DEIM Forum 2019 P2-118

擬似アノテーションにもとづく日本語ツイートの極性判定

小橋　賢介†　　酒井　哲也†

† 早稲田大学基幹理工情報理工学科　〒 169–8555 東京都新宿区大久保 3-4-1
E-mail: †lelouch-ken@ruri.waseda.jp, ††tetsuyasakai@acm.org

あらまし　Twitter を対象とした感情推定は，企業や商品の評判分析や，人的トラブルの検出・防止など，実用的価値が高い．しかし，機械学習による感情推定は大量の学習データを必要とする．そこで本研究では，機械学習によりツイートの極性判定を行う際に，人手によるアノテーションの代わりに感情語辞書に基づく擬似アノテーションを利用する方法を検討する．結果は，SVM を用いた時，人手によるアノテーションを用いた場合の精度と比較して，擬似アノテーションを用いた場合の精度が大きく下回る結果となり．擬似アノテーションは人手によるアノテーションに代わることは実用的ではないと判断した．実用的にするには，擬似アノテーションの精度を向上させる必要がある．

キーワード　Twitter, 感情語辞書, 感情推定, 擬似アノテーション

1 はじめに

ブログ，ホームページ，SNS (Social Networking Service) と情報発信の手段は増加してきている．特に SNS はスマートフォンの普及に伴い，若年層を中心に情報の発信・拡散やコミュニケーションをとるためのツールとして広く使われている．

SNS の中でも Twitter[1]は，ユーザがツイートと呼ばれる 140 文字以内の短文を投稿するという特徴がある．その為，ユーザが気軽に情報発信が可能になり，ユーザは今起こったことや日常で感じたこと，体験したことを投稿している場合が多く，ユーザの感情が現れやすい．感情が表れやすい Twitter を対象とした感情推定は，企業や商品の評判分析や，人的トラブルの検出・防止など，実用的価値が高い．そのため，ツイートの感情推定を行う研究が数多く行われている．

しかし，ツイートに含まれる単語がツイッター特有の表現や省略して文章を短くすることによる文法の乱れ，ツイート内の単語が感情語辞書に登録されていないことにより高い精度を出すことができない．

さらに，トレーニングデータが少ないことで高い精度が出すことができないと考える．トレーニングデータが少なくなるのはポジネガのラベル付けをアノテータが行っていることが原因である．人手で正解データを作成するとコストがかかり，トレーニングデータが必然的に少なくなる．

本研究では，人手による正解データの作成するコストをできるだけ少なくし，トレーニングデータを作成することを目的としている．そこで，トレーニングデータを人手によるアノテーションの代わりに感情語辞書に基づく擬似アノテーションを行う．感情語辞書には高村ら [1] の単語感情極性対応表を使用した．ツイートの形態素解析を行い，形態素ごとに単語感情極性対応表と一致する単語がある場合，スコア付けを行うことで，ツイート全体の感情スコアを算出しポジティブ・ネガティブのラベルづけを行う．人手によるアノテーションから作成したトレーニングデータを用いて機械学習を行う場合と精度を比較して，擬似アノテーションの有効性を検証する．

以降，第 2 章では関連研究を挙げ，第 3 章では提案手法の詳細を説明する．第 4 章では提案手法を実データを用いて実験を行い，結果について報告し，第 5 章で考察したあと，第 6 章で結論を報告する．最後に第 7 章で今後の展望について述べる．

1 : http://twitter.com/

2 関連研究

本節では，本研究に関連する研究について以下に示す．

2.1 感情語辞書

感情語辞書とは，感情を表す単語ごとに感情極性値が付与された辞書である．

感情語辞書には東山ら [2] の日本語評価極性辞書 (名詞編) と小林ら [3] の日本語評価極性辞書 (用言編)，高村ら [1] の単語感情極性対応表がある．

東山らの日本語評価極性辞書 (名詞編) は約 8,500 語の名詞，小林らの日本語評価極性辞書 (用言編) は約 5,000 語の用言でそれぞれ構成されている．東山らと小林らはともに人手による極性情報を付与して辞書が作成された．辞書には単語に対してポジティブ・ネガティブで極性分類し，主観・客観のタグが付与されている．

それに対して，高村らの単語感情極性対応表は語釈文，シソーラス，コーパスによって構築された語彙ネットワークを用いて算出された −1 から +1 までの感情極性値が単語ごとにスコア付けされた．単語感情極性対応表には約 55,000 単語の名詞 (49,000 単語)，動詞 (4,256 単語)，副詞 (1,207 単語)，形容詞 (665 単語) が含まれる．

