6R-5

単語重み付き文章類似度による感情推定

宇原章浩 †　　加藤昇平 †

† 名古屋工業大学

1　はじめに

近年，テキストから感情を推定する研究が行われている．この研究の応用として感性情報を考慮した文章検索やコンピュータと人とのコミュニケーションを目的とした自動対話システム上でのユーザの感情推定などが考えられる．三品らの研究 [1] ではあらかじめ感情ラベルを付与した感情別の文章コーパスを作成し，各コーパス文と未知文との類似度を計算することにより感情推定を行っている．本研究では文章間の類似度を計算する際に単語毎に重みを付けした新しい類似度計算手法を提案することで感情推定手法の精度の向上を目指す．

2　SCCS を用いた文章感情推定

2.1　感情推定システム

提案システムの流れを図 1 に示す．提案システムでは感情ラベルを付与したラベル別の文章コーパス中の文章と未知文との類似度を計算することにより感情推定を行う．コーパスは感情ラベル毎に用意する．入力文に対し感情ラベル付き文章コーパス中の全文章との類似度を計算し，類似度が大きい上位 N 文章についてそれぞれの感情ラベル毎に類似度の値を集計する．集計結果が最大となる感情ラベルを入力文の推定感情として出力する．類似度計算には IMPACT[2] を改良した SCCS （**S**imilarity for **C**ategorization using Weighted **C**ommon **S**ubsequence）を用いる．IMPACT は機械翻訳システムの自動評価手法として提案されたもので 2 文章間の類似度を計算する．推定において入力文及びコーパス中の文章に含まれるすべての単語は基本形に変換した．文章からの感情推定において感情的な要素を含む単語に着目する．そのため単語の感情判断に対する重要度の度合いを示した辞書を用意し類似度計算時に重みとして利用することにした．今回この辞書には高村らが作成した単語感情極性対応表 [3] を用いた．単語感情極性対応表は単語がポジティブかネガティブかをスコアで表したものであるが，単語のポジティブ（または ネガティブ）の度合いが高ければ感情への有効性も高くなると考え重みとして用いた．

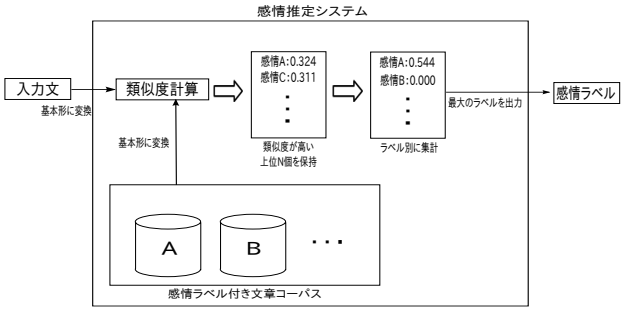

図 1: 提案システムの流れ

2.2　SCCS

SCCS は 2 文章間の類似度を算出する手法であり，まず 2 文章間の最長共通部分列（Longest Common Subsequence，LCS）を定め，LCS 中に含まれている単語を 1 つの要素として扱う．また，ここで定めた LCS に含まれている単語が 2 文章上で共に連続していた場合，1 つの要素としてまとめて扱う．例えば次のような 2 文章を比較するとき

噛みつく **よう に** 、 病父 が **怒鳴っ た** 。
黒岩 は 、 とうとう 爆発 し た **よう に 怒鳴っ た** 。

LCS は「よう，に，怒鳴っ，た,。」と定められ,「よう，に」と「怒鳴っ，た,。」の 2 要素として扱う．そして，2 文章から LCS を除き，除いた 2 文章から LCS を再び定める．この処理を LCS が存在しなくなるまで繰り返す．求めた LCS から式 (1)〜式 (4) を用い，再現率 R，適合率 P を求め，これにより類似度 SCCS を求める．

$$SCCS(X,Y) = \frac{\left(1+\gamma^2\right)R \cdot P}{R+\gamma^2 P} \tag{1}$$

$$R = \left(\frac{\sum_{i=0}^{RN} \left(\alpha^i \sum_{c \in LCS} \left(TFIDF(c) Weight(c) length(c)^{\beta} \right) \right)}{m^{\beta}} \right)^{\frac{1}{\beta}} \tag{2}$$

$$P = \left(\frac{\sum_{i=0}^{RN} \left(\alpha^i \sum_{c \in LCS} \left(TFIDF(c) Weight(c) length(c)^{\beta} \right) \right)}{n^{\beta}} \right)^{\frac{1}{\beta}} \tag{3}$$

$$\gamma = \frac{P}{R} \tag{4}$$

*Emotion Estimation Using Sentence Similarity Calculated by Semantic Orientations Weighted by Words, Akihiro UHARA and Shohei KATO, Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan, {uhara, shohey}@katolab.nitech.ac.jp

Copyright ©2013 Information Processing Society of Japan.
All Rights Reserved.

