災害時における Twitter のつぶやきデータに対するテキストマイニング

1180336　宍戸　海士　【 セキュリティシステム研究室 】

1　はじめに

東日本大震災や熊本地震では，Twitter などの SNS を用いて情報伝達が行われていた．Twitter では被災者自身が情報発信を行なっており，被災地の状況がリアルタイムに把握することができる．しかし，Twitter のデータは膨大なため解析する必要がある．そのため，データの解析に時間や手間が掛かるといった課題が指摘されていた．既存方式 [1] はツイートデータの解析としてツイート分類を行なっており，Support Vector Machine(以下 SVM) と単純ベイズを用いて分類を行なっている．分類するツイートデータは “水” を含むツイートであり，飲み水に関するツイートかそれ以外かで分類を行なっている．分類の精度は SVM が 63.3%，単純ベイズが 70%であった．これは膨大なツイートデータを分類するのに十分な分類精度とはいえない．そこで本研究では Recurrent Neural Networks(以下 RNN) を使用し，文書中の単語の順序に注目してツイート分類を行う．そして，ツイート分類には SVM や単純ベイズより RNN が分類精度が高い事を実証する．

2　提案方式

本研究ではツイートデータの単語が現れる順序に注目し，単語の順序を時系列データとして扱い，ツイート分類を行う．ツイートは既存方式と同じく “水” を含むツイートを集め，“水” が飲み水に関する水か，それ以外の水かの分類を行う．

3　実験

学習させるデータは東日本大震災・熊本自身の発生時である下記の期間から収集したツイートデータ 2400 個を使用する．収集するツイートは “水” 含みハッシュタグ#jishin，#tsunami などの地震に関係のあるツイートである．

- 2011 年 3 月 10 日 ～ 2011 年 4 月 4 日
- 2016 年 4 月 13 日 ～ 2016 年 5 月 13 日

3.1　学習アルゴリズム

学習させるアルゴリズムとして RNN を使用する．RNN は時系列データを扱い，自然言語処理等の分野で使用されている．

3.2　データの準備

ツイートデータが飲み水に関するツイートかそれ以外かのラベル付けを行う．ツイートデータの分かち書きを行い，総データ数の 7 割を学習用データ，残りの 3 割をテスト用データとする．

3.3　学習

ツイートデータの学習を行った結果，学習時のロスとテスト時のロスの関係は図 1 の様な関係になった．図 1 の epoch は学習回数，loss は分類器の損失率を表す．

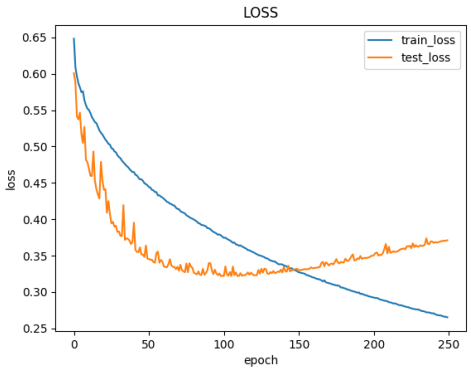

図 1　モデルの LOSS

図 1 からテスト時のロスの最小値は 100epoch 回周辺とわかる．150epoch を越えるとテストのロスが大きくなり，過学習が起きていることがわかる．そのため，分類に使用するモデルとして，100epoch のモデルを使用する．

4　評価

作成したモデルを用いて評価を行なう．モデルの評価にはツイートデータ 200 個を用いる評価の結果は表 1 である．

表 1　モデルの評価

実際＼予測結果	関係あり	関係なし	合計
関係あり	89	11	100
関係なし	22	78	100
合計	111	89	200

表 2　分類精度の比較

	単純ベイズ	SVM	RNN
分類精度	70%	63.3%	83.5%

評価の結果モデルの分類精度は 83.5%である．既存方式との比較を表 2 示す．既存の単純ベイズや SVM よりもツイートデータの分類において分類精度が高いことが分かった．

5　まとめ

文書の単語の並びに注目し，ツイート分類を行うことによって既存方式より高い分類精度でツイート分類を行うことが出来た．そのため，ツイート分類において，単純ベイズや SVM より RNN を用いることが望ましい．

参考文献

[1] 坂巻英一, 亀井悦子, “Twitter 上のつぶやきに関するテキストマイニングの事例研究- 大規模災害発生時の被災地における現状把握への応用 - ”, 日本経営工学会論文誌 65 巻，p.39-50，2014-2015.