

マイクロブログにおける感動詞との共起を利用した 検索語の抽出

湯沢 昭夫^{1,a)} 小林 亜樹^{2,b)}

受付日 2018年12月10日, 採録日 2019年4月1日

概要: 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. 災害ごと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が, 「地震」などの代表語からの既存のクエリ拡張による手法では, 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, 有効に機能しない場合がある. そこで本稿では, 新しい情報であるなど人が有効と感じる情報を入手できたときにマイクロブログ上に投稿される感謝などの語彙をきっかけとした手法を, 検索語集合抽出方式として提案する. 提案手法は, マイクロブログサービス上で感動詞が人と人と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, 代表語だけでなく感動詞との共起関係も利用して検索語集合を抽出する. 全般的な統計的情報を用いる多くのクエリ拡張手法に対して, きっかけとなる投稿から候補語を拾い上げる手法としたために, おおむね数分から十数分程度で検索語集合を更新し続けることを可能とした. 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, 災害被害の発生状況に追従した検索語集合が得られていることを示す.

キーワード: Twitter, マイクロブログ, 感動詞, 災害, 検索語

Extracting Query Words Using Co-occurring Relation of Interjections on Microblog

AKIO YUZAWA^{1,a)} AKI KOBAYASHI^{2,b)}

Received: December 10, 2018, Accepted: April 1, 2019

Abstract: Query words for extracting disaster information from microblog posts are manually selected in most studies. This selection is difficult during a disaster because it is thought that proper query words differ for each disaster. Therefore, it is necessary to automatically determine a set of query words. We propose a method of extracting a set of query words using the co-occurring relation of frequently used words and that of interjections. We focus on interjections contained in posts of microblog that suggest salutary conversation among people. With our method, a set of query words can be updated every few minutes because our method extracts candidate words from a trigger tweet compared to many other query-expansion methods that use general statistical information. We applied our method to actual tweets posted during recent different severe disasters. We experimentally show that our method can extract a set of query words following changing situations of disasters.

Keywords: Twitter, microblog, interjections, disasters, query words

¹ 工学院大学大学院工学研究科電気・電子工学専攻
Electrical Engineering and Electronics Program, Graduate
School of Engineering, Kogakuin University, Shinjuku,
Tokyo 163-8677, Japan

² 工学院大学情報学部情報通信工学科
Department of Information and Communications Engineering,
Faculty of Informatics, Kogakuin University, Shinjuku,
Tokyo 163-8677, Japan

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.

a) cm17051@g.kogakuin.jp

b) aki@cc.kogakuin.ac.jp

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数を記録している^{*1}。

マイクロブログの代表といえる Twitter は情報インフラとしても重要視される。たとえば、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、知人の安否の確認など、リアルタイムに情報を伝達する手段の 1 つとして活用された [1]。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 を使い、市内の漏水箇所やデマ情報の打ち消しに活用された [2]。このように Twitter は重要な情報インフラとして扱われている。

そこで、Twitter を情報源として災害発生時の状況分析が研究されている [3], [4], [5], [6]。これらの研究は、フィルタリングの第 1 段階として災害に関連する投稿を検索するために、人手によって用意した検索語群を事前に設定し、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ごとに大きく異なるため、災害から相当日時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。

マイクロブログサービス上の投稿から、ある事象に関連する単語を自動的に抽出することは広く検討されてきた [11], [12], [13], [14], [15]。これらの研究は、事象ごとの特徴を細かくとら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、その調整方法は明かでない。

そこで本研究では、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。品詞が感動詞に該当する単語を含む投稿文の例を図 1 に示す。これは北海道で地震が起きた際につぶやかれた tweet であり、

@● 電気止まってしまいましたけど、無事ですー！●さんの方も大丈夫ですか？！ ありがとう ございます！まだ余震ありそうなので、お互い注意しましょう…

図 1 品詞が感動詞に該当する単語を含む投稿文の例

Fig. 1 Examples of tweets include “Interjection”.

tweet 本文中には品詞が感動詞に該当する「ありがとう」が含まれている。このように本研究では、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を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。Funayama ら [3] は、単語間の共起関係を可視化し、災害時に有用な情報を含む tweet を抽出する手法を提案した。Middleton ら [4] は、避難所や危険地域をリアルタイムで地図上へ地理可視化する手法を提案した。Sakaki ら [5] は、Twitter を利用しているユーザをセンサと見なし、「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。Maru ら [6] 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。

これらの研究は、人手で検索語群を事前に設定し、それらの検索語を含む投稿を分析対象としている。検索語群は災害の種類や被害状況ごとに大きく異なり、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。また、SVM といった教師あり学習を用いる手法は、後日の分析には向くものの、現場でのリアルタイムでの運用には適さない。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、災害にともな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に関する研究は多く存在する。中島ら [7] 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。高木ら [8] 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。大石ら [9] は、テキスト本文中に現れる単語の距離に着目し、Web 記事を対象にクエリ拡張手法を用いて単語に対して評価し、クエリの関連語を生成する手法を提案した。松尾ら [10] は、対象とする文章のみから、語の共起関係を基に統計的な指標で

^{*1} <http://www.viewproxy.com/Twitter/2018/AnnualReport2017.pdf>

ある χ^2 検定を改良したアルゴリズムを用いて、入力した文章に関連するキーワードを抽出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 [11], [12], [13]。

青島ら [11] 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。斎藤ら [12] 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、単語間の相互情報量で重み付きグラフを生成し、そのグラフから独自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た。新田ら [13] 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ごとの特徴を細かくとら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、検索クエリ拡張ととらえることができる。宮西らによるマイクロブログ上でのクエリ拡張手法 [14] 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（複数の単語の組合せ）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。本手法は、マイクロブログサービス上で感動詞が人と人と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い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。たとえば、坂本らの社会事象を抽出する手法は LDA (Latent Dirichlet Allocation) [19] を用いたトピックを利用して [15]、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 [16]。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。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 [18]、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。

スロット

提案手法で 1 処理単位とする tweet 集合を指す。実際の処理では一定の文章量が必要であるため、tweet 収集部において別に定める tweet 件数ごとに区切って処理単位とする。提案手法では、このスロット単位ごとに独立して検索語抽出を行い、発災以後に投稿された tweet が一定件数収集されるごとに 1 スロット、2 スロット、…と順に呼ぶ。

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。たとえば、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ごとに異なる。そのため、tweet 自身から自動的に抽出されるべきである。本研究では関連語抽出で多用される、語の共起関係の利用を災害語との間で用いるのに加えて、発災後の情報のやりとりがあったことを示唆する、人と人との間の関係を示す表現である感動詞との共起にも着目する。

発災後の tweet 集合自身から統計的に検索語を抽出するためには一定の文章量が必要である。一方で、時々刻々と変化する状況をとらえるためには短期間の tweet 集合を処理対象とすることが望ましい。提案手法では、スロットと呼ぶ（重複 tweet 除去の影響を除いて）一定の tweet 数に

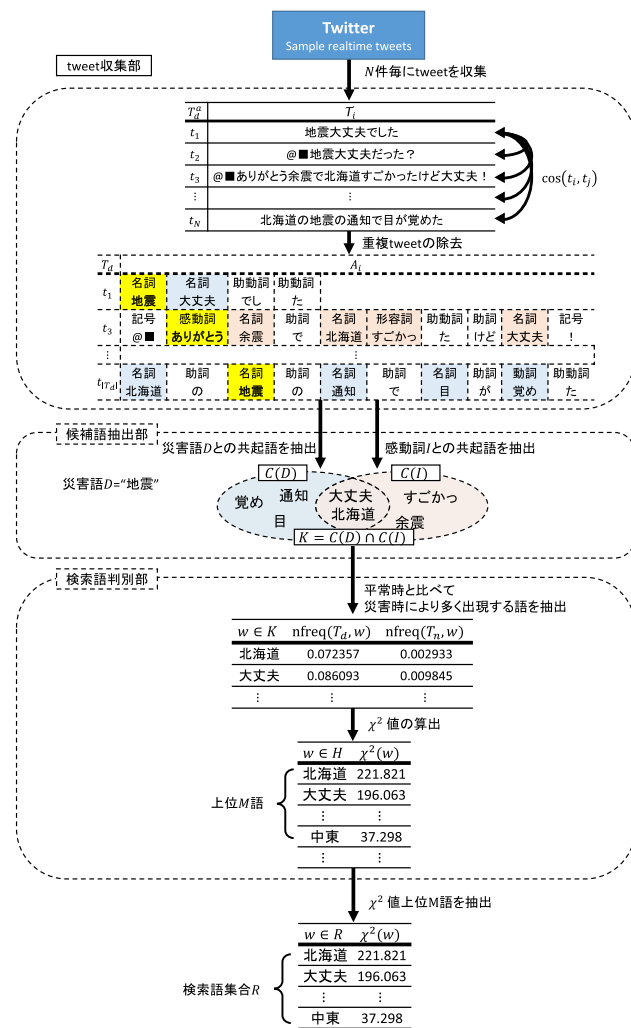


図 2 本システムの概要図

Fig. 2 Overview of proposal system.

