マイクロブログにおける感動詞との共起を利用した検索語抽出システム

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　マイクロブログからの災害状況抽出では，初期段階で適当な検索語群によるフィルタリングが用いられることが多いが，この語群は人手で設定されている．災害毎に異なる検索語群を現場でリアルタイムに思いつくことは難しく，自動化が求められる．筆者らは，マイクロブログサービス上で感動詞が人と人の間での価値ある情報のやりとりの存在を示唆する品詞であることに着目し，代表語だけでなく投稿中の感動詞の共起語など複数の共起関係も利用して検索語集合を抽出する研究を行なっている．しかし，検索語の候補語集合を得る際に，単純に語集合の積を取っており，偶然に代表語や感動詞と共起したため，災害とは関係のない語が選別された．本稿では，語の出現頻度に加えて共起頻度を用いて共起語の関連度を算出することで，尤もらしい災害と関連する語のみの抽出を試みる．最近の異なる種類の大規模災害時の実 tweet データに対して本手法を適用し，妥当な検索語の抽出を検証する．

キーワード　Twitter, マイクロブログ, 感動詞, 災害, 検索語

1. はじめに

マイクロブログの一種である Twitter は，tweet と呼ばれる 140 文字以内の短い文章を投稿するサービスである．その手軽さからユーザ自身の状況やニュースといった実世界の情報がリアルタイムに投稿されている．2017 年末の時点で，3 億 3,000 万人の月間アクティブユーザー数を記録している (注1)．

マイクロブログの代表といえる Twitter は災害時の情報インフラとしても重要視される．例えば，2011 年 3 月 11 日に発生した東日本大震災では，避難場所や被災状況の共有，知人の安否の確認等，リアルタイムに情報伝達する手段の 1 つとして活用され [1]，Twitter は重要な情報インフラとして扱われている．

そこで，Twitter を情報源とする災害発生時の状況分析について研究されている [2]- [3]．これらの研究は，フィルタリングの第一段階として災害に関連する投稿を検索するために，人手によって検索語群を事前に用意し，それらの語を含む投稿のみを分析対象としている．検索語群の選定の作業は災害の種類や被害状況毎に大きく異なるため，災害から相当時日の経過した事後の分析では問題とはならなくとも，発災後の混乱期に語群を適切に準備することは困難である．そのため，よりリアルタイムに近い形で自動的に検索語群を得ることができれば，実運用上の活用につながる可能性がある．

本研究では，マイクロブログサービスの特性であるリアルタイムでの情報配信を活かし，発災後間もない時間帯での被害状況を含めた情報収集を視野に入れた，災害に関連する投稿を検索する検索語群を自動的に抽出することを目的とする．提案手法では，災害を代表する 1 語のみを入力としてマイクロブログサービス上の投稿を逐次処理することで，検索に役立つと考えられる検索語集合を抽出する．マイクロブログサービスは人と人とのコミュニケーション手段であるため，有意義な情報を得たと感じた人がその反応として品詞が感動詞に該当する単語を含む投稿文を投稿する点に着目する．これは災害発生時のコミュニケーションメディアとして，草の根の情報交換の場として機能している Twitter などのマイクロブログの投稿文ならではの特徴的な語彙の利用方法であるといえる．

著者らの従来法 [4] では，代表語との共起語集合を抽出する際に，同一 tweet 内で代表語と語 w が同時に出現するか否かを二値判断する．しかし，同一 tweet 内で代表語の有無によって語 w の出現頻度が偏るような語のほうが，代表語との共起語として尤もらしいと考えるのが妥当である．そこで，統計的に有意に代表語と共起しやすい語であるか否かを判断するために，対数尤度比検定 (Log-Likelihood Test) の考え方を流用する．帰無仮説として，「代表語を含む tweet における単語 w の出現頻度と代表語を含まない tweet における単語 w の出現頻度は等しい」を設定し，この帰無仮説を棄却された単語を，有意に代表語と共起する語として抽出する．このように本稿では，関連語抽出で多用される語の共起関係に加えて，感動詞との共起関係も利用して検索語集合の抽出する手法を提案する．

2. 関連研究

マイクロブログサービス上の投稿記事から，災害発生時の情報を得ようとする研究は多く存在する．例えば，Sakaki ら [2] は，Twitter を利用しているユーザをセンサーとみなし，「地震」または「揺れ」を含む tweet を収集し SVM（Support Vector Machine）を用いて，地震や台風が発生したことに言及している tweet であるかを判別し，台風の進路や地震の震源地を特定する手法を提案した．また，Maru ら [3] は，Twitter による集合知に基づいたネットワーク制御を自動的/自律的に行なうことを目的に，通信障害を検知し障害の発生場所や電話の接続の可否を抽出している．

(注1)：http://www.viewproxy.com/Twitter/2018/AnnualReport2017.pdf

これらの研究は，人手で検索語群を事前に設定しており，その問題点は前述した．これに対し，本研究の特徴は，人手による入力を災害発生時の災害代表語のみとしても，災害に伴う投稿状況を自動的に分析して，適切な検索語群を抽出できるようにする点にある．

一般の文書集合からの関連語抽出手法の一つとして，中島ら [5] は，趣味や興味に基づいたコミュニティ間での話題の広がりに着目し，ブログ記事データを対象に流行語候補を早期発見する手法を提案した．また，高木ら [6] は，対象とする特定分野に関する詳細な辞書の作成を目的に，Word2vec で学習した単語の分散表現を用いて，対象分野に関する単語を拡張する手法を提案した．

これらの研究は，ブログ記事や Web 記事など比較的長い文章であれば，高い精度を示している．これに対し，Twitter に代表されるマイクロブログサービス上では短文や文法的に整っていない文が多く，通常の文書に対する手法の有効性は劣るとされ，マイクロブログに特化して関連語抽出を行う研究が別に存在する [7]- [8]．

青島ら [7] は，同一の話題について言及する tweet 間のつながりを発見するために，単語の共起関係と時系列上の出現間隔に着目した単語間のつながりの度合いを算出する手法を提案した．新田ら [8] は．与えられたクエリに関連する実世界の情報を抽出するために，長期的・短期的の 2 つの局面での単語間の関係性に着目し単語間の相互情報量を用いて，tweet に対してクエリとの関連度を算出する手法を提案した．