日本語評価極性辞書と比較して，単語感情極性対応表は単語数が多い．さらに，日本語評価極性辞書はポジティブ・ネガティブの 2 値のみで構成されているのに対し，単語感情極性対応表は-1 から+1 までの感情極性値が算出されるので，ポジ

ティブ・ニュートラル・ネガティブを判定しやすい．なので，本研究では高村らの単語感情極性対応表を使用した．

2.2 日本語ツイートに対する感情推定

日本語のツイートに対して感情分析を行う研究は数多く行われている．堀宮ら [4] は人間の人間に対する推測能力に着目し，ユーザの発言に対する反応であるリプライ機能を利用して感情推定を行っている．TF-IDF (Term Frequency-Inverse Document Frequency) 法を用いて 1,800 個のツイート-リプライセットの特徴ベクトルを生成し，SVM (Support Vector Machine) を用いて学習を行った．松林ら [5] は Word2Vec [6] を用いてツイートのテキストから特徴ベクトルを生成し，ランダムフォレストを用いた感情の分類器を構築することで「喜・怒・哀・楽・無感情」の 5 つの感情推定を行った．池上ら [7] は顔文字の感情極性とツイートのテキスト部分の感情極性の組み合わせが，ツイート全体の観光極性に与える効果を分析し，顔文字を考慮してツイートの感情極性を推定する手法を提案した．山本ら [8] らは，顔文字や特有の言い回し (「やったあああああ」のように「やった」という喜びを強調する言い回し) で感情が強調，弛緩，転換することでツイート全体の感情が変化すると考え，それらを考慮した感情抽出手法を提案した．感情推定のために定量化された感情語辞書と顔文字辞書も作成した．

3 提案手法

感情推定の学習に使う正解データとして，擬似アノテーションにより生成したラベルを使う手法を提案する．感情を推定するために，ツイートの特徴ベクトル生成を行う．機械学習のモデルには SVM とランダムフォレストを用い, 推定する感情は「positive」「neutral」「negative」の 3 クラスである．2 値分類する機械学習のモデルを使ったので，positive, neutral, negative それぞれを正解とするモデルを 1 つの手法に対して 3 つ作っている．

本研究では，以下に述べる 2 つの擬似アノテーション手法を提案する．

3.1 辞書平均スコア付け

以下に，擬似アノテーション (辞書平均スコア付け) の流れを示す．

（1） ツイートを形態素解析ツール MeCab[2] で形態素ごとに品詞・読み方・漢字を出力

（2） 感情語辞書と一致する単語に対して感情スコア付け

（3） (2) を形態素ごとに繰り返し，ツイート全体の感情スコアを算出

（4） 定めた閾値に従い，感情スコアから感情を決定

まず，(1) でツイートからの単語抽出には形態素解析ツールである MeCab を使用する．MeCab を用いることで品詞・読み方・漢字が出力される．

次に，(2) では感情語辞書と一致する単語に対して感情スコア付けを行う．本研究では，高村らの単語感情極性対応表を使用した．この単語感情極性対応表の単語と形態素が一致した場合，その形態素にスコア付けを行う．一致しなかった場合，スコアには ‘not found’ と定義される．

(3) では，(2) の操作を繰り返し，スコアの合計を単語感情極性対応表の単語と形態素が一致した単語数で割ることでツイート全体のスコアを算出する．単語全てが not found と出力される場合，ツイート全体の感情スコアは 0 となる．図 3.1 にスコア算出の様子を示す．図 3.1 よりツイート「結論，どっちにしろ呪怨は怖い.」が「結論」,「する」,「呪」,「怖い」の 4 単語が単語感情極性対応表と一致してそれらの単語がスコア付けされ，4 単語のスコアの合計をその単語数で割った値が出力された．

(4) では，定めた閾値に従い，算出されたスコアから感情を決定する．以下にスコアの閾値を示す．

$$0.2 < score \leqq 1.0 : positive$$

$$-0.2 \leqq score \leqq 0.2 : neutral$$

$$-1.0 \leqq score < -0.2 : negative$$

3.2 辞書平均ラベル付け:閾値変更

本節の手法は，3.1 節で述べた手法から閾値を変更を行う．ここで，少量のツイートデータの positive，negative，neutral の比率がツイートデータ全体の比率と類似すると仮定して手法を提案する．

