E-056

ウェブ上の因果関係を用いたユーザ入力文からの感情情報の推測
Emotive Information Discovery from User Textual Input Using Causal Associations from the Internet

施 文翰† ジェプカ ラファウ† 荒木 健治†
Wenhan Shi Rafal Rzepka Kenji Araki

1. はじめに

現在，WWW（World Wide Web の略）の利用が非常に盛んであり，電子メールをはじめ，チャット，ブログなどといった大量のテキスト情報がやり取りされている．ウェブのアクセスを行う際には，パソコンや携帯の端末を利用するのが一般的である．そして，相手が書いたテキストの感情を読み取るには，コンピュータが人間の発話を理解する必要が存在する．しかし，実際に人間が対面して会話を行うときには，テキスト情報だけではなく，表情の変動，声の抑揚などのテキストで表現できないパラ言語情報が多く存在している．このようなコンピュータが扱いにくい情報の数値化を行って，コンピュータがユーザの手助けをするシステムは非常に有益であると思われる．

近年，文書の極性の判定や電子メールの文書から感情表現の抽出を行う研究が多く行われており，それらは主に 2 種類のものに分けられる．一つは表層表現に着目するもので，結合価パターン[1]または文末表現[2]を用いて感情の処理を行う．もう一つは Web マイニングを用い，テキストの関連語を検索し処理を行うものである[3]．本手法では，ユーザの発話内容を感情情報のリソースと位置づけて，Web 上にすでに存在している大量のテキスト情報を利用し，因果関係を用いて感情の推測を行う手法を提案する．このことにより本手法では発話の中に感情の表現がない場合でも，発話に対する感情の推測を行うことが可能となる．

2. 感情とは

2.1 感情の定義

人間の感情は五感と価値基準との総合として捉えられる複雑な情報であり[4]，その内容の判断を行うことには困難をともなう．例えば，「安いプリンタを買う」という文は，値段が安いので嬉しいと感じる人もいるが，その逆に値段が安いので，壊れやすいと心配する人もいる．このように同じ発話に対して，全く逆の感情を持つ人間がいる場合が多く見られる．また，自分自身の発話に対して，はっきりと感情を判断することが難しいと感じる人も少なくない．

2.2 コンピュータで人間の感情を扱う

人間の感情が多様で複雑な一方，コンピュータが扱いやすい情報は一義的で単純な数値である．現在の技術において，人間の発話がコンピュータに理解されるためには，人間の感情のカテゴリ化を行うことが有効であると考えられる．実際に人間の感情をカテゴリに分けて数値化すれば，コンピュータが人間の感情を扱うことも可能になると考えられる．

† 北海道大学大学院情報科学研究科 Graduate School of Information Science and Technology Hokkaido University

2.3 感情の分類

本研究では，感情表現を網羅的に収めている『感情表現辞典』[5]を利用して人間の感情のカテゴリ化を行った．ここで感情分類は，著者である中村が〈喜，怒，哀，怖，恥，好，厭，昂，安，驚〉の 10 分類に収集を行ったものである．感情表現辞典[5]には 806 編の日本の小説などの作品の中に存在する全 2,167 語が記載されている．本手法では日本語テキストで感情の推測を行うため，この中村による分類を利用する．

3. システムの概要

3.1 文の入力とフレーズ抽出

フレーズとは文法上の単語の集まりである．本手法においては元のユーザ入力文の意味を代表するフレーズを対象として抽出を行う．ユーザが一つの文を入力すると，Web 検索のヒット数と精度を考慮した上で，入力文を N-gram によって幾つかのフレーズに分けて検索を行う．フレーズの例は図 1 に示す．元の入力文の意味を代表する適切なフレーズの作成を行うため，以下のように抽出ルールを定める：

- 3-gram からフレーズを作成する
- フレーズの先頭は助詞ではない
- フレーズの最後は必ず動詞又は形容詞である
- 符号，顔文字は無視する

入力文：今朝も気温が低くて，寒かった．
3-gram: 気温が低い
4-gram: 気温が低くて寒い
5-gram: 今朝も気温が低い
6-gram: 今朝も気温が低くて寒い

図 1. 入力文からのフレーズ作成

3.2 フレーズへの接続詞の付与

Web で検索を行う際フレーズの後ろに感情表現が出現しやすくなるように，3.1 での作成を行ったフレーズの後ろに順接の役割を担う付属語を付加する．日本語における順接の意味を持つ助詞は「て，と，ので，から，ば」の 5 つで，助動詞は「たら，なら」の二つである．また，2 つのひらがなの組み合わせで順接の役割を持つ単語は「のは，のが，ことが，ことは」の 4 つで全部で 11 個である．すべての助詞に対して付与を行うのは効率ではないため，各接続助詞の頻度の調査を行った．その方法はまず感情データベースを Web 上で検索を行った際のヒット数でソートを行い，上位 10 個の単語の前に接続詞を付与して検索を行う．付与の結果を表 1 に示す．

表 1．接続詞ヒット数の割合

単語	て	ので	たら	なら	ことが	のは
結果	41.97%	7.20%	5.94%	1.17%	0.35%	2.30%
単語	と	から	ば	のが	ことは	
結果	31.97%	6.32%	3.19%	2.15%	0.30%	

