JIWC: クラウドソーシングによる日本語感情表現辞書の構築

柴田大作 若宮翔子 伊藤薫 荒牧英治

奈良先端科学技術大学院大学 情報科学研究科

{shibata.daisaku.rr8, wakamiya, kito, aramaki} @is.naist.jp

1. はじめに

自然言語処理と心理学をはじめとした人文学系研究との距離は遠い．自然言語処理技術の発展により，いまや様々なテキストを解析する需要が高まっている．例えば，Twitter による呟きや通販サイトのレビュー，ひいては疾患をもつ患者の発話を書き起こしたものなど様々である．

しかし，これらのテキストの中身の分析については，評価表現の解析など 2 値定量化できる指標の算出が中心で，書き手の感情の分析や読み手の心理状態の分析など，質的な内容の議論が困難な状態である．

一方，海外では，内容を分析するための言語リソースとして Linguistic Inquiry and Word Count（以降，LIWC と呼ぶ）が存在し，活発に用いられている（英語[1]．スペイン語[2]，中国語[3]）．LIWC は，感情などのカテゴリごとに単語を集計した辞書であり，心理学者や医学研究者が容易に研究利用可能なリソースである．ただし，この日本語版は広く利用な状態ではない[4]．

既存の日本語の感情表現辞書である「感情表現辞典」[5]では，人間の感情を 10 種類に分類している．しかし，感情の種類が多く，その判断が難しい[6]こと，またソーシャルメディアなどで若者が多く使用するスラングに対応していないなどの問題点がある．

本研究では，テキストの質的な内容分析を可能にする日本語感情表現辞書を構築する．辞書の構築に必要となる膨大な感情表現の収集には，クラウドソーシングを用いる．クラウドソーシングを用いて大量のデータを収集したり分類したりする研究はこれまでに多数報告されており，重要なニュースの判別[7]や機械学習のために必要となる正解データの作成[8]などがある．

本稿では，Yahoo!クラウドソーシングを用いて，エピソードを自由記述で回答させるアンケートタスクを実施し，不特定多数の人々が用いているリアルな感情表現を網羅的に収集する．幅広い年齢層からの感情表現を取得できるため，若者のスラングやトレンドなどにも対応した辞書を構築する．

2. LIWC とは

ソーシャルメディアからの性格推定[9]や認知症，自閉症スペクトラム患者の発話を書き起こしたテキストから語彙を抽象化してカテゴリ化する[10][11]ために，Linguistic Inquiry and Word Count [12]が主に利用されている．LIWC は 6,000 以上の単語が 125 カテゴリに分類されている辞書であるが，日本語版は未だ広く利用できる状態ではない．そのため，これを利用する際は，LIWC もしくはコーパスのどちらかを翻訳する必要があり，これにかかるコストは非常に大きい．また，日本語には存在しないカテゴリ（例: 冠詞）や日本語翻訳により曖昧になる単語 (while や during) などが存在する．そのため，翻訳した辞書は曖昧性を含む可能性があり好ましくないという問題がある．

3. クラウドソーシングによる感情表現の収集

3.1 タスク設定

クラウドソーシングを用いて，不特定多数の人々が用いているリアルな感情表現を収集する．タスクとしては，表 1 に示すように「今までで〜だった」エピソードについての自由記述アンケートを設定する．これは，「悲しい」出来事の回答には「悲しい」を表現する単語が含まれると仮定したためである．

タスクの感情カテゴリ（以降，感情と呼ぶ）は Plutchik の感情の輪（図 1）を参考に設定している．感情表現は基本的に中レベル（図中で内側から 2 つ目の円）の語を選んだが，一部質問に当てはめた時に日本語として不適切になるものは別のレベルの語に変更した（「心配」→「不安」）．今回は過去の出来事に関する質問であるため，未来に対する感情である「予測」は不適切であるため除外する．「喜び」については既に別の研究においてデータを収集しており，回答が重複する可能性が大きいため今回は除外する．また，「受容」は普段聞きなれない単語であるため，混乱を避けるために同様の感情を意味する「信頼感」に変更する．

