マイクロブログにおける浸水害情報の解析

朝倉 康伸 * †　萩行 正嗣 ‡　小町 守 †

† 首都大学東京　‡ 株式会社ウェザーニューズ

1　はじめに

近年，マイクロブログの普及に伴い，災害情報などがリアルタイムでネットに投稿されるようになってきた．2011 年の東日本大震災の際には，被災状況や津波警報の共有，安否確認などにマイクロブログが活用された．このようにマイクロブログで多くの情報が発信される中，それらの情報を解析する研究が注目されている．

しかしながら，マイクロブログから災害情報を抽出する場合，災害に関するキーワードを含む投稿を集めるだけでは実際にそこで災害が発生しているのか，そうでないのか判断できない．

災害に関する投稿は以下のようなものがある．

(1) 家の前が 浸水 しています。

(2) この辺りはよく 冠水 する。

(3) 去年の 浸水 は大変だった。

(4) 道路 冠水 に注意です。

災害に関するキーワードを含む投稿には，(1) のように実際起きている災害の報告とともに，(2) のようなある地域における一般的な災害に関する話題や，(3) のような過去に起きた災害についての言及，(4) のような注意喚起している投稿などが含まれている．そこで，テキスト中に記述されているイベント（出来事）が実際に起きているか，そうでないかを判定する解析器が必要になる．

現在マイクロブログからさまざまな災害情報の抽出に取り組まれている一方，浸水害に関する情報抽出には取り組まれていない．また，マイクロブログから浸水害に関する情報を抽出するためには，正確な位置情報が付与されている必要がある．

そこで，本研究では約 98%の投稿に正確な位置情報が付与されているマイクロブログ（ウェザーニュース [1]）のユーザが投稿したウェザーリポートから浸水害に関する投稿を取り出し，それらの投稿が実際に浸水害が起きている時の投稿か，そうでないかをアノテーションし，浸水害に関するデータを作成した．

また，作成したデータを基に，機械学習を用いて浸水害が起きているかそうでないかを判定する分類器を作成した．分類器としてサポートベクターマシン（SVM とする）を用い，複数の素性で比較・実験を行った．

実験の結果，キーワードを含む文節とその文節と直接係り受け関係にある文節の Bag of Ngrams と，日本語拡張モダリティ解析器 Zunda[2] の出力の 2 つを素性として用いた際に最も高い分類精度を示した．

本研究での貢献を以下に示す．

- ウェザーニュースのリポートデータから浸水・冠水に関するリポートを収集し，それぞれのリポートに対して，浸水害が起きているか，そうでないかをアノテーションし，浸水害についてのデータを作成した．
- 作成したデータを用いて浸水害が起きているか，そうでないかを判定する分類器を作成し，複数の素性で比較・実験を行い，F 値で 83.2 を示した．

2　関連研究

近年，マイクロブログの普及に伴い，マイクロブログから災害情報を抽出する研究が盛んに行われている．

2011 年の東日本大震災の際には，ツイッター [3] を利用した被災地の人々の安全に関する情報抽出システムの構築が取り組まれた [3]．この取り組みでは収集したツイート内の人の名前や地名にラベルを付与しツイッターコーパスを作成し，単語分割，固有表現認識，安否に関するトピック分類を行うことで，安全に関する情報抽出システムを構築している．他にも災害時のトラブルに関するツイートとそれに対する支援ツイートの抽出を行った研究 [5] がある．この研究では，SVM を用いてトラブルに関するツイートと支援ツイート

*asakura-yasunobu@ed.tmu.ac.jp

[1] http://weathernews.jp

[2] https://code.google.com/p/zunda/

[3] https://twitter.com/?lang=ja

を抽出し，そこから得られたツイートから対応するツイートを抽出することで，問題解決するシステムを構築している．

本研究と類似した研究として，地震に関するツイートに対し，実際に地震が起きている時の投稿か，そうでないかを分類し，災害位置を推定し地震報告をするシステムを作成した研究 [4] がある．位置情報はツイートの GPS データを使用し，GPS データが付いていないツイートは登録してある location データを使用している．しかし，GPS データが付いたツイートは非常に少なく，正確な位置情報を扱うのは困難である．そのため，本研究では正確な位置情報の扱えるマイクロブログから浸水害コーパスを作成した．

