

言語処理学会第23回年次大会

Japanese linguistic Inquiry and Word Count JIWC : クラウドソーシングによる 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





背景：NLPの発展

- 自然言語処理技術の発展
 - テキスト解析の需要が増加
 - 商品レビュー
 - 疾患をもつ患者の発話
- ある軸に沿って定量化できる指標の算出が中心
 - 書き手の感情や心理分析などは困難
- 一方、海外ではLIWCを用いた質的な分析が盛ん
 - 英語・中国語・スペイン語が存在

背景：LIWCとは

LIWC: Linguistic Inquiry and Word Count

- 海外のテキスト分析に幅広く利用
- 社会学・心理学者などにより開発
- 語彙をカテゴリ化するツール
- 6,000語以上の単語が125カテゴリに分類

<Posemo>	<Negemo>	<Anx>	<Sad>	<Family>
accept	abandon	afraid	abandon	baby
excel	diss	confuse	cry	bro
humour	ignore	fear	doom	dads
prouder	protest	horrible	dull	fam
well	tragic	guilt	gloom	folks

背景：LIWCの使用例

- 海外での使用例
 - 海外での使用例 (ACL Anthology: 266件)
 - 感情分析 [Chowdhury+ 2013]
 - 認知症の検出 [Jarrold+ 2014]
 - メンタルヘルスの分析 [Coppersmith+ 2014]
- 日本での使用例
 - ASDの検出 [Tanaka+ 2014]
 - 性格推定 [那須川+ 2015]
 - 認知症の検出 [Shibata+ 2016]

背景：LIWC日本語化の問題点

- LIWCまたはコーパスの翻訳が必要
 - コストが大きい
- 日本語には存在しないカテゴリ
 - 冠詞
- 日本語に翻訳困難な語が存在
 - like: <Posemo>, <Filler>, <Time>

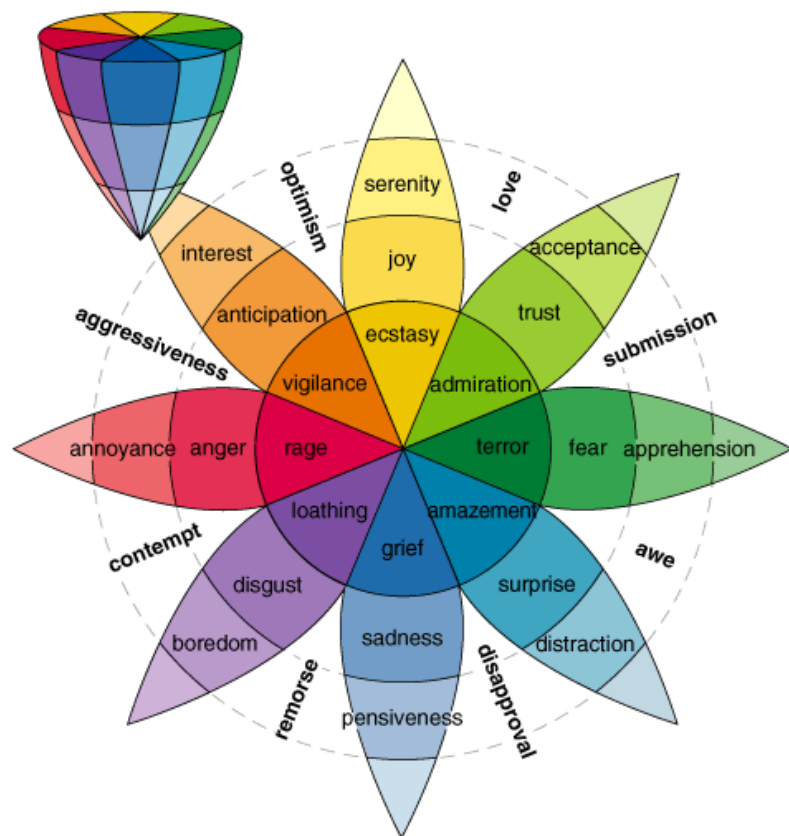
日本語で幅広く利用可能な**言語リソース**が必要
LIWC：感情表現辞書の構築

感情表現の収集：タスク設定

- Yahoo!クラウドソーシングを用いて人々の感情表現を収集
- タスク内容：「今まで～だった」エピソードについての自由記述アンケート
 - 「悲しい」エピソードの回答には「悲しい」を表現する単語が含まれているはず
- 日時：2016年12月31日20時から2017年1月11日15時
- 参加者数：1,785名

感情表現の収集：感情の決定

収集する感情は感情の輪[※]を参考に設定



プルチックの感情の輪

感情の輪の7つの感情

喜び
信頼
心配
驚き
悲しみ
嫌悪感
怒り
予測

最終的な7つの感情

楽しい
信頼感
不安
驚き
悲しい
嫌悪感
怒り
(削除)

