ニュース記事を対象とした 感情表現の抽出・分析方式

桂 凜堂† 清木 康‡

†慶應義塾大学環境情報学部 〒252-0011 神奈川県藤沢市遠藤5322

‡慶應義塾大学政策・メディア研究科 〒252-0011 神奈川県藤沢市遠藤5322

E-mail: †t11240rk@sfc.keio.ac.jp, ‡kiyoki@ sfc.keio.ac.jp

あらまし 本稿では, ニュース記事を対象として, 記事に内在する感情を抽出するシステムの実現方式を示す. 近年, 人工知能という言葉が世の中を騒がせて久しい. 商業用ロボットも出現してきている中, 人と機械とのコミュニケーションの重要性は高まってきていると言える. また, 機械とのコミュニケーションにおいては, 機械が感情を認識し, 理解して表現すること期待されている. 本システムでは, Wordnetとニュース記事の本文中に含まれる単語に対して手動で感情のアノテーションを行い, 感情語データベースを構築した. 感情はベクトルデータとして表現されており, 記事の本文中に含まれる単語をもとに感情の計算を行うことが可能になる. 本システムでは, 自然言語処理をもちいてニュース記事の本文から感情表現を分析・抽出することを目指す. 本稿では, 文化・エンタメのカテゴリに絞ったニュース記事について感情表現を分析・抽出する実験システムを構築し, その他カテゴリに対しての拡張も想定したシステムの実現可能性を検証した.

キーワード 感情解析, 自然言語処理, ニュース

Extraction and Analysis Method of Emotional Expressions for News Articles

Rindo KATSURA† Yasushi Kiyoki‡

†Faculty of Environmental Information, Keio University, Endo 5322, Fujisawa, Kanagawa, JAPAN

‡Graduate School of Media and Governance, Keio University, Endo 5322 Fujisawa, Kanagawa, JAPAN

E-mail: †t11240rk@sfc.keio.ac.jp, ‡kiyoki@sfc.keio.ac.jp

Keyword Emotion Analyze, Natural Language Processing, News Article

1.はじめに

近年, 情報システムとロボティクスを統合した, 新たなコミュニケーションを行うシステム環境の構築が実際的な段階に進展している. 例えば, 内閣府が提案しているSociety 5.0では, 社会全体が効率的に機能することでこれまでの課題を解決しようとしており, 次世代の社会において, 人とロボットのコミュニケーションはこれまで以上の効率化・発展が求められている. 産業では, 受付・案内などのコミュニケーションを人と行うロボットも多く見受けられるようになった. しかし, ロボットが案内などの際に声を発するときにはどうしてもトーンが一定になりがちであり, より感情豊かな表現ができるようなシステムが求められているように思う. 本稿では, ロボットがニュース記事を読む際に感情を表現できるためのシステムを提案する. 本手法では, ニュース記事に内在する感情表現を分析・抽出することを目指す.

2.基本方式

本システムは, (1)感情の定義, (2)記事データの取得, (3)感情語の定義, (4)感情抽出の4つのプロセスによって構成される.

2.1 感情の定義

本システムで用いる感情の定義には, プルチックの感情の輪

を使用する. プルチックの感情の輪において, 人間の基本的な感情は喜び, 信頼, 心配, 驚き, 悲しみ, 嫌悪感, 怒り, 予測の8つであると言及している[1]. 本システムでは, 感情を8次元のベクトルとして定義する. 基本感情とそれらを組み合わることにより, 33通りの感情が表現可能である. また, 感情の強度などを考慮すると, 感情の表現はさらに柔軟にすることができる. 本システムにおける感情のデータ構造を図1に示す.

図1 : プルチックの感情の輪を元に定義された
感情のデータ構造

2.2 記事データの取得

本手法では, ニュース記事はプログラムを用いてNHKのRSSから取得した. 取得した記事はニュース記事データベースに保存される. データベースにはURL, カテゴリ, 記事タイトル, 記事要約, 公開日の項目が保存される. 今回利用したNHKのRSSではニュースは, 9つのカテゴリに分類されている. 本手法では文化・エンタメのカテゴリに解析対象を絞り解析を行った. ニュース記事のデータ構造を図2に示す.

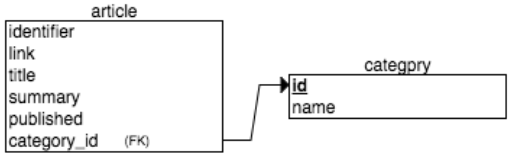

図2 : 記事データのデータ構造

2.3 感情語の定義

本システムでは, (1)Wordnet / Wordnet Affect, (2)記事データの二つを用いて感情語を獲得している[2][3]. 感情語のデータ構造を図3に示す.

(1) Wordnet Affectからの感情語の獲得

Wordnetとは単語の概念辞書であり, 単語はsysnetと呼ばれる同義語グループに分類される. Wordnet Affectとは, Wordnetの感情に関係する単語に対して感情のラベルを付与したものである. 本方式では, Wordnet Affectのラベルを本システムの感情に分類することによって, それに紐づく単語に対して, 本システムで使用している感情のアノテートを行った. これにより, 約4000語の感情語を獲得することができた.

