

単語 1-gram を用いた短文からの感情推定

小川 拓貴†

松本 和幸‡

任 福継‡

†徳島大学知能情報工学科

‡徳島大学大学院ソシオテクノサイエンス研究部

1 はじめに

ユーザとの親和性の高い対話システムを実現するにはユーザの情緒面を考慮に入れて対応をする必要がある [1]. 我々はユーザのつぶやきや一言からユーザの感情を推定し、また対話を始めるきっかけとすることでユーザの状態を考慮した自然な対話を行うエージェントを実現できると考える. しかし、つぶやきや一言は単語数が少なく、含まれる情報が少ないため感情推定は難しい. またこれまでつぶやきや一言のような短文に特化した感情推定手法は存在しない. そこで、本研究では短文に含まれる単語や記号の持つ言語的意味と文章に表れる感情との関係に着目し、単語 1-gram を素性としてコーパスに Web サービス“えもにゅ”*を利用して統計的手法により感情推定を行う.

2 提案手法

2.1 使用コーパス

“えもにゅ”の Web サイトから“楽しい”、“嬉しい”、“悲しい”、“辛い”のいずれかの感情マークの付いた投稿文 3820 文を収集し、各感情ごとに投稿文を分類してコーパスとして使用した. “えもにゅ”とは図 1 のように一言メモと感情マークをセットにして投稿する Web サービスである. この投稿記事をコーパスとして用いることで、短文の書き手自身が適切だと判断した感情を直接コーパスに反映させることができ、また感情タグ付けのコストを削減することが出来る.

2.2 使用素性

文章を分割する際の素性の種類には単語 N-gram や文字 N-gram などがあるが、文章の文字数が少ない短文のための素性について、

- 文章の文字数が少ない短文において、書き手は人々が特定の感情を抱きやすい単語や記号の言語的意

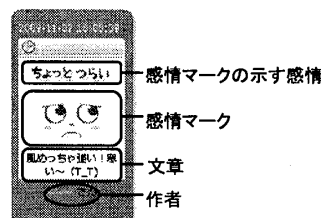


図 1: えもにゅの記事例

味を用いることで自身の感情を表現することが多い.

- 1 文あたりに含まれる素性の数が少ない短文から全ての素性について十分な素性の出現頻度を得るためには素性数を抑えることで 1 素性あたりの出現頻度を上げる必要がある.

以上の 2 点が考えられるので、本研究では短文に含まれる単語や記号の持つ言語的意味を重要視し、また素性の数を抑えられる単語 1-gram を素性として採用した.

2.3 文からの特徴ベクトル作成方法

SVM を用いて学習、分類をするためにコーパスにおける各文を特徴ベクトルに変換する必要がある. 特徴ベクトルの次元には素性の種類を使用するため、まずコーパス中に出現した素性全種類に対して通し番号を振り、コーパス中での各素性の出現頻度を素性出現頻度データベースに格納した. その後、図 2 の流れで文から特徴ベクトルを作成した.

特徴ベクトルの次元の値の計算方法について、ある感情コーパス A の文 s に含まれる素性 α に対応した次元の値は以下の式によって計算される.

$$\begin{aligned} & \text{感情 A での素性 } \alpha \text{ に対応した次元の値} \\ &= \frac{\text{感情 A での素性 } \alpha \text{ の出現頻度}}{\text{全感情での素性 } \alpha \text{ の出現頻度}} \quad (1) \end{aligned}$$

次元の値として正規化した出現頻度を採用することで、出現頻度を採用する場合に比べてある感情でのみ多く用いられる、感情の特徴となりやすい素性に対応する次元の値をその他の次元より大きく付けることが出来る.