これらの研究は，対象とするイベント毎の特徴を細かく捉えるために多数のパラメータを持ち，運用のためにはこれらを調整する必要があるものの，その調整方法は明らかではない．これに対して，本研究は，災害を代表する 1 語のみを入力としてマイクロブログサービス上の投稿を逐次処理することで，検索に役立つと考えられる検索語集合を抽出する．

代表語を入力として検索語を得る枠組みは，検索クエリ拡張と捉えることができる．宮西らによるマイクロブログ上でのクエリ拡張手法 [9] では，ユーザの意図に適合する文書をより多く検索することを目的に，文書中の単語やコンセプト（複数の単語の組み合わせ）の出現頻度と単語の出現頻度の時間変化に着目し，疑似適合フィードバック手法の考えを用いている．この手法は，マイクロブログサービスがコミュニケーションをもたらすサービスである視点が盛り込まれておらず，十分にマイクロブログを活用しているとはいえない．また，語の共起にのみ着目しており，日常的に使われているような語が結果として得られてしまう可能性が考えられる．これに対して，本手法は，マイクロブログサービス上で感動詞が人と人の間での価値ある情報のやりとりの存在を示唆する品詞であることに着目し，「地震」などの代表語だけでなく感動詞との共起関係も利用して検索語集合を抽出している点で異なる．

マイクロブログ上の投稿から話題を表すトピックを見出そうとする研究も，トピックを話題を示す単語群で表すことから，代表語の関連語集合を得られる可能性を指摘できる．例えば，坂本らの社会事象を抽出する手法は LDA [10] を用いたトピックを利用しており [11]，また，山本らは実生活の局面に対応する複数ラベル名の抽出に LDA によるトピックとそれを表す単語との関係性を用いている [12]．しかし，トピックは複数得られ，また，その個数の設定法は知られておらず，さらに，いずれのトピックが目的とする代表語に関連するトピックであるかは，単純にそのトピックを表す単語群だけで識別できるものではないため，本研究の目的に適うものではない．

3. 提 案 手 法

3.1 用 語 定 義

災害語（D : Disaster word）

「地震」といった 1 語またはごく少数の語集合を指す．本研究では，この災害語のみをシステムへの利用者による入力とするため，少数に限っている理由は，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．

感動詞（I : Interjection word）

挨拶や応答といった「ありがとう」「こんにちは」のような語が属する品詞である．本研究は，災害発生時に人と人とのやりとり（Reply）が増加する傾向にあり [13]，その Reply の本文中には感動詞が含まれていることが多く観測された経験則に基づき，感動詞を利用する．

3.2 概　　要

災害に関連する投稿の検索に有効な語集合を検索語集合と呼び，検索語集合を見つけ出すことが本研究の目的である．特に，災害語のみを用いて，多数の共起語集合の中から検索語集合を自動的に抽出する手法である点を特徴とする．

検索語は，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，災害発生前に検索語集合を準備することは困難である．例えば，災害語「地震」に対応する検索語集合は，津波発生の有無で「津波」を含むか否かが変わりうるといえるし，被災地を示す地名なども当然に災害毎に異なる．また，検索語を抽出するためには，十分な文書量を必要とする一方，災害発生時に時々刻々と変化する状況をその場で分析したい欲求を満たす時間経過の短期間性との兼ね合いが必要である．

そのため，分析に十分な文書量が得られる最短時間での発災以後に投稿された tweet 集合を対象として処理することが望ましい．また，同一災害であっても時間経過と共に被害などの局面は変動していくため，この変化に追従して適切な検索語を抽出することが望ましい．

そこで本システムは，発災以後に投稿された tweet を一定件数まで収集する毎に，その対象 tweet 群から検索語集合を抽出することとした．発災以後に投稿された tweet が一定件数になるところまでを一処理単位として，これを slot と呼び，一定件数まで tweet を収集する毎に，1slot，2slot，…と順に呼ぶ．

本システムは，slot 毎に独立に処理し，検索語集合の抽出を行なう．本システムの概要を図 1 に示す．

本システムの処理の流れとして，まず，発災以後の全 tweet(注2)

(注2)：災害時の運用で言えば発災時刻から現在時刻までの tweet となる

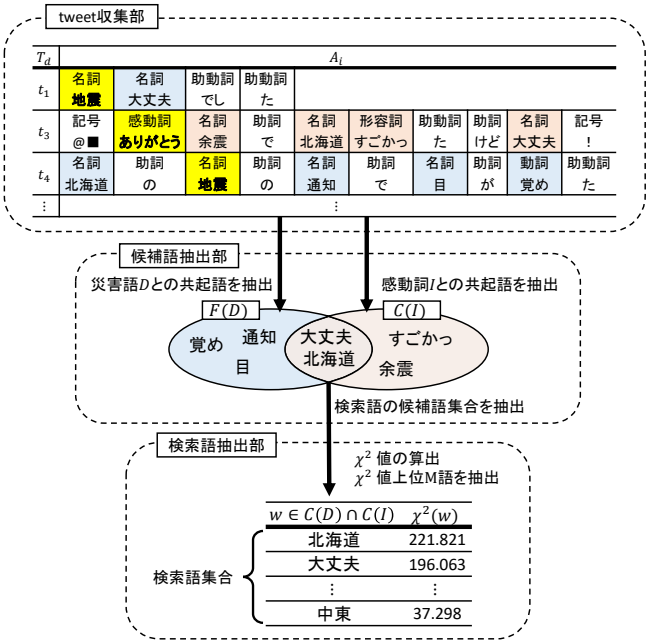

図 1　本システムの概要図

から，重複 tweet を tweet 間の類似度を用いて除去する．次に，この tweet 集合より，tweet 単位でシステムに入力された災害語との共起語と，感動詞との共起語それぞれの語集合を得る．災害語との共起語集合を対象に，災害語との単語間の関連度を対数尤度比 (LLR : Log-Likelihood Ratio) [14] を用いて算出し，統計的仮説検定の枠組で，出現頻度に有意差のある災害語との共起語集合を得る．そして，両者の交わりから検索語の候補集合を得る．この候補より，検索語スコア（χ^2 値；3.5.2 節参照）を算出して高スコアを得た語の集合を抽出結果の検索語集合とする．