最終的に，7 つの感情（「怒り」「不安」「嫌悪感」「信頼感」「楽しい（喜び）」「悲しい」「驚き」）を用いる．

3.2 投入設定

本タスクは 2016 年 12 月 31 日 20 時から 2017 年 1 月 11 日 15 時にわたって実施され，1,785 名のクラウドワーカーが参加した．

4. 感情表現辞書の構築

4.1 前処理

取得した回答データの前処理としてノイズの削除を行う．

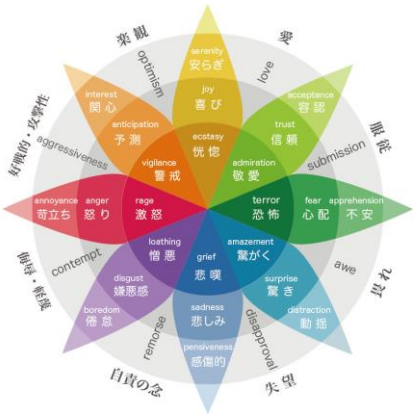

図 1 Plutchik の感情の輪

表 1 クラウドソーシングのタスク内容

タスク内容
Q1. 今までで，最も悲しかった出来事を教えて下さい.
Q2. 今までで，最も不安だった出来事を教えて下さい.
Q3. 今までで，最も怒りが湧いた出来事を教えて下さい.
Q4. 今までで，最も嫌悪感を抱いた出来事を教えて下さい
Q5. 今までで，最も信頼感を抱いた出来事を教えて下さい.
Q6. 今までで，最も驚いた出来事を教えて下さい.

クラウドソーシングの欠点の一つとしてクラウドソーシングの作業結果は低品質であるという問題が挙げられる[13][14]. 本研究においても同様の結果が見られ，意味不明な回答（本稿ではノイズと呼ぶ）がしばしばみられる．作業結果におけるノイズの例としては，「[oga[o[haf;a」や「i'piea'pgka/;lakh/」など適当に入力したと考えられる回答である．このような回答は人手により削除した．

4.2 作成手順

以下の手順で辞書の作成を行う．

Step1: 形態素解析

回答データの形態素を取得する．形態素解析器には Janome[1]を使用する．

Step2: 単語頻度・割合の算出

全回答データの形態素と設問ごとの回答データの形態素を比較し，各感情表現の単語頻度を算出する．この際，各単語の合計頻度が閾値以下の単語は削除する．各感情の単語頻度を各単語の合計頻度で除すことで単語割合を算出する．

Step3: データ整理

算出した単語頻度・割合のデータから，単語頻度が 10 回以下の単語については削除する．また，どの文章にも出現する品詞（助詞・助動詞）や記号（句読点・感嘆符）に関しても同様に削除する．

Step4: 単語スコアの算出

Step2 で算出した単語頻度，単語割合を用いて単語スコアを式(1)で算出する．

$$\frac{W_{ij}}{W_i^*}\log_2(W_{ij}+1) \quad (1)$$

ここで，w_{ij} は単語 w_i の感情 j での頻度，w_{i*}は単語 w_i の全ての感情での頻度である．

Step5: 感情表現の抽出

単語ごとに単語スコアが最も大きくなる感情を求め，その感情の感情表現として単語を辞書に加える．なお，1 つの単語が複数の感情に属することも許容する．

4.3 辞書の更新

辞書の更新は四半期ごとを予定している．これは単語に対する感情が時期によって変化する可能性があり，また時期ごとの特徴的な単語（トレンド単語や季節的な単語など）に対応することを可能にするためである．

4.4 配布場所

本研究により作成した感情表現辞書は国立大学法人奈良先端科学技術大学院大学ソーシャル・コンピューティング研究室のホームページ[2]で公開予定である．

5. 構築した感情表現辞書

4 章で示した手順に沿って感情表現辞書を構築した．なお，4.2 節 Step2 における閾値は 10 とした．表 2 に感情ごとに抽出された感情表現の単語数を示し，表 3 に抽出された感情表現（単語スコアが 1-20 位 20 単語，100-120 位 20 単語，200-220 位 20 単語）を示す．構築した辞書と LIWC で対応するカテゴリに属する単語数を比較したところ（表 3），「怒り」，「不安」，「悲しい」の感情については我々の感情表現辞書の単語数が多く，「楽しい」に関しては LIWC の方が多かった．

