2D-03

中国語感情語辞書と中国語感情表現分析システム

黄永輝† 楊海洋† 冉升† 卜廷君† 李星鋭† 成思遠† 趙鑫† 安達由洋†
東洋大学総合情報学部総合情報学科†

1. はじめに

中国語文に対して、文をポジティブとネガティブに分類するセンチメント・アナリシス[1]、あるいは喜び、哀しみ、怒りなど数種類の感情カテゴリを用いた粗粒度の感情分析の研究報告と実用的サービスが見受けられるが、より細粒度の感情分析に関する研究は見当たらない。例えば、ソーシャルメディアあるいはEC サイト上での世論、商品・サービスに関するレビューなどから細粒度の感情検出ができると、世論の分析、商品の改善や新商品の開発などに大いに参考となる。特に、我々が提案した‘望’感情[2]を抽出すると、商品やサービスあるいは授業改善についての非常に役立つ情報を収集することができる。

本研究では、中国語文を細粒度（11 感情カテゴリ）感情分析するために非常に豊富な語彙を持つ感情語辞書を作成し、その辞書を用いて比較的精度よく高速に中国語文の感情を分析するシステムを実装した。

2. 中国語感情分析システム

本研究では、感情語約 13,000 語を収録した東洋大学版中国語感情語データセット(Toyo University Chinese Emotion-word Dataset, TU-CED)を作成した。そして、TU-CED から中国語感情語辞書(Chinese Emotion Word Dictionary, CEWD)を作成する中国語感情語辞書管理システム(Chinese Emotional-word Dictionary Management System, CEDMS)と、作成された辞書を用いて高速に自由記述文の感情分析をする中国語感情分析システム(Chinese Emotional Expression Analysis System, CEEAS)を開発した。

2.1 中国語感情語データセット

中国語感情語データセットの作成には、大連理工大学から公開されている“情感词汇本体第 2 版”[3]を参考にした。このデータセットは約 25,000 語の感情語を収録しているが、既に使われなくなった単語を多数含んでいる。本研究では、使われなくなった感情語を除外するとともに新しい感情語を追加して TU-CED を作成した。TU-CED は感情語、属する感情カテゴリ、感情強度からなる表（Excel シート）である。

2.2 中国語感情語辞書管理システム

CEDMS は、TU-CED から感情語の見出し語とその語が属する 11 感情カテゴリ［乐,怒,哀,怖,耻,喜,厌,激,安,惊,望］を表す 11 次元ベクトルの対からなる感情語辞書 CEWD を作成する。例えば、“乐”感情カテゴリに属する感情語に対して、感情カテゴリベクトルは[1,0,0,0,0,0,0,0,0,0,0]と表現される。なお、CEWD には、中国語形態素解析器 Jieba[4]と LAC[5]のそれぞれに対応する版を作成した。

CEDMS の感情語辞書生成手順を以下に示す：
手順 1：TU-CED Excel データを入力
手順 2：感情語を形態素解析器 Jieba(LAC)で解析して、単語の原形のリストで表現
手順 3：手順 2 で求めた単語の原形リストを連結して辞書の見出し語を生成
手順 4：見出し語と感情カテゴリベクトルを組みにして CEWD Excel ファイルに出力

上記手順により、原形 1 語からなる見出し語の数は 11,409 (8,001)語、原形 2 語は 928(2,675)語、原形 3 語は 567(1,668)語、原形 4 語は 104(565)語、原形 5 語は 52(103)語、原形 6 語は 28(50)語、原形 7 語は 33(48)語、原形 8 語は 12(14)語、原形 9 語は 5(13)語、原形 10 語は 2(2)語、原形 11 語は 0(0)語、原形 12 語は 1(1)語、原形 13 語は 0(1)語の合計 13,141 (13,141)語の見出し語を持つ CEWD を作成した。

2.3 CEEAS

CEEAS は Python の辞書オブジェクト dict 表現を用いて CEWD に基づき高速に自由記述文の感情を分析する。

CEEAS の感情分析手順を以下に示す：
手順 1：CEWD を読み込み辞書オブジェクトに格納（見出し語をキー、感情ベクトルをバリューとする）
手順 2：分析対象の中国語文を Jieba(LAC)で形態素解析して単語の原形のリストで表現
[以下の手順 3 と 4 を、n を 12(13)から 1 まで減少させながら繰り返す]
手順 3：手順 2 で得た文の単語原形リストを先頭から最後まで順に n 語スライスし、切り出した部分リストを連結して CEWD の見出し語と照合する。照合した見出し語に対応する感情ベクトルを抽出
手順 4：見出し語と照合した感情語の前あるいは後ろに否定語／程度副詞があるかリストの照合で調べる。否定語／程度副詞がある場合はルールに基づいて感情カテゴリベクトルを修正
手順 5：入力された文から抽出された感情ベクトルのリストを合成（各対応する成分ごとに最大値を取る）
手順 6：手順 5 で得た感情ベクトルから感情カテゴリ語に変換
手順 7：手順 5 で得た感情ベクトルから極性値に変換

なお、CEEAS は対話処理モードとファイル処理モードがある。対話処理モードは CUI 上で中国語文を入力すると瞬時に分析結果を出力する。一方ファイル処理

Chinese emotional word dictionary and Chinese emotional expression analysis system
† Yonghui HUANG, Haiyang YANG, Sheng RAN, Tingjun BU, Xingrui LI, Siyuan CHENG, Xin ZHAO, Yoshihiro ADACHI・Toyo University