ここで, X, Y はそれぞれ類似度を比較する 2 文であり, n は X の単語数, m は Y の単語数である. RN は再帰的に求めた LCS の数である. c は LCS 中の要素であり, $length(c)$ は c に含まれる単語数を示す. α と β はパラメータである. $TFIDF(c)$ は c に含まれる単語 の tf-idf 値の合計でこれによりコーパスでの単語の偏りを重みとして利用できる.

$$TFIDF(c) = \sum_{x \in c} (tf(x) idf(x) + \delta) \tag{5}$$

$$tf(x) = \log\left(\frac{n_{xj}}{\sum_{i \in j} n_{ij}} + 1\right) \tag{6}$$

$$idf(x) = \log \frac{|C|}{df_x} \tag{7}$$

ここで, n_{ij} は感情 j のコーパス中に単語 i が含まれている数を表す. $|C|$ はコーパスの総数とし, df_x は単語 x が含まれているコーパス数とする. tf-idf 値が 0 になった場合, 要素 c の他の情報も失われてしまうため微小な正数 δ により補正することで情報の欠損を防いでいる. $Weight(c)$ は要素 c 中に含まれる単語に対する重みである.

2.3 単語感情重み

本稿では $Weight(c)$ として c に含まれる単語の感情判断における重要度の合計値を用いる.

$$Weight(c) = \sum_{x \in c} (|value(x)| + \varepsilon) \tag{8}$$

ここで, $value(x)$ は単語感情極性対応表 [3] における単語 x の極性値であり, 同表中に単語 x が存在しない場合は 0 とする. ε は $Weight$ 値が 0 になることによる情報の欠損を防ぐための微小な正数である.

3 感情推定実験

3.1 評価方法

本稿では, 文学小説中の感情表現を収録した感情表現辞典 [4] を文章コーパスとして採用し,「喜」「好」「怒」「悲」の 4 感情についてそれぞれ 100 文を抜き出し感情別に 4 つのコーパスを作成した. 類似度が大きい上位 5 文章の値を感情ラベル別に合計値をとる. 提案手法の有効性を確認するために Leave-one-out 法による交差検定を行った. 提案手法のパラメータ α, β の値は 0.1, 1.1 と設定し δ, ε は 0.0001 とした. また比較手法として文章の比較に用いられるコサイン類似度, 既存手法である IMPACT[2], IMPACT と同様の機械翻訳システムの自動評価手法である Bleu[5], Bleu を感情推定に適用した手法である RECARE[1] を用意した. また提案手法に用いるの重みを $TFIDF$ のみ*1 または $Weight$ のみ*2 にしたものをそれぞれ用意した.

表 1: 感情推定正答率 (%)

	喜	好	怒	悲	全体
提案手法	56	59	80	62	**64.25**
$Weight$*2	53	54	72	51	57.5
RECARE[1]	75	44	57	54	57.5
$TFIDF$*1	44	54	69	53	55
IMPACT[2]	36	37	38	33	36
cos 類似度	43	31	46	24	36
Bleu[5]	4	8	0	6	4.5

3.2 実験結果

各手法の正答率を表 1 に示す. 提案手法が 64.25%と他の手法と比較し正答率が高いことが分かる. また提案手法は「怒」の感情ラベルの正答率が他の感情ラベルに比べ高いことが分かる. これは $TFIDF$*1 と $Weight$*2 それぞれ単体の時にも同様の傾向がある. 他の手法でも「怒」のラベルの正答率が高くなる傾向はあるがこの 3 手法ではこの傾向が顕著に現れている.

4 まとめ

本稿では文章からの感情推定を目的とし, 単語を重みとして利用した類似度計算手法を提案した. 評価実験の結果, 文学小説中の表現から構築した文章コーパスにおいて提案手法の優位性が確認された. 今回使用したコーパスは規模が小さいため, 今後はより大きなコーパスを使用した評価実験を行う必要がある. また今回の手法では単語のみの重みしか考慮していないため, 文法情報などを取り入れていく必要が考えられる.

参考文献

[1] 三品賢一, 土屋誠司, 鈴木基之, 任福継. コーパスごとの類似度を考慮した用例に基づく感情推定手法の改善. 自然言語処理, Vol. 17, No. 04, pp. 91–110, 2010.

[2] Hiroshi Echizen-ya and Kenji Araki. Automatic evaluation of machine translation based on recursive acquisition of an intuitive common parts continuum. *Proceedings of the Eleventh Machine Translation Summit*, pp. 151–158, 2007.

[3] 高村大也, 乾孝司, 奥村学. スピンモデルによる単語の感情極性抽出. 情報処理学会論文誌ジャーナル, Vol. 47, No. 02, pp. 627–637, 2006.

[4] 中村明. 感情表現辞典. 東京堂出版, 1993.

[5] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu:a method for automatic evaluation of machne translation. *In Proc. of the ACL2002*, pp. 311–318, 2002.

Copyright ©2013 Information Processing Society of Japan.
All Rights Reserved.