

TFIDFを用いた災害・危機に関する言語資料体からの キーワード自動抽出手法の外的妥当性の検証

The External Validity of an Algorithm Using TFIDF to Extract the Set of Keywords of
Corpus about Disasters and Crises

佐藤 翔輔¹, 林 春男², 牧 紀男², 井ノ口 宗成¹

Shosuke SATO¹, Haruo HAYASHI², Norio MAKI² and Munenari INOBUCHI¹

¹ 京都大学大学院 情報学研究科

Graduate School of Informatics, Kyoto University

² 京都大学 防災研究所

Disaster Prevention Research Institute, Kyoto University

In this study, the external validity of the automatic keyword detection method using TFIDF (Term Frequency/Inversed Document Frequency) which we developed for identifying chronological changes in keywords from a total of 2,623 newspaper articles on the 2004 Niigata-ken Chuetsu earthquake. We applied this keyword extraction method for the 19 corpuses of the digital news text data taken from 13 crisis events including both natural hazards and social hazards. As a result, a total of at least 500 news articles should be acquired as the corpus to be analyzed before this method detect a “meaningful set of keywords”, which can be automatically determined with a 89.5% accuracy by taking into account the maximum number of keywords to be detected and the threshold of TFIDF deviation value.

Key Words : keyword detection, disaster social phenomenon, disaster intelligence, web news highlighting disasters, text mining

1. はじめに

(1) 社会現象としての災害・危機に関する情報資料の重要性と問題点

防災の世界では、災害の自然現象としての側面のほか、被災者・対応従事者・被災地外の人々を含む社会の対応や災害からの復興問題などに着目する災害の社会現象としての側面への関心が高まってきている。災害の自然現象を取り扱う領域では、地震動のゆれ、降雨量、河川の水位などを常時観測し、現況把握を迅速におこなったり、蓄積された膨大な観測結果を解析することにより、防災力の向上が図られてきた。自然科学の分野の例にならえば、災害の社会現象に着目する場合でも、社会現象に関する情報資料を迅速に収集し、災害・危機が発生した状況をリアルタイムかつ的確に捉えつづけて、過去の災害・危機の情報資料を利用しやすいかたちで蓄積していく必要がある。

社会現象としての災害・危機に関する情報資料の例としては、被害・対応速報、新聞記事、関係機関が公開するウェブページ、体験談、刊行資料などが挙げられる。これらの情報資料は、災害時には現況把握のために用いられ、平時には過去の危機対応の分析や危機対応訓練のさいの災害現象の再現に用いられる。これらの情報資料を蓄積・検索・利用しようとする場面では、データの形

態が言語データ（テキスト）であることと、データの量が膨大であることが原因となり、情報資料を蓄積するさいに行うキーワードの付与と情報資料の閲覧に多大な時間と労力を要してしまう。情報資料へのキーワードを付与することは、課報活動の場面においても重要な手続きとして位置づけられており、情報資料を管理する上で、またトレンドを分析する上でも欠かすことのできない手続きとなっている¹⁾。言語データの場合には、強震動の加速度波形の時系列変化、雨量の時系列変化と累積雨量など、自然科学の分野に見られる観測結果を定型的に可視化する方法は存在しないため、情報資料を閲覧しようとする場合には、多くの時間を要することになる。

(2) TFIDFを用いた災害・危機に関する言語資料体からの キーワード自動抽出手法の提案

佐藤ら²⁾は以上のような問題を受けて、情報資料の管理に要する労力の軽減を図ることと、膨大な量の文書から情報資料がもつ特徴を概観できることをねらいとし、災害・危機に関する言語資料体からキーワードを自動的に抽出する手法を開発した。手法を開発するに当たり、災害・危機に関する言語資料体は、時間の経過とともに増加するという特徴をもつことから、このような言語資料体に対応できる仕組みが必要となった。これを受け、キーワード抽出に用いられる言葉に重みを与える既存の指標に時間のパラメータを導入し、時々刻々と増加する

言語資料体に対しても、言葉に重みを与えることを可能にした。また、統計学の分野で用いられている残差分析手法を組み合わせることで、任意の時間幅で増加した言語資料中の語とそれ以前にコーパス中にあった語を比較しながらキーワードを抽出する手法を開発した。