よる集合ごとに独立に検索語を抽出できるようにすることでこの要求に対応する。

本システムの概要を図 2 に示す。まず、発災以後の全 tweet *2 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 D との共起語集合 $C(D)$ と、感動詞との共起語集合 $C(I)$ を得て、両者の交わりから検索語の候補集合 $K = C(D) \cap C(I)$ を得る。この候補より、災害発生前後での語の出現頻度が災害発生後の方が高い語集合 H を対象に、検索語スコア (χ^2 値; 3.5.2 項参照) を算出して高スコアを得た上位 M 件の語集合を検索語集合 R とする。このように、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

*2 災害時の運用でいえば発災時刻から現在時刻までの tweet とする。

tweet 間の類似度を算出し、高類似度となった 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 [17] 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両者と共起する語を出力する。この出力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。また、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 [18], その Reply の投稿文内には感動詞に該当する語が含まれていることが多いという経験則に基づき、感動詞を利用している。

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であることから、 χ^2 値を算出する。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 tweet の収集

本システムの tweet 収集部が tweet を収集し続けている。このうち、発災前の tweet を平常時 tweet、発災後の tweet を災害時 tweet 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とらえられる。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よいと考えている。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。

Twitter の Streaming API を用いて、リツイート (引用リツイートを含む) を除外して収集した平常時 tweet 集合を T_n^a 、災害時 tweet 集合を T_d^a とする。ただし、災害時 tweet 集合は処理単位となる 1 スロット分についての表記であり、 T_d^a の tweet 数 (要素数) は N である。

また、tweet t_i を、 i 番目の tweet について、

$$t_i := (T_i, A_i, S_i, B_i) \quad (1)$$

に示す 4 つの情報を含むと定義する。各表記の説明を表 1

表 1 t_i の表記の説明
Table 1 Description of t_i notation.

Notation	Meaning
T_i	元の投稿文 (テキスト).
A_i	T_i を形態素解析した全品詞の語の集合.
S_i	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	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.

T_i	@■ ありがとう余震で北海道すこだったけど大丈夫!									
A_i	記号 @■	感動詞 ありがとう	名詞 余震	助詞 で	名詞 北海道	形容詞 すこ	助動詞 かつ	助詞 けど	名詞 大丈夫	記号 !
S_i	名詞 余震	名詞 北海道	名詞 大丈夫	形容詞 すこ						
B_i	{1, 1, 1, 0, 0, 0, 0, ..., 0}									

図 3 t_i 中の要素例Fig. 3 Elements of t_i .

に示し, 具体例を図 3 に示す. 提案手法では, 品詞が名詞, 動詞, 形容詞のいずれかに該当する語を処理対象とする. ただし, “@” 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. B_i は tweet 集合に含まれる名詞の種類数を次元数とし, 1 tweet に含まれる単語の出現頻度を要素としたベクトル (Bag of Words) である.

3.3.2 重複 tweet の除去

予備的手法 [17] では, 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分かっている. そこで, 文章が重複した tweet を 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 どうして $\cos$ 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 と表す.

3.4 候補語抽出部

処理対象とする品詞選定は, 一般的な SVM による分類器での災害 tweet 判定を行う予備的調査の結果, 感動詞を導入することとした. この調査ではまず, MeCab 内品詞 13 種類のうち「記号」「その他」を除いた, 11 種類の品詞の組合せについてすべて組み合わせ, すなわち $2^{11} - 1$ 通りについてそれぞれ分類実験を行った. その精度が高い結果となる組上位 10% には, 一般的な名詞, 動詞, 形容詞以外では感動詞を含む組合せがほとんどであった. また, 感動詞のみによる検索結果ではほとんど有意義な結果を得ることができなかった. そのため, 感動詞自体を検索語候補とするのではなく, 感動詞との共起語に着目

することとした. 一般に利用される語の頻度分布は Zipf の法則に従い, 感動詞においても実際に利用されている感動詞はごく一部にとどまるが, 頻度上位となる語を事前に知ることとはできないため, 提案手法では感動詞すべてを用いることとしている. この経験則に基づき, T_d 内で災害語, 感動詞いずれとも共起する語を検索語候補とする.

tweet を単位とした語 w との共起語集合は,

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

のように表される. ここで, 共起語となる c は表 1 の S_i に示すとおり, 名詞, 動詞, 形容詞のいずれかである.

w を語集合 W とした場合は,

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

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

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

$$K = C(D) \cap C(I) \quad (4)$$

と表す.

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. そこで, 検索語の候補語集合 K 中で, 頻度が高くなったかどうかを判定するために, χ^2 検定の考え方を流用する. 災害発生前後の tweet 集合中で, 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 χ^2 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,

$$\chi^2(w) = \sum_{a \in \{T_d, T_n\}} \sum_{b \in \{w, \bar{w}\}} \frac{(n_{ab} - E_{ab})^2}{E_{ab}} \quad (5)$$

と, 式 (5) のように計算される. 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,

$$E_{ab} = \frac{n_{a*} \cdot n_{*b}}{Z} \quad (6)$$

である. ここで, 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, Z 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 w は, 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 における出現頻度を $\text{freq}(T, w)$ とするとき, 単語 w の 1 tweet あたりの相対頻度として $\text{nfreq}(T, w)$ を,

$$\text{nfreq}(T, w) := \frac{\text{freq}(T, w)}{|T|} \quad (7)$$

のように, T に含まれる tweet 数 $|T|$ を用いて正規化することで定義する.

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,

$$H = \{w \in K \mid \text{nfreq}(T_d, w) > \text{nfreq}(T_n, w)\} \quad (8)$$

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,

$$R = \{w \in H \mid \chi^2(w) \geq \chi^2_M(H)\}. \quad (9)$$

ここで, $\chi^2_M(H)$ は H 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

提案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 $N = 5,000$ 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. これは, T_d^a としての 5,000 tweet 目までを収集するごとにその対象 tweet 群から検索語集合が次々に得られるということである. 試作システムは, Twitter 社の Streaming API (statuses/sample)*³ を用いて災害時 tweet 集合 T_d^a を収集する. T_d^a から重複 tweet を除外するために, T_d^a の各 tweet の投稿文 T_i を対象に形態素解析をし B_i を得る. T_d^a の各 B_i どうして $\cos$ 類似度を算出し, $\cos$ 類似度が 1 の tweet を T_d^a から除外し T_d を得る. T_n^a 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 T_i の形態素解析結果より 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. 検索語の候補集合 K は, $K = C(D) \cap C(I)$ と両語集合の積集合で求める.

この候補語集合 K より, 検索語集合 R を抽出する. 候補語集合 K の各語を対象に, 式 (5) に基づき χ^2 値を算出し, 式 (8), (9) より上位 M 件の語を検索語集合 R とする.

4.1 tweet 集合の収集

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 1 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.

4.2 重複 tweet の除去

重複 tweet 除去用の BoW ベクトル B_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, 形態素解析器には, MeCab 0.996*⁴ を, MeCab のシステム辞書に, Mecab-ipadic-NEologd*⁵ 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ともに被害などの局面は変動していく. 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, 主に各手法ごとに設定した検索語で災害時 tweet 集合 T_d^a 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. 処理対象の tweet は Streaming API の使用における仕様のため, 全体の 1% しかないが, 全データからのサンプリングによって検証していると考えれば問題はない.

比較手法は,

- (1) 災害語 D のみを検索語とする手法 (災害語手法)
- (2) 災害語との共起語 $C(D)$ の χ^2 値上位 M 語群を検索語とする手法 (災害語 χ^2 手法)
- (3) 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 (災害語頻度手法)
- (4) 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)
- (5) LDA により得られたトピックを構成する語群による

*³ <https://developer.twitter.com/en/docs/tweets/sample-realtime/api-reference/get-statuses-sample>

*⁴ <http://taku910.github.io/mecab/>

*⁵ <https://github.com/neologd/mecab-ipadic-neologd>

手法 (LDA 手法)

の 5 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.

災害語 χ^2 手法

T_d^a 内で災害語と共起する語を対象に, 災害発生後の方が出現頻度が高い語のうち χ^2 値上位 M 件を検索語とする手法である. これは, 提案手法における検索語の候補語集合 K を災害語と共起する語集合 $C(D)$ とした手法といえる.

災害語頻度手法

T_d^a 内で災害語と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_d^a 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 1 tweet を 1 文書とした T_d^a データとなる. ただし, T_d^a は名詞, 動詞, 形容詞のみで構成した. 学習モデルは CBOW, その他パラメータは gensim.models.word2vec.Word2Vec^{*6} のデフォルト値としている. なお, 学習コーパスを Wikipedia などに求めた場合, 規模の点では有利だが固定された類似語集合しか得られないため, 状況の変化を反映させたい本研究には適さない.

LDA 手法

Word2vec 手法と同様のコーパスから LDA トピックモデルを構築し, 得られた各トピックを構成する語 M 件をそれぞれ検索語集合とする手法である. ただし, LDA では一般に複数のトピック数を指定してトピックを得るが, いずれのトピックが目的に合う災害に関するトピックであるかは明らかでない. LDA の主要パラメータであるトピック数の与え方も明らかではない. 本比較実験では, いくつかの設定中で良い結果に思われたトピック数 10 とした場合に, 人手で最も良いトピックを 1 つ選定してそのトピックを構成する語群を検索語集合と見なした.

LDA 実装は gensim (ver.2.2.0) 内の実装を用いた. 学習用コーパスは 1 tweet を 1 文書とした T_d^a データとなる. ただし, T_d^a は名詞, 動詞, 形容詞のみで構成した. 各種パラメータは gensim.models.ldamodel.LdaModel^{*7} のデフォルト値とした.

5.1 実験条件

本実験に用いた 4 つの災害を表 2 に示す. 表 2 は, 災害の名称 (名称), 災害が発生した日時 (災害発生時刻),

表 2 災害一覧

Table 2 Disasters list.

名称	災害発生時刻	種類	場所
北海道地震	2018/09/06 03:07:59.3	地震	北海道胆振地方
大阪地震	2018/06/18 07:58:34.1	地震	大阪府北部
長野地震	2018/05/25 21:13:42.2	地震	長野県北部
平成豪雨	2018/07/06 17:10:00.0	豪雨	西日本全域

表 3 実験環境

Table 3 Environment of the experiment.