(2) 記事データからの感情語の獲得

本研究の特徴的な機能として, (1)により獲得した感情語が多くない場合において, ニュース記事内に含まれる単語を抽出し, それらに対して感情のアノテーションを行った. この際, ニュース記事の単語は各カテゴリに紐づいており, 同じ単語でもカテゴリが異なる場合には別の感情をアノテートすることができる.

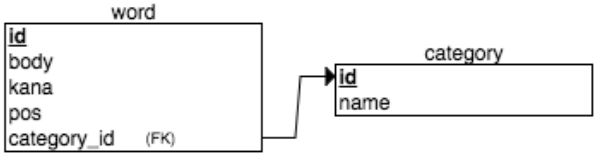

図3 : 感情語のデータ構造

2.4 感情抽出

本手法では, ニュース記事の本文全体に対して解析を行い, 1つの文章ベクトル $\vec{s}$ を計算している. 感情抽出・解析の計算の手法は3つのステップにより行われる.

STEP1: ニュース記事の本文に対して形態素解析を行い, 文章を単語に分割し, 基本形に直す.

STEP2: 各単語に対して, 感情語データベースから検索を行い, 単語の感情ベクトル$\overrightarrow{w}$を取得する.

STEP3: 得られた単語ベクトルに対して感情解析の計算を行い, 文章ベクトル$\vec{s}$を獲得する. 文章ベクトル$\vec{s}$は総単語数をnとすると, 感情解析の計算式は以下の数式で表される.

$$\vec{s} = \sum_{k=1}^{n} \overrightarrow{s_k}$$

3.実現方式

基本方式において示した感情語・記事の感情語を用いて感情抽出方式によって, 本システムを実現した.

3.1 システム概要

本システムの概要を図4に示す．本システムの開発にはPythonのWebフレームワークであるDjangoを利用した[6]．各モジュールもPythonで実装されている．システムは文章解析モジュール，感情抽出モジュール，ニュース取得モジュールからなる．感情語はデータベースセットアップ時に保存され，記事データベースはn取得時に記事データが随時データベースに保存される．

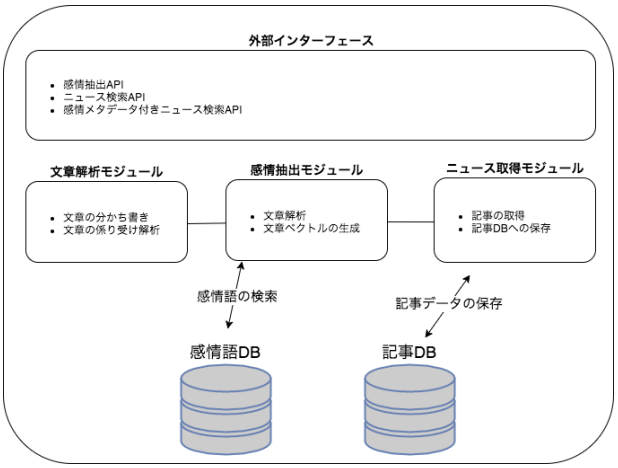

図4 : Djangoを用いた感情抽出・解析，
ニュース記事取得システム概要図

3.2 感情語の獲得

本節では，言語リソースからの感情語の獲得の手法を示す．

(1) Wordnetからの感情語の獲得

Wordnet Affectは英語の資源であるので，まず日本語Wordnetを用いて日本語への変換を行った[5].

Wordnet Affectでは各単語に対してカテゴリが振られており，本手法では各カテゴリを定義した感情に対して分類を行った．

(2) 記事に含まれる単語からの感情語の獲得

ニュース記事からの感情語の獲得はMeCabを用いて記事の本文から名詞，形容詞を抽出し，単語のリストを生成した．その各単語に対して，システムで定義した感情に対しての分類を手動で行った．

3.3 ニュース取得モジュール

ニュース記事はNHKのRSSより取得する．RSSから得られるXMLをプログラムにより解析し，所定の項目を抽出した上で記事データベースに保存した．また，RSSから取得できる記事の本文が短かったため，本文のみ別途Webスクレイピングを行い記事ページから本文を抜き出して記事データベースに保存した．

3.4 文章解析モジュール

文章解析はMeCabを用いて，以下のステップで行われる[7] ．(1) 形態素解析を行い，文章を単語に分割する (2) 各単語の基本形，品詞，読み仮名を取得する．基本形，読み仮名はMecabの出力から得られるもの，品詞はMeCabの品詞を独自に変換したものを利用している．品詞は名詞，形容詞，動詞のみに限定しておりそれ以外のものは無視される．例えば，文章解析モジュールを用いて「今日はいい天気です。」という文章の解析を行った場合には名詞である「今日」と「天気」，形容詞である「いい」が検出される。解析結果をまとめたものを図5に示す．