本システムは検索語を抽出するために，tweet 収集部と候補語抽出部と検索語判別部の 3 つの処理部により構成されている．

tweet 収集部では，災害発生時に Twitter 上に投稿された tweet を収集する．災害発生時刻を入力し，災害発生時刻以降に投稿された tweet 集合を出力する．このとき，各 tweet 間の類似度を算出し，高類似度 tweet は処理対象から除去する．これは，主に公的 Web サイト情報の引き写しとなる tweet や bot などにより，ほぼ同内容の tweet が多数存在し，検索語抽出に悪影響を及ぼすことが予備実験により判明した [4] ためである．

候補語抽出部では，tweet 収集部で得られた tweet 集合を対象に，検索語の候補となる語集合を抽出する．災害語を入力し，入力された語と感動詞との共起する語を出力する．このとき，同一 tweet 内で災害語の有無によって語 w の出現頻度が偏る語は，災害語と語 w に何かしらのつながりがあると考えられる．そこで，災害語と共起する語集合を対象に，災害語との単語間の関連度を対数尤度比を用いて算出し統計的仮説検定を行なうことで，有意差のある災害語との共起語集合を得る．そして，有意差のある災害語との共起語集合と，感動詞との共起語集合との両者と共起する語を候補語集合と呼ぶ (3.4 節参照)．これは，災害語と感動詞との両者と共起する語が検索語にふさわしいことが多く観測された経験則に基づき，両者と共起する語を検索語の候補語集合としている．

検索語判別部では，候補語集合を対象に，検索語を判別する．候補語集合を入力とし，検索語を出力する．経験則として，平常時と比べて災害時により多く出現する語は検索語として適切であることがほとんどであった．しかし，単純に語の共起頻度や出現頻度の上位語を抽出すると，日常的に使われるような語が抽出されてしまう問題が生じる．そこで，災害発生前後での語の出現頻度に差がある語を抽出するために，カイ二乗検定の考え方を採り入れ，χ^2 値による判定で検索語の抽出を行なう．語の出現頻度，χ^2 値をもとに，語の出現頻度が平常時と比べて災害時により多く出現する語のうち，χ^2 値で降順に並べた際の上位 M 件を検索語として選ぶ．

3.3 節では tweet 収集部について，3.4 節では候補語抽出部について，3.5 節では検索語判別部について述べる．

3.3　tweet 収集部

3.3.1　前　処　理

本システムでは，発災以後に投稿された tweet を一定件数まで収集する．ここで，対象とする災害発生後に収集した tweet 集合を災害時 tweet 集合と呼び，対象とする災害発生前に収集した tweet 集合を平常時 tweet 集合と呼ぶ．

平常時 tweet 集合は，災害時 tweet 集合との語の出現頻度の比較対象として収集しておく必要がある．典型的には災害発生直前や発生前 1 日，あるいは 1 週間前の同曜日などが平常時と捉えられる．実用上は，tweet を常時観測，収集しておいた適当な期間のデータであれば良いと考えている．本手法では災害後の tweet を Streaming API を用いて収集するシステムを想定しており，この仕組みを平常時にも運用しておけば十分であると考えているため，これ以上の詳細な議論は行わない．

slot 毎に災害時 tweet 集合と平常時 tweet 集合を，別に定める N 件まで Twitter の Streagming API を用いて収集する．ただし，リツイート・引用ツイートは，語の出現頻度と共起頻度に影響を及ぼすため，除外した上で収集する．

収集した災害時 tweet 集合を T_d^a，平常時 tweet 集合を T_n^a として表わす．また，検索語の抽出に十分な文章量として，T_d^a と T_n^a の tweet 件数を N として表わす．

また，Tweet t_i を，i 番目の tweet について次の 4 つの情報を含むと定義する．各表記の説明を表 1 に示す．

$$t_i := (\mathcal{T}_i, A_i, S_i, B_i) \tag{1}$$

表 1　t_i の表記の説明

Notation	Meaning
$\mathcal{T}_i$	元の投稿文（テキスト）.
A_i	$\mathcal{T}_i$ を形態素解析した全品詞の語の集合.
S_i	$\mathcal{T}_i$ を形態素解析した品詞が名詞，動詞，形容詞のいずれかに該当する語の集合.
B_i	$\mathcal{T}_i$ を形態素解析した品詞が名詞に該当する語による Bag of words ベクトル.

提案手法では，処理対象とする語が，“@”などの記号，英数字のみから構成される語，URL は不要語として除外する．ま

た，動詞，形容詞に該当する語は活用形のまま用いる．

3.3.2 重複 tweet の除去

予備的手法 [4] による分析の結果，災害時 tweet 集合内にほぼ同文章の tweet が多く存在し，結果として災害とは無関係な語の出現頻度が上昇し，検索語として過剰に抽出され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ている．そこで，文章が重複した tweet を，収集したのみの tweet 集合である T_d^a と T_n^a から自動的に除去する．

T_d^a 内に存在する文章が重複した tweet を除去するために，T_d^a の各 B_i 同士で cos 類似度を算出し，類似度が 1 の tweet を T_d^a から除去する．T_n^a も同様に高類似度の tweet を除去する．

T_d^a，T_n^a から高類似度の tweet 除去後の tweet 集合をそれぞれ T_d，T_n と表わす．

3.4 候補語抽出部

候補語抽出部では，災害時 tweet 集合 T_d を対象に，検索語の候補となる語集合を抽出する．災害語を入力し，有意に災害語と共起する語と感動詞との両者と共起する語を検索語の候補語として出力する．

tweet を単位とした語 w との共起語集合は，

$$C(w) := \{c \in S_i \mid \forall t_i \in T_d, w \in A_i, c \neq w\} \tag{2}$$

のように表される．ただし，$c \in S_i$ は tweet t_i の形態素解析した語列中の品詞が名詞，動詞，形容詞のいずれかに該当する語 c を表わし，$w \in A_i$ は tweet t_i の形態素解析した語列中の全品詞の語 w を表わす．

