D-023

語の複数の共起関係と文章間の類似度を利用した災害情報抽出システムの提案
Proposition of System to Extract Disaster-Related Information Utilizing Multiple Co-occurring Relations and Similarity of Sentences

湯沢 昭夫 † 市川 博彬 † 小林 亜樹 ‡
Akio Yuzawa Hiroyoshi Ichikawa Aki Kobayashi

1. はじめに

災害時において，Twitter 上から情報を得ようとする研究は多く存在し，収集した tweet を分析する応用の研究報告が相次いでいる [1]-[3]．tweet の収集には，多数の関連語集合を事前に定義しておいて，それらを含むような tweet を収集するというのがほとんどである．

しかし，災害ごとに関連語集合は変わるため，実運用において，関連語集合をどのように得るのかは課題となっていた．著者らは，この関連語集合を自動的に tweet 自身から抽出する手法を示し，全体として，自動的に災害に関連する tweet の抽出を行うシステムをこれまでに提案してきた [4]．

従来法で自動的に得た関連語集合にみられた一部の問題について，原因を特定し，問題を解決する手法について提案し報告する．

2. 提案手法

2.1. 概要

本稿では，災害関連 tweet を抽出するために，自動的に tweet 自身から手がかり語集合を抽出することが目的である．本システムの概要を図 1 に示す．

災害時 tweet 集合は，災害発生後の一定時間範囲内に存在する tweet 集合である．

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 は災害語との共起語であり，w_d の語集合を $C(d)$ と示す．

感動詞 i は,「ありがとう」のような挨拶や応答で現れる感動詞である．w_i は感動詞との共起語であり，w_i の語集合を $C(i)$ と示す．

$C_k = C(d) \cap C(i)$ であり，災害語と感動詞との両者と共起する語が災害に関連する語 (手がかり語) であることが多く観測された経験則に基づいている．

手がかり語 C_k を利用して，災害時 tweet 集合を対象に，tweet の抽出を行う．したがって，手がかり語集合の精度が最終的な抽出精度に大きく依存する．従来の手法 [4] では，災害時 tweet 集合内に，ほぼ同文章の tweet が多く，結果として災害とは無関係な語の出現頻度が上昇し，手がかり語集合に残ってしまう問題があったために，最終的に無関係な tweet が抽出されることがわかった．

そこで，文章が重複した tweet を自動的に除去する手法を提案する．

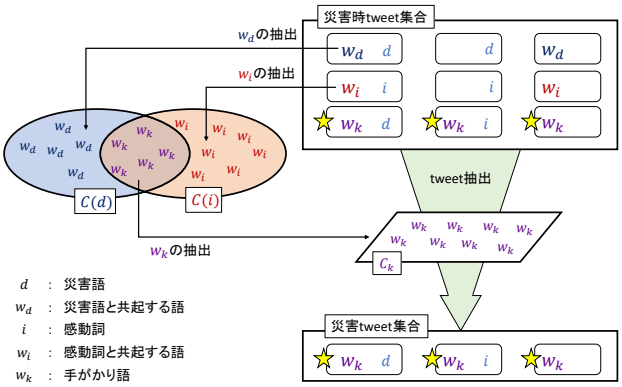

図 1: 本システムの概要図

2.2. 重複 tweet の除去

文章が重複した tweet を除去するために，各 tweet 間の類似度を算出し，高類似度となった tweet は処理対象に追加しないこととする．具体的には，新規 tweet 受信時に，既存の災害時 tweet 集合内の各 tweet との類似度を求め，別に定める閾値 α 以上となった場合には，災害時 tweet 集合への追加を行わないこととする．

2.3. 手がかり語の抽出

2.2 節で得た tweet 集合 S を対象に，災害語と共起する語集合 $C(d)$ と感動詞と共起する語集合 $C(i)$ を抽出する．そして，$C(d)$ と $C(i)$ の積集合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．

$$\chi^2 = \sum_{a=1}^{r} \sum_{b=1}^{c} \frac{(n_{ab} - E_{ab})^2}{E_{ab}} \quad (1)$$

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 b の偏りの程度を示すため，単語 b と単語 b 以外の単語を対象とし $c = 2$ とする．n_{ab} は tweet 集合 a における単語 b の出現頻度である．E_{ab} は tweet 集合 a における単語 b 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 b の出現頻度の比率を，tweet 集合 a に乗ずることによって，tweet 集合 a において単語 b がどの

