Twitter による評判分析を目的とした 評価対象-評価表現データセット作成

栗原 理聡 † 水本 智也 ‡ 乾 健太郎 † ‡
† 東北大学 ‡ 理研 AIP
{masakuri,inui}@ecei.tohoku.ac.jp, tomoya.mizumoto@riken.jp

1 はじめに

企業や自治体といった組織において，大衆からの評判をその時々で素早く把握することは組織の運営を円滑に進める上で重要である．そのため，テキストから評判を分析する研究が盛んに行われており，targeted sentiment analysis（TSA），aspect extraction（AE），opinion extraction（OE）など様々なタスクが設定され，手法が提案されている．

これらのタスクは評判分析を行う上で一定の効果があるが，不十分な点も存在する．TSA は入力されたテキスト内の指定した対象に対する評価極性を同定するタスクであり，希望の対象の評価（ポジティブ/ネガティブ）を得られるという点では効果があるが，評価の根拠，内容といった情報は得られない．例えば，図 1 で評価の欲しい対象が「伊丹空港」であるとき，TSA ではポジティブであるという結果しか得られない．しかし，実際に評判を分析する際は単純な評価だけでなく，どのような評価内容であるかまで同定したい．

このような問題を解決するために、AE や OE などの研究が行われている．AE，OE は何らかの評価が記述されたテキストに対して，それぞれ評価対象，評価内容を同定するタスクである．しかし，レビューサイトなどの評価が述べられたテキストが大量に取得可能であるという前提で行われているものが多く，ドメインに偏りがある．公共交通機関や自治体といった，利用に選択性 *1が乏しい対象ではレビューサイトが充実しておらず，評価が書かれたテキストを大量に収集すること自体が課題である．

評判分析に関するデータセットも公開されているが，評価極性のみに焦点を当てているもの [8] [7] [1] [3] が多い．評価表現に着目したものとして京都観光ブログの評価情報付与データ，[10] のデータがあるが，評価や意見が書かれているテキストが前提でドメインも制限されており，ドメインに関係なく評価対象-評価表現を抽出できるデータはない．

そこで本研究では Twitter*2を利用して，日本語評価

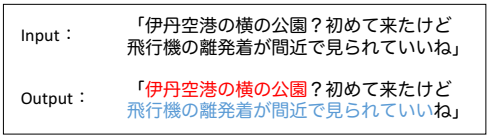

図 1 Twitter を利用した評価対象-評価表現抽出．
（赤：評価対象，青：評価表現）

対象-評価表現データセットの作成を行った．Twitter は対象が縛られず様々なトピックに関して情報が飛び交うため様々なドメインに関する評価対象-評価表現を含むツイートを抽出することが考えられる．本研究でデータセットを作成した手法で，任意の対象に対する詳細な評判分析を可能とするデータセットを作ることも可能である．本稿では，このデータセット作成の手順について報告する．また，作成したデータセットにおける評価対象と評価表現を抽出するタスクのベンチマークとして 3 つのモデルで実験した．そこで得られた結果を元にデータセット作成手順と評価対象-評価表現抽出の課題について報告する．作成したデータセットは公開予定である *3．

2 データセット作成

評価対象-評価表現データセット作成について述べる．データセット作成には Yahoo!クラウドソーシング *4を利用する．本研究では，評価対象-評価表現データセットを二段階に分けて作成した．

1. ツイートが評価情報を含むかを付与
2. 上で評価情報を含むとなったツイートに対して，評価対象-評価表現を付与

Twitter では様々なトピックが飛び交っており，そのままだと評価情報が含まれるツイート数の割合は小さい．そのため，直接評価対象-評価表現を付与してもらうと，ほとんどが評価対象-評価表現を含まないツイートとなり非効率である．本稿では，1 で作るデータを評価情報分類データ，2 で作るデータを評価対象-評価表現データセットと呼ぶ．

*1 ここで言う選択性とは，人が特定の目的の遂行に際して利用するものを，様々な情報から選択する余地のことを指す．
*2 https://twitter.com/
*3 http://www.cl.ecei.tohoku.ac.jp/index.php?Open Resources
*4 https://crowdsourcing.yahoo.co.jp/