さらに，構築した辞書と LIWC で対応するカテゴリにおいて，上位 20 単語の一致率を求めた（表 2）ところ，「怒り」の一致率は 0.4 (=8/20)と最も高く，「不安」「悲しい」の一致率は 0.05 (=1/20) と低い結果となった．

表 2 LIWC との比較

感情	単語数	LIWC単語数（カテゴリ）	LIWCとの一致率
怒り	691	230 (Anger)	0.40
不安	597	116 (Anx)	0.05
嫌悪感	685	–	–
信頼感	646	–	–
楽しい	492	620 (Posemo)	0.20
悲しい	703	136 (Sad)	0.05
驚き	675	–	

[1]http://mocobeta.github.io/janome/

[2]http://sociocom.jp/jiwc.html

6. 考察

表 3 の各感情の感情表現について考察する．

- **「驚き」に関する感情表現**

  「驚き」で注目すべき感情表現は「アメリカ」や「トランプ」である．これは去年のアメリカの大統領選挙において，大多数の予想に反してトランプ氏が当選したことによるものであると思われる．クラウドソーシングを用いて構築することで時事的な単語を補うことが可能である．

- **「楽しい」に関する感情表現**

  「楽しい」の感情表現にはお正月やクリスマスの休日に家族や友達と過ごしたであろうと思われる単語が多く含まれている．これはクラウドソーシングを行った時期が 2016 年の 12 月末であったことに起因していると思われる．

- **「不安」に関する感情表現**

  「不安」の感情表現には試験や就職活動に関する単語が多く含まれている．クラウドワーカーが学生の場合は「試験」や「受験」の結果に最も不安を感じたと推測できる．また，「就職」に関しては現在の若者が就職に対して大きな不安感を抱いていることを示唆しており，現在の日本の現状[3]とも一致する．

- **「嫌悪感」に関する感情表現**

  「嫌悪感」には「セクハラ」や「痴漢」など実際に嫌悪感を覚える単語が含まれている．また，「男」や「おじさん」という単語が含まれることから，回答者に女性が多い可能性を示唆しており，性別によるバイアスを防ぐ処置が必要である．

- **「悲しい」に関する感情表現**

  「悲しい」には「亡くなる」や「死」といった人やペットの生死に関する単語が多く含まれている．一見ポジティブな意味を表す「大好き」という単語が含まれるが，これは「大好きな○○○が亡くなったとき」といった回答が多いためである．

- **「信頼感」に関する感情表現**

  「信頼感」には「助ける」や「支える」といった助け合いに関する単語が多く含まれている．

- **「怒り」に関する感情表現**

  「怒り」には「裏切り」や「騙す」といった不誠実な単語が多く含まれている．また，回答には不倫に関するものが多く，「驚き」の感情表現と同様に時事的な単語まで補っている．

7. まとめ

本稿ではクラウドソーシングによる感情表現辞書の構築の可能性を示すことができた．

ただし，サンプリング時期に依存した結果がみられるため，今後，定期的に複数回の採取の必要性が明らかとなった．本リソースが，今後の日本語感情分析に貢献することを祈念している．

謝辞

本研究の一部は，JSPS 科研費 JP16H06395, JP16H06399, JST, ACT-I, および厚生労働省科学研究費補助金（課題番号：H28-ICT-一般-008）の支援を受けたものです．

参 考 文 献

[1] Robbins ML, Mehl MR. Innovative Real-World Assessment of Health-Relevant Social Processes: The Ear and Liwc in Psychosomatic Medicine, Psychosom Med. 75(3): A6-A, 2013

[2] Salas-Zarate MD, Lopez-Lopez E, Valencia-Garcia R, Aussenac-Gilles N, Almela A, Alor-Hernandez G. A study on LIWC categories for opinion mining in Spanish reviews, J Inf Sci., 40(6), pp. 749-60, 2014

[3] Zhao N, Jiao DD, Bai ST, Zhu TS. Evaluating the Validity of Simplified Chinese Version of LIWC in Detecting Psychological Expressions in Short Texts on Social Network Services, PloS one. 2016;11(6).