CPU	Intel(R) Core(TM) i5-3550 CPU @ 3.30 GHz
RAM	16 GB
OS	Ubuntu 16.04.2 (64 bit)

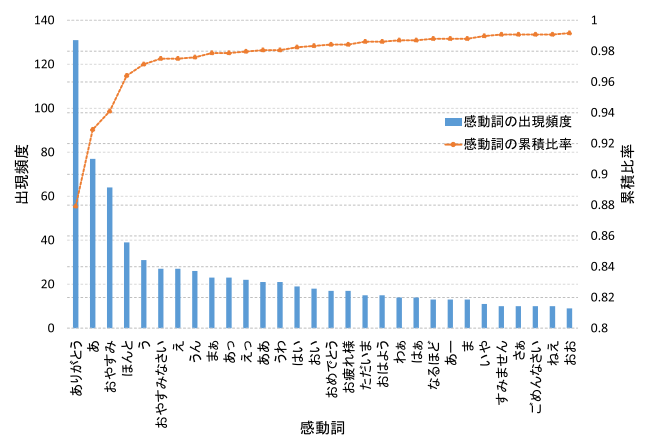


図 4 感動詞に該当する上位 30 語の出現頻度 (北海道地震)

Fig. 4 Frequent words of interjection (case: Hokkaido earthquake).

災害の種類 (種類), 災害が発生した地点 (場所) の 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.

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 は災害発生時刻の 24 時間前から災害発生時刻までの 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. 災害語 D として, 対象とする災害が地震の場合は「地震」, 豪雨の場合は「大雨」とした. また, 上位 M 語を $M = 10$ とした. 本実験で利用した実験環境を表 3 に示す.

提案手法で利用する感動詞は, MeCab のシステム辞書の Mecab-ipadic-NEologd に登録されている全感動詞 4,953 語である. 3.4 節で述べたように, 実際に相当数として観測される感動詞はこのうちのごく一部である. 北海道地震発災直後に投稿された 5,000 tweet を例に感動詞の出現頻度上位 30 語を図 4 に示す. 図 4 は, 横軸は感動詞の語, 左軸はその出現頻度, 右軸はその累積比率である. 出現頻

^{*6} <https://radimrehurek.com/gensim/models/word2vec.html>

^{*7} <https://radimrehurek.com/gensim/models/ldamodel.html>

表 4 各災害による状況の変化が起きた時間帯

Table 4 Time codes when the situation changes for each disaster.

名称	時刻	発災経過時間	備考
北海道地震	2018/09/06 03:07:59	0:00:00	北海道胆振地方で震度 6 強の地震
	2018/09/06 13:35:00	10:27:01	砂川発電所 3 号機の運転開始
	2018/09/06 15:30:00	12:22:01	気象庁が震度 7 を観測したことを発表
大阪地震	2018/06/18 07:58:34	0:00:00	大阪府北部で震度 6 弱の地震
	2018/06/18 16:31:47	8:33:13	大阪府北部で震度 3 の地震
	2018/06/19 00:31:25	16:32:51	大阪府北部で震度 4 の地震
長野地震	2018/05/25 21:13:42	0:00:00	長野県北部で震度 5 強の地震
	2018/05/25 22:00:12	0:46:30	長野県北部で震度 1 の地震
	2018/05/25 22:56:02	1:42:20	長野県北部で震度 3 の地震
平成豪雨	2018/07/06 17:10:00	0:00:00	福岡, 佐賀, 長崎に大雨特別警報
	2018/07/06 19:50:00	2:40:00	広島, 岡山, 鳥取に大雨特別警報
	2018/07/06 23:00:00	5:50:00	京都, 兵庫に大雨特別警報
	2018/07/07 13:10:00	20:00:00	岐阜に大雨特別警報

度上位 4 位までで累積比率は約 95%を超えており、適切な感動詞のみを用いることでさらに提案手法を改良できる可能性はある。一方で頻出上位語自体は災害ごとによって異なるとも考えられ、感動詞の選定自体は提案手法では行っていない。

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、各手法ごとに得られた語集合に災害語 D を加え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 かどうかを判別し、検索結果の精度によって評価する。また、各手法で得られた検索語集合によって、どの程度災害に関連する投稿を検索できるかを調査するために、検索結果から災害語を含む tweet を除いた場合での精度も評価する。

全検索結果の tweet 数は多数となるため、各手法、各スロットごとの検索結果 tweet 数が 50 tweet を上回るスロットについては、ランダムに選択した 50 件のみを人手による評価対象とした。この結果、評価対象の全 26 スロットでの実評価 tweet 数は、正味 6,405 件である。この人手による評価は原則として 3 人の独立した実験参加者の多数決による。ただし、一部に著者単独での評価を含んでいるため、評価対象とする tweet を独立に選択したために 3 人の多数決と著者単独の双方で評価することとなった tweet の判別結果を比較し、すべての tweet でその判断結果が一致していることを確かめることで、評価結果の妥当性を担保している。

同一災害で変化していく局面をとらえた検索語集合が得られているかの検証には、各災害ごと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 を比較することで行う。

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 4 に示す。表 4 は、対象とする災害の名称 (名称)、状況の変化が起きた時刻 (時刻)、対象とする災害後から経過した時間 (発災経過時間)、どのような変化が起きたのか (備考)、の 4 項目を示している。これらの時間帯に該当するスロットを評価対象とした。

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、検索語集合抽出スロット自身の tweet 集合を検索対象とする場合と、その次のスロット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。

5.3 実験結果

5.3.1 tweet 収集に要した時間

T_d^a で 5,000 tweet を 1 スロットとしたため、その時間は各スロットにより異なる。各災害ごとの発災時刻を基点として、スロット時間長の推移を示したのが図 5 である。左端を発災時刻として経過時間に対して、その時刻帯でのスロット時間長を縦軸に示している。

結果を見ると、いずれの災害も発災直後は 5~15 分程度で 1 スロットを構成する tweet が投稿されたことが分かる。これは、規模の大きな災害に対して利用者の関心が高く、Twitter がコミュニケーションツールとしてよく使用され

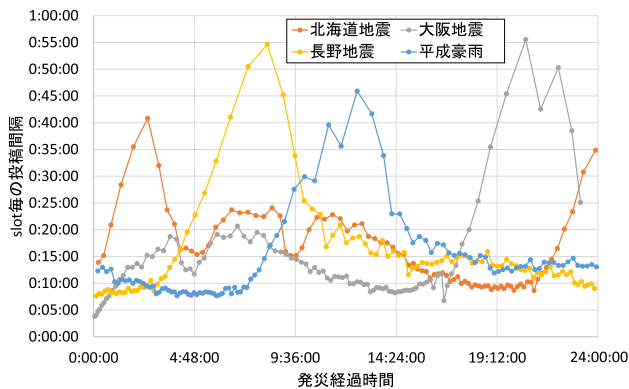


図 5 tweet 収集に要した時間

Fig. 5 Tweet collection time per slot.

ていることを示している。その後は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、おおむね 10~25 分程度のスロット時間長である。すなわち、提案手法はこの程度の時間ごと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従できることを示唆している。一方、最大 1 時間程度までスロット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 としているため分かりにくいだが、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、ここで注意すべきは、本実験で用いた試作システムでの tweet 収集部は、無料ライセンスの範囲内で Streaming API を用いた結果、全体の 1% のサンプリング tweet のみの件数を観測しているという点である。すなわち、より短スロット長を望むならば、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 5,000 tweet の条件を満たすことになる。したがって、このスロット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。また、本試作システムでの 1 スロットの処理にかかる時間は約 1 分程度であり、処理速度が問題となることも考えにくい。なお、ごく短時間のスロット長でも提案手法が有効であるかは検証していないため、本実験で検証した数分程度未満のスロット長での有効性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った場合と、各比較手法で検索を行った場合との抽出精度の結果を表 5 に示す。また、表 5 の結果から災害語を含む tweet を除外した抽出精度の結果を表 6 に示す。この評価対象に選んだスロットは 5.2 節で説明したとおり、表 4 に示す状況の変化が起きた時間帯に対応するスロットである。加

えて、各手法によって得られた検索語集合を表 7、表 8、表 9、表 10 に示す。検索語集合は表 7~表 10 で、対象とする災害の名称 (名称)、評価対象としたスロット番号 (スロット)、提案手法および各比較手法 (手法)、各手法で得られた検索語集合上位 10 語 (検索語) を示している。

表 5、表 6 は、対象とする災害の名称 (名称)、評価対象としたスロット番号 (スロット)、提案手法および各比較手法 (手法)、各手法で得られた検索語集合を用いて抽出された tweet 数 (件数)、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*8。精度、再現率より F-measure (F 値) も算出している。また、現スロットは評価対象としたスロット番号であり、次スロットは評価対象としたスロットの次の番号である。

表 5、表 6 で取り上げたスロットは、表 4 で示した状況の変化が起きた時間帯に対応した 13 スロットであり、次スロットを含めると検証に取り上げたスロットは計 26 スロットである。総合的な性能を示す F 値はこのうち 12 スロットで提案手法が最良であり、この値は災害語 χ^2 手法と同一で手法間で最大であった。この結果は、いずれも災害語を検索語として含む場合の各手法の総合的な数値としての比較であって、共通部分である災害語による検索結果の影響を受けている。そこで、表 6 で災害語を除いた検索結果の同様な評価を示す。これは、各手法が独自に拡張した検索語集合による性能評価であるといえ、この部分での F 値は、提案手法が 1 位となるスロットが 12 スロットで最大 (第 2 位は災害語 χ^2 手法で 8 スロットで 1 位) であり、提案手法の有効性を示している。

また、この結果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ともに、提案手法の必要性を示しているものといえる。一方で、長野地震においては、内閣府によると被害がなかった*9と報告されており、そもそも元 tweet 内に有効なデータがほとんど存在しなかったため、各手法において全体的に精度が低いことが確認できる。

災害語頻度手法、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。災害はふだん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

*8 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。

*9 http://www.bousai.go.jp/updates/h30jishin_nagano/pdf/300525_jishin_nagano01.pdf

表 5 各災害各手法による tweet 検索結果

Table 5 Result of extracting tweets for each disaster.