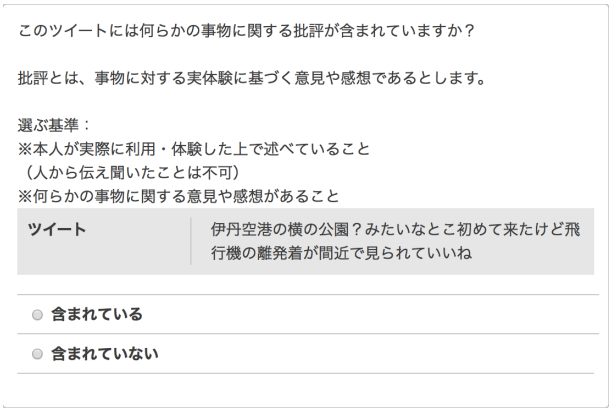

図 2　評価情報分類用アノテーション画面.

p(z = 1) ツイート数	<0.5 32,620	<0.6 940	<0.7 849	<0.8 684	<0.9 553	≤1.0 3,954
票数 (z = 1) ツイート数	0-2 31,226	3 4,011		4 2,797		5 1,556

表 1　（上段 2 段）: GLAD においてラベル z が「含まれている」(z = 1) である確率毎のツイート数，（下段 2 段）: ラベル z が「含まれている」(z = 1) であるとワーカーが判断した投票数毎のツイート数.

2.1　評価情報分類データ

評価分析対象を含むツイートのみをクラウドソーシングに投げ，1 ツイートあたり 5 人のワーカーにアノテーションしてもらう．得られたデータに対して多数決により真のラベルを決定することも考えられるが，ここでは GLAD[9] を利用する．GLAD はデータセット作成にクラウドソーシングを利用する際に問題となる，アノテーションデータの質向上を目的とした手法である．ワーカー i の能力 α_i，タスク j の難易度 β_j の 2 パラーメータを，実際に得られたワーカーのアノテーション結果から EM アルゴリズムを用いて導出し．真のラベルを推定する．本研究では，GLAD によって真のラベルが「含まれている」である確率値が 0.5 以上のツイートを評価情報が含まれるツイートとする．

2.2　評価対象-評価表現データセット

評価対象-評価表現のアノテーションデータを作成するために，ラベル付けに Hanawa ら [2] による，Yahoo! クラウドソーシングとアノテーションツール brat [*5] を連携したアノテーションシステムを利用する．このシステムを使うことで，クラウドソーシングによって特定の文字列に対してラベルを付与することができる．

図 3 のような画面で，1 ツイートあたり 5 人のワーカーに REVIEW（R ラベル），TARGET（T ラベル），R ラベルが存在しない場合は NONE ラベル（N ラベル）の付与を行ってもらう．このタスクでのクラウドソーシングの質の調査のために 1,000 件のツイートに対し，1 ツイートあたり 10 人でラベルをつけたデータを作成し

[*5] http://brat.nlplab.org/

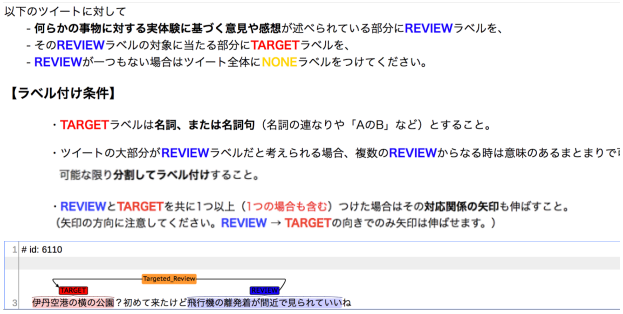

図 3　評価対象-評価表現アノテーション画面.

た．このデータに対し一致率（式 1）0.5 以上のラベルを同一とみなした時の 1 ツイートあたりのラベルつきデータ数を調査すると，R ラベルに対し平均 3.13 個のラベルを得た．一方 1 ツイートのワーカーあたりのラベル数の平均は R ラベルに対し 0.92 である．つまり，1 ワーカーは 1 ラベル程度しかつけていないが，約 3 ラベル得られるということは．ワーカー間のアノテーションに揺れがあることを意味する．

そのため，本研究ではワーカーが付けたラベルを 1 つの gold ラベルに統合することを行う．統合手順は以下の通りである（図 4 も参照）.

