Word Mover's Distanceを用いたコーパス拡張による感情推定精度向上の検討

藤野 尚也　松本 和幸　吉田 稔　北 研二
徳島大学

{matumoto;mino;kita}@is.tokushima-u.ac.jp

1 はじめに

近年，深層学習をはじめとする機械学習はテキストマイニングや画像認識などの様々な研究分野で利用されている．今後の技術発展において必要不可欠であり，莫大な費用を投じている企業も少なくない．また，感情推定技術は人間とコンピュータ同士の円滑なコミュニケーションの実現に重要な役割を果たす．そういった中，よく問題点として挙げられるのが訓練データ量の不足や質の低下である．データ量の不足，質の低下は機械学習そのものの質の低下に直結する．本研究では，代表的なSNSの一つであるTwitterに投稿されるテキストである「ツイート文」に着目し，そこに投稿されるツイート文を感情推定の対象とする．また，コーパス拡張が感情推定モデルの精度にどのような影響を及ぼすかを分析し考察する．

2 関連研究

藤野らの研究 [1] では，ユーザの属性が感情表現に特徴として現れるという仮定から，感情推定モデルを属性別に作成する手法の提案を行っている．しかし，訓練データ量の不足・偏りが推定精度にも大きく影響する結果となっている．ただ，いたずらに訓練データ量を増やすだけでは，アノテーションコストの増大を引き起こし，アノテーションの揺れ等を考えると期待した効果が得られるとも限らない．また，藤野らの別の研究 [2] ではデータ量の偏りをなくすために新たなコーパスの拡張も検討されている．しかし，各ツイート文の絵文字のみに着目しているため，文の意味内容を考慮していないラベル付けとなっている．人手でつけた高品質なラベル付けとは異なり，不正解ラベルが多く混入している．同様に，西本らの研究 [3] では，語順並び替えや類似語置換によるコーパス拡張を行っている．しかし，この手法では類似文に対しては効果のある手法と考えられるが，汎用性が低く他の新しい文章に十分対応できるものとはいえない．

本研究では，これらの問題点を解決するために人手でつけた高品質なラベル付きデータを元データとしてWord Mover's Distance(WMD) を用いたコーパス拡張を行う．WMDはKusnerらによって提案された手法 [4] である．最適化問題の一つである輸送問題を求めるアルゴリズムEarth Mover's Distance(EMD)[5] を用いて，単語アライメントを行い文章間の距離を求める手法である．word2vec[6] による分散表現を用いており，対応付けのコストが最も低い場合のコストの総和を距離として算出している．

3 提案手法

本研究では，フィッシャーの感情系統図 [7] に従い，4大感情である「楽しさ」「驚き」「怒り」「悲しみ」を基本感情として採用する．以下に具体的なラベル付けの例を表1に示す．

表 1: ラベル付の例

感情ラベル	ツイート文
楽しさ	無事大会優勝できました！
驚き	締切り忘れてた！ほんとあぶない!
怒り	蚊に起こされた… かゆい…
悲しみ	先生に怒られた… へこむ…

また，元データとして用いる人手でつけた高品質なラベルデータ29,838文をもとに，ラベルの付いてい

ないデータ 1,300,451 文との文章間の距離を WMD を用いて求める．元データとの距離が最も近かったデータの距離の総和を取り，その平均距離を閾値として設定し，閾値より距離の近いツイート文を追加コーパスとして採用する．

拡張したコーパスをもちいて，Starspace[8] による学習を行う．Starspace は，Facebook 社が開発した自然言語処理ツールであり，テキスト分類が可能で，分散表現を高速かつ高精度に学習することができる．

4 実験と考察

4.1 実験方法

提案手法の妥当性を確認するために，評価実験を行う．元データである人手でラベル付けを行ったツイート文，各感情 2,200 文，合計 8,800 文と WMD を用いて作成した追加コーパス，各感情 21,722 文，合計 86,888 文を訓練データとして用いる．また，訓練データとは別に人手でラベル付けを行った各感情 219 文，合計 872 文をテストデータとして用い，適合率，再現率，f 値を算出する．訓練データ，テストデータ共にツイート文中には，感情表出に関連が深いと考えられる絵文字が各ツイート文中に 1 回以上含まれるデータとなっている．

