語の複数の共起関係を利用した災害tweet抽出システム

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　〒163–8677 東京都新宿区西新宿 1–24–2
†† 工学院大学情報学部情報通信工学科　〒163–8677 東京都新宿区西新宿 1–24–2
E-mail: †cm17051@ns.kogakuin.ac.jp, ††aki@cc.kogakuin.ac.jp

あらまし　災害発生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．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．筆者らは，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．本稿では，災害発生時の 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手がかり語を判別した．手がかり語集合を用いた tweet 抽出を試み，提案手法の有効性を検証する．

キーワード　Twitter, SNS, 情報抽出, 共起

1. はじめに

マイクロブログサービスの一種である Twitter は，誰でも手軽に情報をリアルタイムに発信や受信することができる．そのため利用者同士でのコミュニケーションや情報の共有が盛んに行われている．2012 年 10 月の時点で，全世界で 1 日 5 億を超える tweet が投稿されており [1]，全世界では約 3 億人が利用している [2].

一方で，2011 年 3 月 11 日に発生した東日本大震災の際において，Twitter は避難場所や被災状況の共有，知人の安否の確認などに活用された [3]．また，2016 年 4 月 14 日に発生した熊本地震の際においては，市長自らが Twitter を用いて，市内の漏水箇所やデマ情報の打ち消しに活用された [4]．さらに，2017 年 7 月 5 日から 6 日にかけて発生した福岡・大分豪雨の際においては，日田市のＪＲ久大線の鉄橋が流失したことを伝えたのは Twitter が最初であったという [5].

しかし，こうした被災地の状況について言及している情報の多くは，人手によって収集が行われていたのが現状であり，災害に関する投稿は増加傾向にあり，さらに困難となっている [6].

本研究では，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発生時には増加する [8] 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．これらの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) を判別する．そして手がかり語集合を基に，災害情報の抽出する手法を提案し，提案手法に基づき実際のツイート判別を行い目的のツイートを抽出するシステムを構築する．

Twitter を対象とした情報を検索，抽出する試みは多数存在する [9] [10] [11]．坂巻ら [10] は，震災時において被災地で今何が求められているのか，という需要の把握を目的に，単純ベイズ分類器を用いて震災に関連する投稿を抽出する手法を提案している．本研究では tweet の教師データを用いていないため，教師データを要する点で異なる．斎藤ら [9] は，災害の情報と，その災害のなすコンテキスト情報を得ることを目的に，語の共起の相互情報量で重み付きグラフを生成し，そのグラフから独立ソース度という指標より災害のコンテキストの情報を抽出する手法を提案している．1 日分のデータを用いて精確さを指向しており，処理対象となる tweet の選別に複数の語集合を想定している．Sakaki ら [11] は，Twitter のリアルタイム性に着目しており，つぶやきをセンサー代わりに，Twitter 上で地震などのイベントをリアルタイムに検出することを目的としている．

これらに対して，本研究では，災害語 1 語のみから，比較的短時間の tweet 集合を用いて，最初に検索対象とするべき関連語を選別することを主目的としており，現段階では問題へのアプローチが異なる．

本論文の構成として，第 2 章で提案モデルについて説明する．第 3 章で提案モデルを実現するための試作システムについて説明をする．第 4 章で本手法の有効性を確認するために実際の tweet を用いて実験と評価を行う．第 5 章で実験結果についての考察を行い，第 6 章でまとめと今後の課題について述べる．

2. 提案手法

2.1 災害情報の定義

災害が発生した際，被災地において何が起きているかという被災地の状況把握につながる情報を入手することが希求されている [4] [5].

これらの情報は，被災地にいる投稿者によって投稿される．「揺れた！」「緊急地震速報きた」といったような被災地に居る投稿者が揺れたことを伝える情報や，「揺れたけど大丈夫だった」「震源地いるけど無事です」といった投稿者の身の安全を伝える内容を投稿した投稿者は，被災地にいる可能性が高く，被災地の状況把握につながる情報を投稿する可能性がある．

本研究では，こうした点を踏まえて「被災地の状況把握につながる情報」に加えて，被災地の状況把握につながる情報を得るための候補として，「投稿者が揺れたことを伝える情報」，「投稿者の安否を伝える情報」をまとめて「災害情報」と呼ぶ．

2.2 概　　要

本研究では，複数の共起関係を利用して手がかり語を判別し，手がかり語を用いて災害情報の抽出する手法を提案する．

地震が起きた際に「揺れ」のような語を災害に関連する語と想定し，これを手がかり語と呼ぶ．手がかり語を含む tweet には，災害情報が含まれていると推定し，手がかり語を含む tweet(災害 tweet) の抽出を目的としている．

このとき，手がかり語は災害の発生時刻や場所によって表現が異なり，予測することは難しく，事前に手がかり語集合を準備することは困難である．そのため，tweet 自身から自動的に抽出されるべきである．