[4] 那須川哲哉，上條浩一，山本眞大，北村英哉，日本語における筆者の性格推定のための言語的特徴の調査，言語処理学会年次大会発表論文集，pp.1181-1184, 2016.

[5] 中村明，感情表現辞典，東京堂出版，1993

[6] 山本湧輝，熊本忠彦，灘本明代，ツイートの感情の関係に基づく Twitter 感情軸の決定，DEIM Forum, 2015.

[7] 高濱隆輔，馬場雪乃，清水伸幸，藤田澄男，鹿島久嗣，クラウドソーシングによる重要ニュース選択，人工知能学会全国大会, 2016.

[8] 山下達雄，東野進一，商品レビューに含まれるストア言及の抽出，情報処理学会全国大会, 2016.

[9] 北村英哉，那須川哲哉，上條浩一，山本真大，日本語における筆者の性格推定のための言語特徴の調査，言語処理学会第 22 回年次大会予稿集，pp.1181-1184, 2016.

[10] 柴田大作，若宮翔子，荒牧英治，“アルツハイマーの発症に伴う代名詞の増加”，第 51 回 UBI・第 5 回 ASD 合同研究発表会, 2016.

[11] 田中宏季，サクリアニサクティ，グラムニュービック，戸田智基，中村哲，物語発話からの自閉症スペクトラム障害児と定型発達児の語彙と韻律の特性分析, 日本音響学会春期大会, pp. 1487-1490, 2014.

[12] W. Pennebaker, R.J. Booth, R.L. Boyd, and M.E. Francis. Linguistic inquiry and word count: LIWC2015, 2015.

[13] A. Kittur, E. Chi, and B. Suh. “Crowdsourcing user studies with mechanical turk”, In Proc. of CHI, 2008.

[14] J.S.Downs, M.B.Holbrook, S.Sheng and L.F.Cranor, Are your participants gaming gaming the system?: screening Mechanical Turk wprkers, In Proc. of CHI, 2010.

[3] http://www.sankei.com/economy/news/160627/prl1606270057-n1.html

表 3 構築した感情表現辞書の一部．*は LIWC に近い概念の英語が含まれる単語を示す．
括弧内はスコアを示す．それぞれ 1-20 位， 100-120 位， 200-220 位の単語スコアから抜粋