語集合 $W = \{w\}$ とした場合は，

$$C(W) := \bigcup_{w \in W} C(w), \tag{3}$$

と，各語に対する共起語の和集合を表わすこととする．これを用いると，災害語 D との共起語集合は $C(D)$，感動詞 I との共起語集合は $C(I)$ と表せる．

災害語 D との共起語集合 $C(D)$ から，有意に災害語と共起する語を抽出には，対数尤度比検定の考え方を流用する．帰無仮説として「災害語を含む tweet における単語 w の出現頻度と災害語を含まない tweet における単語 w の出現頻度は等しい」を設定し，対数尤度比検定を行なう．そのために，災害語と共起する語集合 $C(D)$ の各語を対象に，検定統計量として対数尤度比 (LLR : Log-Likelihood Ratio) [14] を計算する．

災害語と共起する語集合 $C(D)$ を対象に，語 w の対数尤度比 $LLR(w)$ は，T_d に出現する災害語 D と語 w の出現頻度，災害語 D と語 w との共起頻度を用いて (4) 式のように計算される．

$$LLR(w) = 2 \sum_{u \in \{D, \bar{D}\}} \sum_{h \in \{w, \bar{w}\}} O_{uh} \log \frac{O_{uh}}{E_{uh}} \tag{4}$$

ただし，u は災害語 D，災害語 D 以外の単語 $\bar{D}$ を示し，h は単語 w，単語 w 以外の単語 $\bar{w}$ を示している．O_{uh} は語 u と語 h との共起頻度，E_{uh} は語 u と語 h との共起頻度の期待値であり，(5) 式のように計算される．

$$E_{uh} = \frac{(O_{uh} + O_{\bar{u}h}) \cdot (O_{uh} + O_{u\bar{h}})}{(O_{uh} + O_{u\bar{h}} + O_{\bar{u}h} + O_{\bar{u}\bar{h}})}. \tag{5}$$

$LLR(w)$ が，自由度 1 の χ^2 分布に従い，有意水準 α%の限界値 β 以上であれば帰無仮説を棄却し，有意に災害語との共起語として抽出する．有意に災害語との共起語集合 $F(D)$ は，

$$F(D) := \{w \in C(D) \mid LLR(w) > \beta\} \tag{6}$$

と表せる．検索語の候補語集合は，$F(D)$ と $C(I)$ の両者ともに含まれる語によって構成し，

$$K = F(D) \cap C(I) \tag{7}$$

と表わす．

3.5 検索語判別部

3.5.1 概　　要

経験則として検索語としてふさわしい語は，災害発生後に投稿中に含まれる頻度が高くなった語であることがほとんどであった．しかし，単純に語の出現頻度や共起頻度の上位語を抽出すると，日常的に使われるような語が抽出されてしまう問題がある．そこで，平常時と比べて災害時にのみ語の出現頻度が高い語を検索語として抽出する．

ここでは，災害発生前後での語の出現頻度に差があるかどうかを判定するために，カイ二乗検定の考え方を流用する．災害発生前後の tweet 集合内で，判定対象の語の出現頻度が異なるかを，語の出現が互いに独立である前提下で検定するのが本来のカイ二乗検定の使われ方である．しかし，ここでは，出現頻度の異なり方がより大きいといえる語を検出することを目的に χ^2 値の大小のみを利用した順位付けを行う．

すなわち，検索語の候補語集合のうち，災害発生前後での出現頻度が災害発生後の方が高い語について，χ^2 値の高い順に一定順位までの語を検索語として抽出する．

3.5.2 検索語の判別

検索語の候補語集合 K を対象に，検索語を判別するために，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測るために χ^2 値を計算する．

語 w の χ^2 値 $\chi^2(w)$ は，災害時 tweet 集合 T_d と平常時 tweet 集合 T_n の，各 tweet 集合に出現する語の出現頻度を用いて (8) 式のように計算される．

$$\chi^2(w) = \sum_{a \in \{T_d, T_n\}} \sum_{b \in \{w, \bar{w}\}} \frac{(n_{ab} - E_{ab})^2}{E_{ab}} \tag{8}$$

ただし，a は各 tweet 集合を示し，b は単語 w，単語 w 以外の単語 $\bar{w}$ を示している．また，処理対象とする語は品詞が名詞，動詞，形容詞のいずれかに該当する語とする．

n_{ab} は tweet 集合 a における単語 b の出現頻度，E_{ab} は tweet 集合 a における単語 b の出現頻度の期待値で，

$$E_{ab} = \frac{n_{a*} \cdot n_{*b}}{N} \tag{9}$$

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の総和とする．

$\chi^2(w)$ 値が十分大きい場合，T_d または T_n のどちらかに偏っ

て語の出現頻度が高い語 w であるといえる．ここでは，T_d に偏って出現する語が望まれる．そのため，χ^2 値が十分大きく，語の出現頻度が T_n と比べ T_d により多く出現する語を検索語として判別する．単語 w の tweet 集合 T における出現頻度 $\mathrm{nfreq}(T,w)$ を

$$\mathrm{nfreq}(T,w) := \frac{\mathrm{freq}(T,w)}{\mid T \mid} \tag{10}$$

として定義する．

$\mathrm{nfreq}(T,w)$ は，tweet 集合 T における 1tweet あたりの語 w の出現頻度を表わす．また，$\mathrm{freq}(T,w)$ は T に出現する単語 w の出現頻度，$\mid T \mid$ は T の tweet の件数を表わす．