そこで本研究では，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．また，人と人とのやりとりが災害発生時には増加する [4] 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．これらの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) を判別し，手がかり語集合を基に，災害情報の抽出を行う．本手法の全体像を図 1 に示す．

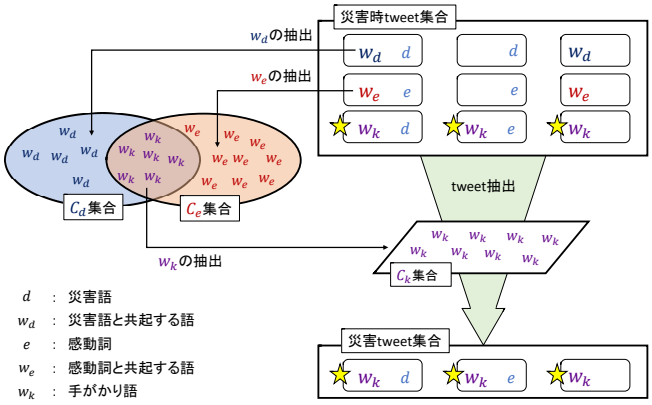

図 1　提案手法の概要図

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，図 1 に描かれている角丸四角形は tweet としている．

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」「こんにちは」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．

手がかり語集合 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．

これらの状態をもとに，災害時 tweet 集合を対象に，災害情報を含むと推定される手がかり語を含む tweet を抽出を行う．これを災害 tweet と呼ぶ．

2.3 災害時 tweet の収集

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 集合としているが，その範囲内に存在する全ての tweet を得るのは困難である．無償利用できる Twitter の streaming API を利用しても全体の 1%しか得られない．そこで，災害時 tweet 集合の近似集合を得るための工夫を行う．

Twitter の streaming API を用いて得られる全体の 1%の tweet 集合を災害時部分 tweet 集合 T_d と呼ぶ．災害時部分 tweet 集合 T_d の各 tweet t_d を対象に，災害語 d を含む tweet t_q を投稿した投稿者は，それ以降に災害情報を投稿する可能性が高いと考え，災害語を含む tweet t_q を投稿した時刻よりも後に投稿された tweet t_s を災害時部分後続 tweet 集合 T_s と呼び，収集を行う．

災害時 tweet 集合，災害時部分 tweet 集合，災害時部分後続 tweet 集合は図 2 のような関係となる．図 2 は，横軸を投稿者，縦軸を時間であり，全ての投稿者の TimeLine に見立てている．また，白色四角形は tweet であり，時刻 at_1 は災害発生前の任意の時刻を，時刻 at_2 は災害発生後の任意の時刻を表している．緑色四角形は災害時 tweet 集合であり，青色四角形は災害時部分 tweet 集合，赤色四角形は災害時部分後続 tweet 集合である．

しかし，災害時部分後続 tweet 集合を充分に活かせるだけの分析結果が得られなかったため，以後の処理では混ぜ，災害時部分 tweet 集合と災害時部分後続 tweet 集合の和集合を，災害時 tweet 集合の近似集合とみなして扱う．

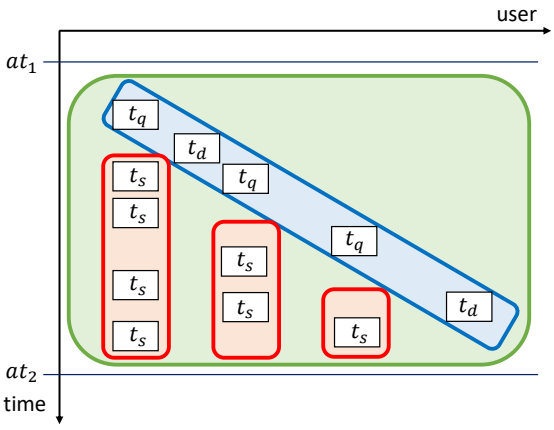

図 2　各 tweet 集合間の関係

2.3.1 災害時部分 tweet の収集

Twitter の streaming API を用いて，災害発生時の一定時間範囲内で Twitter に投稿された tweet を収集する．

収集した tweet 集合を T_d とし，各 tweet を t_d として表す．

$$T_d = \{t_d \mid \text{createdtime}(t_d) \in [at_1, at_2]\} \qquad (1)$$

ここで，createdtime(t) は tweet t が投稿された時刻を得る関数である．時刻 at_1 は災害発生直前の任意の時刻を，時刻 at_2 は，災害発生後の任意の時刻であり，別に設定される．

2.3.2 災害時部分後続 tweet の収集

tweet 集合 T_d を対象に，tweet 本文中に災害語 d を含む tweet を抽出する．抽出した tweet 集合を T_q とし，各 tweet を災害語 d を含む tweet t_q として表わす．

$$T_q = \{t_q \mid t_q \in T_d, d \in t_q\} \tag{2}$$

同一の投稿者によって，災害語を含む tweet $t_q \in T_q$ の投稿時刻よりも，時間的に後に投稿された tweet を収集する．これを災害時部分後続 tweet と呼ぶ．