以下に，擬似アノテーション (辞書平均ラベル付け:閾値変更) の流れを示す．

（1） 3.1 節の手法を用いてツイートに感情スコア付け

（2） 少量のツイートデータを人手でアノテーション

（3） positive:neutral:negative の比率を求める

（4） (3) 得られた比率を用いて閾値を変更

（5） (4) で定めた閾値に従って，感情を決定

まず，3.1 節で述べた手法と同様にツイートに感情スコア付けを行う．次に，少量のツイートデータを人手でアノテーションを行う．アノテーションされたデータから positive，neutral，negative の比率を求める．そして，(3) で得られた positive，neutral，negative の比率を用いて，閾値を変更する．例えば，$positive : neutral : negative = 40 : 25 : 35$ の場合，感情スコアの高い順番にソートし，上位 40% を positive と判定，positive 以外上位 25% を neutral と判定，positive と neutral と判定されるツイート以外を negative と判定するように閾値を変更する．

4 実験と結果

ここでは, 第 3 章の提案手法を基に実験を行った．データセットと比較手法，評価方法，得られた結果を以下に示す．本研究では，Word2Vec [6] で事前学習された単語埋め込みベクトルを利用した．Word2Vec[3] の学習にはツイートデータを使用した．

2 : http://taku910.github.io/mecab/

3 : https://github.com/sugiyamath/word2vec_japanese_twitter

```
文章:結論、どっちにしろ呪怨は怖い。
{'POS1': '名詞',     'BaseForm': '結論',  'PN': -0.583387}
{'POS1': '記号',     'BaseForm': '、',    'PN': 'notfound'}
{'POS1': '名詞',     'BaseForm': 'どっち','PN': 'notfound'}
{'POS1': '助詞',     'BaseForm': 'に',    'PN': 'notfound'}
{'POS1': '動詞',     'BaseForm': 'する',  'PN': -0.602913}
{'POS1': '名詞',     'BaseForm': '呪',    'PN': -0.9936290000000001}
{'POS1': '名詞',     'BaseForm': '*',     'PN': 'notfound'}
{'POS1': '助詞',     'BaseForm': 'は',    'PN': 'notfound'}
{'POS1': '形容詞',   'BaseForm': '怖い',  'PN': -0.997999}
{'POS1': '記号',     'BaseForm': '。',    'PN': 'notfound'}
感情スコア:-0.7944820000000001
```

図 1　スコア算出の様子

4.1　データセット

本研究では，2014 年 1 月 1 日から 2014 年 12 月 31 日のツイート 2,270 件を取得した．ツイートデータは人手によりアノテーションを行った．アノテーションは 20 代男性と 20 代女性の 2 人で行い，ツイートを positive，neutral，negative に分類した．ツイートに対し，2 人の分類結果が異なっていた場合は，以下のように処理した。

- (positive，negative) の場合:ツイートをデータセットから削除
- (positive，neutral) の場合:positive
- (negative，neutral) の場合:negative

処理を行った結果，ツイートデータは 2,136 件となった．データセットの分布を表 1 に示す．

表 1　データセットの分布

データセット	サイズ
positive	1,153
neutral	345
negative	638

ツイートをトレーニング，バリデーション，テストセットに分割した．トレーニングセットはモデルの学習に用いる．バリデーションセットはパラメータを最適化するために用いる．テストセットはモデルの評価のために用いる．

各データセットのサイズを表 2 に示す．

表 2　データセットとサイズ

データセット	サイズ
トレーニングセット	1,282
バリデーションセット	320
テストセット	534

4.2　比 較 手 法

比較手法として，人手によるアノテーションを用いた手法を挙げる．

4.3　人手によるアノテーションを用いた機械学習

人手によるアノテーションを用いた手法について説明する．4.1 節で述べたトレーニングデータを用いて学習器に学習させる．その際，正解データとして，人手により付加されたアノテーションを用いる．

4.4　ランダムに推定した感情ラベルを用いた機械学習

3.2 節で述べた 3 感情の比率になるように，感情のラベル付けをランダムに行う．

4.5　評 価 方 法

4.1 節で述べたテストデータを用いて以下に述べる評価指標を各手法について計算する．評価指標として，Precision，Recall，Accuracy，F-measure，Cohen's Kappa [9] を用いた．「positive」，「neutral」，「negative」の各感情について，ツイートがその感情であるか (P) とその感情でない (N) かの 2 クラスで判定する．そのために Confusion Matrix はその感情である (P) とその感情でない (N) という軸と予測した感情と実際の感情が正しかった (T) か間違っていた (F) かという軸を用いる．本研究における Confusion Matrix を表 3 に示す．