事実性解析を用いた災害予測に取り組んでいる研究には文献 [1, 2] がある．ARAMAKI ら [1] はツイッターを用いたインフルエンザの流行検出に初めて取り組み，Kitagawa ら [2] はインフルエンザの流行検出を行うために “インフルエンザ” を含むツイートに対し，ツイートした本人もしくはその周りの人がインフルエンザに感染しているか否かの事実性解析を行った．この研究では素性として日本語機能表現辞書つつじ [4] の機能表現意味 ID と日本語拡張モダリティ解析器 Zunda の出力を使用することで分類精度の向上が見られたと報告している．

3 データとタスク設定

本研究では，浸水害のイベントが起きているかそうでないかを判定する分類器の作成に取り組んだ．データは，ウェザーニュースという天気予報アプリケーションでユーザが投稿したウェザーリポートのコメントに “浸水”，“冠水”，“水に浸” を含む投稿（579 件）を収集した．収集期間は東日本大震災が発生した 2011 年 3 月 11 日からの 4 日間や 7 月や 2 月の台風や降雪が多い時期の計 17 日間の投稿を収集した．また，本研究で使用したデータは，1 リポートあたり平均 5 文であり，1 文あたり平均 5 文節，85 単語である．

また，収集した投稿で浸水害が起きているかそうでないかをアノテーションした．

アノテーションの基準としては，以下の条件を全て満たしたものを正例（起きている）とした．

- 投稿者のいる位置によらず，どこかで浸水害が起きていると判断できる．
- リポートの中の 1 つ以上の文で浸水害が起きていると判断できる．
- リポートの投稿時に浸水害が起きていると判断できる．

アノテーションの例を以下に示す．

(5) 東京駅で 浸水 していたが、こちらは無事でした。 正例

(6) ルーフバルコニーの排水追い付かず 冠水 気味です。 負例

(7) 三里木の旧道はいつも 冠水 するけど、まだちょっとたまってるくらいです。 負例

(5) の場合，投稿者のいる位置では浸水は起きていないが，東京駅では浸水が起きていると判断できるため，正例のラベルを付与している．(6) の場合，ルーフバルコニーにおいて冠水は起きているが，浸水害は起きいないと判断できるため，負例のラベルを付与している．(7) の場合，リポート時には冠水は起きていないと判断できるため，負例のラベルを付与している．アノテーションは理系大学院生 2 人で行い，κ 係数は 0.870 であった．

4 機械学習による浸水害情報の分類

本稿では浸水害情報解析を教師あり学習による 2 値分類問題として定式化し，以下に書く素性の説明及び (8) を例とした場合の具体的な素性を示す．

(8) 今日は大雨なので 浸水 する予定。

BoN：リポート文章中のキーワードを含む文の 1-gram から 4-gram の Bag of N-grams

→今日 は 大雨 な ので 浸水 する 予定 。(1-gram の例)

Dep：キーワードを含む文節と，キーワードを含む文節の係り先，係り元の文節のみを取り出した文に対する BoN の素性

→大雨_[dep] な_[dep] ので_[dep] 浸水_[dep] する_[dep] 予定_[dep] 。_[dep] 大雨_な_[dep] な_ので_[dep] ので_浸水_[dep] 浸水_する_[dep] する_予定_[dep] 予定_。_[dep] (1，2-gram の例)

Zunda：日本語拡張モダリティ解析器で，文中のイベント（動詞や形容詞，事態性名詞など）に対して，時制や仮想性（仮定の話かどうか），真偽判断（イベントが起こったかどうか），などの解析が可能である．“浸水”，“冠水” は事態性名詞であり，実験で使用するリポートには “浸水”，“冠水” が含まれているため，素性としてそれらキーワードに対する Zunda の出力のうち，時制（未来，非未来），仮想（条件，帰結，0），

[4] http://kotoba.nuee.nagoya-u.ac.jp/tsutsuji/

表 1: 線形カーネル SVM の分類精度

素性	適合率	再現率	F1
BoN	75.9	85.9	80.4
Dep	**79.2**	84.7	81.6
Zunda	62.5	**96.8**	76.0
BoN+Dep	76.7	85.9	80.8
BoN+Zunda	76.1	86.2	80.7
Dep+Zunda	78.4	88.5	**83.1**
All	76.5	87.8	81.7

表 2: 多項式カーネル SVM の分類精度

素性	適合率	再現率	F1
BoN	75.6	85.0	79.8
Dep	**79.2**	83.7	81.1
Zunda	62.4	**96.5**	75.8
BoN+Dep	78.0	85.6	81.3
BoN+Zunda	75.5	86.9	80.7
Dep+Zunda	78.7	88.5	**83.2**
All	77.5	88.2	82.4