先行研究では、代表的なポータルサイト³⁾上に配信された平成16年新潟県中越地震を取り上げたウェブニュース記事(全2,623件)に対して提案手法を適用した。その結果、命を守る活動、社会のフローを回復させる活動、社会のストックを再建する活動が、それぞれ10~100時間、100~1,000時間、1,000時間以降を中心として同時並行で展開するという災害過程の理論モデル⁴⁾と符合することが確認され、手法の一定の有効性が示された。

(3)研究の目的

提案されたキーワードを自動的に抽出する手法についての議論は、外的妥当性の検証に至っていない。どのような災害・危機に関する言語資料体(コーパス)に対しても、キーワードを適切に抽出できるかは明らかになっていない。また、提案手法は解析結果から得られた残差値の大きいものから、操作的に上位3つをキーワードとして選定しているが、残差値上位3つという数の過不足についても未確認である。そこで本研究では、災害・危機を取り上げた複数のウェブニュースを提案手法への適用を通して、キーワード自動抽出手法の外的妥当性を検証することを目的とする。具体的には、提案手法がもつ特徴や問題点を把握し、手法の一般性の確認と手法の精緻化を目指す。

2. キーワード自動抽出手法の概要

佐藤ら²⁾は、時系列的に増加する災害・危機に関する言語資料体(コーパス)から、災害過程が時々刻々と変化する様子をキーワードを用いて可視化し、大量の文書がもつ情報量を縮約するテキストマイニング手法を提案した。本章では、提案手法の概要を述べる。

まず、テキストデータベースからテキスト群を形態素解析⁵⁾によって形態素への分かち書きをおこない、品詞情報にもとづいて、非自立と接尾以外の名詞・動詞・副詞・形容詞を抽出する。抽出した形態素について、単位文書ごとに(記事ごと)、単語の出現頻度 TF と、時間増加型 $TFIDF$ ⁽¹⁾ という単語⁽²⁾の重み付け指標(式[1])を算出する:

$$\text{時間増加型}TFIDF(t_i, d_j) = TF(t_i, d_j) \cdot \log \frac{N_j}{DF(t_i, d_j)} \quad [1]$$

t_i : 時間増加型 $TFIDF(t_i, d_j)$ を算出する対象となる単語。 i はその識別子を表す添え字。

d_j : 時間増加型 $TFIDF(t_i, d_j)$ を算出する単語 t_i が含まれている単位文書。文書の単位は文章、段落、文など任意に定める。この研究では、記事を文書単位とする。 j は識別子を表す添え字。

N_j : 文書 d_j (記事)が出現した時点での文書(記事)の総数。

$TF(t_i, d_j)$: Term Frequencyの略。単語 t_i が記事 d_j に出現した回数。

$DF(t_i)$: Document Frequencyの略。キーワード t_i が出現する文書数(文書頻度)。

ある時点 $t - \Delta t$ から t の間におけるコーパス $C_{t-\Delta t}$ (コー

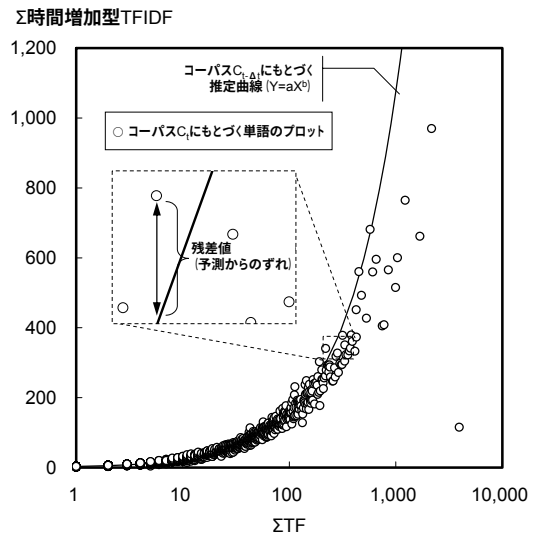


図1 キーワード自動抽出手法の概要
(ΣTF と Σ 時間増加型 $TFIDF$ の関係)

パスの増分) からキーワードを抽出するために、時点 $t - \Delta t$ までのコーパスにもとづく TF の累積値 (ΣTF) と時間増加型 $TFIDF$ の累積値 (Σ 時間増加型 $TFIDF$) の関係式を求める。佐藤らは、 ΣTF と Σ 時間増加型 $TFIDF$ の間には、強い曲線関係があることを明らかにした。これを受け、 ΣTF を独立変数 X 、 Σ 時間増加型 $TFIDF$ を従属変数 Y にした $Y=aX^b$ (a, b は定数) の曲線のあてはめをおこなった推定式を用いて、時点 t の ΣTF を入力値とした時点 t の Σ 時間増加型 $TFIDF$ の推定値を求める。時点 t で実際に計測された Σ 時間増加型 $TFIDF$ の実測値から、ある時点 t から Δt さかのぼった時点のコーパス $C_{t-\Delta t}$ の特性から予測される Σ 時間増加型 $TFIDF$ の推定値を減じたもの(残差値)を予測からのずれとして捉え、単語の特異性として数量的に評価する(図1)。先行研究では、残差値が大きい上位3つを単語をキーワード(特異語⁽³⁾)として抽出する。