表 3　本研究における Confusion Matrix

		予測した感情 N	予測した感情 P
実際の感情	N	TN	FP
	P	FN	TP

Confusion Matrix 内の 4 つの事象 (TP，FP，TN，FN) を以下に定義する．

3 クラスの感情のうち 1 つの感情について予測した場合，

TP (True Positive) はその感情であると予測し，正解データの感情が一致した場合の数を表している．

FP (False Positive) はその感情であると予測し，正解データの感情が一致しない場合の数を表している．

TN (True Negative) はその感情でないと予測し，正解データの感情が一致した場合の数を表している．

FN (False Negative) はその感情でないと予測し，正解データの感情が一致しない場合の数を表している．

それぞれの評価指標の式は以下に示す．

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F\text{-}measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

F-measure は Precision と Recall の調和平均をとった値である．

次に，Cohen's Kappa の計算方法について説明する．全データ N に対して正解ラベルが第 i カテゴリ ($i = 1$ のとき N，$i = 2$ のとき P) に分類され，かつ予測ラベルが第 j カテゴリ ($j = 1$ のとき N，$j = 2$ のとき P) に分類に分類されたときのアイテムの件数を O_{ij} とする。Confusion Matrix と比較すると，O_{11} は TN，O_{12} は FP，O_{21} は FN，O_{22} は TP である．正解ラベルが第 i カテゴリに分類したアイテムの件数を $O_{i\bullet}$，予測ラベルが第 j カテゴリに分類したアイテムの件数を $O_{i\bullet}$ と表す．その際に，正解ラベルと予測ラベルが独立して分類された仮想的なデータは以下のような式で表す．

$$C_{i_j} = \frac{O_{i\bullet} O_{\bullet j}}{N} \tag{4}$$

正解ラベルと予測ラベルの一致度 P_0，独立して分類された仮想的なデータの下での期待値から求めた一般値 P_c はそれぞれ以下のように表される．

$$P_0 = \frac{\sum_{i=1}^{2} O_{ii}}{N} \tag{5}$$

$$P_c = \frac{\sum_{i=1}^{2} O_{ii}}{N} \tag{6}$$

以上の式を用いて，偶然以上の一致を表す Cohen の κ は以下のような式で表せる．一致度 κ の範囲は $[-1, 1]$ である．

$$\kappa = \frac{P_0 - P_c}{1 - P_c} = \frac{\sum_{i=1}^{2} O_{ii} - \sum_{i=1}^{2} C_{ii}}{N - \sum_{i=1}^{2} C_{ii}} \tag{7}$$

4.6 評　　価

以下の流れで評価を行う．

（1） 提案手法の擬似アノテーションの精度を比較

（2） 比較手法と提案手法の精度を比較

ともに，4.5 節で述べた方法で評価する．

4.6.1 擬似アノテーションの精度

第 3 章で述べた擬似アノテーション 2 つの手法を行い，極性判定を行う．

3.2 節の辞書平均スコア付け:閾値変更では，トレーニングデータの中から 100 件のツイートデータを用いて人手のアノテーションを参照した．positive，neutral，negative の比率は以下のようになった．

$$positive : neutral : negative = 59 : 13 : 28$$

ツイートの比率を用いて閾値の変更を行った．変更した閾値を以下に示す．

$$-0.297 < score \leqq 1.0 : positive$$

$$-0.448 \leqq score \leqq -0.297 : neutral$$

$$-1.000 \leqq score < -0.448 : negative$$

擬似アノテーションの評価結果を結果を表 4 に示す．また，Cohen's Kappa において各感情の最も高い値を下線で示す．

Cohen's Kappa において，「辞書平均スコア付け:閾値変更」は 3 感情のうち 2 感情は精度が高い．したがって，「辞書平均スコア付け:閾値変更」が 「辞書平均スコア付け」より，わずかに優れている結果となった．閾値変更を行うことで，擬似アノテーションの精度が向上していることが確認できた．

4.6.2 提案手法と比較手法の精度の比較

提案手法で述べた 2 つの手法とと比較手法で述べた 2 つの手法の精度の比較を行う．「辞書平均ラベル付け:閾値変更」の設定は 4.6.1 節に述べた方法を用いた．そして，「ランダムに推定した感情ラベルを用いた機械学習」では，4.6.1 節で述べた比率を用いた．

学習器に SVM を用いた場合を表 5，学習器にランダムフォレストを用いた場合を表 6 に示す．また，Cohen's Kappa において各感情の最も高い値を下線で示す．

学習器に SVM を用いた場合の結果を述べる．Cohen's Kappa において，「人手によるアノテーション」は 3 感情全てでの精度が最も高い．「人手によるアノテーション」の精度は，他の 3 つの手法の精度と大きく差がある．