$\forall t_q \in T_q$ の災害時部分後続 tweet 集合を

$$\mathrm{succT}(t_q, \mathrm{createdtime}(t_q) \in [at_1, at_2]) \tag{3}$$

とする．

$\forall t_q \in T_q$ についての災害時部分後続 tweet 集合の和集合を

$$T_s = \bigcup_{t_q \in T_q} \mathrm{succT}(t_q, \mathrm{createdtime}(t_q) \in [at_1, at_2]) \tag{4}$$

とする．

2.3.3 災害時 tweet 集合

(1) 式と (4) 式より，収集された災害時部分 tweet 集合 T_d と災害時部分後続 tweet 集合 T_s の和集合を災害時 tweet 集合として，

$$T = T_d \cup T_s \tag{5}$$

とする．

2.4 手がかり語の判別

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_e の積と取ることで災害に関連する語 (手がかり語) が得られると仮定し，C_d と C_e の積集合を得る．

また，災害発生時と平常時とを比較して災害発生時によく出現する語は手がかり語であると仮定し，よく出現する語であるか否かの判定に χ^2 値を用いる．

災害時 tweet 集合と平常時 tweet 集合とで出現する語に対して，各 tweet 集合の語の出現頻度を用いて，災害時 tweet 集合と平常時 tweet 集合との間の χ^2 値を求める．このとき，χ^2 値が十分大きな値となる場合，災害発生時か平常時のどちらかかでよく出現する語であると言える．

なお，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指標として用いている．そのため，背景にある分布などを無視している．

手がかり語の基準として，語の出現頻度，χ^2 値をもとに，

- 災害時において，語の出現頻度が平常時と比べて高い語
- χ^2 値で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選ぶ．

2.4.1 災害語と共起する語の収集

手がかり語を得るために，対象とする tweet 集合を T_x とする．ただし，tweet 集合 T_x は任意の tweet 集合とする．

tweet 集合 T_x の各 tweet t_x について，tweet 本文中に災害語 d を含む tweet を対象に形態素解析を行い，災害語 d 以外の災害語と共起する語 w_d を抽出する．

抽出した災害語と共起する語 w_d の語集合を

$$C_d = \{w_d \mid w_d, d \in t_x, w_d \neq d\} \tag{6}$$

とする．

2.4.2 感動詞と共起する語の収集

tweet 集合 T_x の各 tweet t_x について，tweet 本文中に感動詞 e を含む tweet を対象に形態素解析を行い，感動詞 e 以外の感動詞と共起する語 w_e を抽出する．

抽出した感動詞と共起する語 w_e の語集合を

$$C_e = \{w_e \mid w_e, e \in t_x, w_e \neq e\} \tag{7}$$

とする．

2.4.3 平常時 tweet の収集

災害時 tweet 集合との語の出現頻度の比較対象として，対象とする災害が発生していない時 (平常時) の tweet を収集する．

災害発生前日の一定時間範囲内で Twitter に投稿された tweet を収集し，平常時 tweet と呼ぶ．

平常時 tweet 集合を T_{nd} とし，各 tweet を t_{nd} として表す．

$$T_{nd} = \{t_{nd} \mid \mathrm{createdtime}(t_{nd}) \in [at_3, at_4]\} \tag{8}$$

時刻 at_3，at_4 は，災害発生前日の任意の時刻で別に設定される．

2.4.4 手がかり語集合

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_e の積集合 $C_d \cup C_e$ を対象に，手がかり語 C_k の判別を行う．

$$C_k = \{w_k \mid C_d \cup C_e, \mathrm{cond.}\} \tag{9}$$

条件 cond. は，

（1） $\mathrm{normalize}(T_x, w_k) > \mathrm{normalize}(T_{nd}, w_k)$

（2） 単語 w_k の χ^2 値を降順に並べた際の上位 M 件

ここで，M は語数であり，別に設定される．

$\mathrm{normalize}(T, w)$ は，1tweet あたりの単語 w の出現頻度を得る正規化関数である．1tweet あたりの出現頻度へと tweet 数 $|T|$ を用いて正規化する．(10) 式で定義される．

$$\mathrm{normalize}(T, w) = \frac{\mathrm{freq}(T, w)}{|T|} \tag{10}$$

$\mathrm{freq}(T, w)$ は，tweet 集合 T に出現する単語 w の出現頻度を得る関数である．

ある語の出現頻度が災害前後で統計的な差異がどの程度であるかを見るために χ^2 値を用いる．単語 w の χ^2 値は，w の災害前後の出現頻度と，災害前後の全ての語の出現頻度とを用いて (11) 式に示すとおり定義される．

