」と「怒る—喜ぶ」の 2 種類を用いることにする．

2. 要求仕様

提案手法には，コストと実用性の観点から，任意の感情尺度に対し，単語と感情尺度との対応関係を明示的に示す正解データを用いないで，感情辞書を自動構築できることが要求される．

感情辞書を手作業で構築する方法は，高コストである上，作業者間で判断基準が異なる，作業者の性格，体調，気分によって判断基準が変動する，感情尺度の追加・変更といったメンテナンスが容易でない，といった問題を生じることから，辞書の自動構築は必須と言える．一方，抽出すべき印象の種類は，応用分野やその時々の状況によって異なることから，任意の感情尺度を設定できることが必要とされる．また，任意の感情尺度に対し，あらかじめ正解データを用意することはできないので，正解データを必要としない教師なし学習の実現が要求される．

3. 印象抽出手法の提案

本章では，新聞記事データベース中の記事に現れる単語の感情尺度値と重みを算出する手法，ならびに Web ニュース記事の感情尺度値を決定する手法を提案する．

3.1 設計方針

任意の感情尺度に対し，正解データを用いないで，感情辞書を自動構築するためには，何らかのヒューリスティックな手法を設計する必要がある．そこで，「印象語 e を含む記事からはその印象語が表す印象を受ける」という仮定のもと，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）は感情尺度を構成する 2 つの印象語のどちらと，より高い確率で共起するか，という観点で手法の設計を行う．

3.2 感情辞書の自動構築

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．感情尺度値は，単語が感情尺度を構成する 2 つの印象語のどちらと共起する確率が高いかを示す指標であり，以下のようにして求められる．

表 1 感情辞書に登録された単語の数

感情尺度	悲しい—うれしい	怒る—喜ぶ
サ変名詞	6,742	7,013
普通名詞	23,114	23,675
動詞	13,503	14,228
形容詞	3,244	3,263
カタカナ	17,683	17,750
合　計	64,286	65,929

まず，日経新聞全文記事データベース（1990 年版〜2001 年版）から感情辞書構築の材料となる記事を抽出した．各年版には 17 万前後の記事（約 200MB）が含まれており，全部で 200 万強の記事が得られた．

y 年版に掲載された記事のうち，印象語 e を含む記事の数を $N(y,e)$，印象語 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，印象語 e が現れたときに，対象語 w も現れる確率 $P(y,e,w)$ は，

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

と表される．ここで，対象語 w の印象語 e_1 に対する $P(y,e_1,w)$ と印象語 e_2 に対する $P(y,e_2,w)$ の比 $R(y,e_1,e_2,w)$ を対象語 w が印象語 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し，次式で求める．

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

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，次式を用いて平均することにより，対象語 w の感情尺度「e_1—e_2」における評価値 $S(e_1,e_2,w)$ を求める．

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

オリンピック関連用語など，出現する年は限られているが，出現する場合には特定の印象語との結びつきが強い単語が存在する．そこで，$N(y,e_1\&w)+N(y,e_2\&w)>0$ のときは $T=1$，それ以外のときは $T=0$ となる関数 T を導入し，対象語 w が出現しなかった年を分母から除外した．

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登録単語数ならびに登録例を表 1，表 2，表 3 に示す．なお，感情尺度には「悲しい—うれしい」,「怒る—喜ぶ」の 2 種類を用いた．また，表 2，表 3 には，感情尺度に対する値が 0.8 以上の単語及び 0.2 以下の単語の中から，重みの最も大きい単語を抜き出した．

Extracting Impressions from Newspaper Accounts on the Web, Tadahiko Kumamoto† and Katsumi Tanaka†,††, † National Institute of Information and Communications Technology, †† Kyoto University

表 2　感情尺度「悲しい—うれしい」と共起の高い単語

対象語		感情尺度値	重み
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
五輪	普通名詞	0.159	1.212
悲鳴	サ変名詞	0.049	1.228
販売する	動詞	0.189	1.104
うれしい	形容詞	0.068	1.858
パー	カタカナ	0.129	0.843