Emotion estimation for short sentence using word 1-gram as the feature
†Takuki OGAWA ‡Kazuyuki MATSUMOTO ‡Fuji REN
†Dept of Information Science Intelligent Syst. Faculty of Eng. University of Tokushima
‡Institute of TechFnology and Science, The University of Tokushima
*http://emonyu.jp/

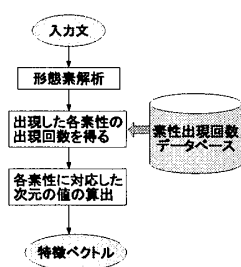


図 2: 特徴ベクトル作成の流れ

表 1: 素性別, 感情毎の再現率 (R), 適合率 (P), F 値

	単語 1-gram			単語 2-gram		
	R	P	F 値	R	P	F 値
楽しい	10.2	43.2	16.4	5.8	46.7	10.2
悲しい	41.6	71.3	52.4	24.0	74.7	36.3
嬉しい	15.1	51.8	23.2	8.9	57.4	15.1
辛い	31.6	62.6	41.8	15.8	66.6	25.4

3 推定実験

提案手法の有効性を確認するために, 単語 1-gram を素性として感情推定実験を行い, 比較のために単語 2-gram を素性とした感情推定実験も合わせて行った.

3.1 実験方法

Web サービス“えもにゅ”から“悲しい”, “楽しい”, “辛い”, “嬉しい”の 4 感情いずれかの感情マークが付いた投稿文をそれぞれ約 950 文, 合計 3820 文を収集してコーパスとし, 各コーパスをそれぞれ 10 分割して 9 割を学習コーパス, 1 割をテストコーパスとした. そして, 単語 1-gram を素性として各感情コーパスを特徴ベクトルに変換し, SVM により one vs rest 法で学習, 感情モデルを構築, テストコーパスを特徴ベクトルに変換したものを, 感情モデルを用いた SVM へ入力し再現率, 適合率, F 値を用いて評価した. 実験は学習コーパス, テストコーパスに含まれる文章を逐次変更して 10 回行い, その平均を結果とした. 提案手法の比較実験として, 素性を単語 2-gram に変更して単語 1-gram と同様の方法で評価を行った.

3.2 実験結果

推定精度の評価基準には再現率 (Recall) と適合率 (Precision) の調和平均を取った F 値を用いる. 表 1 に各素性での実験で得られた F 値, 再現率 (R), 適合率 (P) を示す.

4 考察

F 値を用いて比較すると, 単語 1-gram を素性とした場合の方が全ての感情において単語 2-gram を素性とした場合より「楽しい」で 6.2 「悲しい」で 16.2, 「うれしい」で 8.1, 「辛い」で 16.4 高い値を示している. 表 1 より F 値を決定する要素別に調べると, 再現率は単語 1-gram を素性とした場合の方が単語 2-gram を素性とした場合に比べ 2 倍近く高い結果が出ている. 逆に適合率は単語 2-gram を素性とした場合の方が高い結果となっている. この結果より, 単語 1-gram を素性とした場合, 単語 2-gram に比べて再現率が大きいので短文の表現している感情をより高確率で検出することができ, また適合率が低いので誤推定をする率が高いと解釈できる. また, 再現率と適合率はトレードオフの関係にあるので, これら 2 つの調和平均である F 値で総合的に判断すると単語 1-gram を素性とした場合の方が短文からの感情推定においては優位と言える.

単語 1-gram を素性とした感情推定実験の誤り分析をしたところ, 出現頻度 1 回の素性について出現頻度を正規化すると正規化した値が計算上取りうる最大値 1 となる. そのため出現頻度 1 回の, 信頼性の低い素性が推定する際の大きな特徴となってしまう推定誤りを引き起こしていた. また, 文章内に推定の特徴となるような値を持つ素性が無かったために推定誤りを起こしているものもあった.

5 まとめ

本研究では, Web サービス“えもにゅ”をコーパスとして用い, 単語 1-gram を素性として SVM を使用して短文からの感情推定を行った. 推定実験を行った結果, F 値を評価基準とすると単語 1-gram を素性とした場合, 単語 2-gram を素性とした場合より全ての感情で高い値となり, 本手法の有効性を示すことが出来た. 今後, 素性に対応した次元の値の計算方法の改良, コーパスを増やすことで素性の出現頻度を上げて推定の特徴となる素性を増やすことで推定精度を上げたい.

参考文献

- [1] 徳久良子, 徳久雅人, 乾健太郎, 岡田直之. 相手の情緒を理解する対話システムの構築に向けて: 情緒タグつき対話コーパスの構築. 電子情報通信学会技術研究報告. TL, 思考と言語, Vol. 99, No. 237, pp. 13-20, 1999.