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \tag{11}$$

ここで，r は tweet 集合の個数を示し，T_x と T_{nd} の 2 つの tweet 集合を対象とするため，$r = 2$ とする．c は単語種類数であり，異なる tweet 集合間で単語 j の偏りの程度を示す．単語 j と単語 j 以外の単語を対象とするため，$c = 2$ とする．n_{ij} は tweet 集合 i における単語 j の出現頻度であり，$\mathrm{freq}(T, w)$ と同義である．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 (12) 式で算出を行う．

$$E_{ij} = n_{i.} * \frac{n_{.j}}{N} \tag{12}$$

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．

2.5 災害 tweet の抽出

災害時 tweet 集合 T を対象に，手がかり語 w_k を含む tweet を抽出し，これを災害 tweet と呼ぶ．

災害 tweet 集合を

$$T_k = \{t_k, t_k \in T, w_k \in t_k\} \tag{13}$$

とする．

3. 試作システム

3.1 概　　要

本研究では提案手法を適用し，災害時 tweet 集合を対象に手がかり語を含む tweet の抽出を行う．

災害語 d をシステムに入力すると，災害時部分 tweet 集合 T_d を対象に形態素解析を行い，災害語 d と共起する語集合 C_d，感動詞 e と共起する語集合 C_e を得る．そして，C_d と C_e の積集合を対象に χ^2 値を求め，手がかり語 w_k を判別する．そして，災害時 tweet 集合 T を対象に，手がかり語集合 C_k の中に含まれる語 w_k のいずれかを含む tweet の抽出を行う．

3.2 災害時 tweet 集合の収集

災害時部分 tweet 集合 T_d と平常時 tweet 集合 T_{nd} を収集するために Twitter の Streaming API(statuses/sample) [14] を用いた．また，日本語で書かれている tweet のみを対象とした．Streaming API(statuses/sample) は全 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．

災害時部分後続 tweet 集合 T_s は Twitter の REST API(statuses/user_timeline) [15] を用いて収集を行い，日本語で書かれている tweet のみを対象とした．

また，リツイート・引用リツイートは単語の共起頻度に影響を及ぼす可能性を考慮して除外した．

3.3 手がかり語集合の収集

手がかり語集合 C_k を得るための tweet 集合 T_x は，災害時部分 tweet 集合 T_d を対象とした．

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_e を抽出するために，形態素解析器として MeCab [12] を使用した．システム辞書は mecab-ipadic-NEologd [13] を使用し，2017 年 12 月 13 日に更新したものを利用した．また，ストップワードとして，URL や“@”をはじめとする英字や記号で構成されるものは不要とし除外した．また，品詞は { 名詞，動詞，形容詞 } のいずれかに該当するものとした．

4. 評 価 実 験

4.1 実 験 目 的

本手法の有効性を明らかにするために，tweet の抽出精度で評価を行う．比較対象となる各手法として，

（1） 災害共起頻出語手法
（2） 災害共起 χ^2 語手法
（3） 感動詞共起 χ^2 語手法
（4） 災害共起非含意語手法
（5） 感動詞共起非含意語手法

の 5 種類の手法を用いて提案手法との比較を行う．

これらの各比較手法は，図 1 に示す提案手法の災害時 tweet 集合を対象に，災害 tweet を抽出する際の語集合を変更した手法である．また，各比較手法の語集合の関係を図 3 に示す．各比較手法についての説明を述べる．

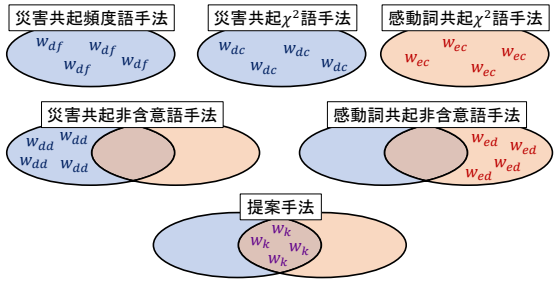

図 3　各比較手法における語集合の関係

災害共起頻出語手法

災害語と共起する語 w_d の語集合 C_d を対象に，$\mathrm{freq}(T_x, w_d)$ より災害語と共起する語 w_d の出現頻度を求める．w_d の出現頻度を降順に並べた上位 M 件の語を，災害共起頻出語 w_{df} と呼び，その語集合を C_{df} とする．そして，災害時 tweet 集合 T を対象に，災害共起頻出語 w_{df} を含む tweet の抽出を行う．

災害共起 χ^2 語手法

災害語と共起する語 w_d の語集合 C_d を対象に，(11) 式より災害語と共起する語 w_d の χ^2 値を求める．w_d の χ^2 値を降順に並べた上位 M 件の語を，災害共起 χ^2 語 w_{dc} と呼び，語集合を C_{dc} とする．そして，災害時 tweet 集合 T を対象に，災害共起 χ^2 語 w_{dc} を含む tweet の抽出を行う．