他の 3 つの手法のみで比較を行った場合，Cohen's Kappa において 「辞書ラベル付け:閾値変更」が 3 感情のうち 2 感情の精度が最も高い．したがって，「人手によるアノテーション」の結果に一番近い結果が得られた手法は 「辞書ラベル付け:閾値変更」であることがわかる．

次に，学習器にランダムフォレストを用いた場合の結果を述べる．Cohen's Kappa において 「人手によるアノテーション」は 3 感情のうち 1 感情の精度が最も高い．表 5 の SVM の場合と比較してランダムフォレストの場合は，4 つの手法のともほとんど大きな差は見られない．neutral について着目すると，4 つの手法とも各評価指標において同じ結果が得られている．他の 3 つの手法のみで比較を行った場合，Cohen's Kappa において「辞書ラベル付け:閾値変更」が 3 感情のうち 1 感情の精度が最も高く，F-measure において 3 感情のうち 2 感情の精度が最も高い．したがって，「人手によるアノテーション」の結果に一番近い結果が得られた手法は 「辞書ラベル付け:閾値変更」であることがわかる．しかし，先ほど述べたようにどの手法も精度の違いがあまり見られない．

表 4　擬似アノテーションの精度

手法	感情	Precision	Recall	Accuracy	F-measure	Cohen's Kappa
擬似アノテーションのみ	positive	0.892	0.120	0.537	0.211	0.101
辞書平均スコア付け	neutral	0.257	0.200	0.755	0.225	0.082
	negative	0.357	0.926	0.515	0.515	0.134
擬似アノテーションのみ	positive	0.685	0.308	0.569	0.425	0.154
辞書平均スコア付け:閾値変更	neutral	0.159	0.116	0.734	0.134	−0.018
	negative	0.375	0.785	0.536	0.508	0.162

表 5　提案手法と比較手法の精度 (SVM)

手法	感情	Precision	Recall	Accuracy	F-measure	Cohen's Kappa
人手によるアノテーション	positive	0.737	0.732	0.727	0.735	0.453
を用いた機械学習 (SVM)	neutral	0.541	0.211	0.828	0.303	0.226
	negative	0.646	0.190	0.721	0.294	0.180
ランダムに推定した感情ラベル	positive	0.517	1.000	0.517	0.681	0.000
を用いた機械学習 (SVM)	neutral	0.172	0.116	0.743	0.138	−0.006
	negative	0.329	0.319	0.594	0.324	0.034
擬似アノテーション	positive	1.000	0.011	0.489	0.022	0.011
を用いた機械学習 (SVM)	neutral	0.600	0.032	0.824	0.060	0.043
辞書平均スコア付け	negative	0.347	0.914	0.449	0.503	0.109
擬似アノテーション	positive	0.649	0.362	0.569	0.465	0.151
を用いた機械学習 (SVM)	neutral	0.148	0.084	0.751	0.107	−0.025
辞書平均ラベル付け:閾値変更	negative	0.368	0.865	0.506	0.516	0.154

表 6　提案手法と比較手法の精度 (ランダムフォレスト)

手法	感情	Precision	Recall	Accuracy	F-measure	Cohen's Kappa
人手によるアノテーション	positive	0.579	0.866	0.605	0.694	0.195
を用いた機械学習	neutral	−	0.000	0.822	0.000	0.000
	negative	0.484	0.190	0.691	0.273	0.122
ランダムに推定した感情ラベル	positive	0.516	0.996	0.515	0.680	−0.004
を用いた機械学習	neutral	−	0.000	0.822	0.000	0.000
	negative	0.500	0.006	0.695	0.012	0.005
擬似アノテーション	positive	1.000	0.004	0.485	0.007	0.004
を用いた機械学習	neutral	0.000	0.000	0.820	0.000	−0.004
辞書平均スコア付け	negative	0.321	0.994	0.356	0.485	0.044
擬似アノテーション	positive	0.776	0.138	0.534	0.234	0.092
を用いた機械学習	neutral	−	0.000	0.822	0.000	0.000
辞書平均ラベル付け:閾値変更	negative	0.362	0.969	0.468	0.527	0.148

表 7　擬似アノテーション 2 つの手法対の p 値 (一致度)

システム対	positive	neutral	negative
辞書平均スコア付け・辞書平均:閾値変更: (1)	$p = 0.5483$	$p = 0.7175$	$p = 0.0678$

表 8　SVM を学習器とした場合の各手法対対の p 値 (一致度)