感情表現の収集：タスク内容

6種類の設問を設定

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.
-

回答にはその「感情」を表す「感情表現」
が含まれているはず

感情表現の収集：タスク画面（1）

終了

【経験談募集】「最も～だった」エピソードについての自由記述アンケート

全員ポイント進呈 ヒント

タスクの種類

アンケート

もらえるポイント

 10

タスクにかかるめやす時間

10分00秒

残り件数

 215件

(全2,000件)

タスク実施中のユーザー：0人

89%完了

 実施できる件数の上限

1件

 制限時間 ヒント

60分00秒

 残り日数

終了

(2017/01/11 15:25)

 オーナー

非公開

 このタスクって、どんな内容？

【なにをするの?】

このタスクでは、あなたの「**今まで最も～だった**」エピソード(例:今までで**最も嬉しかった**エピソード)についての体験談をお伺いします。

【なにに使うの?】

研究利用

【回答内容の取り扱いについて】

回答の送信をもって、この調査の趣旨本研究に**ご同意**いただけたものとさせていただきます。また、回答は無記名ですので、このタスクへの参加によって個人が特定されることはありません。**回答内容は研究の目的以外で使用されることは無く、データは研究終了後に消去いたします。**調査結果につきましては、報告書を作成したり、情報系・医療系の学会や専門誌などで発表する可能性があります。

感情表現の収集：タスク画面（2）

プレビュー表示切り替え：●パソコン用 ●スマートフォン用

次の設問へ

設定した設問ID：1

今までで、最も悲しかった出来事を教えて下さい。

入力フォーム

エントリーする

感情表現の収集：解答例（1）

- 最も悲しかった出来事
 - 祖母の死
 - 愛犬が死んだ時
- 最も驚いた出来事
 - 東日本大震災
 - トランプが当選したこと
- 最も不安だった出来事
 - 仕事の面接
 - 大学受験

感情表現の収集：解答例（2）

- 最も怒りが湧いた出来事
 - 裏切り
 - 理不尽なことで怒られたとき
- 最も信頼感を抱いた出来事
 - 親友に困っているとき助けてもらったとき
 - 結婚したこと
- 最も嫌悪感を抱いた出来事
 - 上司のパワハラ
 - 会社でのセクハラ発言

JIWCの構築：前処理

取得したデータのノイズを削除

- CSの作業結果は低品質 [A.Kittur + 2015]
- 本研究においても同様の結果

Q1. 最も悲しかった出来事はなんですか

Ans. ;jag;ja;jajpofjamelj

Q2. 最も不安だった出来事はなんですか

Ans. k[tih[kgaf;

適当に入力したと思われる回答は人手で削除

JIWCの構築：作成手順（1）

- **Step1: 形態素解析**
 - JUMAN[※]++による形態素解析
 - 代表表記を取得
- **Step2: 単語頻度・割合の算出**
 - 単語ごとの単語頻度を算出
 - 単語ごとの割合を算出

$$\text{Ratio of Word} = \frac{w_{ij}}{w_i^*}$$

w_{ij} は単語 w_i の感情 j における頻度

w_i^* は全ての感情での頻度

JIWCの構築：作成手順（2）

- **Step3: データ整理**
 - 単語頻度が10以下の単語は削除
 - 助詞・助動詞や記号も同様に削除
- **Step4: 単語スコアの算出**
 - 単語頻度・割合より単語スコアを算出

$$Score = \frac{w_{ij}}{w_i^*} \log(w_{ij} + 1)$$

w_{ij} は単語 w_i の感情 j における頻度

w_i^* は全ての感情での頻度

JIWCの構築：作成手順（3）

Step5: 感情表現の分類

単語スコアが最も大きい感情
をその感情の感情表現として追加

単語	驚き	楽しい	不安	嫌悪感	悲しい	信頼感	怒り
会社	0.38	0.11	1.18	1.15	0.38	0.75	1.36

「会社」の単語スコアは感情「怒り」で最も大きい

感情「怒り」に感情表現「会社」を分類

現在配布中のJIWC（2017年1月版）では未対応
感情表現の数が増加しだい、対応予定

JIWCの構築：感情表現

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)

JIWCの構築：LIWCとの比較

上位20単語の英訳とLIWCとの一致率

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	-	-

怒り：0.40 (=8/20)

不安：0.05 (=1/20)

楽しい：0.20 (=4/20)

悲しい：0.05 (=1/20)