感動詞共起 χ^2 語手法

感動詞と共起する語 w_e の語集合 C_e を対象に，(11) 式より感動詞と共起する語 w_e の χ^2 値を求める．w_e の χ^2 値を降順に並べた上位 M 件の語を，感動詞共起 χ^2 語 w_{ec} と呼び，語集合を C_{ec} とする．そして，災害時 tweet 集合 T を対象に，感動詞共起 χ^2 語 w_{ec} を含む tweet の抽出を行う．

災害共起非含意語手法

災害語と共起する語 w_d の語集合 C_d と感動詞と共起する語 w_e の語集合 C_e との差集合 $C_d \cup \overline{C_e}$ を対象に，(11) 式より χ^2 値を求める．χ^2 値を降順に並べた上位 M 件の語を，災害共起 diff 語 w_{dd} と呼び，語集合を C_{dd} とする．そして，災害時 tweet 集合 T を対象に，災害共起非含意語 w_{dd} を含む tweet の抽出を行う．

感動詞共起非含意語手法

災害語と共起する語 w_d の語集合 C_d と感動詞と共起する語 w_e の語集合 C_e との差集合 $\overline{C_d} \cup C_e$ を対象に，(11) 式より χ^2 値を求める．χ^2 値を降順に並べた上位 M 件の語を，感動詞共起 diff 語 w_{ed} と呼び，語集合を C_{ed} とする．そして，災害時

表 1　各手法による tweet 抽出結果

手法	合計	正解数	正解割合
提案手法	2228	348	0.156
災害共起頻出語手法	4959	368	0.074
災害共起 χ^2 語手法	816	17	0.021
感動詞共起 χ^2 語手法	2333	334	0.143
災害共起非含意語手法	218	11	0.050
感動詞共起非含意語手法	263	0	0

tweet 集合 T を対象に，感動詞共起非含意語 w_{ed} を含む tweet の抽出を行う．

4.2　実 験 条 件

2016 年 6 月 16 日 14 時 21 分頃に北海道函館市で発生した震度 6 弱の地震を対象とする．

災害時部分 tweet 集合として，地震発生 1 分前の 14:21:00 から 15:59:59 の間に投稿された 29670 件を収集．災害時部分後続 tweet 集合として，地震発生 1 分前の 14:21:00 から 15:59:59 の間に投稿された 1019 件を収集．合計 30689 件の tweet を災害時 tweet 集合とする．

また，地震発生前日の 2016 年 6 月 15 日 14:21:00 から 15:59:59 の間に投稿された 28792 件の平常時 tweet を収集．これらの tweet 集合を実験に用いる．

各パラメータとして，災害語 $d =$ “地震”，語数 $M = 10$，at_1 の時刻を 2016 年 6 月 16 日 14:21:00，at_2 の時刻を 2016 年 6 月 16 日 15:59:59 とした．

精度での比較を行うために，著者 1 名が災害情報であるか否かを人手で判断を行い，人が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．

4.3　実 験 結 果

4.3.1　各手法の推定精度の比較

提案手法で災害 tweet の抽出を行った場合と，各比較手法で災害 tweet の抽出を行った場合とでの抽出精度の結果を表 1，図 4 に示す．

表 1 は，各手法における，抽出された tweet 数 (合計)，うち人手により災害情報を含む正解とされた tweet 数 (正解数)，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．

図 4 は，表 1 の横軸を正解 tweet 数，縦軸を正解割合としてプロットしたものである．この図 4 の縦軸は検索精度を表している．一方，横軸は再現率に相当する軸として用意した．ただし，本実験では一部の tweet のみが分析対象であるため，真の正解集合全体を得ることはできない．そのため，実験データセット中に見いだされた正解 tweet 数を横軸にとっている．

この結果，図 4 のグラフにおいては右上にいくほど良い結果であると言える．

表 1，図 4 より，提案手法が正解割合において 0.156 と最も良い結果を得た．正解数では，提案手法は 348 件であるのに対して，災害共起頻出語手法では 368 件と 20 件多く抽出される結果となった．

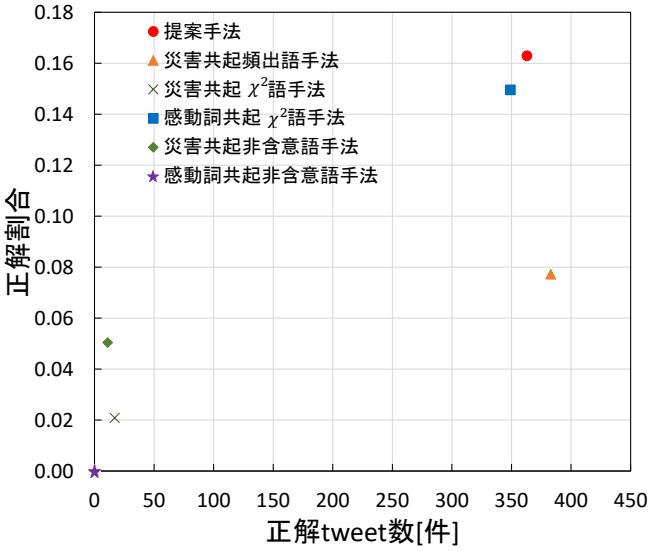

図 4　各手法における正解 tweet 数とその精度

5.　考　　察

5.1　各手法における推定精度の考察

提案手法および各比較手法において，表 1，図 4 の結果を詳しく分析するため，災害 tweet を抽出する際に用いた語集合の内訳を表 2〜7 に示す．

表 2〜7 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 T を対象に該当する語を含む tweet が抽出された tweet 数 (w を含む tweet 数)，うち人手により災害情報を含む正解とされた tweet 数 (正解数)，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．