システム対	positive	neutral	negative
人手によるアノテーション・辞書平均スコア付け: (1)	$p < 0.01$	$p = 0.9987$	$p < 0.01$
人手によるアノテーション・ランダムに推定した感情ラベル: (2)	$p < 0.01$	$p = 0.0042$	$p = 0.0001$
人手によるアノテーション・辞書平均:閾値変更: (3)	$p < 0.01$	$p = 0.0115$	$p < 0.01$
辞書平均スコア付け・ランダムに推定した感情ラベル: (4)	$p = 0.7809$	$p = 0.0070$	$p < 0.01$
辞書平均スコア付け・辞書平均:閾値変更: (5)	$p = 0.0345$	$p = 0.0183$	$p = 0.2316$
ランダムに推定した感情ラベル・辞書平均:閾値変更: (6)	$p = 0.2918$	$p = 0.9906$	$p = 0.0162$

表 9　ランダムフォレストを学習器とした場合の各手法対対の p 値 (一致度)

システム対	positive	neutral	negative
人手によるアノテーション・辞書平均スコア付け: (1)	$p = 0.0004$	$p = 0.9998$	$p < 0.01$
人手によるアノテーション・ランダムに推定した感情ラベル: (2)	$p = 0.0167$	$p = 1.0000$	$p = 0.9992$
人手によるアノテーション・辞書平均:閾値変更: (3)	$p = 0.0898$	$p = 1.0000$	$p < 0.01$
辞書平均スコア付け・ランダムに推定した感情ラベル: (4)	$p = 0.7583$	$p = 0.9998$	$p < 0.01$
辞書平均スコア付け・辞書平均:閾値変更: (5)	$p = 0.3789$	$p = 0.9998$	$p = 0.0006$
ランダムに推定した感情ラベル・辞書平均:閾値変更: (6)	$p = 0.9272$	$p = 1.0000$	$p < 0.01$

4.7　統計的検定

本研究では，統計的有意性を確かめるために，Tukey HSD 検定を行った．

各ツイートに対して，システムの評価は以下のように行った．

- システムから出力された感情と正解データが一致:1
- システムから出力された感情と正解データが不一致:0

以上に示す評価指標を 「一致度」とする．この評価指標を用いて検定を行った．

各システム対の差に関する p 値を算出した．p 値はシステムが同一のものだと仮定した場合，そのシステム間の差データが得られる確率を示している．p 値が 0.05 以下の場合は有意水準 5% においてシステム間の結果が有意であるとする．p 値が 0.05 以下の場合は本節に示す表で下線で示す．

まず，4.6.1 節で用いた擬似アノテーション 2 つの手法対の差に関する p 値を算出し，表 7 に示す．

表 7 について論ずる．システム対 (1) から，いずれの感情においても有意な差が得られないため，「辞書平均:閾値変更」の精度が最も優れている結果は統計的に有意であるとはいえない．

次に，表 5 で SVM を用いた時の，提案手法と比較手法の精度を求めたときに用いた各システム対の差に関する p 値を算出し，表 8 に示す．

表 8 について論ずる．評価の結果では，各手法の中では 「人手によるアノテーション」が最も優れていると述べた．その結果が統計的に有意であることを示すために，その手法ともう一つの手法のシステム対に着目する．システム対 (1) では，neutral 以外において有意な差が得られたため，neutral 以外の感情において，「辞書平均スコア付け」より 「人手によるアノテーション」の精度が優れていることは統計的に有意である．システム対 (2) と (3) では，全ての感情において有意な差が得られたため，「ランダムに推定した感情ラベル」や 「辞書平均:閾値変更」より 「人手によるアノテーション」は精度が優れていることは統計的に有意である．つまり，neutral 以外では，人手によるアノテーションが最も優れていることは統計的に有意である．さらに，結果では他の 3 つの手法では「辞書平均:閾値変更」が優れていると述べた．その結果が統計的に有意であることを示すために，その手法ともう一つの手法のシステム対に着目する．システム対 (5) では，negative 以外において有意な差が得られた．したがって，negative 以外において「辞書平均スコア付け」より「辞書平均:閾値変更」の精度の方が優れていることは統計的に有意である．システム対 (6) では，nagative のみ有意な差が得られた．したがって，negative において「ランダムに推定した感情ラベル」より「辞書平均:閾値変更」の精度の方が優れていることは統計的に有意である．