怒り（重み）		不安（重み）		嫌悪感（重み）		信頼感（重み）		楽しい（重み）		悲しい（重み）		驚き（重み）	
怒り*	(5.18)	不安*	(6.85)	嫌悪	(5.55)	くれる	(8.22)	クリスマス	(5.72)	亡くなる	(7.46)	驚く	(5.44)
理不尽	(4.93)	受験	(5.91)	痴漢	(3.72)	助ける	(6.26)	忘年会*	(5.59)	死	(6.81)	宝くじ	(5.12)
湧く	(4.22)	発表	(5.82)	電車	(3.54)	困る	(5.62)	楽しい*	(5.57)	死ぬ	(5.77)	当選	(4.59)
れる	(3.74)	どう	(5.45)	セクハラ	(3.43)	信頼	(4.91)	正月	(5.42)	飼う	(5.44)	当たる	(4.39)
騙す*	(3.17)	就職	(5.21)	悪口	(3.16)	相談	(4.70)	お正月	(5.37)	祖母	(5.44)	解散	(3.99)
つかれる	(3.10)	入試	(4.59)	不倫	(3.07)	支える	(4.64)	過ごせる	(4.86)	悲しい*	(5.38)	大統領	(3.91)
裏切り	(3.06)	結果	(4.46)	座る	(2.96)	親身	(4.58)	帰省	(4.80)	ペット	(5.16)	トランプ	(3.88)
裏切る*	(3.02)	将来	(4.42)	男	(2.93)	励ます	(4.15)	久しぶり	(4.78)	祖父	(5.10)	芸能人	(3.54)
喧嘩*	(2.96)	試験	(4.38)	嫌	(2.91)	もらう	(4.09)	初詣	(4.64)	愛犬	(4.98)	懸賞	(3.22)
盗む	(2.83)	検査	(3.79)	平気	(2.86)	手伝う	(3.73)	遊ぶ*	(4.62)	他界	(4.89)	アメリカ	(3.19)
られる	(2.82)	大学	(3.71)	男性	(2.84)	優しい	(3.72)	温泉	(4.58)	大好き	(4.47)	偶然	(3.14)
扱い	(2.72)	無事	(3.66)	人	(2.78)	味方	(3.63)	旅行	(4.57)	おじいちゃん	(4.32)	結婚	(3.02)
破る*	(2.66)	活動	(3.59)	態度	(2.73)	フォロー	(3.53)	過ごす	(4.47)	ばあちゃん	(4.22)	津波	(2.88)
お前	(2.66)	迷子	(3.58)	抱く	(2.72)	アドバイス	(3.42)	飲む	(4.30)	亡くす	(4.12)	受かる	(2.74)
浮気	(2.64)	手術	(3.45)	おじさん	(2.66)	かばう	(3.42)	買い物	(4.25)	別れる	(4.10)	誕生	(2.48)
怒る	(2.51)	決まる	(3.43)	隣	(2.60)	悩み	(3.42)	会	(4.21)	犬	(4.09)	妊娠	(2.46)
暴言*	(2.49)	面接	(3.38)	発言	(2.60)	愛情	(3.42)	面白い*	(4.17)	失恋	(3.93)	成長	(2.34)
責任	(2.47)	なかなか	(3.36)	女性	(2.59)	落ち込む	(3.42)	家族	(4.14)	別れ	(3.78)	婚	(2.32)
詐欺*	(2.44)	これから	(3.22)	感	(2.56)	守る	(3.32)	食べる	(4.10)	しまう	(3.57)	大震災	(2.26)
嘘*	(2.40)	出産	(3.19)	韓国	(2.50)	どんな	(3.27)	集まる	(4.09)	死去	(3.54)	同級生	(2.13)
否定	(1.12)	予定	(1.30)	うそ	(1.17)	必要	(1.20)	行う	(1.10)	好き	(1.10)	朝	(0.91)
義理	(1.12)	不良	(1.30)	ママ	(1.17)	いろいろ	(1.20)	誕生	(1.09)	帰る	(1.09)	知人	(0.91)
姑	(1.11)	テスト	(1.24)	すぎる	(1.16)	得る	(1.20)	時間	(1.07)	向かう	(1.08)	弟	(0.91)
自身	(1.11)	お金	(1.20)	会社	(1.15)	成功	(1.20)	仲間	(1.06)	難病	(1.08)	家	(0.89)
給料	(1.11)	なくなる	(1.20)	者	(1.14)	行動	(1.19)	ネット	(1.05)	事故	(1.07)	っ子	(0.89)
くる	(1.08)	あと	(1.20)	知り合い	(1.14)	チーム	(1.17)	できる	(1.05)	日	(1.05)	なる	(0.87)
指導	(1.08)	会社	(1.18)	続ける	(1.13)	やる	(1.17)	好き	(1.03)	時	(1.01)	とき	(0.87)
注意	(1.08)	日	(1.18)	員	(1.12)	大変	(1.17)	親戚	(0.98)	今	(1.01)	みる	(0.86)
夫	(1.07)	費	(1.17)	店員	(1.12)	認める	(1.17)	達	(0.93)	まだ	(1.00)	近所	(0.85)