提案手法における手がかり語集合 C_k の内訳である表 2 より，震源地であり地名を示す「北海道」「函館」，投稿者の安否確認や被災地を心配していることを伝える「大丈夫」「心配」，投稿者が揺れを感知したことを伝える「揺れ」といったような語が得られた．これは，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_e の積集合を取ることで，日常の文脈で出てくるような語 (表 3 の「あっ」や表 5 の「ええ」) を除去できたため，災害に関連するような語のみが得られた．その結果，提案手法は各比較手法よりも正解割合を上昇できたと考えられる．

また，災害共起頻度語手法の内訳である表 3 は，「あっ」「ない」「てる」のような語が得られた．これは，単純に災害語と共起する語の頻出上位 10 件を取ったためだと考えらる．「てる」といったような語は，通常の日常の文脈で使われているような語であるため，災害情報以外の tweet を多く抽出したため，正解割合を低下させたのだと考えられる．

一方で，災害共起 χ^2 語手法の内訳である表 4 より，χ^2 値が「おく」「クラス」「画像」「避難」のような語が得られた．これは，平常時 tweet 集合 T_{nd} に災害語 d(地震) を含む tweet が 9 件とかなり少なかったのが原因であると考えられる．平常時 tweet 集合 T_{nd}，(11) 式である χ^2 の見直しを行う必要があると考えられる．

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	正解数	正解割合	χ^2
北海道	767	44	0.057	580.674
大丈夫	930	110	0.118	390.111
函館	399	75	0.188	304.927
揺れ	407	220	0.5	274.975
震度 6 弱	173	19	0.110	148.057
震度 6	120	15	0.125	95.966
津波	120	19	0.158	94.900
心配	235	38	0.162	76.897
震度	81	14	0.173	63.659
余震	87	21	0.241	58.277

表 3　災害共起頻出語手法による語集合

単語 w_{df}	w_{df} を含む tweet 数	正解数	正解割合	$\mathrm{freq}(T_d, w_{df})$
北海道	767	44	0.057	264
大丈夫	930	110	0.118	203
あっ	449	31	0.069	102
函館	399	75	0.188	74
揺れ	407	220	0.5	51
心配	235	38	0.162	41
ない	866	24	0.028	39
震度 6 弱	173	19	0.110	37
怖い	158	20	0.127	34
てる	2129	46	0.022	34

表 4　災害共起 χ^2 語手法による語集合

単語 w_{dc}	w_{dc} を含む tweet 数	正解数	正解割合	χ^2
おく	30	1	0.033	17.287
クラス	27	1	0.037	17.287
画像	183	1	0.005	17.287
避難	13	5	0.385	17.287
地震情報	22	0	0	12.490
周辺	21	1	0.048	12.490
あたり	56	3	0.054	12.490
多い	161	3	0.019	11.290
昨日	152	1	0.007	9.619
ツイート	220	2	0.009	9.619

表 5　感動詞共起 χ^2 語手法による語集合

単語 w_{ec}	w_{ec} を含む tweet 数	正解数	正解割合	χ^2
北海道	767	44	0.057	51.216
揺れ	407	220	0.5	34.573
函館	399	75	0.188	26.433
大丈夫	930	110	0.118	25.783
心配	235	38	0.162	13.809
うさ	18	0	0	10.162
震度 6 弱	173	19	0.110	10.162
当日	24	0	0	10.162
ええ	53	0	0	9.486
うち	197	25	0.127	9.422

表 6　災害共起非含意語手法による語集合

単語 w_{dd}	w_{dd} を含む tweet 数	正解数	正解割合	χ^2
地震情報	22	0	0	3.078
東部	23	0	0	0.803
21 分	15	2	0.133	0.455
函館市	38	4	0.105	0.455
内浦湾	98	3	0.031	0.455
深さ	82	1	0.012	0.398
最大震度	59	0	0	0.398
防災	13	0	0	0.398
東北	17	2	0.118	0.341
北海道函館市	9	0	0	0.341

