7K-05

感動詞の共起に着目した災害tweet抽出手法

湯沢昭夫 †　　小林亜樹 ‡

† 工学院大学大学院工学研究科 電気・電子工学専攻 ‡ 工学院大学情報学部情報通信工学科

1 はじめに

災害時の被災地の状況を知るのに，Twitterなどの SNS からの情報が有効である [1]．しかし，多数の投稿からの自動分類に教師あり学習を用いることは難しく，単純な入力のみで動作することが望まれる．著者らは，災害を示す災害語と共起する語集合を用いて SNS 上の投稿を分類し，災害に関連する投稿を抽出する研究を行なっている [2]．

本稿では，災害時の SNS 上では，通常よりも多くの人のやりとりが発生している点に着目し，投稿中の感動詞の共起語など複数の共起関係を利用して災害に関連する語 (手がかり語) 集合を生成する手法を提案する．

2 提案手法

本研究では，災害に関連する投稿の抽出を行うために，災害に関連する語 (手がかり語) の抽出を目的とする．

「地震」のような災害語を含む tweet を検索すれば良いのであれば単純な部分一致検索で十分である．また，「揺れ」「津波」のような関連語集合を準備できるのであれば，検索の和集合で対応できる．しかし，関連語の予測は難しく，tweet 自身から自動的に抽出されるべきである．そこで，代表語として「地震」をシステムに入力すると，tweet 内での語の共起関係を用いて関連語集合を得ることとした．人と人とのやりとりが災害時には増加する [3] ことから，挨拶に用いられる感動詞との共起語集合も用いることとした．

本手法の全体像を図 1 に示す．

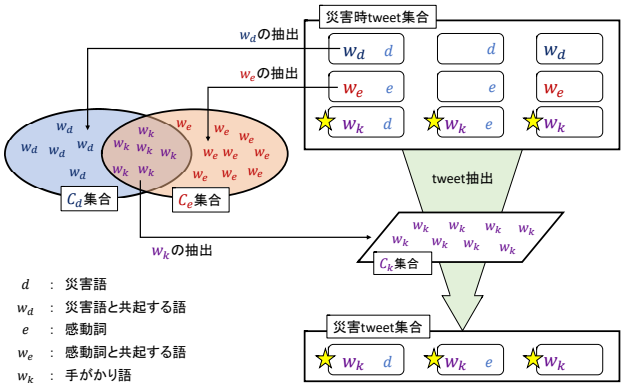

図 1: 提案手法の概要図

災害時 tweet 集合とは，災害発生後の一定時間範囲内に存在する tweet 集合であり，角丸四角形で描かれているものは tweet である．

災害語 d は，「地震」といった 1 語またはごく少数の語集合であることを想定している．これは，発生した災害を代表すると思われる語を想起し入力する部分のみが人手であるため，その負担を抑制しようとする意図である．w_d は災害語と共起する語であり，w_d の語集合を C_d と示す．

感動詞 e は，挨拶や応答といった「ありがとう」のような品詞が感動詞に該当する語である．w_e は感動詞と共起する語であり，w_e の語集合を C_e と示す．

C_k は，C_d と C_e の積集合であり，災害語と共起する語と感動詞と共起する語の積と取ることで災害に関連する語 (手がかり語) が得られるのではないかという仮定のもとで，積集合としている．

これらの状態を前提条件として，C_k を対象に手がかり語 w_k を選ぶ．その基準として

- 単語 w_k の出現頻度が平常時と比べて高い語
- 単語 w_k の χ^2 値を降順に並べた際の上位 M 件

の 2 つの条件を満たす語を手がかり語として抽出する．

単語 w_k の χ^2 値は，w_k の災害前後の出現頻度と，災害前後の全語の出現頻度とを用いて (1) 式に示すとおり定義される．

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-E_{ij})^2}{E_{ij}} \tag{1}$$

ここで，r は tweet 集合の個数を示し，災害時 tweet 集合と平常時 tweet 集合の 2 つを対象とするため $r=2$ とする．平常時 tweet 集合とは，災害が発生していない時の tweet 集合である．c は単語種類数を示し，異なる tweet 集合間で単語 j の偏りの程度を示すため，単語 j と単語 j 以外の単語を対象とし $c=2$ とする．n_{ij} は tweet 集合 i における単語 j の出現頻度である．E_{ij} は tweet 集合 i における単語 j の期待値であり，各 tweet 集合における全単語の出現頻度に対する各 tweet 集合における単語 j の出現頻度の比率を，tweet 集合 i に乗ずることによって，tweet 集合 i において単語 j がどの程度出現するかを定める．E_{ij} は (2) 式で算出を行う．

$$E_{ij} = n_{i.} * \frac{n_{.j}}{N} \tag{2}$$

このとき，$n_{i.}$ を tweet 集合 i における総単語数，$n_{.j}$ を各 tweet 集合の単語 j の出現頻度，N を各 tweet 集合における総単語数とする．

本研究では χ^2 値を統計学的な検定手法として用いるのではなく，単純に偏りの度合いを示すための尺度として用いている．そのため，背景にある分布等を無視している．

Tweet Discovery for Disaster information using Co-occurrence words of interjections.