1. 1 ワーカーの複数ラベルを別ラベル化
2. 一致率が 0.8 以上のラベルの統合
   ラベル l_1 と l_2 の R ラベル，T ラベルそれぞれの一致率（式 1）が 0.8 以上のラベルは同一とみなし，R ラベルが長い方を小さいラベルに統合する（同長の場合 T ラベル長の小さい方）．ここで $LCS\,(l_1, l_2)$ は l_1，l_2 間の最長共通部分文字列とする．

$$\text{一致率} = \frac{LCS(l_1, l_2).length}{\max(l_1.length, l_2.length)} \qquad (1)$$

3. 多数決で最上位のデータ数が 1 ラベルの場合，gold ラベル確定.
4. 複数のラベルが存在する場合，以下.
   (a) N ラベルのデータ除外（R ラベル有を優先）
   (b) T ラベルが文頭に近いデータを優先
   (c) T と R ラベルの距離が近いデータを優先

2.3　アノテーション結果

上記の手順をもとに，本研究ではレビューサイトが充実していない対象として公共交通機関（鉄道駅，空港）を対象とし，2016 年 8 月に投稿されたツイートに対し，クラウドソーシングを利用した評価対象-評価表現データの作成を試みた．評価情報分類データ収集では 39,600 件のツイートをクラウドソーシングに投げ，「含まれている」のタグ付きツイートデータを 6,980 件得た．ここで，GLAD を本データに適応したときの真のラベルが「含まれている」である確率毎のツイート数を表 1 に示す．上記で収集した評価表現を含む可能性の高いデータ

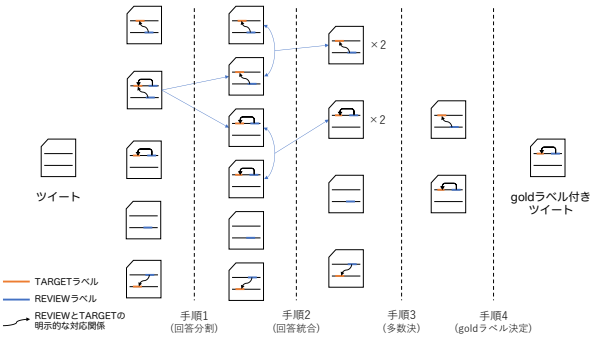

図 4　5 ワーカーのラベルから 1gold ラベルへの統合手順．(1 列目)：評価情報を含むツイート．(2 列目)：5 ワーカーのアノテーション．(3 列目)：1 ワーカーが 2 つの R ラベルを付けている結果をそれぞれ別ラベル化．(4 列目)：得られた 6 ラベルで，一致率 0.8 以上のラベルを統合．(5 列目)：ここまでで存在するラベル間で多数決．(6 列目)：残ったラベルに対し手順 2.4 のルールに基づき最終的な gold ラベル決定．

6,980 件を使って，評価対象-評価表現データセットを作成した．一段階目で，評価表現を含むツイートを抽出したが，それでも R ラベルのつかないツイートが 1,083 件あった．評価対象-評価表現データ中の平均ツイート長は 57 文字（SD: 34.92）であった．そのツイートに対して REVIEW のついたフレーズの長さの平均は 16 文字（SD: 13.25），最大 125 文字，最小 1 文字であった．同様に TARGET は，平均 7 文字 (SD: 5.93)，最大 80 文字，最小 1 文字であった．

3　実験

2 節で作成したデータセットを元に，与えれらた入力テキスト（ツイート）に対して，評価情報が含まれている場合，評価対象-評価表現のペアを抽出する系列ラベリングのタスクに取り組み，データセットの性質についての分析を行う．データは REVIEW と TARGET のラベルのついた 6,980 件のデータを train:dev:test = 8:1:1 の割合で分割し，train および dev では R ラベル，T ラベルがどちらも付与されていない（NONE ラベル）データを除き，学習に利用した．

3.1　モデル

今回使用したモデルは以下の 3 種類である．

■CRF:　T ラベルと R ラベルを一つの CRF モデルで同定する．実装は sklearn-crfsuite[*6] を使用した．

■二段階 CRF (n-CRF):　本研究で設定したタスクでは R ラベルが存在して初めて T ラベルが存在するという，ラベル間の強い依存関係がある．そこで，まず R ラベルを同定するモデルを構築する．その REVIEW モデルの予測結果を素性として加えて T ラベルを同定するモデルを学習するという 2 段階でラベル予測を行う．ラベルごとにモデルを構築することで，素性選択にも自由度が生まれ，CRF よりも性能が向上することが期待される．実装は sklearn-crfsuite を使用した．

■biLSTM-CNN-CRF (NN-CRF):　ニューラルネットワークモデルとして，Ma ら [5] の手法を使った．このモデルは NER や POS tagging のタスクで当時 state-of the-art を達成した．実装は deep-crf[*7] を利用した．