すると，検索語の候補語集合で災害発生後のほうが出現頻度の高い語の集合 H は，

$$H = \{w \in K \mid \mathrm{nfreq}(T_d,w) > \mathrm{nfreq}(T_n,w)\} \tag{11}$$

となり，このうち χ^2 値上位 M 件を検索語集合として抽出する．すなわち，求める検索語集合 R は，

$$R = \{w \in H \mid \chi^2(w) \geqq \chi^2_M(R)\}. \tag{12}$$

ここで，$\chi^2_M(R)$ は R 内の語の χ^2 値の上位 M 番目の χ^2 値を示す．

4. 試作システム

関連語抽出手法，災害有意語手法に基づく検索語抽出システムを試作した．

検索語の分析に十分な文書量として，試作システム上では経験的に N=5000 とし，一処理単位として検索語集合 R を逐次的に計算することとした．すなわち，T^a_d の 5000tweet 目までを収集する毎にその対象 tweet 群から検索語集合が次々に得られることになる．

試作システムは，Twitter の Streaming API(statuses/sample)[(注3)] を用いて災害時 tweet 集合 T^a_d を収集する．なお，平常時 tweet 集合 T^a_n については，継続的に収集しておいた適当な tweet 集合から発災時刻から発災時刻より 24 時間前の期間に投稿された tweet を充てる．

T^a_d から重複 tweet を除外するために，T^a_d の各 tweet の投稿文 $\mathcal{T}_i$ を対象に形態素解析をし B_i を得る．T^a_d の各 B_i 同士で cos 類似度を算出し，cos 類似度が 1 の tweet を T^a_d から除外し T_d を得る．T^a_n も同様に重複 tweet を除去し T_n を得る．

検索語の候補語集合を得るために，災害時 tweet 集合 T_d の各 tweet の投稿文 $\mathcal{T}_i$ を対象に形態素解析し，A_i と S_i を得る．A_i に試作システムに入力された災害語 D が含まれていた場合，S_i に含まれる災害語 D 以外の語を抽出し，災害語 D と共起する語集合 $C(D)$ を得る．感動詞と共起する語集合 $C(I)$ も同様に得る．次に，$C(D)$ の各語 w に対して対数尤度比 $LLR(w)$ を算出し，(6) 式に基づき有意差のある災害語と共起する語集合 $F(D)$ を得る．検索語の候補語集合として，$F(D)$ と $C(I)$ との積集合から候補語集合 $K = F(D) \cap C(I)$ を得る．

この候補語集合 K の各語を対象に，(8) 式に基づき χ^2 値を算出し，(11)(12) 式より上位 M 件の語を検索語集合 R とする．

tweet の収集には，Twitter の Streaming API 経由で，無料ライセンスの範囲で動作させた．したがって，全投稿 tweet の 1%が得られるはずである．収集 tweet の属性のひとつである language プロパティが `'ja'`，すなわち日本語と設定されている tweet のみを使用し，他の tweet は破棄している．また，リツイート（引用ツイート含む）も除外している．

形態素解析器には，MeCab 0.996[(注4)] を，MeCab のシステム辞書には，Mecab-ipadic-NEologd[(注5)] を，重複 tweet 除去用の BoW ベクトル W_i の作成には，Python 上のトピックモデルライブラリである gensim(ver.2.2.0) を，cos 類似度の算出には，Python 上の汎用的な数値演算ライブラリである NumPy(ver.1.12.1) を使用している．

5. 評価実験

提案手法が，異なる災害に対して，それぞれ異なる適切な検索語を抽出できるかを，他の手法との比較を通じて評価する．また，同一災害であっても時間経過と共に被害などの局面は変動していくため，提案手法がこの変化に追従して適切な検索語を抽出できているかについても評価する．評価対象は，主に各手法毎に抽出した検索語で災害時 tweet 集合 T^a_d 内を検索した結果を用いる．提案手法の実行には作成した試作システムを用いた．

比較手法は，

（1） 災害語 D のみとする手法（災害語手法）

（2） 災害語との共起語 $C(D)$ の出現頻度上位 M 語群を検索語とする手法（災害共起語手法）

（3） word2vec を用いて災害語と類似する上位 M 語群を検索語とする手法（word2vec 手法）

の 3 手法を用いる．各比較手法について説明する．

災害語手法

検索語集合を災害語として入力した単語のみとする手法である．本実験では，「地震」といった 1 語を用いている．

災害共起語手法

T^a_d 内で災害語との共起する語のうち頻出上位 M 件を検索語とする手法である．これは，主題語との共起を扱う一般的なクエリ拡張の手法に相当するといえる．

Word2vec 手法

T^a_d をコーパスとして構築した word2vec モデルを用いて，災害語 D との最類似語上位 M 件を検索語とする手法である．word2vec 実装は gensim(ver.2.2.0) 内の実装を用いた．学習用コーパスは 1tweet を 1 文書とした T^a_d データとなる．ただし，処理対象とした語は品詞が名詞，動詞，形容詞のいずれかに該当する語とした．学習モデルは CBOW，各種パラメータは

(注3)：https://developer.twitter.com/en/docs/tweets/sample-realtime/api-reference/get-statuses-sample

(注4)：http://taku910.github.io/mecab/

(注5)：https://github.com/neologd/mecab-ipadic-neologd

gensim.models.word2vec.Word2Vec(注6) の既定値とした．

5.1 実験条件

本実験に用いた 3 つの災害を表 2 に示す．

表 2　災害一覧

災害名称	災害発生時刻	種類	災害語 D
北海道地震	2018/09/06 03:07:59	地震	地震
平成豪雨	2018/07/06 17:10:00	豪雨	大雨
SB 障害	2018/12/06 13:39:00	通信障害	ソフトバンク

表 2 は，災害の名称 (災害名称)，災害が発生した日時 (災害発生時刻)，災害の種類 (種類)，災害を代表する語 (災害語 D)4 項目を示している．ただし，発災日時は気象庁が提供している情報に準拠した．豪雨の災害発生時刻は，気象庁から長崎，福岡，佐賀の 3 県に大雨特別警報が発令された時刻とした．通信障害の災害発生時刻は，ソフトバンク社から通信障害が発生したと報告された時刻とした．