[入力]
「今日はいい天気です。」

[出力]
今日 0 きょう
いい 2 いい
天気 0 てんき

図5: 文章解析モジュールを用いて文章の解析を行った結果

3.5 感情抽出モジュール

感情抽出は，文章解析の結果をもとに行う．文章解析の結果，文章中の単語から本システムでの感情抽出に使われる単語群が得られる．この単語群に対してデータベースに対して検索クエリを投げてそれに紐づく感情ベクトルを獲得する．

4.実験

本システムの感情抽出モジュールを使って， 以下の3つの文章を対象に感情解析を行なった．

1. 歌手AのNKH紅白歌合戦への出場の記事

歌手AのNHK紅白歌合戦への出場の記事に対して，感情分析を行った．解析結果を表1に示す．記事の内容としては歌手Aのこれまでの活躍とNHK紅白歌合戦への出場を紹介している記事である． 記事内に喜びの感情キーワードが多く含まれていたため，このような解析結果となった．また，悲しみの単語は全く含まれていないためFear，Sadness，Disgustの感情は0.0になっておりきちんと解析できていると言える．

表1 : 歌手Aの記事の感情解析を既存の言語資源のみ(上)と記事感情語も含めた場合(下)の解析結果

Joy	Trust	Fear	Surprise	Sadness	Disgust	Anger	Anticipation
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
12.0	2.0	0.0	3.0	0.0	0.0	1.0	1.0

2. 傷害事件関連の記事

傷害事件関連の記事に対して，感情分析を行った．解析結果を表2に示す． 記事には傷害事件の詳細が記述されていたため，Fear，Sadnessなどの感情が大きくなっているのがわかる．また，警察などの単語に反応してTrustの感情も含まれている．喜びの感情は全く含まれていない．

表2 : 女子幼児の記事の感情解析を既存の言語資源のみ(上)と記事感情語も含めた場合(下)の解析結果

Joy	Trust	Fear	Surprise	Sadness	Disgust	Anger	Anticipation
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	5.0	8.0	2.0	5.0	0.0	0.0	1.0

3. 韓国の歌手Jの逝去の記事

韓国の歌手Jの逝去の記事に対して，感情分析を行った．解析結果を表3に示す． 記事の内容としては韓国の歌手Jの逝去を知らせる内容とこれまでの活躍を紹介する内容であった． 解析結果を見ると， ほぼ全ての感情に分散していることがわかる． これは逝去の内容と活躍の内容が含まれているからだと考えられる．

表3 : 韓国歌手Jの記事の感情解析を既存の言語資源のみ(上)と記事感情語も含めた場合(下)の解析結果

Joy	Trust	Fear	Surprise	Sadness	Disgust	Anger	Anticipation
0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
6.0	5.0	3.0	4.0	2.0	0.0	2.0	3.0

5.考察

記事内の単語が感情後データベースに含まれている場合には，高い精度で文章の解析ができていることがわかった．記事に含まれる単語をカテゴリ別に手動で感情のアノテーションを行うことで， カテゴリに最適化された単語の感情による解析結果が有意な結果として得られた．

しかし， ニュース記事に含まれる単語数は非常に膨大であり，手動でアノテーションを行うことは非常に大変である．また流行りなどから新出単語なども頻繁に出現していまい， 継続してシステムによる解析を高い精度で行うには日々感情語を運用していくことが必要であると言える．

6.結論および今後の展望

本稿では，ニュース記事を対象として，記事に内在する感情を抽出するシステムの実現方式を示した． 感情語に，既存の言語資源を用いるだけでなく，ニュース記事に含まれる単語をカテゴリ別に感情をアノテートすることで，既存の言語資源を用いた場合よりも高い精度での解析が可能となった．本稿で適用したカテゴリだけでなく，他のカテゴリに対してもアノテーションを行うことによりニュース記事の感情抽出・解析が可能になる．

今後は実際にロボットを使用して，本システムとの連携部分について進め実現可能性を検証する．

参 考 文 献

1. Robert Plutchik, “The Nature of Emotions”, http://www.emotionalcompetency.com/papers/plutchiknatureofemotions%202001.pdf.
2. Victoria Bobicev, Victoria Maxim, Tatiana Prodan,Natalia Burciu, Victoria Anghelu, “Emotions in words: developing a multilingual WordNet-Affect”, 2010.
3. George A. Miller, Richard Beckwith, Christiane Fellbaum,
4. Derek Gross, and Katherine Miller, “Introduction to WordNet: An On-line Lexical Database”, 1993
5. Yoshimitsu Torii, Dipankar Das, Sivaji Bandyopadhyay, Manabu Okumura, “Developing Japanese WordNet Affect for Analyzing Emotions”, 2011
6. The Web framework for perfectionists with deadlines | Django, https://www.djangoproject.com/
7. Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto: Applying Conditional Random Fields to Japanese Morphological Analysis, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230-237 (2004.)