結婚式	(1.07)	無くなる	(1.17)	義理	(1.12)	良い	(1.16)	出来る	(0.93)	原因	(1.00)	世の中	(0.84)
問題	(1.07)	直後	(1.17)	差別	(1.11)	頑張る	(1.15)	地元	(0.92)	トイレ	(1.00)	企業	(0.84)
内	(1.07)	がん	(1.17)	つかれる	(1.10)	的	(1.14)	無い	(0.92)	余命	(0.99)	選ぶ	(0.84)
テロ	(1.06)	待ち	(1.17)	時	(1.09)	ところ	(1.14)	事	(0.89)	半年	(0.99)	社内	(0.84)
つく	(1.05)	いく	(1.16)	ある	(1.08)	旦那	(1.13)	日	(0.88)	前	(0.99)	待ち	(0.84)
動物	(1.05)	中	(1.15)	違う	(1.08)	為	(1.12)	合う	(0.85)	代	(0.99)	勉強	(0.84)
人	(1.05)	職	(1.15)	度	(1.05)	約束	(1.11)	全員	(0.84)	万	(0.99)	転勤	(0.84)
さ	(1.04)	数	(1.14)	ない	(1.05)	変わる	(1.10)	部屋	(0.84)	それ	(0.96)	テレビ	(0.84)
ある	(1.04)	子供	(1.13)	意見	(1.05)	無い	(1.09)	親しい	(0.84)	元気	(0.95)	出る	(0.82)
貸す	(1.03)	わかる	(1.11)	方	(1.05)	新人	(1.08)	迎える	(0.84)	頑張る	(0.95)	友達	(0.81)
すぎる	(1.02)	する	(1.11)	お金	(1.04)	指導	(1.08)	東京	(0.81)	入る	(0.94)	不倫	(0.81)
母親	(0.00)	生まれる	(0.64)	家	(0.63)	たち	(0.63)	ママ	(0.29)	弟	(0.64)	倒れる	(0.51)
働く	(0.67)	分かる	(0.64)	ところ	(0.62)	すべて	(0.62)	呼ぶ	(0.29)	生まれる	(0.64)	来る	(0.51)
発覚	(0.67)	入る	(0.64)	関わる	(0.62)	迷う	(0.62)	か月	(0.29)	受験	(0.64)	チケット	(0.50)
戻る	(0.67)	際	(0.63)	携帯	(0.62)	状況	(0.62)	両親	(0.28)	悪い	(0.63)	難病	(0.50)
家	(0.67)	交通	(0.62)	せる	(0.61)	進む	(0.62)	妹	(0.27)	子	(0.63)	長女	(0.50)
知人	(0.67)	状況	(0.62)	夫	(0.61)	好き	(0.62)	姉	(0.27)	よう	(0.63)	救急	(0.50)
いきなり	(0.66)	決める	(0.62)	毎日	(0.60)	嫁	(0.62)	健康	(0.27)	いく	(0.63)	高い	(0.50)
殺人	(0.66)	警察	(0.62)	覚える	(0.60)	一番	(0.62)	いる	(0.27)	中学	(0.62)	一番	(0.50)
犯罪	(0.66)	センター	(0.62)	父親	(0.60)	経験	(0.62)	犬	(0.27)	長い	(0.62)	大きい	(0.49)
事故	(0.66)	止まる	(0.62)	そう	(0.60)	信用	(0.62)	高い	(0.26)	中	(0.62)	道	(0.49)
そう	(0.66)	地	(0.62)	わかる	(0.60)	覚える	(0.60)	上がる	(0.26)	止まる	(0.62)	れる	(0.49)
兄	(0.66)	親	(0.62)	先輩	(0.59)	みんな	(0.60)	はず	(0.26)	小さい	(0.62)	忘れる	(0.48)
性	(0.65)	いる	(0.62)	最も	(0.58)	受ける	(0.60)	話す	(0.26)	すべて	(0.62)	もの	(0.48)
意見	(0.65)	やめる	(0.62)	傷つける	(0.58)	父	(0.59)	朝	(0.26)	行く	(0.62)	母	(0.47)
弟	(0.64)	ない	(0.61)	動物	(0.58)	連れる	(0.58)	回	(0.26)	経験	(0.62)	気づく	(0.46)
財布	(0.64)	全く	(0.60)	姿	(0.58)	思いつく	(0.58)	高校	(0.26)	信用	(0.62)	時期	(0.46)
気	(0.64)	二	(0.60)	時代	(0.57)	達	(0.57)	彼女	(0.25)	学校	(0.60)	性	(0.46)
最も	(0.62)	ため	(0.60)	良い	(0.57)	体調	(0.57)	夜	(0.25)	続ける	(0.60)	後	(0.46)
報告	(0.62)	父	(0.59)	ん	(0.56)	電話	(0.56)	とき	(0.25)	姉	(0.59)	翌日	(0.46)
意味	(0.62)	家	(0.59)	周り	(0.56)	もの	(0.56)	電話	(0.24)	欲しい	(0.58)	余命	(0.46)