Vol.2016-NL-227 No.4
2016/7/29

Character を用いた SNS 文章分類

清水隆範[†1]　劉 牧[†1]

概要：近年, SNS は顧客の声を直接的に理解する上で重要なツールとなっている. SNS 分析の中に, Sentiment, Emotion など分類する言語処理技術が必要である．普通的に，日本語の分類はまず形態素分析を行い単語分割し，後は Lexicon それども学習した分類モデルを利用して分類する．本稿では，形態素分析を行わず文章から単語を分離せずに，文章から文字を分離しベクトル化して，文字ベースの文章特徴化である. 評価実験では，日本語, 英語ツイートの Sentiment 分類，中国語 Weibo データの subjective と objective 分類，日本語文章歪み耐性実験を行い，文字ベースの有効性を示し，その適用が期待できる．

Character Based Social Media Text Classification

TAKANORI SHIMIZU[†1]　MU LIU[†1]

1. 序論

1.1 背景

Web2.0 により，消費者がウェブサイトに書き込むことが容易になり，Twitter や Facebook に代表されるようなソーシャルメディアに自由にかつ気楽に書き込み，発信できるようなった．SNS は顧客の声を直接的に理解する上で重要なツールとなっている．そのため，近年，SNS の文章を分析する手法が盛んに研究されている．

英語の分析においては，近年，Word Vector[1]や Paragraph Vector[2]のような新しい Word Embedding 手法により，単語や文章の空間距離を活かした分析方法が提案されている．日本語や中国語の分析は，単語と単語の間にセパレータが無いため，形態素分析を行い単語分割した後に，文章分析を行うのが一般的である．しかし，SNS に書き込まれる文章は，整形された文章ではないため，スペルミスがあったり，文法が正しくなかったり，単語のスペルの一部を書き換えたスラング的な単語を使われる．そのため，形態素分析が正しく行えず，文章分類する際に精度を低下させる原因の一つである．

1.2 関連研究

現在，SNS 文章分析の研究対象は，形態素分析関連と Feature 化関連がある．

(1) 形態素分析に関する関連研究

形態素分析を利用し SNS 文章を単語に分解し，文章を単語で Feature 化している．手法 1 は，テキストデータをクリーニングし，形態素分析を利用して，文章から単語を分離する．分離した単語は BOW などを利用してベクトル化する．その後は単語ベクトルを利用して文章の Feature を計算する．文章の Feature は例えば各単語ベクトルのベクトル和やその平均など方法で計算する．この手法の問題として, スペルと文法の正しい文章は問題なく分析できるが, SNS のような文章は正しく単語分割することができなく，それが分析精度を下げる可能性がある．手法 2 は，SNS 文章に対する形態素分析精度低い問題を解決するため，SNS 向け形態素分析を改良する手段が提案されている．具体的には以下の 3 つの手法があり，それぞれ課題がある．

- SNS 顔文字辞書/ルール顔文字は数万種類あり，全てを網羅することが容易でない.
- SNS 専用表現辞書・SNS 重複表現辞書/ルール辞書にない未知の表現に対応できない.
- SNS データを利用して SNS 形態素分析ツールを SNS データで再訓練するには訓練データに依存する.

(2) Feature 化関連研究

単語を分離した後，BOW や Word embedding を利用して単語をベクトル化する．BOW は，単語の共起性を利用しベクトル化する．Word embedding は，低次元で単語をベクター化する技術で，2013 年に Mikolov は Word2Vec[1]の提案を皮切りに, 盛んに研究されている. Word embedding は, BOW に比べ，スパース性の問題，次元の問題，Semantic 情報欠落を改善できる手法である．

1.3 目的

これまでの手法と，後述する新しい手法を，SNS 文章群を二分類し，分類精度を比較する．また，SNS 文章にノイズを加え擬似的に, より SNS のように歪んだ文章を生成し, 各手法の SNS 文章歪み耐性について評価する．最後に，今回提案する手法は，形態素分析を使わないため，言語依存性がないため, 日本語と同様にセパレータがない中国語と，セパレータがある英語についても，既存方法と提案方法とで文章分類精度を比較する．

2. 提案手法

2.1 手法

今回の提案は，文章を Feature 化する際に従来の方法のような単語での Feature 化ではなく, 文字で Feature 化する．形態素分析を行わず文章から単語を分離せずに，文章から