表 1 より上位の五つの付属語「て，と，ので，から，たら」が全体の 90%以上を占めることが確認できた．したがって，この五つの接続詞のみをフレーズと組み合わせることで，9 割以上の情報の収集を行うことができる．

3.3 Web からの結果節の抽出

ここでは，入力文から得られたフレーズを，先述の付属語と繋いで Web から入力文と因果関係が存在する文を得る．検索エンジンには Google を利用した．最初に，フレーズに接続詞を付けて生成を行った新たなフレーズをクエリとして検索を行う．検索結果として得られたスニペットをテキスト情報として保存を行う．ここでは，一つのクエリごとに 100 個のスニペットをデータとして扱っている．つまり一つのフレーズに対して，500 個のスニペットが保存される．

3.4 関連文データの分析

保存されたテキストデータに，フレーズが存在する文の検索を行う．得られた文の中に存在する，フレーズより後方に存在する一文を感情推測のデータとして保存を行う．保存する際，以下の二つのルールの適用を行う．

- 後の文に“しかし”などのような後ろに日本語における逆接の意味を持つ単語が存在する場合，その単語の後ろの文をデータとして保存を行う．
- 後の文に“ない”などのような否定の意味を持つ単語が存在する場合，この文は保存しない．

保存されたデータに対し形態素解析を行う．解析ツールには Mecab を利用し，解析を行った単語に対し感情データベース（2.3 参照）とのマッチングを行う．感情のカテゴリが 10 種類存在するので，各カテゴリにマッチングする単語の数をそのカテゴリの点数とする．

すべての単語に対してマッチングを行った後，各感情のカテゴリの点数を見て，点数が高い順にソートを行う．推測は，全体の上位 60%以上占める感情をユーザに対して（一つまたは複数）出力することにより行う．

4. 評価実験

4.1 実験方法および結果

事前にユーザから一文の自由な発話文と文に対する感情をアンケート形式によりデータの収集を行った．入力文に対するシステムの感情推測結果とその文が持つ感情との比較を行った．入力文が複数の感情を持っている場合，システムの予測結果がいずれかに該当すれば成功とみなす．実験の例と結果を表 2，3，4 に示す．

表 2．自由発話文評価実験結果

成功	失敗	合計
23	16	39
58.9%	41.1%	

表 3．自由発話感情推測成功例

ユーザ入力文	ユーザ感情	システム推測
今日新しいプリンタを買った．	喜	驚，喜
とても元気だった人が病気で亡くなってしまった．	哀，驚	哀，驚，喜

表 4．自由発話感情推測失敗例

ユーザ入力文	ユーザ感情	システム推測
外が激しく吹雪いていることに気付いた．	厭，驚	好
資料用意の時間もギリギリだ．	昂	なし

4.2 考察

評価実験における失敗の原因として，発話に含まれるフレーズの抽出を行うルールが不完全であるということが考えられる．例えば失敗例の一文目は「吹雪いていることに気付く」が有効なフレーズであるが，「ことに気付く」というフレーズが抽出され，良い事と悪い事に気付く両方の感情が抽出された．そこで次に人手でフレーズを抽出して実験を行った．実験データは発話評価実験と同じ 39 文を用いた．実験結果を表 5 に示す．発話評価実験の結果より精度が 18%向上した．

表 5．フレーズ再抽出追加実験結果

成功	失敗	合計
30	9	39
76.9%	23.1%	

5. まとめ

本論文では，Web 上の情報を用いて，人間の発話の感情の推測を行うするシステムの提案を行った．人手で作成されたルールなどを使用せず，Web 上に存在する大量のテキスト情報を用いることにより，日常の発話について 58.9%の精度で感情の推測が可能であることを示した．したがって，感情推測システムにおいて，Web という知識源が有効であることが確認できた．しかし，元の発話文の意味を表すフレーズだけの抽出を行うことが課題として残されている．また，動詞，形容詞以外の単語への対応とリアルタイムでの結果の出力が今後の課題である．本手法とは別に，Ptaszynski らによる研究[6]では顔文字，符号などを用いて感情推測を行っている．この研究によるシステムは Ptaszynski らのシステムと統合することが可能であり，精度のさらなる向上が見込まれる．

参考文献

[1]田中努，徳久雅人，村上仁一，池田悟．”結合価パターンへの情緒生成情報の付与”，言語処理学会第 10 回年次大会発表論文，pp.345-348,(2004)

[2]横野光．“情緒推定のための発話文の文末表現の分類”, 情報処理学会自然言語処理研究会報告, p1-6, (2005).

[3]熊本忠彦，田中克己．”Web ニュース記事からの喜怒哀楽抽出”，自然言語処理研究会報告, (2005)

[4]伊藤正男, 他，『認知科学 6 情動』，岩波書店 1994

[5]中村明，『感情表現辞典』，東京堂出版，1993

[6]Michal Ptaszynski, Pawel Dyabla, Rafal Rzepka, Kenji Araki "Effective Analysis of Emotiveness in Utterances Based on Features of Lexical and Non-Lexical Layer of Speech." NLP 2008 Conference