そして，表 6 でランダムフォレストを用いた時の，提案手法と比較手法の精度を求めたときに用いた各システム対の差に関する p 値を算出し，表 9 に示す．

表 9 について論ずる．評価の結果では，SVM を用いた場合と比較して精度の差は小さいが，「人手によるアノテーション」は最も優れていると述べた．その結果が統計的に有意であることを示すために，その手法ともう一つの手法のシステム対に着目する．先ほど，表 8 について述べたときと同様の方法で有意差を示す．その方法を用いると，人手によるアノテーションは他の手法と比較して最も優れているという結果は，統計的に有意ではない．さらに，結果では他の 3 つの手法では 「辞書平均:閾値変更」が最も優れていると述べた．その結果が統計的に有意であることを示すために，その手法ともう一つの手法のシステム対に着目する．システム対 (5) と (6) ともに negative の時のみ，提案手法の中で 「辞書平均:閾値変更」が最も優れていることは統計的に有意である．しかし，negative 以外の感

情はシステム対 (5) と (6) ともに有意的な差が得られなかったので，今回の評価からは結論が得られなかった．

5 考　　察

SVM を用いた場合では，「人手によるアノテーション」の精度が最も優れており，他の 3 つの手法の精度と大きく差をつける結果となった．つまり，人手によるアノテーションの代わりに擬似アノテーションを用いて機械学習を行うことは適切ではないことがわかった．そこで，擬似アノテーションを用いた機械学習の精度が低くなった原因の考察を行った．

5.1 感情語辞書にない単語

ツイートの形態素全てにおいて感情語辞書の単語と一致しなかった場合，辞書平均スコアが 0 になる．本研究で用いたツイート 2136 件に対して辞書平均スコア付けを行い，スコアが 0 であるツイートのサイズを以下の表 10 に示す．

表 10　辞書平均スコア 0 のツイート

実際の感情	サイズ
positive	67
neutral	26
negative	49
合計	142

表 10 より辞書平均スコア 0 のツイートは全体の 6.6% 存在する．「辞書平均スコア付け」の場合，辞書平均スコアが 0 のツイートは positive や negative のツイートが誤って neutral と判定される．さらに，「辞書平均:閾値変更」の場合，辞書平均スコアが 0 のツイートは neutral や negative のツイートが誤って positive と判定される．これが精度が低くなった原因の 1 つである．この問題を解決するためには，トレーニングセットから辞書平均スコアが 0 のツイートを少量アノテーションし，positive，neutral，negative の比率を求め，辞書平均スコア 0 のツイートをその比率になるように感情をランダムに設定することで改善できると考えられる．

5.2 学習済みの Word2Vec に存在しない単語

機械学習のトレーニングセットの前処理の際に，ツイートのベクトル化を行う．ツイートの単語が Word2Vec に含まれていない場合，未知語のベクトルが与えられる．

表 11 に，単語ベクトルが与えられなかったツイート例を示す．

これらのツイートは同じベクトルとして扱われ，学習器が同じ感情のツイートとして誤って学習する．これも精度が低くなった原因の一つであると考えられる．

5.3 副詞，否定語の処理

本研究では，副詞や否定語の処理は行っていない．それにより，誤った感情がラベル付けされる．これが精度が低くなった原因の一つであると考える．

まず，副詞について誤ったラベル付けを示す．図 2 のスコアは，「辞書平均スコア付け」の場合，neutral と誤って判定される．さらに，「辞書平均:閾値変更」の場合，誤って positive と判定される．そこで，副詞の後に続く形容詞を強調するように，形容詞と同じスコアを 「とても」に与えることで，「とても」という単語は後に続く形容詞・動詞を強調する表現として使うことができる．例えば，図 2 のスコアは 0.26084075 となり，positive と判定された．この方法で，擬似アノテーションの精度を改善できると考えられる．

表 11　単語ベクトルが与えられなかったツイート例

ツイート	辞書平均スコア	実際の感情
ダウト！	0.0	negative
イカチェ	0.0	negative
トゥルトゥルトゥルトゥル	0.0	neutral
スゴイノ？	0.0	positive
おおきに	0.0	positive
アスパラベーコン	−0.460286	neutral
$\kappa\alpha\lambda\eta\nu\upsilon\chi\tau$	0.0	neutral
ユアーウェルコメ	0.0	positive
レゴブロック	0.0	neutral
タービュランス	0.0	neutral
凰壮 ！！！！！！！！！！！	0.0	neutral
なでなで	−0.744753	positive
オメデトウゴザイマス！！	0.0	positive