3.2　素性

CRF および n-CRF におけるモデルの性能向上が期待できると考え利用した素性を以下に示す．

■表層形 (surf):　時刻 t，t-1，t+1 の単語の表層形．

■n-gram(ng):　時刻 t の単語を先頭とする原形の uni-gram，bi-gram，tri-gram．

■品詞 (pos):　時刻 t，t-1，t+1 の単語の品詞．

■接尾辞 (suff):　時刻 t，t-1，t+1 の単語の接尾辞．

■辞書 (dic):　R ラベルを同定するにあたり，時刻 t の単語，または n-gram が日本語評価極性辞書（用言編）[4] の「評価」の評価極性（ポジティブ/ネガティヴ）が付与された単語（列）に含まれているかのバイナリ

■R ラベル (rv):　T ラベルを同定するにあたり，時刻 t の単語が R ラベルが付与された単語かのバイナリ

■直接係り受け (dd)　T ラベルを同定するにあたり，時刻 t の単語が R ラベルが付与された単語と直接係り受けの関係にあるかのバイナリ．係り受け解析には CaboCha [*8] を利用した．

■類似度 (ws):　T ラベルを同定するにあたり，時刻 t の単語分散表現と R ラベルが付与された単語列における単語分散表現の平均とのコサイン類似度．分散表現の学習には word2vec を利用し，次元数は 100 とした．

3.3　評価

比較的長いスパンのラベルを同定するタスクの性能評価として，Nguyen ら [6] が行なっている評価を参考に，R ラベル，T ラベルが存在しないことにより精度の計算が不可能となることを避けるため以下のようにマイクロ平均で全体の性能を評価する（F1 値は調和平均）．

$$\mathrm{Pre} = \frac{\sum_i^n (\#\text{true positive words for data } i)}{\sum_i^n (\#\text{words in the predicted span for data } i)}$$

$$\mathrm{Rec} = \frac{\sum_i^n (\#\text{true positive words for data } i)}{\sum_i^n (\#\text{words in the gold span for data } i)}$$

3.4　結果・分析

実験の結果を表 3 に示す．R ラベル，T ラベル共にニューラルネットワークのモデルが最も高い性能を出したが，それでも F1 値で 50 ポイント程度であり，現状の系列ラベリングモデルでは高い性能を発揮できていない．一方で評価の仕方を再考する必要性も見受けられた．表 4 の (1) の出力例の場合，評価方法を 3.3 節に

[*6] https://github.com/TeamHG-Memex/sklearn-crfsuite/

[*7] https://github.com/aonotas/deep-crf/

[*8] https://taku910.github.io/cabocha/

モデル	ラベル	素性選択						REVIEW			TARGET			
		base	suff	dic	rv	dd	ws	Pre	Rec	F1	Pre	Rec	F1	F1(上界)
CRF	A	✓	✓					45.19	36.68	40.05	58.14	45.98	51.35	
n-CRF	R	✓		✓				44.14	33.11	37.84				
	T_p	✓	✓		✓						63.60	46.45	53.69	54.22
		✓	✓		✓	✓					64.17	44.99	52.89	56.20
		✓	✓		✓		✓				64.23	45.10	53.00	53.20
		✓	✓		✓	✓	✓				**65.76**	46.68	54.60	56.42
	T_g	✓	✓		✓	✓	✓				53.00	46.97	49.80	63.02
NN-CRF	A							**57.12**	**50.33**	**53.51**	63.55	**52.62**	**57.57**	

表 2　モデルごとの評価対象-評価表現抽出精度．surf，ng，pos はベース素性として base と表記している．n-CRF における T_p は 2 段階目の学習時に 1 段階目の REVIEW の予測結果を利用したモデル．T_g は 2 段階目の学習時に REVIEW の gold ラベルを利用したモデル．F1（上界）は，テスト時に REVIEW の gold ラベルを利用した精度である．（2 列目の A は R ラベル，T ラベルの同時推定，R は R ラベルの推定を表す．）

ならうと，R ラベルに関してのモデルの精度は低くなるが，モデルの出力ラベルは不自然ではない．この例に対する 5 人のワーカーのアノテーション結果を見ると，一人のワーカーは n-CRF/NN-CRF と同じラベルづけをしていることもわかった．しかし， 2.2 節の手順 2.4 により，gold ラベルとして採用されなかった．このことからデータセット作成のラベルの統合に関してはまだ改善の余地があると言える．