†1 (株)ソニー・インタラクティブエンタテインメント

文字を分離しベクトル化して，文字ベースの文章特徴化する．文字として分離した後，Bag Of Character(BOC)と Character To Vector(C2V)を利用して，文字でベクトル化する．この文字ベクトルを利用して，unigram および bigram に，文章を Feature 化する．

C2V の場合, ある Social media 文書 D を Character 列 $C_1,C_2 \cdots C_n$ としたとき，D に対する文書ベクトル S_{c2v} は，次式により与えられる．Character2vec は word2vec の Skip-gram アリゴリズムを利用して Character のベクトルを計算する．

$$S = (\sum_{i=1}^{n} Character2vec(C_i))/n$$

BOC の場合，ある Social media 文書 D を Character 列 $C_1,C_2 \cdots C_n$ として，D に対する文書ベクトル S_{BOC} は，次式により与えられる．LSI(Latent Se-mantic Indexing) は Character ベクトルの次元を圧縮する．

$$S = (\sum_{i=1}^{n} LSI(C_i))/n$$

人為歪み化の場合，Character のリスト C を作ります．リスト C は“あいうえおかきくけこ…ABCD…-*+”など Character の列です．ある Social media 文書 D を Word 列 $W_1,W_2 \cdots W_n$ として，Word 列から単語 W_r を Randomly ピックアップして，そして C 列から C_r を Randomly ピックアップして，W_r の後に C_r を追加して Wr’になり，D の中の W_r を新構成した W_r’に入れ替わる．上記処理を各 D に 1, 3, 5, 7, 9, 11 回掛けて，各アルゴリズムの文章歪み耐性を検証する．

提案手法の利点は，形態素分析が不要になり，形態素分析に起因する問題も無くなり，かつ，全体のプロセス時間も削減できる可能性がある．また，システム化を考えたとき，形態素分析が多言語対応するときに課題となるが，今回の提案は，文字での embedding のため言語依存性が無くなり, 同一アルゴリズムで全言語対応できる可能性がある．日本や中国語においては，動詞を表す文字と時制を表す文字で文章が構成されるため，ステミングに対応できる可能性もある．また，記号も embedding できるため，顔文字や絵文字も対応できる．

2.2 実験手順

この提案は，文章を文字ベクトルで表現することであり，文章特徴化方法，分類モデル選択方法，そのパラメータ選択方法は既存の手法を使用する．検証と比較ために文章特徴化方法は平均化で，分類モデルは SVM を使用する．Embedding のベクトル数は 200 で，このベクトルを SVM 分類器を利用して，文章群を二分類し，分類精度(F-measure)を評価する．以下の 3 点を確認する．

- 既存方法と提案方法の二分類結果比較.
- 評価した上位三位方法で，文章歪み耐性比較.
- 中国語と英語の二分類性能比較による多言語対応性確認.

(1) 日本語評価データ

また，提供するテンプレートファイルは，**エラー! 参照元が見つかりません。**に示す通り，2 つのセクションから構成している．

具体的な収集方法は, 日本語 Twitter から収集したカスタマイズした顔文字リストと Sentiment lexicon 日本語単語両方付き[4] Tweet 40 万件を使用し, 内 90%を学習データとする．顔文字リストは Twitter によく使う顔文字を収集しリストを作成した．Sentiment lexicon は Negative 表現を収集しリストを作成した．データセットは全てマニュアルチェックできないため，ランダムで 100 件抽出しマニュアルチェックした．精度は 85%程度である．Positive と Negative の二分類を九回検証し，毎回 10 万件の文章を用意し，90% を学習データとする．

(2) 中国語評価データ

中国語のデータは，中国語 Weibo データセット[5]を使用した．emotion tag あり(subjective) と emotion tag ない(objective)の二分類を行った．9732 件の文章を用意し，90%を学習データとする．

(3) 英語評価データ

英語のデータは Stanford Large Movie Review Dataset[6]を使用した．Positive と Negative の二分類を行った．2 万件の文章を用意し，90%を学習データとする．

2.3 分析方法

評価対象は，以下の Character ベースで，比較対象を以下の Word ベースとし，F-measure を測定する．Word ベースの際の形態素分析は，Mecab を使用した．W2V,P2V および C2V の次元は 200 とし,BoW および BoC は,LSI で 200 次元に削減した.