†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

程度出現するかを定める．E_{ab} は (2) 式で算出を行う．

$$E_{ab} = n_{a.} * \frac{n_{.b}}{N} \quad (2)$$

このとき，$n_{a.}$ を tweet 集合 a における総単語数，$n_{.b}$ を各 tweet 集合の単語 b 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．

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，従来手法 [4] と比較し手がかり語の抽出精度で評価を行う．

3.2. 条件

2017 年 10 月 6 日 23:56:41 に福島県楢葉町で起きた震度 5 弱の地震を対象とする．

災害時 tweet 集合を，地震発生 1 分前の 23:55:41 から 00:05:41 の 10 分間に日本語を用いて投稿された，合計 10596 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 2017 年 10 月 5 日の 23:55:41 から 00:05:41 の 10 分間に日本語を用いて投稿された，合計 6018 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．

各パラメータとして，災害語 d=“地震”，語数 M=20，$\alpha = 1.0$ とした．また，著者 1 名が手がかり語が災害に関連するか否かの判断を行い，災害に関連すると判断した語を正解，それ以外の語を不正解とした．

3.3. 結果と考察

実験結果を表 1 に示す．各手法で得られた手がかり語集合を表 2 に示す．

表 1 は，手がかり語のうち人手により正解とされた語 (正解)，合計に含まれる正解の割合 (正解割合) の 2 項目を示している．

表 2 は，各手法によって得られた χ^2 値を降順に並べた際の順位 (順位)，χ^2 値上位 20 件の語 (単語) の 2 項目を示している．

対象の災害について，従来手法を適用すると，表 2 より，上位に「モンスト」といった災害と無関係な単語がいくつか抽出された．同様な基準で提案手法を適用したところ，それらの単語の出現頻度が大幅に下がったため，上位から除去することに成功した．そのため，正解割合において提案手法は従来手法よりも高い値が得られた (表 1)．最終的な tweet 抽出に及ぼす影響についての評価などは今後の課題である．

本システムの実行には，当時の災害語入力を除いて人手の介入が不要である．今回のデータセットでは，災害発生から 10 分間の計 10596tweet 分を，形態素解析から本手法による除外などを含めた全体での処理時間は概ね 70 分で実行できた．

現時点での実行速度は，Python のループを用いた cos 類似度による類似度を求めるところで概ね 60 分程度掛かっており，実行速度の改善の余地がある．

表 1: 各手法による手がかり語の抽出精度

	合計	正解	正解割合
提案手法	20	19	0.95
従来手法	20	17	0.85

表 2: 各手法による手がかり語集合

	提案手法	従来手法
順位	単語	単語
1	揺れ	揺れ
2	緊急地震速報	緊急地震速報
3	大丈夫	大丈夫
4	福島	ゆれ
5	ゆれ	福島
6	地震速報	地震速報
7	怖い	モンスト
8	でかい	びっくり
9	こわい	怖い
10	長い	地震だ
11	びっくり	でかい
12	震度 5 弱	長い
13	揺れる	こわい
14	ぁぁ	震度 5 弱
15	心臓	揺れる
16	震源	おっぱい
17	津波	速報
18	警報	ぁぁ
19	速報	心臓
20	震度 3	津波

4. おわりに

本論文では，災害に関連する投稿を抽出するために，文章が重複した tweet を自動的に除去する手法を提案した．福島県楢葉町で起きた地震を対象に実験を行い，本手法の有効性を確認した．

より厳密な評価については今後の課題である．

参考文献

[1] T. Funayama, Y. Yamamoto and O. Uchida, “Development of visualization application of tweet data for extracting information in case of disaster,” 2017 15th International Conference on ICT and Knowledge Engineering (ICT&KE), Bangkok, 2017, pp. 1-5. doi: 10.1109/ICTKE.2017.8259620

[2] S. E. Middleton, L. Middleton and S. Modafferi, “Real-Time Crisis Mapping of Natural Disasters Using Social Media,” in IEEE Intelligent Systems, vol. 29, no. 2, pp. 9-17, Mar.-Apr. 2014. doi: 10.1109/MIS.2013.126

[3] T. Sakaki, M. Okazaki and Y. Matsuo, “Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors” Proc. 19th International Conference on World Wide Web (WWW 2010), pp.851-860, Apr. 2010.

[4] A. Yuzawa, H. Ichikawa and A,Kobayashi, “Tweet Discovery for Disaster Information using Multiple Co-occurrence Relation of Words” Proc.IEEE International Workshop on Big Data and IoT Security in Smart Computing(IEEE BITS 2018), pp.1-6, Jun. 2018.