真偽判断（成立，不成立，不成立から成立，成立から不成立，高確率，低確率，低確率からの高確率，高確率からの低確率，0）に関する出力を利用した．

→未来_[時制] 0_[仮想] 高確率_[真偽]

5 実験・評価

本研究では単語分割に MeCab[5]（ver.0.996）を利用し，構文解析には CaboCha[6]（ver.0.68）を利用した．MeCab の辞書は IPADIC（ver.2.7.0）を用いた．

アノテーションした結果，“水に浸”をキーワードとして収集したリポート 12 件は花や農作物に関する言及が多く，浸水害に関するリポートが少なかった．そこで実験では 579 件から “水に浸”をキーワードとして収集したリポートを除く正例 313 件と負例 254 件の 567 件で実験を行った．評価は 567 件のデータで 5 分割交叉検定を行い，適合率，再現率，F 値によって評価した．本研究では線形分類器として線形カーネル SVM，非線形分類器として多項式カーネル SVM を使用し，比較・実験を行った．SVM のツールとして libSVM（ver.1.04）を使用した．ハイパーパラメータはコストパラメータ C とカーネル関数の係数 coef0 について予備実験にてチューニングを行った．多項式カーネルの次元数は 3 とした．

線形カーネル SVM と多項式カーネル SVM の分類精度をそれぞれ表 1，表 2 に示す．

本研究で取り組んだ分類問題では，線形カーネル SVM と多項式カーネル SVM には分類精度の差はなく，各素性の精度の分布も類似した分布を示した．また，比較した素性では Dep と Zunda の 2 つを使用した際に最も高い分類精度を示した．

[5]http://taku910.github.io/mecab/

[6]https://taku910.github.io/cabocha/

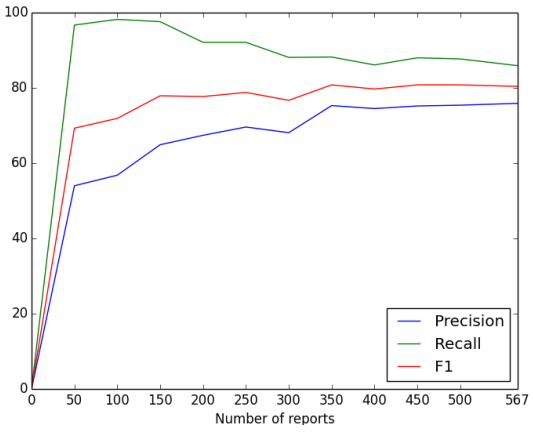

図 1: データ量と分類精度

6 分析

表 1，表 2 から Dep 素性のみを使用した際に適合率が最も高い精度を示していることが分かる．Dep 素性では BoN 素性に比べてキーワードに対し係り受け関係にある文節のみを考慮するため，浸水害の分類に深い関係にある単語が素性になっていると考えられる．BoN 素性を使用した際にはバイアス項が正に大きな影響を与える素性となっており，負例に大きな影響を与える素性がない場合は正例であると判断しており，再現率を上げている一方，適合率を大幅に下げたと考えられる．Dep 素性では BoN 素性のノイズが減るため適合率が上がったと考えられる．

また Zunda 素性ではほとんどの事例に対し正例であると判定しているため高い再現率を示し，適合率は非常に低い値を示している．たとえば，“浸水しない”という入力に対し Zunda の出力ではイベントが起こったかどうかの真偽判断で不成立だと出力する一方，“浸水していない”という入力に対し成立だと出力しており，浸水していない場合の表記の揺れに対し Zunda が対応できていないと考えられる．

適合率と再現率でそれぞれ最も高い精度を示した Dep 素性と Zunda 素性を組み合わせた素性が最も高い F 値を示した．

また線形 SVM において分類に大きく影響した Dep+Zunda 素性について調べた．正負それぞれの素性とその重みについて表 3 に示す．

正に最も大きな影響を及ぼす素性は Zunda の真偽判断の出力である “成立_[真偽]”であった．これは Zunda の真偽判断の素性はすべての事例に対して付随してお

表 3: 線形 SVM の Dep+Zunda 素性の重み

正の素性	重み	負の素性	重み
成立_[真偽]	0.44938	冠水_する	-0.43767
津波	0.32826	に	-0.37203
冠水_と	0.31268	不成立_[真偽]	-0.35858
、_冠水	0.25815	を	-0.34530
冠水_し_て	0.25528	注意	-0.34294
の_冠水	0.23073	心配	-0.34058
冠水_状態	0.22084	道路_の	-0.31980
道	0.20543	冠水_や	-0.30046
し_て_いる	0.20365	し_そう	-0.29838
あちこち	0.20088	そう	-0.28952
バイアス項	0.41139		