4.2 実験結果

表 2: 訓練データ:元データのみ

感情ラベル	再現率	適合率	F 値
楽しさ	0.456	0.458	0.457
驚き	0.378	0.560	0.452
怒り	0.424	0.436	0.430
悲しみ	0.461	0.340	0.391

表 3: 訓練データ:元データ＋追加コーパス

感情ラベル	再現率	適合率	F 値
楽しさ	0.406	0.360	0.381
驚き	0.296	0.377	0.332
怒り	0.401	0.377	0.389
悲しみ	0.356	0.348	0.352

「元データ」を訓練データとして実験を行った場合，「元データ＋追加コーパス」を訓練データとして実験を行った場合の 2 通りの実験結果を表 2 , 表 3 に示す．

4.3 考察

実験結果から，コーパスの拡張によって全体的に精度の低下をもたらしてしまう結果となった．人手でつけた高品質なラベルが占める割合が少なくなったことが原因として挙げられる．また，今回すべてのラベルにおいてデータ数を統一しているにもかかわらず他の感情に比べて「驚き」の精度が低い結果となった.「驚き」という感情は，他の感情と複合的に用いられることが多いため機械学習において推定することが他の感情に比べて難しいと考えられる．具体例を表に示す．

表 4: 複合ラベルの例

感情ラベル	ツイート文
楽しさ＋驚き	なんとなく応募した懸賞に当たってしまった !!!
怒り＋驚き	まさか，あいつがこの前の一件の犯人だとは．許せない
悲しみ＋驚き	訃報を聞いてびっくり．どうか安らかに．

加えて，WMD によるコーパスの拡張は語順を考慮していない拡張となっている．語順を考慮した拡張方法も検討要素の一つである．課題として,「驚き」を除く他の 3 感情での実験，Starspace 以外の手法との比較や，WMD 以外の手法でのコーパス拡張を比較検証したい．

5 おわりに

本研究では，機械学習を行う際によく問題点として挙げられる訓練データの不足・質の低下を解決するために WMD を用いたコーパスの拡張を検討した．しかし，期待した効果は得られず原因として，人手でつけた高品質なラベルの割合の低下や「驚き」という感情は他の感情と複合的に用いられるため感情推定が比較的難しいのではないかという結論に至った．課題としてコーパス拡張手法の変更，学習手法の変更,「驚き」を除く他の 3 感情での比較検証を行いたい．

謝辞

本研究の一部は，科学研究費補助金 （18K11549, 15K16077 ）の補助を受けて行った.

参考文献

[1] 藤野尚也, 松本和幸, 吉田稔, 北研二. ユーザの性別と感情表出傾向との関連. 第 31 回人工知能全国大会発表論文集, 2017.

[2] Naoya Fujino, Kazuyuki Matsumoto, Minoru Yoshida, Kenji Kita. Emotion estimation adapted to gender of user based on deep neural networks. In *Proceedings of The 12th International Conference on Natural Lnguage Processing and Knowledge Engineering*, 2017.

[3] 西本慎之介. データ拡張による感情分析のアスペクト推定. 2017. 奈良先端科学技術大学院大学修士論文.

[4] Matt J.Kusner, Yu Sun, Nicholas I.Kolkin, Kilian Q.Weinberger. From word embeddings to document distances. 2015. In Proceedings of the 32nd International Conference on Machine Learning.

[5] Yossi Rubner, Carlo Tomasi, Leonidas J.Guibas. The earth mover's distance as a metric for image retrieval. 2000. International Jounal of Computer Vision.

[6] Tomas Mikolov, kai Che, Greg Corrad, Jeffrey Dean. Efficient estimation of woed representations in vector space. 2013. In Proceedings of Workshop at ICLR.

[7] Fisher K.W., shaver P., Carnchan P. *A skill approach to emotional development.* Child development today and tomorrow,pp.107-136, 1989.

[8] Ledell Wu, et al. Embed all the things! 2017.