表 7　感動詞共起非含意語手法による語集合

単語 w_{ed}	w_{ed} を含む tweet 数	正解数	正解割合	χ^2
当日	24	0	0	10.726
うさ	18	0	0	10.726
ええ	53	0	0	10.107
グレイ	9	0	0	8.580
スタンド	40	0	0	7.507
くださ	7	0	0	7.507
課金	33	0	0	7.507
泣き	44	0	0	6.979
美味しい	57	0	0	6.886
ポケモン	28	0	0	6.435

感動詞共起 χ^2 語手法の内訳である表 5 より,「北海道」「揺れ」「函館」「大丈夫」「心配」といった提案手法と同じような語が得られた．これは，災害時において人と人とのやりとりが増加する [8] ことから，感動詞である「ありがとう」や「こんにちは」という語と共起したため,「北海道」や「揺れ」といった語が得られたのだと考えらえる．しかし,「うさ」「ええ」「当日」のような災害とは無関係な語が得られていることから，感動詞と共起する語のみでは災害に関連する語のみを得るのは困難であると考えられる．

災害共起非含意語手法の内訳である表 6 より,「函館市」「地震情報」といった語が得られた．これらの語を含む tweet のほとんどは地震速報 bot による各地の震度に関する情報であった．

感動詞共起非含意語手法の内訳である表 7 より,「スタンド」「くださ」といった語が得られた．感動詞と共起するが災害語とは共起しない語は，各語の正解と正解割合を見る限り．災害とは無関係な語であることが分かる．

5.2　抽出された一部の tweet

表 8 は，それぞれの tweet において，ID，災害情報を含むか否かを○×で表し (正解)，その tweet の本文を示す．また，tweet 本文中に共起語を下線で示す，投稿者のユーザ名は黒四角形記号で伏字としている．ID は各 tweet の通し番号の役割を果たす．

表 8 より，ID1-1，1-2，1-3 のように,「揺れ」や「震度 6 弱」といった語は，その語自身が地震に関連する語であったため，

表 8　提案手法により抽出した一部の tweet の内容

ID	正解	tweet 本文
1-1	○	いきなりめっちゃ揺れ てびびった
1-2	○	まず音がすごくて直後に 揺れ て、そのあとに緊急地震速報が来た。震源地から近いと仕方が ない よね。川汲、朝から 揺れ てたね。前震だったのかな。
1-3	○	@■ 震度 6 弱 だったけど 大丈夫 だよ !!! ありがとう !!
1-4	×	揺れ た地域の方 大丈夫 ですか ?? 余震や身の回り気を付けてくださいね…
1-5	×	また地震 !? しかも 北海道ってもう北も南も日本 揺れ すぎてるよ…地下が恐ろしい…。もうすぐ来るのかなぁこっちも…
2-1	○	こちら全然 揺れ てません！揺れ てないけど関東の本社や支社から安否確認のお電話が何件も入ってます！あのね関東のみなさん、北海道 は広いんだよ！
2-2	○	北海道 にいますが旭川なので全然 大丈夫 です
2-3	○	@■ 函館 凄い 揺れ たよ (絵文字) ■■ちゃんファミリー 大丈夫？
2-4	○	@■ @■ 北海道 広すぎて 函館 で地震来ても札幌とか旭川民からしたら何が起きたか分からない現実 (顔文字) でも札幌のテレビはほとんど地震のことで予定変更され てる！
2-5	×	ついに 北海道 にも来たのか。。
3-1	○	地震 あっ たのね。安否確認メール来ててびっくり。
3-2	○	そこそこ人数多い会社な上に通信機器を一人一台以上の割合で持ってる ので、地震速報鳴ると大音量になってビビる…
3-3	×	5 回くらい負け てる んで今日サークル行きません
4-1	○	@■ ■■ちゃん、私の所は余り 揺れ なくてホッとしたよ。全部の警報音が鳴ってオロオロしちゃった。心配 してくれてありがとう (絵文字)
4-2	×	@■ 函館、かなり 揺れ があったそうですが、大丈夫 でしょうか ??
4-3	×	@■ 北海道 で大きな地震があったみたいだけど 大丈夫？
4-4	×	@■ 大丈夫 で良かったです。(T_T) 心配 になりました。(T_T)
4-5	×	@■ ありがとうございます、大丈夫 ですよ～！当日お待ちしております (絵文字)!!

災害に関連する語を含む tweet，つまり災害情報が正しく抽出することができた．しかし，1-4，1-5 のように，「揺れ」という語を含んでいるが，揺れた地域を心配している投稿者による二次災害への注意喚起を伝える内容であったり，いつ投稿者の住んでいる地域に地震が起こるのか不安を伝える内容も抽出された．

また，「揺れ」や「震度 6 弱」といった災害語に対応する語のみならず，「北海道」や「函館」といった震源地である地名を表わす語も自動的に抽出された．これは，災害発生時において，投稿者が「どこ」にいてその投稿者自身の安否を伝える内容や，「どこで」地震が起きたのかを伝える内容の投稿が多かったため，「北海道」や「函館」といった特有の語が得られたのだと考えられる．そのため，ID2-1，2-2，2-3，2-4 のような tweet が得られた．