実験の対象期間は，災害発生時刻から最大 24 時間までとした．また，平常時の tweet 集合 T_n^a は災害時の 24 時間前の tweet を用いた．

MeCab のシステム辞書の Mecab-ipadic-NEologd は，2018 年 5 月 14 日に更新した辞書を使用する．

また，上位 M 語を $M = 10$，有意水準は $\alpha = 5\%$ とした．

5.2 評価方法

提案手法は，マイクロブログ上の投稿から当該災害に関連する投稿を効率的に検索できる検索語集合を得ることを目的としている．そこで slot 毎に，各手法で得られた語集合を検索語集合として，本実験で用いた災害時 tweet 集合 T_d^a から検索語のいずれかが含まれている tweet を検索し，この検索結果 tweet を人手による判断で災害に関連する tweet かどうかを判別することで，当該災害に関連する投稿を効率的に検索できる検索語集合が得られているか，検索結果の精度によって評価する．

また，同一災害で変化していく局面を捉えた検索語集合が得られているかを，各災害毎に大きく局面の変動があったと考えられる時間帯をいくつか選び，その最初の方の時間帯での検索語集合，検索結果 tweet を slot 毎に比較することで検証する．

対象とする災害が発生した後，状況の変化が起きた時間帯を表 3 に示す．表 3 は，対象とする災害の名称 (災害名称)，状況の変化が起きた時刻 (時刻)，どういう変化が起きたのか (備考)，の 3 項目を示している．これらの時間帯に該当する slot を評価対象とした．

提案手法のように，実 tweet 集合から検索語集合を得る場合，得られた検索語集合でどの元 tweet 集合を検索対象とするかには選択の余地がある．データ処理の立場からは，検索語集合を得た元 tweet 集合を検索対象とするべきに思える一方，実運用の立場では，直近データに基づく検索語集合を，その後，「今」投稿されてくる tweet の検索に使えた方が効果的であるとも考えられる．そこで，この両者の場合で結果に差異が生じるかを，検索語集合抽出 slot 自身の tweet 集合を検索対象とする場合と，その次の slot を検索対象とする場合の 2 通りについて結果を比較しながら議論する．

各 slot の全検索結果は多数の tweet に上るため，slot 毎の各検索結果 tweet 集合内の tweet 数が 50 件を上回る場合は，50 件をランダムに選択した上でそれらでの精度を評価した．

人手による判断は，各 slot 各手法での検索結果 tweet 合計 2278 件を著者 1 名が独立に災害に関連するかの判断をした．

表 3　各災害による状況の変化が起きた時間帯

災害名称	時刻	備考
北海道地震	2018/09/06 03:07:59	北海道胆振地方で震度 6 強の地震.
	2018/09/06 15:30:00	気象庁が震度 7 を観測したと発表.
SB 障害	2018/12/06 13:39:00	全国でソフトバンク社の携帯電話サービスにおける通信障害.
	2018/12/06 18:04:00	携帯電話サービスの復旧.
平成豪雨	2018/07/06 17:10:00	福岡，佐賀，長崎に大雨特別警報.
	2018/07/06 19:50:00	広島，岡山，鳥取に大雨特別警報.

5.3 実験結果

5.3.1 tweet 収集に要した時間

T_d^a での 5000tweet を 1slot としたため，その時間は各 slot により異なる．各災害毎の発災時刻を基点として，slot 時間長の推移を示したのが図 2 である．左端を発災時刻として経過時間に対して，その時刻帯での slot 時間長を縦軸に示している．

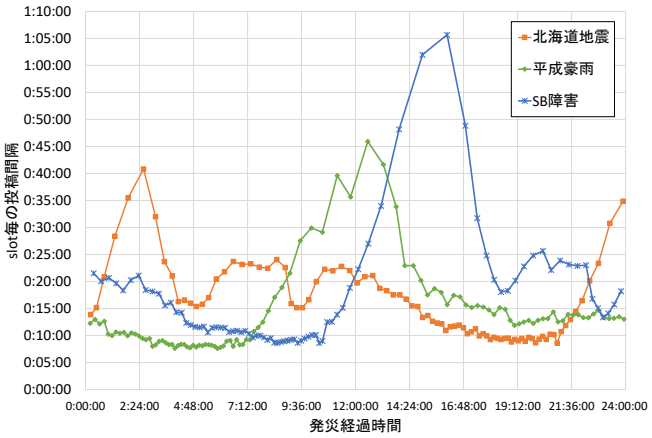

図 2　tweet 収集に要した時間

結果を見ると，時間が大きく伸びる山の形状が観察される時間帯を除くと，概ね 10〜25 分程度で 1slot を構成する tweet が投稿されたことが分かる．すなわち，提案手法はこの程度の時間毎に新たな検索語集合を得ることができ，被害局面の変動にもこの程度の時間粒度で追随できることを示唆している．

一方，最大 1 時間程度まで slot 時間長が延びる山が確認できる．本グラフでは発災時刻を時刻 0 としているためわかりにくいが，これらの山を記録しているのはいずれも未明〜早朝の一般に twitter の利用が低調となる時間帯である．処理対象となるデータ量が少なくなる時間帯では，それなりの時間長を要しないと必要なデータ量が集まらないように見える．

ただし，tweet 収集部では，全体の 1%のサンプリング tweet のみの件数を観測している．すなわち，全 tweet を収集するようにすれば，グラフ上で 1 時間程度必要とする時間帯であっても 1 分以内で 5000tweet の条件を満たすことになる．したがっ

(注6)：https://radimrehurek.com/gensim/models/word2vec.html

て，この slot 長による検索語集合抽出の時間追従性の限界は，tweet 収集の実装に依存するものであり，提案手法そのものの限界を示すものではない．なお，ごく短時間の slot 長でも提案手法が有効であるかは検証していないため，その検証は課題として残されている．

5.3.2 各手法の抽出精度の比較

提案手法によって得られた検索語集合で検索を行なった場合と，各比較手法で検索を行なった場合との抽出精度の結果を表 4，各手法の精度，再現率，F-measure の平均を表 5 に示す．