- Word2Vec アリゴリズムを利用した unigram と bigram の Character2Vec.
- Character base の LDA　P2V.
- Bigram Bag of Characters.
- Unigram と bigram の Word2Vec with cleaning.
- Unigram と bigram の Word2Vec without cleaning.
- Word base の LDA　P2V.
- Bigram の BOW.

最後に二分類に使った文章の特徴ベクトルを，t-distributed stochastic neighbor embedding(t-SNE)を用い 2 次元に写像し，二分類の重心が離れているかを視覚的に確認する．

3. 結果

3.1 各方法との F-measure 比較

F-measure 結果を図 1 と表 1 に示す.

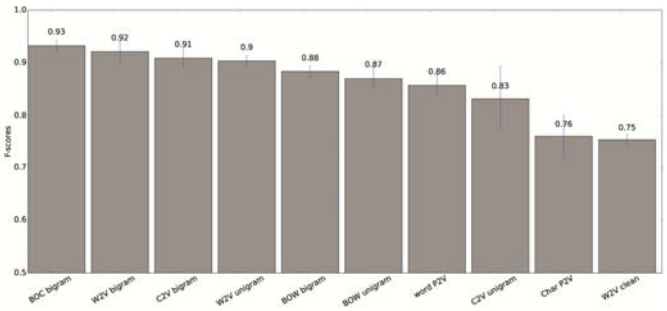

図 2　各方法の分類 F-measure

表 2　各方法の分類 Precision, Recall と F-measure

Method	Precision	Recall	F-measure
W2V unigram	0.90	0.90	0.90
W2V bigram	0.92	0.92	0.92
C2V unigram	0.83	0.83	0.83
C2V bigram	0.91	0.91	0.91
BOW unigram	0.88	0.88	0.87
BOW bigram	0.89	0.88	0.88
BOC bigram	**0.93**	**0.93**	**0.93**
Word P2V	0.86	0.86	0.86
Char P2V	0.78	0.76	0.76
W2V with Cleaning	0.75	0.75	0.75

3.2 文章歪み耐性評価

文章歪み耐性の結果を図 1 に示す.

3.3 多言語対応性評価

多言語対応性評価結果を表 2 に示す.

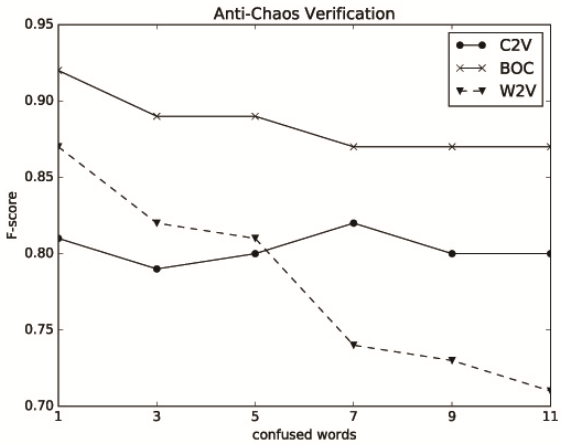

図 2　文章歪み耐性 F-measure

表 2　多言語各方法での分類 F-measure

Method	中国語 F-measure	英語 F-measure
W2V unigram	0.56	0.76
W2V bigram	0.34	0.63
C2V unigram	**0.64**	0.62
C2V bigram	0.35	0.73
C2V trigram	0.34	0.75
BOW unigram	0.63	0.78
BOW bigram	0.63	0.75
BOC bigram	0.41	0.66
BOC trigram	0.46	**0.80**

4. 考察

各手法での F-measure は, BOC, W2V, C2V の順番となり, Character ベースが Word ベースより優れている結果となった．文章歪みの結果は，Word ベースの W2V は歪みが大きくなるほど F-measure が 10%程度に低下したが，Character ベースの BOC および C2V は，歪みが増えても 2%程度の減少に留まった．多言語処理の結果も，英語においては日本語と同様に Character ベースの方が分離能力が 2%程度高い結果となった．中国語においては，Character ベースの方が分離能力が 1%程度高く，Character ベースの優位性はある．処理プロセス面では，Word ベースは日本語，英語，中国語それぞれの処理プロセスであるが，Character ベースは同一の処理プロセスで処理できた．

今回日本語においてもっと F-measure が高かった BOC と，現在もっと多く使われている手法の BOW をそれぞれで文章群を特徴化したものを t-SNE で 2 次元写像したものを以下に示す．見てもわかるとおり，BOC は分類対象の 2 グループの中心距離が BOW より離れている．そのため，SVM での分類精度も高いものと思われる．