2 段階にすることで CRF よりも性能向上を期待した n-CRF だが，F1 値の性能は CRF を下回った．分析した結果，主に 2 つの要因があると考える．まず，上記でも示したものと同様に，表 4 の (2) の CRF の出力は不自然ではないが gold ラベルと差異があった例であり，ラベル統合をより慎重に行う必要がある．2 点目として，(3) に示すような T ラベルのみ付与するエラーが多く見つかった，実際にテストデータ内で T ラベルの付けられたデータ数を見ると，CRF で 192 件，n-CRF で 204 件と悪化していた，また，T ラベルの後に連続して R ラベルが来るものを分析した．その結果，2 つのラベルが連続したものは 121 件あり，CRF は 53 件に対して失敗しており，n-CRF は 71 件失敗していた．2 段階にすることで T ラベルのみの予測は減ると期待していたが，T ラベルのすぐ後に R ラベルが来るようなものは，CRF の方が強いことがわかった．この結果から，R ラベルを当てる場合にも T ラベルがどれか（どれがなり得るか）の情報は必要であると考える．

4　おわりに

本研究では，任意の指定した対象に対する Twitter からの評判分析を可能とするために，クラウドソーシングを利用しての評価対象-評価表現データセットの作成を行なった．作成したデータセットに対し実験を行い，本データセットの分析を行った．ベンチマークの実験結果を分析することで，データセット作成手順とモデルの構築に課題が見えた．今後の課題としては，実際に精度よく評価表現と評価対象を抽出可能なモデルの構築と，データセット作成手順の洗練が考えられる．

(1)	gold	犬吠駅前で Wi-Fi が繋がることに正直驚いている ww
	CRF	犬吠駅前で Wi-Fi が繋がることに正直驚いている ww
	n-CRF/NN-CRF	犬吠駅前で Wi-Fi が繋がることに正直驚いている ww
(2)	gold	柏駅で火事です火事ですってサイレンなってて焦った
	CRF/NN-CRF	柏駅で火事です火事ですってサイレンなってて焦った
	n-CRF	柏駅で火事です火事ですってサイレンなってて焦った
(3)	gold	東京駅混んでる... 新大阪行き満席か... 連休初日?!だからか
	CRF/NN-CRF	東京駅混んでる... 新大阪行き満席か... 連休初日?!だからか
	n-CRF	東京駅混んでる... 新大阪行き満席か... 連休初日?!だからか

表 3　gold ラベルとモデルの出力ラベルの分析．
（赤：T ラベル，青：R ラベル）

謝辞　本研究は JST CREST(課題番号: JPMJCR1301)，および JSPS 科研費 15H01702 の支援を受けたものである．

参考文献

[1] J. Blitzer, J. Blitzer, M. Dredze, M. Dredze, F. Pereira, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. *Annual Meeting-Association for Computational Linguistics*, Vol. 45, No. 1, p. 440, 2007.

[2] Kazuaki Hanawa, Akira Sasaki, Naoaki Okazaki, and Kentaro Inui. A Crowdsourcing Approach for Annotating Causal Relation Instances in Wikipedia. *The 31st Pacific Asia Conference on Language, Information and Computation PACLIC 31*, 2016.

[3] Koichiro Yoshino Satoshi Nakamura. Ikuo Keshi, Yu Suzuki. Semantically readable distributed representation learning for social media mining. Proceedings of the International Conference on Web Intelligence (WI ' 17), p. 716722, 2017.

[4] Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi, and Toshikazu Fukushima. Collecting Evaluative Expressions for Opinion Extraction. *Journal of Natural Language Processing*, Vol. 12, No. 3, pp. 203–222, 2005.

[5] Xuezhe Ma and Eduard Hovy. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1064–1074, 2016.

[6] An T Nguyen, Byron C Wallace, Junyi Jessy Li, Ani Nenkova, and Matthew Lease. Aggregating and Predicting Sequence Labels from Crowd Annotations. *Proceedings of the conference. Association for Computational Linguistics. Meeting*, p. 299, 2017.

[7] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *In ACL*, No. June, pp. 115–124, 2005.

[8] Sara Rosenthal, Noura Farra, and Preslav Nakov. SemEval-2017 Task 4: Sentiment Analysis in Twitter. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 502–518, 2017.

[9] Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. *Advances in Neural Information Processing Systems*, Vol. 22, No. 1, pp. 1–9, 2009.

[10] Cardie Wiebe, Janyce. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, Vol. 39, No. 2, pp. 165–210, 2005.