表 3　感情尺度「怒る—喜ぶ」と共起の高い単語

対象語		感情尺度値	重み
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

3.3　入力記事の感情尺度値の算出

記事 $TEXT$ が入力されたら，汎用日本語形態素解析システム juman[2] を用いて，形態素解析し，記事に含まれる単語（普通名詞，サ変名詞，形容詞，動詞，カタカナ）の種類を調べる．次に，感情辞書を用いて，各単語の感情尺度値 $S(e_1, e_2, w)$ と重み $M(e_1, e_2, w)$ を得，入力記事の感情尺度値 $O(e_1, e_2, TEXT)$ 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，感情尺度値 S の値に依存する傾斜配分であり，感情尺度と関係のない一般的な単語（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，導入された．

4.　性能評価

提案手法の有効性を検証するために，Web 上のニュースサイト（Yahoo ニュース[*1]）から記事 100 件を収集し，各記事に対して被験者 50 人（20 代から 60 代の女性 30 名，男性 20 名）が評価した結果と提案手法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？」という問題を与えた．各被験者は，すべての記事に対し，2 つの評価尺度「悲しそうに／怒りを込めて（5 点）—どちらかといえば悲しそうに／怒った感じで（4 点）—中間，どちらともいえない，どちらでもない（3 点）—どちらかといえばうれしそうに／喜びを込めて（2 点）—うれしそうに／喜びを込めて（1 点）」を用いて，5 段階評価を行った．

一方，3.2 節で述べた手法で感情辞書（感情尺度「悲しい—うれしい」，「怒る—喜ぶ」）を構築し，3.3 節の手法を用いて各記事に対する感情尺度値を求めた．そして，この感情尺度値と被験者 50 人の評価結果とを比較した．但し，提案手法が出力する感情尺度値が 0.570 以上のときを「悲しそうに，どちらかといえば悲しそうに」，「怒りを込めて，どちらかといえば怒った感じで」，0.343 以下のときを「どちらかといえばうれしそうに，うれしそうに」，「どちらかといえば喜びを込めて，喜びを込めて」，それ以外のときを「中間，どちらともいえない，どちらでもない」と 3 段階に設定し，被験者の得点も「5/4 点」，「3 点」，「2/1 点」の 3 段階評価に変換して，比較した．両方の評価結果が一致した数（一致数）とその割合（一致率），ならびに最多クラス（すなわち「中間」クラス）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 4 にまとめる．なお，閾値は実験的に設定した．

表 4　被験者の評価結果と提案手法の感情尺度値の比較

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%

表 4 から，感情尺度「怒る—喜ぶ」に対する一致率は，チャンス率に比べ 12 ポイント高く，単語レベルの出現確率，共起確率を用いた比較的単純な手法にしては，高い性能を得ていることがわかる．一方，感情尺度「悲しい—うれしい」に対する一致率は，チャンス率をわずかに上回っているにすぎず，良好な結果とは言えない．しかしながら，いずれにせよ，実用レベルには程遠い．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，単に印象抽出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（性格やし好，興味など），状態（気分や体調など），あるいは購読環境（場所や時間帯，購読履歴など）に応じた処理が必要と考えられる．今後の課題とする．

5.　まとめ

本稿では，Web ニュースサイトから得られる記事を対象に，任意の感情尺度に対して評価値（0〜1 の実数値）を算出する手法を提案した．すなわち，単語（普通名詞，サ変名詞，形容詞，動詞，カタカナ）と感情尺度の対応関係を示す感情辞書を新聞記事データベースから自動構築するための手法と，この感情辞書を用いて，入力記事の感情尺度値を決定する手法を提案した．

今後は，提案手法の高精度化に加え，ユーザへの個人適応を視野に，記事と印象との対応関係に及ぼす要因を調べていきたい．特に，同じ記事でも読む順番（例えば，悲しい記事の後に読む場合とうれしい記事の後に読む場合など）や話題の新規性（続報か第一報か，など）によって，印象に違いが生じる可能性があるという点に着目し，ユーザの購読履歴を管理し，活用する手法を設計していきたい．また，提案手法をベースに，共感を伝えるニュースリーダやテキストの印象を考慮した Web 検索システム等の開発を進めていきたい．

参考文献

[1] 日経全文記事データベース DVD-ROM 版，1990–1995 年版，1996–2000 年版，2001 年版，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，http://pine.kuee.kyoto-u.ac.jp/nl-resource/juman.html，1999.

*1　http://dailynews.yahoo.co.jp/fc/