名称	スロット	手法	現スロット					次スロット				
			検索	正解	精度	再現率	F 値	検索	正解	精度	再現率	F 値
北海道地震	1	提案手法	2,145	48	0.96	0.86	0.91	1,953	48	0.96	0.91	0.94
		災害語 χ^2	2,145	48	0.96	0.86	0.91	1,953	46	0.92	0.88	0.90
		災害語頻度	2,140	44	0.88	0.78	0.83	1,983	40	0.80	0.77	0.79
		災害語	710	46	0.92	0.27	0.42	408	47	0.94	0.19	0.31
		Word2vec	1,418	43	0.86	0.51	0.64	1,242	25	0.50	0.30	0.38
		LDA	1,482	37	0.74	0.46	0.56	1,000	37	0.74	0.36	0.48
	28	提案手法	405	37	0.74	0.43	0.55	413	35	0.70	0.45	0.55
		災害語 χ^2	405	34	0.68	0.40	0.50	413	30	0.60	0.39	0.47
		災害語頻度	630	22	0.44	0.40	0.42	615	25	0.50	0.48	0.49
		災害語	84	42	0.84	0.10	0.18	87	35	0.70	0.10	0.17
		Word2vec	767	14	0.28	0.31	0.29	766	21	0.42	0.50	0.46
		LDA	767	9	0.18	0.20	0.19	697	8	0.16	0.18	0.17
	35	提案手法	404	37	0.74	0.48	0.59	349	37	0.74	0.41	0.53
		災害語 χ^2	387	35	0.70	0.44	0.54	330	42	0.84	0.44	0.58
		災害語頻度	651	18	0.36	0.38	0.37	623	23	0.46	0.45	0.46
		災害語	97	35	0.70	0.11	0.19	85	30	0.60	0.08	0.14
		Word2vec	587	15	0.30	0.29	0.29	589	14	0.28	0.26	0.27
		LDA	821	13	0.26	0.35	0.30	830	12	0.24	0.32	0.27
大阪地震	1	提案手法	2,626	46	0.92	0.87	0.90	2,421	45	0.90	0.87	0.89
		災害語 χ^2	2,626	46	0.92	0.87	0.90	2,421	44	0.88	0.85	0.87
		災害語頻度	2,556	48	0.96	0.89	0.92	2,495	43	0.86	0.86	0.86
		災害語	1,026	47	0.94	0.35	0.51	796	46	0.92	0.29	0.44
		Word2vec	1,339	41	0.82	0.40	0.53	1,253	35	0.70	0.35	0.47
		LDA	2,388	44	0.88	0.76	0.82	2,258	42	0.84	0.76	0.80
	39	提案手法	625	34	0.68	0.39	0.50	572	25	0.50	0.31	0.38
		災害語 χ^2	625	35	0.70	0.41	0.51	572	36	0.72	0.44	0.55
		災害語頻度	1,187	17	0.34	0.37	0.36	1,185	10	0.20	0.25	0.22
		災害語	181	39	0.78	0.13	0.22	195	37	0.74	0.15	0.26
		Word2vec	1,487	16	0.32	0.44	0.37	1,453	12	0.24	0.37	0.29
		LDA	583	29	0.58	0.31	0.41	551	28	0.56	0.33	0.42
	85	提案手法	1,779	48	0.96	0.91	0.93	933	37	0.74	0.66	0.70
		災害語 χ^2	1,779	44	0.88	0.83	0.86	933	37	0.74	0.66	0.70
		災害語頻度	1,766	38	0.76	0.71	0.74	1,196	20	0.40	0.46	0.43
		災害語	337	41	0.82	0.15	0.25	245	37	0.74	0.17	0.28
		Word2vec	1,036	26	0.52	0.29	0.37	1,052	20	0.40	0.40	0.40
		LDA	1,549	39	0.78	0.64	0.71	1,064	26	0.52	0.53	0.53
長野地震	1	提案手法	510	37	0.74	0.43	0.54	265	35	0.70	0.31	0.43
		災害語 χ^2	502	45	0.90	0.51	0.65	264	38	0.76	0.34	0.47
		災害語頻度	541	35	0.70	0.43	0.53	326	25	0.50	0.27	0.35
		災害語	189	47	0.94	0.20	0.33	105	49	0.98	0.17	0.29
		Word2vec	719	14	0.28	0.23	0.25	659	7	0.14	0.15	0.15
		LDA	801	15	0.30	0.27	0.29	627	11	0.22	0.23	0.23
	7	提案手法	44	11	0.25	0.19	0.22	32	10	0.31	0.16	0.21
		災害語 χ^2	41	13	0.32	0.22	0.26	37	15	0.41	0.24	0.30
		災害語頻度	256	2	0.04	0.18	0.07	243	2	0.04	0.16	0.06
		災害語	13	10	0.77	0.17	0.28	12	10	0.83	0.16	0.27
		Word2vec	314	0	0	0	0	308	0	0	0	0
		LDA	296	0	0	0	0	271	2	0.04	0.18	0.07
	12	提案手法	45	14	0.31	0.10	0.15	21	1	0.05	0.10	0.06
		災害語 χ^2	25	15	0.60	0.10	0.18	8	1	0.13	0.10	0.11
		災害語頻度	298	2	0.04	0.08	0.05	256	0	0	0	0
		災害語	10	10	1.00	0.07	0.13	2	1	0.50	0.10	0.16
		Word2vec	339	2	0.04	0.09	0.06	305	0	0	0	0
		LDA	1,542	0	0	0	0	1,510	0	0	0	0
平成豪雨	1	提案手法	369	41	0.82	0.36	0.50	245	34	0.68	0.26	0.38
		災害語 χ^2	370	41	0.82	0.36	0.50	219	31	0.62	0.22	0.32
		災害語頻度	935	16	0.32	0.35	0.33	857	13	0.26	0.35	0.30
		災害語	54	48	0.96	0.06	0.11	38	37	0.97	0.06	0.11
		Word2vec	871	12	0.24	0.25	0.24	897	11	0.22	0.31	0.26
		LDA	503	12	0.24	0.14	0.18	437	5	0.10	0.07	0.08
	15	提案手法	341	38	0.76	0.32	0.45	252	29	0.58	0.26	0.36
		災害語 χ^2	261	37	0.74	0.24	0.36	184	24	0.48	0.16	0.24
		災害語頻度	770	8	0.16	0.15	0.16	669	12	0.24	0.29	0.26
		災害語	31	29	0.94	0.04	0.07	31	27	0.87	0.05	0.09
		Word2vec	990	6	0.12	0.15	0.13	958	6	0.12	0.21	0.15
		LDA	735	6	0.12	0.11	0.11	589	4	0.08	0.08	0.08
	37	提案手法	144	32	0.64	0.20	0.30	125	33	0.66	0.17	0.26
		災害語 χ^2	144	31	0.62	0.19	0.29	115	36	0.72	0.17	0.27
		災害語頻度	1,026	7	0.14	0.31	0.19	1,072	10	0.20	0.43	0.27
		災害語	27	23	0.85	0.05	0.09	19	17	0.89	0.03	0.07
		Word2vec	288	12	0.24	0.15	0.18	270	12	0.24	0.13	0.17
		LDA	294	14	0.28	0.18	0.22	249	12	0.24	0.12	0.16
	86	提案手法	177	10	0.20	0.15	0.17	146	9	0.18	0.12	0.14
		災害語 χ^2	186	14	0.28	0.22	0.24	139	16	0.32	0.20	0.24
		災害語頻度	964	3	0.06	0.24	0.10	1,010	4	0.08	0.36	0.13
		災害語	19	11	0.58	0.05	0.09	19	12	0.63	0.05	0.10
		Word2vec	315	6	0.12	0.16	0.14	333	3	0.06	0.09	0.07
		LDA	535	3	0.06	0.13	0.08	577	6	0.12	0.31	0.17

表 6 各災害各手法による災害語を含む tweet を除いた検索結果

Table 6 Result of extracting tweets for each disaster (Except tweets including disaster word).