モードでは、Excel ファイル（またはテキストファイル）から分析対象の文データを読み込み、一括処理した分析結果を Excel ファイルに出力する。

2.4 CEEAS による対話的処理の実行例

図 1 に CEEAS による対話的処理の実行例を示す。

```
对对话内容进行情感分析。
请输入文本内容（输入q结束）
>> 我喜欢这件衣服的设计

情感词汇: [['喜欢']]
情感向量: [0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0]
情感分类: ['好']
极性值: 1、极性: positive

请输入文本内容（输入q结束）
>> 我不喜欢这件衣服的设计

情感词汇: [['不', '喜欢']]
情感向量: [0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0]
情感分类: ['厭']
极性值: -1、极性: negative
```

図 1．対話的処理の実行例

3. Jieba 版 CEEAS の分析精度と速度

中国語のニュースサイトやレビューサイトから収集した 3,784 文を 7 人でそれぞれ感情カテゴリベクトルを付け、それらの多数決を取って教師ラベルとした CEEAS 評価用データセットを用意した。

3.1 CEEAS の精度評価

中国語文合計 3,784 文の CEEAS 評価用データセットに対する感情分析精度を表 1 に示す。この表の中で、文数とは各感情カテゴリに属する文の数を表す。

表 1．CEEAS の感情分析精度

	文数	accuracy	precision	recall	Flscore
乐	1,186	0.88	0.77	0.87	0.82
怒	496	0.88	0.55	0.67	0.60
哀	265	0.96	0.77	0.55	0.64
怖	25	0.99	0.34	0.84	0.48
耻	4	1.00	0.23	0.75	0.35
喜	492	0.91	0.60	0.85	0.70
厌	1,174	0.85	0.79	0.72	0.75
激	14	0.99	0.40	0.71	0.51
安	94	0.97	0.45	0.82	0.58
惊	44	0.97	0.28	0.68	0.40
望	315	0.96	0.69	0.91	0.78

3.2 CEEAS の速度評価

図 2 に、CEEAS 評価用データセットに対する計算速度の測定結果を示す。データを切り出したり、コピーして増やしたりして、1,000 文から 10,000 文の分析時間を測定した。測定環境は Intel Core i7-1068NG7 CPU @2.30.GHz、16 GB memory である。図 2 に示すように Jieba 版 CEEAS は 10,000 文の感情ベクトルを約 0.6 秒で計算する。

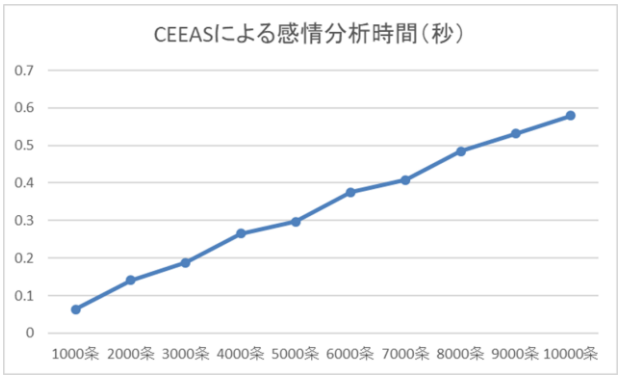

図 2．CEEAS による感情分析時間

4. まとめ

本研究では、精度よく高速に中国語文の感情分析をする中国語感情語辞書管理システム CEDMS と、中国語感情表現分析システム CEEAS を開発した。

CEEAS は、授業中に実施した学生アンケート回答文を感情分析してより良い授業の支援をする、講演会などでの参加者アンケートから感情状態を分析して講演内容に反映する、SNS や Web 上で発信される意見やユーザーレビューを感情分析して世論やトレンドなどを反映した商品・サービスを開発する、アンケートを使った心理カウンセリングに利用するなど多くの応用が考えられる。

今後の課題として、より大規模に中国語文を収集して CEEAS の精度評価をするとともに、感情語辞書 CEWD の語彙を増やすことが挙げられる。また、感情語は時代の変化や、SNS などのデジタル文化の隆盛とともに変化を続けており、新しい感情カテゴリの追加も大きな課題である。

参考文献

[1] Microsoft, “感情分析事前構築済みモデル, 感情分析”, https://docs.microsoft.com/ja-jp/ai-builder/ prebuilt-sentiment-analysis (2022/01/04 アクセス).
[2] 瀬山透矢, Amilcare Astremo, 安達由洋, “ソーシャルメディアおよび EC サイトでのレビュー分析のための‘望’感情の抽出”, FIT2021 (2021).
[3] 徐琳宏, 林鸿飞, 潘宇, 任惠, 陈建美, “情感词汇本体的构造”, 情报学报, 27(2), pp.180-185 (2008).
[4] jieba, https://github.com/fxsjy/jieba (2022/01/06 アクセス).
[5] Jiao Zhenyu, Sun Shuqi and Sun Ke, “Chinese Lexical Analysis with Deep Bi-GRU-CRF Network”, https://arxiv.org/abs/1807.01882 (2022/01/06 アクセス).