表 4 は，対象とする災害の名称 (災害名称)，評価対象とした slot 番号 (slot)，提案手法および各比較手法 (手法)，各手法で得られた語集合を用いて検索し，その検索結果 tweet の件数 (検索)，うち 50 件をランダムサンプリングした tweet を人手により正解とされた tweet の件数 (正解)，ランダムサンプリングした 50 件の tweet に含まれる正解の割合 (精度)，プーリングに基づき人手による判定を行った正解 tweet に対する割合 (再現率) を示している (注7)．精度，再現率より F 値（F-measure）も算出している．また，現 slot は評価対象とした slot であり，次 slot は評価対象とした slot の次である．

表 4，表 5 より，現 slot 検索，次 slot 検索いずれにおいても，災害語手法を除き，提案手法によって得られた検索語を用いた検索結果が最も良い精度を得たことがわかる．また，正解件数では提案手法が比較手法をほとんどの場合で上回っている．

災害語手法は，災害語自体を含んだ tweet を検索するため，精度は非常に高い．しかし，再現率は低い．これは，単に災害語だけの検索結果では，十分な情報源となる tweet を検索できないことを示しており，tweet などからの災害情報抽出研究で複数の語を検索語としていることを支持している結果であると共に，提案手法の必要性を示しているものといえる．

災害共起語手法および word2vec 手法は，災害語との共起を扱う一般的なクエリ拡張の手法に相当するといえる．災害は普段起きない事象であるため，新たに起こったことを説明する「こと」や「いる」「てる」という語の頻度が大きく上昇し，検索語として抽出された．これらの語は，災害以外の状況を表すことにも使われる，ごく一般的な語句であり，その時間帯の災害とは無関係な tweet の検索結果にも多数含まれたため，精度が下がる結果を招いた．従って，災害語との共起だけを扱う一般的なクエリ拡張の手法では，災害時の tweet を適切にフィルタリングできないといえ，本研究の優位性が示されている．

これらから，提案手法は比較手法と比較して，災害時の情報収集における検索語集合としてふさわしい語を抽出できているといえ，提案手法の有効性を確認した．

5.3.3 検索語集合の slot 間の順位相関係数

試作システム上で逐次的に計算される検索語集合 R が，連続する slot 間でどの程度の語の入れ替わりが起きているかを，2 つの異なるランキングリストから上位 M 件のみを対象に相関を求める Fagin らの順位相関係数 τ [15] を用いて確認した．隣り合う slot 間の検索語集合の相関の推移を図 3 に示す．

図 3 は，横軸は各災害毎の発災時刻を基点とした slot 番号，縦軸は連続する slot 間の順位相関係数 τ を示している．τ の範囲は 0〜1 であり，$\tau = 0$ は 2 つのリストの完全な不一致を意味し，$\tau = 1$ は完全な一致を意味する．

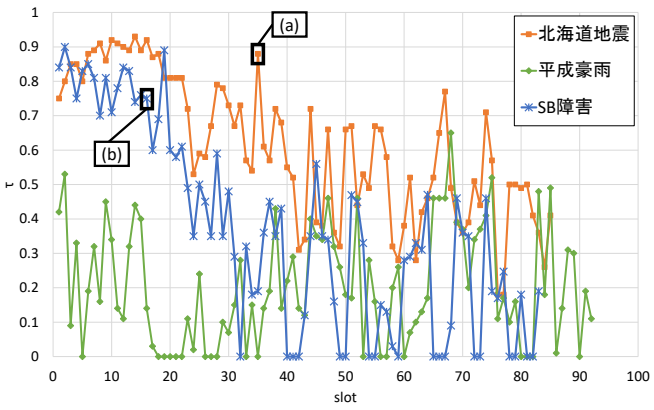

図 3　連続する slot 間の検索語集合の順位相関係数

検索語集合の語の移り変わりを確認するために，数点の slot を抜粋した．その slot を四角で囲み，(a)〜(b) の記号を振って示すこととする．

(a) の slot は，気象庁が北海道厚真町で震度 7 を観測したことを発表した時刻に相当する．発災後から (a) の slot の直前まで，震度 6 強が観測されたと報道されていた．(a) の slot から，気象庁が北海道厚真町で震度 7 を観測したことを発表したため，「震度 6 強」という語から「震度 7」という語に入れ替わりが起きたため，(a) の slot の前後で，τ の沈み込み，すなわち，大きく検索語の集合が入れ替わっていることが観察できる．

(b) の slot は，ソフトバンク社の通信障害が復旧した時刻に相当する．(b) の slot の次 slot で τ は一旦下がるが，(b) の 2,3 つ先の slot の τ は上昇している．これは，検索語集合が通信障害が復旧したことにより，通信障害により携帯が使えなくなったことの報告や不平不満を表すような単語から，復旧を表すような単語に入れ替わりが起きたためである．(b) の 3 つ先の slot で τ は高止まりし，これ以降，τ は下がっていく．これは，ソフトバンク社の通信障害が復旧し日常に戻ったため，検索語集合内から通信障害もしくはその復旧に関する語が，ソフトバンク社の他の話題の語と入れ替わりが起きたためである．

平成豪雨では，全体を通して τ が低く，検索語集合の移り変わりが激しいことが観察できる．平成豪雨は，特別大雨警報による避難指示や，洪水や川の氾濫による浸水報告，浸水の影響で工場が爆発する二次被害等の被害が複数発生した．そのため，特別大雨警報が発令された「岡山」「広島」といった地域名や，「洪水」「氾濫」といった水害を表わすような語が，時間経過とともに検索語集合内で常に入れ替わりが起きたため，全体を通して τ が低い結果となった．

(注7)：ただし，本実験では似通ったアルゴリズムによる 4 手法でのプーリングであるため，真の正解 tweet を十分被覆していない可能性がある点には留意する必要がある．しかし，4 手法間の差異を見るためだけであれば有効であると考えられる．