名称	スロット	手法	現スロット					次スロット				
			検索	正解	精度	再現率	F 値	検索	正解	精度	再現率	F 値
北海道地震	1	提案手法	1,435	32	0.9412	0.1887	0.3144	1,545	37	0.9487	0.1812	0.3042
		災害語 χ^2	1,435	31	0.9394	0.1884	0.3138	1,545	34	0.9444	0.1803	0.3029
		災害語頻度	1,430	36	0.8571	0.1713	0.2855	1,575	33	0.8250	0.1606	0.2689
		Word2vec	708	14	0.6667	0.0660	0.1200	834	11	0.3235	0.0333	0.0605
		LDA	772	13	0.5000	0.0539	0.0974	592	19	0.6552	0.0479	0.0893
	28	提案手法	321	32	0.7273	0.0351	0.0669	326	28	0.7000	0.0395	0.0748
		災害語 χ^2	321	27	0.6279	0.0303	0.0578	326	22	0.5641	0.0318	0.0602
		災害語頻度	546	11	0.2821	0.0231	0.0428	528	18	0.4286	0.0392	0.0718
		Word2vec	683	8	0.1905	0.0196	0.0355	679	15	0.3488	0.0410	0.0734
		LDA	683	3	0.0698	0.0072	0.0130	610	8	0.1633	0.0172	0.0312
	35	提案手法	307	30	0.7500	0.0424	0.0802	264	34	0.8095	0.0439	0.0832
		災害語 χ^2	290	29	0.7436	0.0397	0.0754	245	36	0.8780	0.0442	0.0841
		災害語頻度	554	13	0.3095	0.0316	0.0573	538	16	0.3721	0.0411	0.0740
		Word2vec	490	8	0.1905	0.0172	0.0315	504	8	0.1905	0.0197	0.0357
		LDA	724	11	0.2391	0.0319	0.0563	745	5	0.1190	0.0182	0.0316
大阪地震	1	提案手法	1,600	29	0.8788	0.1850	0.3056	1,625	35	0.8750	0.1602	0.2708
		災害語 χ^2	1,600	22	0.8462	0.1781	0.2942	1,625	30	0.8571	0.1569	0.2652
		災害語頻度	1,530	32	0.9412	0.1894	0.3154	1,699	23	0.7931	0.1518	0.2548
		Word2vec	313	2	0.2222	0.0091	0.0176	457	9	0.3913	0.0201	0.0383
		LDA	1,362	22	0.8800	0.1577	0.2674	1,462	26	0.8125	0.1338	0.2298
	39	提案手法	444	21	0.6563	0.0363	0.0689	377	17	0.4595	0.0165	0.0319
		災害語 χ^2	444	22	0.6471	0.0358	0.0679	377	23	0.6970	0.0251	0.0484
		災害語頻度	1,006	13	0.2955	0.0371	0.0659	990	3	0.0750	0.0071	0.0129
		Word2vec	1,306	10	0.2273	0.0370	0.0637	1,258	5	0.1220	0.0146	0.0261
		LDA	402	18	0.5455	0.0273	0.0521	356	9	0.3333	0.0113	0.0219
	85	提案手法	1,442	39	0.9512	0.1644	0.2804	688	28	0.7179	0.0611	0.1127
		災害語 χ^2	1,442	34	0.8500	0.1469	0.2506	688	25	0.7143	0.0608	0.1121
		災害語頻度	1,429	31	0.7381	0.1265	0.2159	951	16	0.3556	0.0418	0.0749
		Word2vec	699	9	0.3000	0.0251	0.0464	807	8	0.2424	0.0242	0.0440
		LDA	1,212	34	0.7727	0.1123	0.1961	819	14	0.4118	0.0417	0.0758
長野地震	1	提案手法	321	21	0.6176	0.0386	0.0727	160	13	0.4643	0.0130	0.0253
		災害語 χ^2	313	25	0.8621	0.0526	0.0991	159	25	0.6757	0.0188	0.0366
		災害語頻度	352	21	0.5833	0.0400	0.0749	221	11	0.3056	0.0118	0.0227
		Word2vec	530	1	0.0270	0.0028	0.0051	554	0	0	0	0
		LDA	612	5	0.1250	0.0149	0.0266	522	3	0.0732	0.0067	0.0122
	7	提案手法	31	1	0.0323	0.0000	0.0001	20	0	0	0	0
		災害語 χ^2	28	3	0.1071	0.0001	0.0002	25	5	0.2000	0.0002	0.0003
		災害語頻度	243	1	0.0204	0.0001	0.0003	231	0	0	0	0
		Word2vec	301	0	0	0	0	296	0	0	0	0
		LDA	283	0	0	0	0	259	0	0	0	0
	12	提案手法	35	4	0.1143	0.0001	0.0002	19	0	0	0	0
		災害語 χ^2	15	5	0.3333	0.0001	0.0002	6	0	0	0	0
		災害語頻度	288	0	0	0	0	254	0	0	0	0
		Word2vec	329	0	0	0	0	303	0	0	0	0
		LDA	1,532	0	0	0	0	1,508	0	0	0	0
平成豪雨	1	提案手法	315	31	0.7750	0.0376	0.0717	207	26	0.6190	0.0167	0.0326
		災害語 χ^2	316	33	0.7857	0.0382	0.0729	181	27	0.6000	0.0142	0.0277
		災害語頻度	881	12	0.2667	0.0362	0.0637	819	9	0.2000	0.0214	0.0386
		Word2vec	817	10	0.2083	0.0262	0.0466	859	9	0.1875	0.0210	0.0378
		LDA	449	9	0.1915	0.0132	0.0248	399	2	0.0426	0.0022	0.0042
	15	提案手法	310	34	0.7391	0.0273	0.0526	221	25	0.5435	0.0124	0.0242
		災害語 χ^2	230	33	0.7174	0.0196	0.0382	153	18	0.4186	0.0066	0.0130
		災害語頻度	739	7	0.1429	0.0126	0.0231	638	9	0.1915	0.0126	0.0236
		Word2vec	959	4	0.0833	0.0095	0.0171	927	4	0.0833	0.0080	0.0145
		LDA	704	5	0.1042	0.0087	0.0161	558	4	0.0816	0.0047	0.0089
	37	提案手法	117	23	0.5750	0.0122	0.0239	106	26	0.6047	0.0126	0.0247
		災害語 χ^2	117	25	0.5814	0.0124	0.0242	96	30	0.6977	0.0132	0.0258
		災害語頻度	999	6	0.1224	0.0222	0.0376	1,053	8	0.1667	0.0345	0.0571
		Word2vec	261	7	0.1556	0.0074	0.0141	251	5	0.1190	0.0059	0.0112
		LDA	267	10	0.2222	0.0108	0.0206	230	6	0.1364	0.0062	0.0118
	86	提案手法	158	7	0.1667	0.0016	0.0032	127	7	0.1556	0.0010	0.0020
		災害語 χ^2	167	13	0.2826	0.0029	0.0057	120	9	0.2143	0.0013	0.0026
		災害語頻度	945	2	0.0408	0.0024	0.0045	991	4	0.0800	0.0040	0.0075
		Word2vec	296	5	0.1020	0.0019	0.0036	314	1	0.0217	0.0003	0.0007
		LDA	516	3	0.0612	0.0019	0.0038	558	3	0.0667	0.0019	0.0036

きく上昇し、これらも検索語として抽出されていた。しかし、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ている。したがって、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

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ている。したがって、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

表 7 各手法における検索語集合 (北海道地震)

Table 7 A set of search words for each method (Hokkaido earthquake).

名称	スロット	手法	検索語									
			1	2	3	4	5	6	7	8	9	10
北海道地震	1	提案手法 災害語 χ^2 災害語頻度 Word2vec LDA	揺れ 揺れ 大丈夫 自分 北海道	北海道 北海道 大丈夫 いい やばい	大丈夫 大丈夫 揺れ 日本 大丈夫	余震 余震 胆振 なる ゆれ	停電 停電 地方 怖い 地震	震度 6 強 震度 6 強 中東 れる 震度 6	胆振 胆振 深さ てる 怪我	怖い 怖い てる 起き みなさん	地方 地方 台風 ビビっ 震源	怖かつ 怖かつ 停電 部屋 心配
	28	提案手法 災害語 χ^2 災害語頻度 Word2vec LDA	北海道 北海道 北海道 停電 ある	無事 無事 台風 北海道 無事	心配 心配 大丈夫 友達 こと	大丈夫 大丈夫 停電 大丈夫 よかつ	災害 災害 災害 いい 感じ	復旧 復旧 いる なっ 疲れ	電気 震度 6 強 復旧 よう 怖い	震度 6 強 電気 なる つけ いる	停電 余震 起き 電気 バイト	余震 停電 心配 被害 ところ
	35	提案手法 災害語 χ^2 災害語頻度 Word2vec LDA	北海道 北海道 あつ 大丈夫	震度 7 震度 7 震度 7 これ	電気 電気 停電 思っ てる	復旧 復旧 台風 思い 震度 7	停電 停電 てる 心配 やつ	無事 電力 気象庁 大丈夫 北海道	余震 余震 胆振 来る ほしい	気象庁 無事 日本 しれ 違う	札幌 気象庁 情報 地震	日本 札幌 ニュース 日本 停電

表 8 各手法における検索語集合 (大阪地震)

Table 8 A set of search words for each method (Osaka earthquake).

名称	スロット	手法	検索語									
			1	2	3	4	5	6	7	8	9	10
大阪地震	1	提案手法 災害語 χ^2 災害語頻度 Word2vec LDA	揺れ 揺れ 大丈夫 いる 揺れ	大阪 大阪 揺れ そう 大阪	大丈夫 大丈夫 電車 トイレ 震度 6 弱	電車 電車 大阪 でる 震度 6	ゆれ ゆれ きた ください 北部	やばい やばい すぎ もの 緊急地震速報	びっくり びっくり やばい よう 大丈夫	関西 関西 大阪府 こと くつ	緊急地震速報 緊急地震速報 てる なっ めちゃくちゃ	震度 6 弱 震度 6 弱 北部 する 大阪府
	39	提案手法 災害語 χ^2 災害語頻度 Word2vec LDA	揺れ 揺れ 大阪 揺れ 大丈夫	余震 余震 大丈夫 大丈夫 余震	無事 無事 てる 無事 ほしい	大阪 大阪 関西 つけ 怖い	大丈夫 揺れ 良かつ もらっ 結局	心配 心配 あつ ない 結局	関西 関西 こと ある なる	つけ つけ 言っ 名前	怖かつ 怖かつ 今日 これ 地震	起き 起き 怖い 余震 電車
	85	提案手法 災害語 χ^2 災害語頻度 Word2vec LDA	揺れ 揺れ 怖い 思う 感じ	怖い 怖い きた みんな 揺れ	余震 余震 てる あと 寝る	ゆれ ゆれ 揺れ てる 地震	揺れる 揺れる 大丈夫 思っ 心配	震度 4 震度 4 北部 あれ 関西	関西 関西 あつ あり てる	こわい こわい 関西 これ ねれ	大丈夫 大丈夫 起き 友達 怖く	ほんま ほんま 大阪府 思い せい

表 9 各手法における検索語集合 (長野地震)

Table 9 A set of search words for each method (Nagano earthquake).