表 4: 分類に失敗した例

失敗例 1	道路も冠水していません！
失敗例 2	昔ね。我が家も床下浸水した。

り，かつ，ほとんどの事例に対して “成立_[真偽]”が付与され，いくつかの負の事例に対して “不成立_[真偽]” が付与されるため，“成立_[真偽]”がバイアス項よりも大きな正の重みが付いていると考えられる．また負に大きな影響を及ぼす素性には “注意”や “し_そう”など，浸水害の注意喚起に関する素性が多かった．

分類に失敗していた事例について，表 4 に示す．失敗例 1 のように正に大きな影響を及ぼす素性 “冠水して” により負例の投稿を正例だと判定してしまうといった誤分類が多かった．これは，“冠水して”などの高頻度で現れることで大きな重みになっている素性が “いません”などの負に影響を及ぼす素性よりも大きく影響するためだと考えられる．本研究では多項式カーネル SVM を使用したが，データ量が少なく，組み合わせ素性がスパースになり，有効に機能しなかったと考えられる．また，素性としてリポート中のキーワードを含む文内の単語のみを考慮しており，キーワードを含まない文は考慮できないため，失敗例 2 のようにキーワードを含む文だけを見れば正例であるが，文脈を見れば過去の話をしているといった，キーワードを含む文のみでは正しい分類ができないような事例もあった．

データ量による分類精度の比較を図 1 に示す．図 1 から，データ量の分類精度は 350 件から大きく変化していないことがわかる．本研究で使用したキーワードを含む浸水害のリポートでは，少ないデータで大半の事例を学習できている一方，低頻度で分類に重要な単語に対して学習できていないと考えられる．

7 おわりに

本研究では，ウェザーニュースのリポートから浸水害に関する投稿を収集し，浸水害が起きているかそうでないかをアノテーションし浸水害コーパスを作成した．また線形カーネル SVM と多項式カーネル SVM を用いて，浸水害情報の事実性解析を行った．今回のタスクではキーワードを含む文節とその文節と直接係り受け関係にある文節の Bag of Ngram と，Zunda の出力の 2 つを素性として用いることで，最も高い F 値を得ることを示した．

今後の課題として以下のものが挙げられる．

ウェザーリポートの天気報告を利用することで降雨・降雪のイベントについて事実性に関する正解データを自動で生成することが可能である．そこで浸水害リポートとの相関を分析し，分野適応することで，より正確な浸水害情報の抽出を目指す．

また，本研究では，浸水害に関しての事実性を判定し規模の小さな浸水・冠水に関しては負例のラベルを付与した．しかし，まず浸水や冠水に関する事実性を判定してから災害規模を推定することで，浸水害の事実性を正確に判定できると考えられる．

“浸水”，“冠水”のみをキーワードとして浸水害に関するリポートを収集したが，他にも浸水害に関するキーワードは存在する．そのため，今回使用したキーワード以外のキーワードを探し，再現率を上げる必要がある．また，作成したデータは，台風の多い時期や雪の多い時期などを含む 17 日間の投稿から収集したが，より多くのデータを収集する必要がある．

また本研究で使用したデータには GPS による経度緯度情報が約 98 %付いているため，それらの位置情報を用いて浸水害の可視化システムを構築する．

参考文献

[1] Eiji ARAMAKI, Sachiko MASKAWA, and Mizuki MORITA. Twitter catches the flu: Detecting influenza epidemics using Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1568–1576, 2011.

[2] Yoshiaki Kitagawa, Mamoru Komachi, Eiji Aramaki, Naoaki Okazaki, and Hiroshi Ishikawa. Disease event detection based on deep modality analysis. In *Proceedings of the ACL-IJCNLP 2015 Student Research Workshop*, pp. 28–34, 2015.

[3] Graham Neubig, Yuichiroh Matsubayashi, Masato Hagiwara, and Koji Murakami. Safety information mining — what can nlp do in a disaster—. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pp. 965–973, 2011.

[4] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW ’10, pp. 851–860, 2010.

[5] István Varga, Motoki Sano, Kentaro Torisawa, Chikara Hashimoto, Kiyonori Ohtake, Takao Kawai, Jong-Hoon Oh, and Stijn De Saeger. Aid is out there: Looking for help from tweets during a large scale disaster. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 1619–1629, 2013.