ID4-1，4-2，4-3，4-4 のように，投稿者間での安否の確認が行われていることが確認された．そのため，4-2，4-3，4-4 のような tweet を受け取った人は，被災地にいる可能性が高いと考え，このメンション関係は災害情報の精度を高められるのではないかと考えられる．これらのメンション関係の活用については今後の課題である．

一方で，表 3 より，「あっ」「ない」「てる」といった通常の日常の文脈で使われている語を含む tweet は ID3-1，3-2，3-3 となる．これは，「あっ」といった語自身では意味を持たないが，それと共起する「安否確認」「地震速報」といった語が災害語と対応する災害に関連する語であったため，災害情報が得られたのだと考えられる．「あっ」や「ない」といった語の活用については今後の課題である．

5.3　機械判別で目指すべき水準

前提条件的には提案手法よりも楽をしているが，機械判別で目指すべき水準として可能性がある数字の例として挙げる．

Support Vector Machine(SVM) を用いて 2 値分類を行う．SVM の実装として，Python の機械学習ライブラリである scikit-learn の LinearSVC [16] を用いる．カーネルは線形カーネルとして，パラメータはデフォルト値である C=1.0 とした．5 分割交差検証により推定精度の評価を行う．

学習データは，提案手法で抽出した人手で災害情報か否か判別済み 2228 件の tweet を使用する．SVM に与える素性として，tweet を MeCab により形態素解析を行い，品詞が { 名詞，動詞，形容詞 } のいずれかに該当する語の Bag-of-Words とした．ただし，システム辞書は mecab-ipadic-NEologd を使用し，2017 年 12 月 13 日に更新したものを利用した．また，ストップワードとして，URL や “@” をはじめとする英字や記号で構成される語は不要とし除外した．

評価尺度として，precision, recall, f-measure を設ける．

SVM による 5 分割交差検証の結果を表 9 に示す．Precision は約 0.64，Recall は約 0.53，F-measure は約 0.57 という結果が得られた．このことから，precision である約 0.64 が機械判別で目指すべき水準として可能性がある数値となった．

表 9　SVM による判別結果

Precision	Recall	F-measure
0.6362	0.5283	0.5719

6. 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を判別し，手がかり語集合を用いた tweet 抽出システムを提案した．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．

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．

文　　献

[1] CNET : Report: Twitter hits half a billion tweets a day, https://www.cnet.com/news/report-twitter-hits-half-a-billion-tweets-a-day/, (参照 2018-01-08)

[2] THE HUFFINGTON POST:Twitter が国内ユーザー数を初公表 「増加率は世界一」，http://www.huffingtonpost.jp/2016/02/18/twitter-japan_n_9260630.html，(参照 2018-01-08)

[3] 河井 孝仁，藤代 裕之，“東日本大震災の災害情報における Twitter の利用分析”，広報研究 = Corporate communication studies，Vol.17，pp.118-128(2013)．

[4] 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 2018-01-08)

[5] 大分合同新聞：県、ツイッター活用　　幅広く災害情報収集，https://www.oita-press.co.jp/1010000000/2017/07/31/JD0055998437，(参照 2018-01-08)．

[6] 毎日新聞：ツイッター投稿、１週間で２６１０万件，http://mainichi.jp/articles/20160519/k00/00m/040/059000c，(参照 2018-01-08)．

[7] 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)．

[8] Miyabe,M.,Miura,A.,and Aramaki,E.:Use Trend Analysis of Twitter after the Great East Japan Earthquake, Proc. 2012 ACM conference on Computer Supported Cooperative Work(CSCW' 12), pp.175-178,(2012).

[9] 斎藤 翔太，伊川 洋平，鈴木 秀幸，村上 明子，“Twitter を用いた災害情報の早期発見”，電子情報通信学会技術研究報告．NLC，言語理解とコミュニケーション，Vol.114，No.81，pp.7-12(2014)．

[10] 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-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，Vol.65，No.1，pp.39-50(2014)．

[11] Takeshi, S. Makoto, O. and Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors, Proc. 19th International Conference on World Wide Web (WWW 2010), pp.851-860,(2010).

[12] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://taku910.github.io/mecab/, (参照 2018-02-07)．

[13] mecab-ipadic-NEologd: Neologism dictionary for MeCab, https://github.com/neologd/mecab-ipadic-neologd, (参照 2018-02-07)．

[14] Twitter: GET statuses/sample, https://developer.twitter.com/en/docs/tweets/sample-realtime/api-reference/get-statuses-sample , (参照 2018-02-07)．

[15] Twitter: GET statuses/user_timeline, https://developer.twitter.com/en/docs/tweets/timelines/api-reference/get-statuses-user_timeline , (参照 2018-02-07)．

[16] scikit-learn, sklearn.svm.LinearSVC, http://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC, (参照 2018-02-07)．