名称	スロット	手法	検索語									
			1	2	3	4	5	6	7	8	9	10
長野地震	1	提案手法 災害語 χ^2 災害語頻度 Word2vec LDA	揺れ 揺れ 大丈夫 くる 揺れ	長野 長野 長野 みんな てる	大丈夫 震度 5 強 新潟県 くれ 地震	震度 4 長野県 揺れ フォー 言う	震度 新潟県 長野県 ツイート 今日	栄村 大丈夫 あつ いる 心配	情報 北部 新潟 情報 待つ	無事 震度 4 最大震度 しмай 忘れ	地震だ 震度 中越 限定 すごい	珍しい 栄村 予想 さん 了解
	7	提案手法 災害語 χ^2 災害語頻度 Word2vec LDA	長野 長野 みたい どこ 楽しい	おき 小動物 才能 でき 欲しい	回つ 北部 あつ 良い わたし	電車 査定 来る ライブ なん	止まり 週間 みんな よかつ 良い	強行 長野県 長野県 ちゃう つい	合宿 おき 現地 配信中 あれ	鈴井貴之 回つ 赴く プレゼント 地震	長野県上田市 才能 大丈夫 問題 いいね	つき 揺れ ところ 全て フォロー
	12	提案手法 災害語 χ^2 災害語頻度 Word2vec LDA	長野県 長野県 長野県 出来 それ	余震 起こつ 飯山線 買って てる	飯山線 新潟県 余震 時間 する	震度 5 強 予感 苦手 うち ツイート	ごろ 余震 意見 作っ そう	付け 21 時 言う ちゃっ ある	泣ける 覚め うち 良い さん	タイミング 被害 こと 確認 怖い	運転見合わせ 北部 北部 名前 誕生日	得意 揺れ 北海道 あり ない

表 10 各手法における検索語集合 (平成豪雨)

Table 10 A set of search words for each method (Heavy rain of Heisei).

名称	スロ ット	手法	検索語									
			1	2	3	4	5	6	7	8	9	10
平成 豪雨	1	提案手法 災害語 χ^2 災害語頻度 Word2vec LDA	特別警報 特別警報 特別警報 感じ いい	福岡 大雨特別警報 警報 言っ 今日	警報 福岡 大丈夫 そう ここ	佐賀 警報 てる つけ 大雨特別警報	警戒 佐賀 ください いい でき	大丈夫 警戒 影響 思う マン	最大級 大丈夫 警戒 さん 帰宅	避難 最大級 なる (絵文字) マジ	心配 長崎 福岡 思っ めちゃくちゃ	付け 避難 さん ある お気に入り
	15	提案手法 災害語 χ^2 災害語頻度 Word2vec LDA	特別警報 特別警報 ミリ みたい 特別警報	大雨特別警報 大雨特別警報 地震 仕事 しんちゃん	広島 広島 福岡 自分 やばい	避難 避難 つけ ない さん	やばい やばい てる 最近 すごい	つけ つけ なっ 思っ 好き	うまし 19 時 特別警報 れる かわいい	大丈夫 うまし 言い 意味 広島	心配 オウム ください もん 食べ	被害 7 月 6 日 大阪 すぎ フォローバ
	37	提案手法 災害語 χ^2 災害語頻度 Word2vec LDA	特別警報 特別警報 てる よかつ 大雨特別警報	無事 無事 でき 無理 今日	兵庫 兵庫 なっ 買っ 兵庫	広島 広島 大丈夫 (絵文字) 笑笑	京都 京都 明日 わから 京都	皆さん 皆さん さん もん 爆笑	間に合っ こんばんわ 特別警報 通り ゆかしやべ	こんばんわ 間に合っ 降っ 行き 溢れ	台風 願い しれ やばい 終わっ	浸水 台風 こと 忘れ かい
	86	提案手法 災害語 χ^2 災害語頻度 Word2vec LDA	七夕 七夕 七夕 うち やばい	願い事 岐阜 なっ 早く なっ	岐阜県 願い事 いい フォロー ある	被害 岐阜県 無事 楽しみ 大雨特別警報	関西 被害 心配 ところ あなた	どころ 岡山 行く 行く 最高 よう	心配 関西 行く 行く いき	見る どころ よかつ やつ 見え	実家 心配 被害 送っ やめ	アイドル 見る 九州 とき 出来る

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。

LDA 手法は、特定の話題語集合を得ることを目的とはしておらず、いずれのトピック数においても適切な語集合を得ることができない場合もあったため、精度を下げる結果となった。また、トピックを構成する語群のうち、いずれが災害に関するトピックであるかは人手で判断する必要があり、実用は困難であるといえる。

これから、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。続いて、局面の変化に追従して適切な検索語を抽出できているかを検証する。

5.3.3 スロット推移にともなう検索語の推移

表 7～表 10 が各スロットにおける検索語集合である。これらの評価対象としたスロットは、表 4 で示した変化のあった時間帯に対応している。たとえば、表 7 の北海道地震のスロット 28 では電気の復旧にともない「復旧」が、スロット 35 では気象庁による震度 7 への訂正を受けて「震度 7」が入っているなど、提案手法が状況の推移に従って適切な検索語を抽出できている様子がうかがえる。同様に大阪地震でも順に「震度 6 弱」「余震」「震度 4」といった語を代表として状況の変化をとらえられている。平成豪雨の例では地域名が推移している様子としてとらえられる。一方、被害のなかった長野地震ではそこまで明確な傾向は得られておらず、他手法との比較で地震に関連しそうな語が多く見える程度である。これらの結果からは、提案手法と災害語 χ^2 値手法とが、状況の推移に応じた検索語を抽出できているといえる。

試作システム上で逐次的に計算される検索語集合 R が、隣り合うスロット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 [20] を用いて確認した。隣り合うスロット間の検索語集合の相関の推移を図 6 に示す。図 6 は、横軸は各災害ごとの発災時刻を基点としたスロット番号、縦軸は隣り合うスロット間の順位相関係数 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。検索語集合の語の移り変わりを確認するために、数点のスロットを抜粋した。そのスロットを四角で囲み、(a)～(b) の記号を振って示すこととする。本文中で参照するときには (a) のように表記する。

(a) のスロット (16 スロット) は、停電や水道管の破裂の影響による断水が報告された時刻に相当する。これ以前では、災害が発生した周辺地域に住んでいる人の安否確認や注意喚起を表すような「大丈夫」や「お気」「つけ」といった語が比較的安定した順位で検出されていたのに対し、(a) で「断水」が入り、その後 18 スロットで「復旧」、

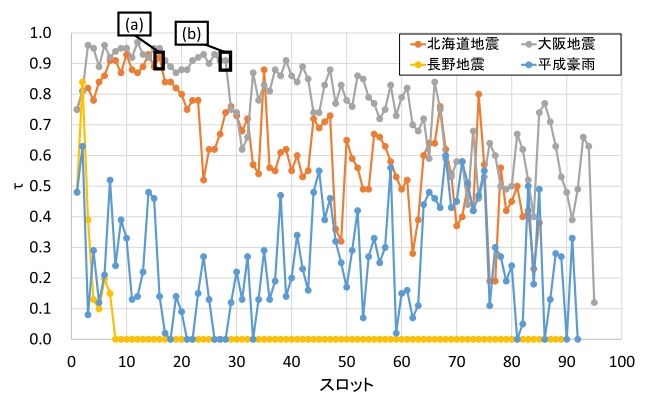


図 6 各災害における隣り合うスロット間の検索語集合の順位相関係数

Fig. 6 Rank correlation coefficient order of a set of query words between adjoining slots for each disaster.

21 スロットで「電気」といった語が出現するなどし、順位が比較的内れ替わりやすい状況になっていることが、 τ の推移からも読み取れる。

(b) のスロットの前後を見ると、発災後から (b) のスロットまで τ は横ばいになっており、検索語集合の語が変動していないことが確認できる。これは、災害が発生した周辺地域に住んでいる人の安否確認や注意喚起が、発災時刻から (b) のスロットに相当する時間帯まで行われており、安否確認や注意喚起を表すような単語が検索語集合につねに入っていたためである。一方で、(b) のスロットの次スロットでは τ が下がっており、検索語集合の入れ替わりが起きていることが確認できる。これは、余震が発生したことにより、地域や余震を表すような単語が検索語集合に入ったため、この区間は大きな変動が見て取れる。

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など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」といった水害を表すような語が、時間経過とともに検索語集合内でつね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。

長野地震では、10 スロット以降から τ が 0 となっている。長野地震は、内閣府によると被害がなかったと報告されており、地震によって日常生活に支障が出なかったため、Twitter 上で話題にならなかった。そのため、10 スロット以降から、災害語である「地震」という語を含む tweet は 1つのスロットで 0～3 件程度となり、災害語との共起語が存在せず、検索語集合の要素数が 0 という状態になったため、 τ が 0 となった。

また、北海道地震を対象に、提案手法によって得られた検索語集合の語の移り変わりが特徴的であった 1, 16 スロットの 2 点を抜粋した。1 スロット目は発災直後、16 ス

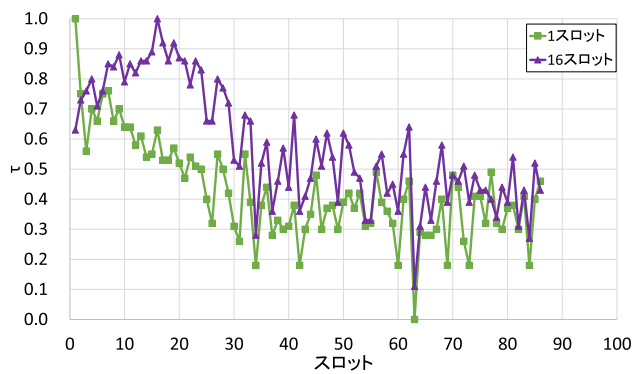


図 7 北海道地震におけるスロット間の検索語集合の順位相関係数
Fig. 7 Rank correlation coefficient order of a set of query words between adjoining slots for Hokkaido earthquake.