†Akio Yuzawa　‡Aki Kobayashi

†Electrical Engineering and Electronics, Kogakuin University Graduate School

‡Department of Information and Communications Engineering, Faculty of Engineering, Kogakuin University

3 評価実験

3.1 目的

本手法の有効性を明らかにするために，災害に関連する投稿の抽出精度で評価を行う．

災害語と共起する語 w_d の語集合 C_d を対象に，災害語と共起する語 w_d の出現頻度を求め降順に並べた上位 M 件の語で災害時 tweet 集合を対象に tweet の抽出を行った場合 (以降，災害共起頻出語手法) と，本手法を用いて取得した手がかり語集合で災害時 tweet 集合を対象に tweet の抽出を行った場合で比較し検証する．

3.2 条件

2016 年 6 月 16 日 北海道函館市で起きた震度 6 弱の地震を対象とし，災害語を“ 地震 ”とする．

災害時 tweet 集合を，地震発生 1 分前の 14:21:00 から 15:59:59 の間に日本語を用いて投稿された，合計 29670 件の tweet を収集した．

平常時 tweet 集合を，地震発生 1 日前である 2016 年 6 月 15 日の 14:21:00 から 15:59:59 の間に日本語を用いて投稿された，合計 19234 件の tweet を収集した．

以上より，得られた tweet 集合を実験に用いる．ただし，リツイート・引用ツイートは除去した．

災害時 tweet 集合および平常時 tweet 集合に streaming API を使っているため，全 tweet 対象にはできないが，検証の目的にはこれらから辿れる一部のサンプルを用いていると理解すれば問題ない．

パラメータとして，$M = 10$ とした．また，著者 1 名が災害に関連する投稿であるか否かの判断を行い，災害情報であると判断した tweet を正解，それ以外の tweet を不正解とした．

3.3 結果と考察

実験結果を表 1 に示す．各手法で得られた語集合を表 2，表 3 に示す．

表 1 は，各手法における，抽出された tweet 数 (合計)，うち人手により正解とされた tweet 数 (正解)，合計に含まれる正解の割合 (正解割合)，の 3 項目を示している．

表 2，表 3 は，各手法によって得られた χ^2 値もしくは出現頻度上位 10 件の語 (単語)，災害時 tweet 集合を対象に抽出された tweet 数 (w を含む tweet 数)，うち人手により正解とされた tweet 数 (正解)，w を含む tweet 数に含まれる正解の割合 (正解割合)，該当する語の χ^2 値もしくは出現頻度，の 5 項目を示している．

表 1 より，正解割合において提案手法は他の手法よりも高い値が得られた．これは，表 2 より，「津波」「震度」「余震」といった，その語自身が災害に関連する情報を含むであろう語が得られたため，正解割合が高い値になったのだと考えられる．また，「北海道」「函館」といった震源地である地名も得られていた．

一方で表 3 は，「あっ」「てる」「ない」といった，日常の文脈で使われている語が抽出された．これらの語自身には，何か重要な情報はないが，それと共起する他の語たちに何か地震に関する情報を含む語なのではないかと考えられる．このような語の取り扱い方については今後の課題である．

表 1: 各手法による tweet 抽出結果

	合計	正解	正解割合
提案手法	2010	315	0.157
災害共起頻出語手法	4664	332	0.071

表 2: 提案手法による手がかり語集合

単語 w_k	w_k を含む tweet 数	正解	正解割合	χ^2 値
北海道	721	41	0.057	580.674
大丈夫	859	98	0.114	390.111
函館	352	59	0.168	304.927
揺れ	337	187	0.555	274.975
震度 6 弱	165	17	0.103	148.057
震度 6	107	11	0.103	95.966
津波	109	18	0.165	94.900
心配	195	29	0.149	76.897
震度	71	13	0.183	63.659
余震	65	16	0.246	58.277

表 3: 災害共起頻出語手法による語集合

単語 w_d	w_d を含む tweet 数	正解	正解割合	出現頻度
北海道	721	41	0.057	3.091
大丈夫	859	98	0.114	3.788
あっ	426	28	0.066	1.866
函館	352	59	0.168	1.346
揺れ	337	187	0.555	0.008
心配	195	29	0.149	0.741
ない	835	17	0.020	0.705
震度 6 弱	165	17	0.103	0.668
怖い	144	17	0.118	0.614
てる	2023	36	0.018	0.244

謝辞

本研究の一部は科研費（26242013）の助成を受けたものである．

4 おわりに

本論文では，災害情報を得るために，投稿中の感動詞と共起する語，災害語と共起する語の 2 つの共起関係を利用して手がかり語集合を生成する手法を提案した．2016 年 6 月 16 日 北海道函館市で起きた震度 6 弱の地震を対象に実験を行い，本手法の有効性を確認した．

今後の課題として，地震以外の他の災害に対して，本手法が有効かどうかの検討が挙げられる．

参考文献

[1] 毎日新聞 : 情報発信でツイッター活用　大西市長に聞く，http://mainichi.jp/articles/20161017/k00/00e/040/121000c，(参照 2018-01-01)

[2] 湯沢 昭夫，小林 亜樹，“ 災害時における現地情報 Tweet 抽出手法 ”，DEIM Forum 2017，3K-01，pp.1-6(2017).

[3] 宮部 真衣，荒牧 英治，三浦 麻子，“ 東日本大震災における Twitter の利用傾向の分析 ”，研究報告グループウェアとネットワークサービス（GN），2011-GN-81，No.17，pp.1-7(2011).