次に，否定語について誤ったラベル付けを示す．「嬉しくないわ！」というツイートは一般的に negative な感情である．しかし，表 3 のスコアは「辞書平均スコア付け」の場合，neutral と誤って判定される．さらに，「辞書平均:閾値変更」の場合，誤って positive と判定される．この問題を解決するために，動詞の後に否定語 「ない」がある場合，その動詞の感情極性値の正負を逆転させる．例えば，「嬉しくないわ！」のスコアは −0.998871 になり，negative と判定される．つまり，動詞の感情極性値の正負を逆転させることで，精度を改善できると考えられる．

6 結　　論

本研究では，機械学習によりツイートの極性判定を行う際に，人手によるアノテーションの代わりに感情極性辞書に基づく擬似アノテーションを利用する方法を提案し，評価を行った．結果は，SVM を用いた時，人手によるアノテーションを用いた場合と比較して，精度が大きく下回る結果となり，人手によるアノテーションの代わりに擬似アノテーションを用いて機械学習を行うことは適切ではないことがわかった．実用的にするには，擬似アノテーションの精度を向上させる必要がある．

7 今後の展望

今後の展望としては，第 5 章で述べたように感情語にない単語，副詞や否定語の処理を適切に行い，擬似アノテーションの精度をあげることである．

さらに，統計的検定を行うため，Tukey-HSD 検定を用いて各システムの任意の 2 値の p 値を算出したが，感情によって

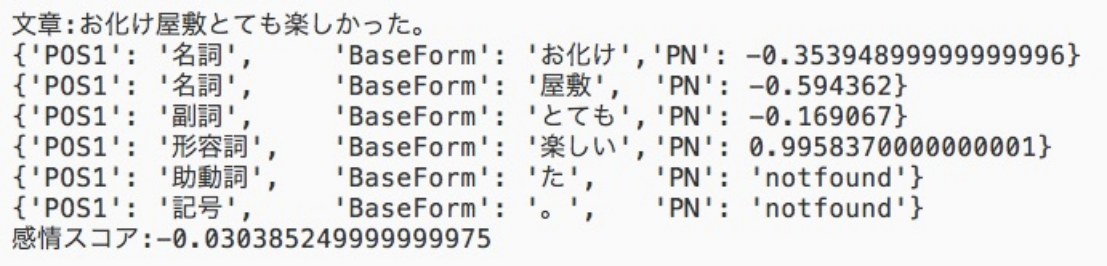

図 2　ツイートに副詞 (とても) がある場合のスコア付けの様子

図 3　ツイートに否定語 (ない) がある場合のスコア付けの様子

は，有意差が得られないものがあった．例えば， 一部感情については有意差がないものが存在した．したがって，さらにデータが必要となるだろう．検定結果は「一部感情の有意差が示せるが，一部感情の有意差がない」といった検定結果がほとんどである．それは，positive，neutral，negative それぞれを正解とする 2 値分類のモデルを使ったからである．その問題の改善点として，3 値分類を同時に行えるような学習器を用いることで評価や検定を簡潔化することができるだろう．

文　　献

[1] 高村大也, 乾考司, 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 Vol.47 No.2, pp.627-637, 2006.
[2] 東山昌彦, 乾健太郎, 松本裕治,”述語の選択選好性に着目した名詞評価極性の獲得”, 言語処理学会第 14 回年次大会論文集, pp.584-587, 2008.
[3] 小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一, ”意見抽出のための評価表現の収集”, 自然言語処理，Vol.12, No.3, pp.203-222, 2005.
[4] 堀宮ありさ, 板野遼平, 佐藤晴彦, 小山聡, 栗原正仁, 沼澤 政信, “Twitter における発話者へのリプライを用いたユーザ感情推定手法,” DEIM Forum 2012 F2-1, 2012.
[5] 松林圭, 五味京祐, 古川和折, 松尾祐佳, 松原良和, 日諸 マルセロ優次, 中村拓哉, 山下晃弘, 松林勝志, “Twitter 上に投稿された文章に基づく感情推定手法とその応用に関する検討,” 情報処理学会第 78 回全国大会, 2016.
[6] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, ”Efficient estimation of word representations in vector space”, Workshop at International Conference on Learning Representations (ICLR), 2013.
[7] 池川知里, 新妻弘崇, 大田学,”顔文字の役割を利用したツイートの感情極性推定,” DEIM Forum 2014 E6-4, 2014
[8] 山本湧輝, 熊本忠彦, 灘本明代,”Twitter 特有表現を考慮したツイートの多次元感情抽出手法の提案”, 2014 年度情 報処理学会関西支部 支部大会, 2014.
[9] 酒井哲也, ”情報アクセス評価方法論 検索エンジンの進歩のために”, pp.198-200. コロナ社, 2015.