嫌悪感・信頼感・驚き是对应するカテゴリが存在しないため考慮しない

JIWCの構築：更新

更新は**四半期**ごとを予定

単語の感情が時期に依存（ニュースや季節）

驚き	
驚く	偶然
宝くじ	結婚
当選	津波
当たる	受かる
解散	誕生
大統領	妊娠
トランプ	成長
芸能人	婚
懸賞	大震災
アメリカ	同級生

不安	
不安	大学
受験	無事
発表	活動
どう	迷子
就職	手術
入試	決まる
結果	面接
将来	なかなか
試験	これから
検査	出産

- ・ **トランプ**や**大統領**は時期で変化する可能性
- ・ **受験**や**就職**は季節に依存している可能性

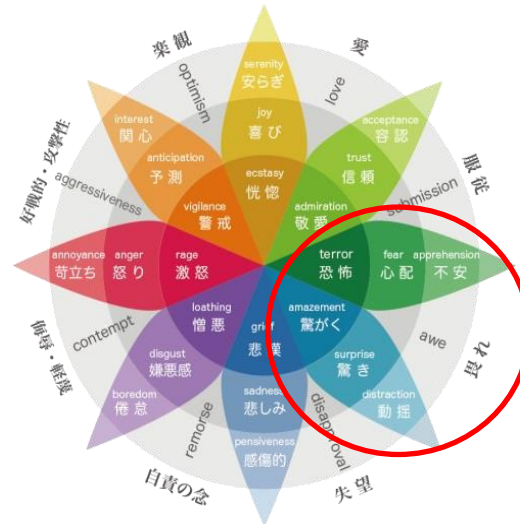




考察：単語スコアについて

複数の感情で単語スコアが同等の単語

単語	驚き	楽しい	不安	嫌悪感	悲しい	信頼感	怒り
震災	0.61	0.00	0.76	0.00	0.07	0.06	0.03

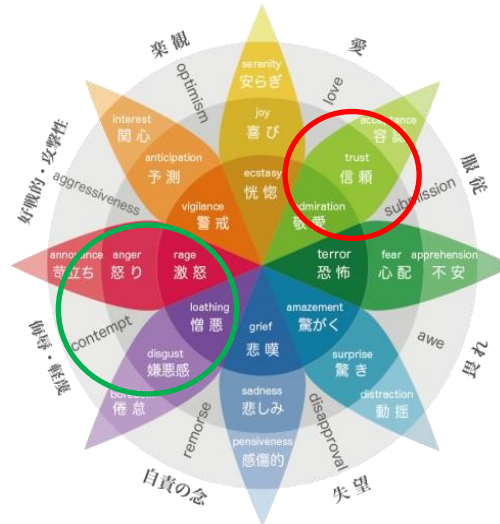


不安と驚きは類似する感情
同じ単語が複数の感情に属することを許容

考察：単語スコアについて

複数の感情で単語スコアが同等の単語

単語	驚き	楽しい	不安	嫌悪感	悲しい	信頼感	怒り
上司	0.03	0.00	0.01	0.60	0.01	0.66	0.60



対極の感情で単語スコアが同等
データ数を増やし再度検討

まとめ

- クラウドソーシングにより人間の感情表現を集めた感情表現辞書（JIWC）を構築
- 時事的な単語を収集することが可能
- 辞書の更新は四半期ごとを予定
- 今後の予定
 - 解析ツールの開発
 - 認知症者の語りに適用 [1]
 - 感情分析による群衆密度の推定 [2]

[1] 音声発話による認知症スクリーニング技術の開発 感情表現辞書を用いた発話内容の質的分析, 柴田大作, 若宮翔子, 木下彩栄, 荒牧英治, 第21回日本医療情報学会春季学術大会（投稿中）

[2] Twitterによる群衆密度の推定：第29回関門海峡花火大会での実証実験, 柴田大作, 若宮翔子, 宮部真衣, 大西正輝, 山下倫央, 野田五十樹, 荒牧英治, DEIM2016



JIWC: 日本語感情表現辞書

はじめに

近年、自然言語処理技術の発展に伴いテキストの内容を分析する研究への関心が高まっています。テキストの内容を分析する際は、Linguistic Inquiry and Word Count（以下、LIWC）を翻訳したものを用いることが多く、日本語で幅広く利用可能な言語ツールが存在しませんでした。そこで我々は日本語の感情表現辞書の構築を行いました。この辞書がテキストの分析などに携わっている方の助けとなれば幸いです。

日本語感情表現辞書（JIWC）

人間の感情表現をクラウドソーシングにより取得し、それをまとめた辞書です。感情は「怒り」、「不安」、「嫌悪感」、「信頼感」、「楽しい」、「悲しい」、「驚き」の7つとなっています。クラウドソーシングを用いることで時事的な単語（2017年2月現在であれば、「驚き」：トランプ など）を取得することが可能となりました。本辞書は年4回（季節ごと）の更新を予定しております。辞書の詳細に関しては[1]をご参照ください。

タスク内容

- Q1. 今までで、最も悲しかった出来事を教えてください。
- Q2. 今までで、最も不安だった出来事を教えてください。
- Q3. 今までで、最も怒りが湧いた出来事を教えてください。
- Q4. 今までで、最も嫌悪感を抱いた出来事を教えてください。
- Q5. 今までで、最も信頼感を抱いた出来事を教えてください。
- Q6. 今までで、最も驚いた出来事を教えてください。

奈良先端大 ソーシャル・コンピューティング研究室
ホームページにて配布中： <http://sociocom.jp/jiwc.html>