表 11 提案手法における検索語集合 (北海道地震: 15-35 スロット)
Table 11 Extracted words list (case: Hokkaido earthquake, slot: 15-35).

スロット	1	2	3	4	5	6	7	8	9	10
15	北海道	無事	余震	大丈夫	心配	停電	災害	起き	つけ	お気
16	北海道	無事	余震	大丈夫	心配	停電	断水	災害	揺れ	続い
17	北海道	無事	心配	余震	大丈夫	停電	災害	断水	家族	備え
18	北海道	無事	停電	余震	災害	大丈夫	心配	充電	携帯	復旧
19	北海道	大丈夫	災害	停電	余震	心配	断水	怪我	揺れ	起き
20	北海道	無事	停電	余震	心配	断水	災害	怪我	揺れ	起き
21	北海道	無事	断水	11時	余震	大丈夫	心配	充電	災害	使用
22	北海道	災害	心配	11時	余震	復旧	断水	大丈夫	起き	連絡
23	北海道	無事	停電	断水	余震	大丈夫	心配	11時	状況	電波
24	北海道	無事	停電	断水	余震	大丈夫	心配	11時	状況	電波
25	北海道	無事	12時	余震	心配	断水	充電	充電	状況	電波
26	北海道	災害	余震	無事	心配	断水	充電	充電	状況	電波
27	北海道	災害	断水	停電	余震	復旧	断水	充電	状況	電波
28	北海道	無事	心配	大丈夫	災害	復旧	断水	充電	状況	電波
29	北海道	停電	余震	災害	復旧	断水	充電	充電	状況	電波
30	北海道	復旧	停電	無事	災害	復旧	断水	充電	状況	電波
31	北海道	電気	復旧	電力	災害	復旧	断水	充電	状況	電波
32	北海道	電気	余震	停電	災害	復旧	断水	充電	状況	電波
33	北海道	電気	心配	余震	災害	復旧	断水	充電	状況	電波
34	北海道	電気	復旧	テク	充電	心配	断水	充電	状況	電波
35	北海道	震度7	電気	復旧	停電	無事	余震	気象庁	札幌	日本

ロット目は停電や水道管の破裂の影響による断水が報告された時刻に相当する。

検索語集合の累積的な変化を見るために、北海道地震における発災直後のスロット 1 と、断水の報告されたスロット 16 それぞれの検索語集合との順位相関係数の推移を図 7 に示す。1 スロット目との r は 0.7 付近まで一気に下がり、その後 30 スロット付近にかけて順次低下している様子が見られる。これは 1 スロット目の「怖い」「揺れ」「胆振」「震度 6 強」といった揺れに対する直接的な語から被害を示す語へと変わり、その後緩やかにそれらの語の種類が増えていった様子を表している。

このことは、スロット 16 との r の推移によっても裏付けられる。この推移グラフは、スロット 16 を頂点として両側に裾を引くような形状となっている。これは、近い時間帯では検索語も似た傾向にあることを示している。また、30 スロット目付近以降は、1 スロット目との r と同様な水準で前後しながら推移している。この区間に対応する検索語リストの一覧を表 11 に示す。状況の推移に沿って現れるいくつかの語には共通の色をつけてある。各スロット間で一貫しているのは「北海道」「無事」「災害」「心配」といった語群であり、これらが、30 スロット目以降でも一定の相

関値を示している理由である。一方、「電気」「復旧」は時間を追うごとに継続的に抽出されるようになっており、状況の変化にともない tweet が変化した結果として、適切な検索語の抽出が行われている様子がうかがえる。

これらの検索語の累積的な変化推移の様子からは、特定の時間帯に対応する検索語集合が得られていることが分かる。また、時間帯間での検索語集合はまったくの独立ではなく、近傍の時間帯では相対的に高い類似性を示していることも分かる。具体的な検索語と突き合わせた結果、被害の拡がりの様子を示す検索語の推移とよく合致していることをふまえると、提案手法が状況の推移に追従した検索語集合が得られていることが確認できた。

5.3.4 提案手法と災害語 χ^2 手法の比較

比較手法の中で、提案手法に最も近い結果を得ている災害語 χ^2 手法との違いについて検証する。両者の比較には、検索語集合の要素が異なる語のみを用いた tweet 検索結果の精度による評価を用いる。評価対象は、北海道地震において震度 7 に気象庁発表の訂正があった時間帯であるスロット 35 以降である。発災後間もない時間帯は、両手法における検索語にほとんど差がないため、図 7 に示したように話題が広がり、さらに震度の訂正という話題までも広がってきたスロット 35 以降で比較する。提案手法と災害語 χ^2 手法で検索した場合との抽出精度の結果を表 12、表 13 に示す。

表 12 は、対象とする災害の名称 (名称)、2 手法で得られた検索語集合の単語が異なるスロット番号 (スロット)、提案手法および災害語 χ^2 手法 (手法)、各手法で得られた検索語集合を用いて抽出された tweet 数 (件数)、検索結果を人手により正解とされた tweet の件数 (正解)、検索結果 tweet に含まれる正解 tweet 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)、精度、再現率より算出した F-measure (F 値) を示している。表 13 は、表 12 より各手法の精度、再現率、F 値の差をとり、その平均を示している。これは提案手法の方が精度が良ければ正の値になり、災害語 χ^2 手法の方が良い場合は負の値になる。

表 12、表 13 より、提案手法によって得られた検索語を用いた検索結果が、多くのスロットで F 値の観点で最も良い結果を得たことが分かる。提案手法において、精度の良い語を見逃している代わりに、漏れなく網羅的に災害に関連する投稿を検索できる検索語が抽出できたため、精度については 2 つの手法間で差はあまり見られないが、再現率については提案手法の方が良い結果を示した。たとえば、48 スロット目では、精度の高い「胆振」「震度 7」という語を見逃した代わりに再現率の高い「安心」「情報」という語を抽出したため、F 値の観点で提案手法の方が良い結果を得ることができた。

提案手法が有利不利両方の状態で連続するスロットの例

表 12 2 手法での異なる検索語のみによる評価

Table 12 Result by the different words between the two methods.

スロット	手法	検索	正解	精度	再現率	F 値
35	提案手法	23	1	0.043	0.167	0.069
	災害語 χ^2	6	5	0.833	0.833	0.833
36	提案手法	25	5	0.200	1	0.333
	災害語 χ^2	2	0	0	0	0
37	提案手法	26	5	0.192	0.313	0.238
	災害語 χ^2	29	11	0.379	0.688	0.489
38	提案手法	34	19	0.559	0.704	0.623
	災害語 χ^2	19	8	0.421	0.296	0.348
40	提案手法	19	1	0.053	0.500	0.095
	災害語 χ^2	8	1	0.125	0.500	0.200
41	提案手法	31	18	0.581	0.900	0.706
	災害語 χ^2	13	2	0.154	0.100	0.121
42	提案手法	8	5	0.625	1	0.769
	災害語 χ^2	0	0	0	0	0
43	提案手法	5	0	0	0	0
	災害語 χ^2	7	2	0.286	1	0.444
45	提案手法	3	2	0.667	1	0.800
	災害語 χ^2	0	0	0	0	0
46	提案手法	44	31	0.705	1	0.827
	災害語 χ^2	8	0	0	0	0
47	提案手法	19	9	0.474	0.750	0.581
	災害語 χ^2	7	3	0.429	0.250	0.316
48	提案手法	36	10	0.278	0.769	0.408
	災害語 χ^2	7	3	0.429	0.231	0.300
49	提案手法	14	0	0	0	0
	災害語 χ^2	10	0	0	0	0
51	提案手法	0	0	0	0	0
	災害語 χ^2	5	0	0	0	0
52	提案手法	23	22	0.957	0.759	0.846
	災害語 χ^2	11	7	0.636	0.241	0.350
53	提案手法	2	0	0	0	0
	災害語 χ^2	2	2	1	1	1
54	提案手法	7	0	0	0	0
	災害語 χ^2	10	1	0.100	1	0.182
55	提案手法	69	4	0.058	0.235	0.093
	災害語 χ^2	22	13	0.591	0.765	0.667
56	提案手法	56	20	0.357	1	0.526
	災害語 χ^2	11	0	0	0	0
57	提案手法	17	10	0.588	0.769	0.667
	災害語 χ^2	17	3	0.176	0.231	0.200
58	提案手法	77	30	0.390	0.811	0.526
	災害語 χ^2	8	7	0.875	0.189	0.311
59	提案手法	19	0	0	0	0
	災害語 χ^2	4	4	1	1	1
60	提案手法	15	0	0	0	0
	災害語 χ^2	6	2	0.333	1	0.500
61	提案手法	21	16	0.762	0.889	0.821
	災害語 χ^2	5	2	0.400	0.111	0.174
62	提案手法	21	21	1	0.955	0.977
	災害語 χ^2	3	1	0.333	0.045	0.080
63	提案手法	209	8	0.038	0.421	0.070
	災害語 χ^2	32	11	0.344	0.579	0.431
65	提案手法	11	5	0.455	0.556	0.500
	災害語 χ^2	5	4	0.800	0.444	0.571
66	提案手法	20	2	0.100	1	0.182
	災害語 χ^2	4	0	0	0	0
68	提案手法	4	3	0.750	1	0.857
	災害語 χ^2	23	0	0	0	0
69	提案手法	40	11	0.275	1	0.431
	災害語 χ^2	29	0	0	0	0
70	提案手法	12	7	0.583	1	0.737
	災害語 χ^2	1	0	0	0	0
71	提案手法	36	6	0.167	0.857	0.279
	災害語 χ^2	4	1	0.250	0.143	0.182
73	提案手法	31	8	0.258	0.889	0.400
	災害語 χ^2	6	1	0.167	0.111	0.133
75	提案手法	31	0	0	0	0
	災害語 χ^2	17	2	0.118	1	0.211
76	提案手法	25	13	0.520	1	0.684
	災害語 χ^2	14	0	0	0	0
77	提案手法	194	6	0.031	0.600	0.059
	災害語 χ^2	7	4	0.571	0.400	0.471
79	提案手法	9	0	0	0	0
	災害語 χ^2	10	1	0.100	1	0.182

表 13 2 手法間の差平均

Table 13 Average of difference by the two methods.