近年，Character ベースの DNN 関連研究が出ている[7]. 私たちも日本語評価データを利用して、Character ベースの LSTM と CNN 二分類器を訓練して、その分類の能力を試した．CNN は Accuracy 88%で収束している．LSTM は Accurary 82%で収束している．いずれも Bigram BOC と Bigram C2V の結果に勝っていない．その原因として、訓練データ不足，NN の Parameter の Tuning 不足などが考えられる．様々な領域で DNN ベースの Text mining はその有効性を証明されているが、Tuning の難易さや、数百万件程度でデータの量が少ない場合においては（例えばある商品に対するコメントなど），DNN ではない方が有効ある可能性が高い。

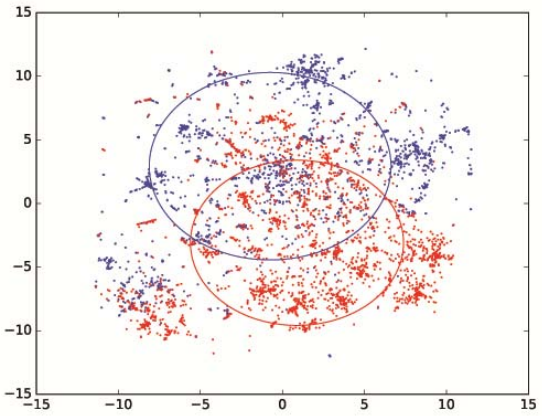

図 3　t-SNE で視覚化した BOC

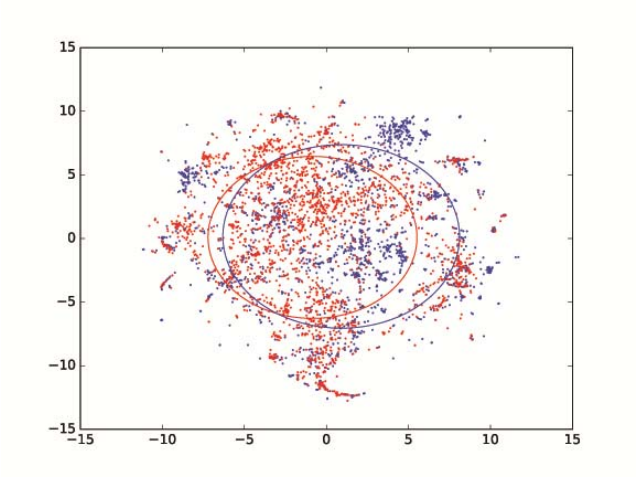

図 4　t-SNE で視覚化した BOW

5. 結論

5.1 まとめ

Character ベースの分類能力の優位性が確認できた．SNS の文章は歪んだ文章であることが多いが，Character ベースは，歪んだ文章でも高い精度で分類できることが確認できた．形態素分析が不要で，Character を直接的に利用するために文章正規化も不要である．Character ベースで，多言語処理できることも確認できた．

5.2 展望

引き続き，Character ベースの検証および最適化を行う．他の領域への応用可能性を探索する．

参考文献

[1] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Distributed Representations of Words and Phrases and their Compositionality. In Advances in Neural Information Processing Systems 26, Curran Associates, Inc, 2013, p. 3111-3119.
[2] Quoc Le and Tomas Mikolov. Distributed Representations of Sentences and Documents. In Proceedings of the 31st International Conference on Machine Learning (ICML 2014), JMLR, 2014, p. 1188-1196.
[3] Bo Han, Paul Cook, and Timothy Baldwin. Lexical normalization for social media text. ACM Transactions on Intelligent Systems and Technology, 2013, 4(1):5.
[4] 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. 情報処理学会関西支部 支部大会講演論文集, 2014, 5p.
[5] “NLP&CC 2013 中文微博情例数据”. http://www. datatang. com/data/44114/, (参照 2016-06-13).
[6] Maas, Andrew L. and Daly, Raymond E. and Pham, Peter T. and Huang, Dan and Ng, Andrew Y. and Potts, Christopher. Learning Word Vectors for Sentiment Analysis. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, p. 142-150.
[7] Xiang Zhang, Junbo Zhao, and Yann LeCun, Character-level Convolutional Networks for Text Classification. In Advances in Neural Information Processing Systems, 2015, p. 649–657.