表 4　各手法による tweet 検索結果

災害名称	slot	手法	現 slot					次 slot				
			検索	正解	精度	再現率	F-measure	検索	正解	精度	再現率	F-measure
北海道地震	1	提案手法	1759	50	1.00	0.71	0.83	1770	47	0.94	0.73	0.82
		災害語	710	50	1.00	0.29	0.45	408	50	1.00	0.18	0.30
		災害共起語	1971	43	0.86	0.69	0.76	1993	37	0.74	0.64	0.69
		word2vec	1525	29	0.58	0.36	0.44	1613	32	0.64	0.45	0.53
	35	提案手法	383	43	0.86	0.32	0.47	318	45	0.90	0.30	0.45
		災害語	97	49	0.98	0.09	0.17	85	48	0.96	0.08	0.16
		災害共起語	637	22	0.44	0.27	0.34	604	17	0.34	0.21	0.26
		word2vec	1502	10	0.20	0.29	0.24	1553	1	0.02	0.03	0.02
平成豪雨	1	提案手法	341	42	0.84	0.39	0.53	223	35	0.70	0.25	0.37
		災害語	54	50	1.00	0.07	0.14	38	38	0.76	0.05	0.09
		災害共起語	918	12	0.24	0.30	0.27	836	10	0.20	0.27	0.23
		word2vec	1200	9	0.18	0.29	0.22	1247	8	0.16	0.32	0.21
	15	提案手法	312	29	0.58	0.33	0.42	229	36	0.72	0.34	0.46
		災害語	31	28	0.90	0.05	0.10	31	28	0.90	0.06	0.11
		災害共起語	756	14	0.28	0.39	0.33	649	8	0.16	0.21	0.18
		word2vec	1229	5	0.10	0.23	0.14	1143	3	0.06	0.14	0.08
SB 障害	1	提案手法	260	47	0.94	0.36	0.52	356	44	0.88	0.44	0.58
		災害語	82	48	0.96	0.12	0.21	153	50	1.00	0.21	0.35
		災害共起語	777	15	0.30	0.35	0.32	829	17	0.34	0.39	0.36
		word2vec	758	10	0.20	0.22	0.21	809	13	0.26	0.29	0.28
	15	提案手法	244	44	0.88	0.34	0.49	266	47	0.94	0.35	0.51
		災害語	109	49	0.98	0.17	0.29	143	48	0.96	0.19	0.32
		災害共起語	552	14	0.28	0.25	0.26	566	18	0.36	0.28	0.32
		word2vec	738	13	0.26	0.30	0.28	767	11	0.22	0.23	0.23

表 5　各手法による tweet 検索結果の平均

手法	現 slot			次 slot		
	精度	再現率	F-measure	精度	再現率	F-measure
提案手法	0.85	0.41	0.54	0.85	0.40	0.53
災害語	0.97	0.13	0.22	0.93	0.13	0.22
災害共起語	0.40	0.37	0.38	0.36	0.34	0.34
word2vec	0.25	0.28	0.26	0.23	0.24	0.23

6. おわりに

本研究では，マイクロブログサービスの特性であるリアルタイムでの情報配信を活かし，発災後間もない時間帯での被害状況を含めた情報収集を視野に入れた，災害に関連する投稿を検索する検索語集合を自動的に抽出する手法を提案した．評価実験において，最近の異なる種類の大規模災害時の tweet に対して本手法を適用した結果，提案手法が，災害被害の発生状況に追従した検索語集合を高い精度で抽出できたことを示した．

ごく単純な災害語の入力だけとはいえ，人手の介在を必要としているのは事実であり，これを完全自動化することは今後の課題である．

文　献

[1] 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)．

[2] Sakaki, T., Okazaki, M., 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).

[3] Maru, C., Enoki, M., Nakao, A., Yamamoto, S., Yamaguchi, S., and Oguchi, M.:Network Failure Detection System for Traffic Control using Social Information in Large-Scale Disasters, ITU Kaleidoscope Conference 2015: Trust in the Information Society, S5.3, pp.1-7(2015).

[4] Yuzawa, A., Ichikawa, H., and Kobayashi, A.:Extracting Tweets related to Disaster Information by using Multiple Co-occurrence Relation of Words, Proc. 2018 IEEE International Conference on Smart Computing (SMARTCOMP 2018), pp.321-326(2018).

[5] 中島 伸介，張 建偉，稲垣 陽一，中本 レン：大規模なブログ記事時系列分析に基づく流行語候補の早期発見手法，情報処理学会論文誌データベース（TOD），Vol.6，No.1，pp.1-15(2013)．

[6] 高木 涼太，風間 一洋，榊 剛史：単語の分散表現に基づく専門用語辞書の拡張法，研究報告ドキュメントコミュニケーション（DC），Vol.2018-DC-110，No.22，pp.1-6(2018)．

[7] 青島 傳隼，坂本 翼，横山 昌平，福田 直樹，石川 博：文脈的なつながりを考慮したツイート群の効果的な抽出・提示手法の実現，情報処理学会論文誌データベース（TOD），Vol.6，No.2，pp.61-84(2013)．

[8] 新田 直子，角谷 直人，馬場口 登：単語間の関係性の経時変化を考慮したマイクロブログからの実世界観測情報の抽出，日本データベース学会論文誌，Vol.13-J，No.1，pp.7-12(2014)．

[9] 宮西 大樹，関 和広，上原 邦昭：コンセプト追跡を用いたマイクロブログ検索，情報処理学会論文誌データベース（TOD），Vol.7，No.2，pp.1-10(2014)．

[10] Blei, D. M., Ng, A. Y., and Jordan, M. I.:Latent dirichlet allocation, Journal of machine Learning research, Vol.3, pp.993-1022, ACM(2003).

[11] 坂本 一磨，中村 健二，山本 雄平，田中 成典：平時と異なる事象に対するソーシャルセンシング技術に関する研究，情報処理学会論文誌，Vol.59，No.10，pp.1866-1879(2018)．

[12] 山本 修平，佐藤 哲司：トピックと局面の対応関係に基づく実生活ツイートのマルチラベル分類，情報処理学会論文誌データベース（TOD），Vol.7，No.2，pp.24-36(2014)．

[13] 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).

[14] Dunning, T.:Accurate Methods for the Statistics of Surprise and Coincidence, Computational Linguistics, Vol.19, No.1, pp.61-74(1993).

[15] Ronald Fagin, Ravi Kumar, D. Sivakumar.:Comparing top k lists, Proc. The fourteenth annual ACM-SIAM symposium on Discrete algorithms, ISBN:0-89871-538-5(2003).