スロット	精度	再現率	F 値
35-79	0.022	0.235	0.119

として示す。スロット 37 での差異語は、3 語ずつで、提案手法のみに含まれる検索語が「中止」「16 時」「本震」、比較手法のみが「震度 7」「余震」「9 月 6 日」である。また、スロット 38 では 2 語ずつの差異語となっており、提案手法のみが「心配」「無事」、比較手法のみが「中止」「16 時」である。表 14 に各スロット各語のみを含む tweet をスロット内時間順で各 3 tweet ずつ例示している。ここでは比較を目的としているため、他の両手法に共通で抽出された語、たとえば「北海道」「電気」などを含む tweet は比較対象から除外している。そのため、目的に適うと考えられる tweet が少ない状況となっている点は注意されたい。

これらの語のうち、「16 時」「9 月 6 日」の日時を示す語はあまりよい結果をもたらしておらず、両スロットで両手法にまたがって抽出されていることから、順位的に境界線上に存在するような状況となっていると考えられる。当該スロットの日時であるため、目的とする tweet も含まれるものの、ノイズとなる関係ない tweet も多い語であるといえる。同様に両手法にまたがって抽出された語には「中止」があげられ、これも境界線上の順位であるためと考えられ、閾値設定に課題を残す結果といえる。地震で中止になったイベントの話題など、語としての寄与はある程度ある。

提案手法独自で抽出されている語群は「本震」「心配」「無事」である。このうち「本震」は比較手法での「震度 7」や「余震」とも対応するような意味合いの語として機能しており、両者の優劣はつけにくい。これに対して、「心配」「無事」といった直接的に相手を思いやるような語は提案手法特有といえ、直接現地につながるアカウントとのやりとりであることが読み取れる。これらの語を含む tweet は目的とする災害に関する情報を含んでいることも相対的に多かった。

提案手法独自で抽出された語に再現率に寄与するような語が多かったのは事実であるが、個別の語の意味についての分析は定量的に行うことは困難で、例示では精度の低さから良さを示すことは難しいかもしれない。抽出された検索語レベルで提案手法独自の語をいくつかあげると「安心」「情報」「大丈夫」などであり、これに対する比較手法では「胆振」「震度 7」などが多かった。この先は意味的な解釈に踏み込む必要があるため断言はできないが、これらの語を並べてみると提案手法のほうが状況を主観視しているときに用いられやすい語が抽出されているように思われる。提案手法を元にして、検索後の tweet の内容分析、意味抽出や、reply など関連 tweet への拡張などとの組合せについては今後に取り組みを進めたい課題である。

これから、提案手法は災害語 χ^2 手法と比較して、適切な検索語集合を抽出できているといえ、提案手法の有効性を確認した。

表 14 2 手法での tweet の実例

Table 14 Examples of tweets by the two methods.

スロット	手法	検索語	tweet 本文
37	提案手法	中止	断水についてデマ情報が流れているので水道局の情報を信じましょう。計画断水は中止と書かれております。
			ファンミ札幌中止に魔剤 3 本賭けた。
			日本人じゃ無い説出たわいいねのやつ事件が発生したため中止になりました
		16 時	16 時になったわよ。
			16 時 5 分です。
			あれおかしいな…今日の 16 時に管理会社の人来るって言ってただけだなあ…
		本震	@■ 本震怖いねそれとしゅんぺーにくびーそうさんに連絡取れとるん？
			@■ うん！今ヘルメット被って本震に備えてる姉御ありがとう (≧▽≦*)
			@■ これが本震じゃないとしたら……………
	災害語 χ^2	震度 7	気象庁 報道発表資料 (URL) 「震央付近の震度データが入電してないから、職員が現地に行って確認する」っていう発表があったけど、それで現地確認したら震度 7 だったってことか…
			震度 7 って…阪神淡路大震災クラスだったというのか
			東日本の時よりやばいじゃん震度 7 って建物全部壊れるみたいな強さじゃないっけ
		余震	@■ 大丈夫？しばらく余震続くかもしれないから気をつけてね
			@■ これから暗くなりますのでどうぞ気をつけて下さい。余震にも充分注意して下さい
			家の中は少しの被害ですんだけど外に出ると道路はぐにやぐにや。傾いた家もある。大きな余震来ないで
		9 月 6 日	2018 年 9 月 6 日風間隆夫です (URL)
			auiPhone8 を MNPau 購入サポート対象へ 9 月 6 日から価格相場変動に注意 (URL)
			2018 年 9 月 6 日 (木) 08:10~08:19—HondaSmileMission—TOKYO FM—(URL) (URL)
38	提案手法	心配	2 ない様子を見せながら「皆さんお久しぶりです。本日は私ごときのために集まっていたいただき、ありがとうございます」とあいさつ。「度重なる事件によっていろんな方を傷つけたりご心配をかけたことをお詫びします」と頭を下げた。
			「Ingress」アニメ化の心配は試写会で吹き飛んだ？ ITmediaPC USER
			@■ 内地で出来た友達から、私の親戚を心配してきた友達には、厚真町から札幌まで直線で 70 キロメートル、帯広まで 100 キロメートルあると答えといた。虻田も、どうやら断水まではしてないよう。
		無事	9 月は全部有給消化。1 ヶ月の休みの使い方は、お世話になった人たちに大切な人たちに、会いに行くって手紙を書くって決めてた。9 月に入って一番最初に苦小牧の友人に手紙を書いた。「5 日に着くって」「5 日仕事だから 6 日受け取るね」って言ったのに。どうか無事でいてください。
			本日、無事に双子の女の子が産まれました。予定より 10 日ほど早かったけど母子共に健康です。甚大な災害が続いたなか生まれてきたこの子達には、常に人を思いやれる優しい子に育ってほしい。そして世界一のママありがとう！
			@■ ご無事で
	災害語 χ^2	中止	ファンミ中止か~~~~~
			みんな中止やなあーハジマザやるかなー
			ギルキスファンミ中止決定。振替公演も検討していただいてるみたいだしいつか開催できることを祈っておこう。(URL)
		16 時	大雨と突風及び落雷に関する京都府気象情報 第 1 号 平成 30 年 09 月 06 日 16 時 18 分 京都府気象台発表 (見出し) 京都府では、7 日未明から 8 日にかけて局地的に雷を伴って激しい雨が降り、大雨となる おそれがあります。土砂災害に十分注意し、浸水害、河川の増水、竜巻などの激しい…
			FC2 ライブで 16 時 26 分から放送開始しました (URL)#fc2live
			16 時 20 分です

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、北海道胆振地方で起きた震度7の地震をはじめとした、最近の異なる種類の大規模災害時に投稿された 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、人手の介入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。

参考文献

- [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] 毎日新聞: 情報発信でツイッター活用 大西市長に聞く, 入手先 (<http://mainichi.jp/articles/20161017/k00/00e/040/121000c>) (参照 2018-12-10).
- [3] Funayama, T., Yamamoto, Y. and Uchida, O.: Development of Visualization Application of Tweet. Data for Extracting Information in case of Disaster, *Proc. 2017 15th International Conference on ICT and Knowledge Engineering (ICT&KE)*, pp.1-5, DOI: 10.1109/ICTKE.2017.8259620 (2017).
- [4] Middleton, S.E., Middleton, L. and Modafferi, S.: Real-Time Crisis Mapping of Natural Disasters Using Social Media, *Proc. IEEE Intelligent Systems*, Vol.29, No.2, pp.9-17, DOI: 10.1109/MIS.2013.126 (2014).
- [5] 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, DOI: 10.1145/1772690.1772777 (2010).
- [6] 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*, pp.1-7, DOI: 10.1109/Kaleidoscope.2015.7383642 (2015).
- [7] 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).
- [8] 高木涼太, 風間一洋, 榎 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).
- [9] 大石哲也, 倉元俊介, 峯 恒憲, 長谷川隆三, 藤田 博, 越村三幸: 関連単語抽出アルゴリズムを用いた Web 検索クエリの生成, 電子情報通信学会論文誌 D, Vol.J92-D, No.3, pp.281-292 (2009).
- [10] 松尾 豊, 石塚 満: 語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム, 人工知能学会論文誌, Vol.17, No.3, pp.217-223 (2002).
- [11] 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).

- [12] 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, Vol.114, No.81, NLC2014-2, pp.7-12 (2014).
- [13] 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).
- [14] 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).
- [15] 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).
- [16] 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).
- [17] 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, DOI: 10.1109/SMARTCOMP.2018.00023 (2018).
- [18] 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, DOI: 10.1145/2141512.2141571 (2012).
- [19] 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).
- [20] Fagin, R., Kumar, R. and Sivakumar, D.: Comparing top k lists, *Proc. 14th Annual ACM-SIAM Symposium on Discrete Algorithms*, Vol.17, No.1, pp.134-160 (2003).



湯沢 昭夫

2017 年工学院大学工学部情報通信工学科卒業。2019 年同大学大学院工学研究科電気・電子工学専攻修士課程修了。同年株式会社 DTS 入社。



小林 亜樹 (正会員)

2000 年東京工業大学大学院理工学研究科電気・電子工学専攻博士後期課程修了。博士 (工学)。同大学院助手, (独) メディア教育開発センター助教授等を経て現在, 工学院大学情報学部情報通信工学科准教授。主に, 情報検索・推薦, ネットワーク, 画像認識, インタラクティブシステムの研究に従事。電子情報通信学会, 日本データベース学会, ACM 等各会員。

(担当編集委員 山本 岳洋)