3. 災害・危機を取り上げた複数のウェブニュースコーパスの構築とキーワード自動抽出手法の適用

先行研究では、新潟県中越地震について報じたウェブニュースをキーワード自動抽出手法に適用し、手法の内的妥当性を確認した。社会を取り巻く災害や危機は、新潟県中越地震のような地震をハザード(外力)とするものや、発生からおおよそ半年で二千件を上回るようなニュース記事が発信されるような社会の関心の高いものに止まらない。地震とは異なるハザードによって発生するもの、社会の関心の高いものから低いものまで、多種・多様な災害や危機が存在する。これらのパラメータが異なる場合においても、提案手法が的確なキーワードを常に抽出できるかについては明らかになっていない。また、提案手法の適用結果から抽出されたキーワードの候補のうち、残差値の大きい上位3つをキーワードとして判定することの妥当性も他の事例では確認されていない。

以上の観点からの検証をねらいとし、多種類のハザードに起因する災害・危機について報じたウェブニュースに着目して、大小様々なコーパスの大きさ(記事数や記事の増加傾向)になることと、また提案手法から得られ

表1 災害・危機を取り上げたウェブニュースから構築したコーパス

イベント	情報源	収集期間	件数
平成16年新潟県中越地震	Yahoo!ニュース	2004/10/23～2006/4/30	2,623
平成17年宮城県沖で発生した地震	Yahoo!ニュース	2005/8/16～2005/9/16	278
平成18年豪雪	YOMIURI ONLINE	2005/12/18～2006/3/22	160
平成18年豪雪	KYODO NEWS	2005/12/14～2006/3/14	148
平成18年豪雪	Asahi.com	2006/1/18～2006/3/20	50
平成18年豪雪	さがけ on the web	2006/1/05～2006/2/26	225
平成18年豪雪	信毎ニュース特集	2005/12/31～2006/3/08	187
平成18年豪雪	新潟日報社ホームページ	2006/1/4～2006/2/08	59
平成18年豪雪	6新聞社*	2005/12/14～2006/3/22	829
平成18年7月豪雨	Yahoo!ニュース	2006/7/18～2007/1/04	805
首都圏大停電	Yahoo!ニュース	2006/8/14～2006/9/21	93
平成18年台風13号, 14号	Yahoo!ニュース	2006/9/10～2007/1/14	349
北朝鮮の核実験	Yahoo!ニュース	2006/10/09～2007/1/20	1,933
北海道佐呂間町の竜巻	Yahoo!ニュース	2006/11/07～2007/1/10	27
ノロウィルス感染	Yahoo!ニュース	2006/11/20～2007/1/14	59
「あるある大事典II」のねつ造問題	Yahoo!ニュース	2007/1/11～2007/3/02	194
不二家の期限切れ原料使用問題	Yahoo!ニュース	2007/1/11～2007/3/02	240
宮崎島の鳥インフルエンザ問題	Yahoo!ニュース	2007/1/12～2007/3/01	34
平成19年能登半島地震	Yahoo!ニュース	2007/3/25～2007/4/30	654

*平成18年豪雪について取り上げた各新聞社のニュースサイトを統合

の残差値の分布が多様になることを期待し、13のイベントについて、全部で19種類の災害・危機について報じたウェブニュースを収集し、ウェブニュースコーパスを構築した(表1)。対象とした災害・危機には、地震以外の自然現象をハザードとする6つの自然災害、事故、武力攻撃、健康を脅かすもの、社会問題をハザードとする6つの危機が含まれている。記事が100件を下回るコーパスが6つ、100件以上200件未満のコーパス4つ、200件以上500件未満のコーパスが4つ、500件以上1,000件未満のコーパス2つ、1,000件以上2,000件未満のコーパスを1つ構築した。平成18年豪雪については、表1に示した全国紙3社と地方紙3社が公開した豪雪災害の特集ページから収集した全ウェブニュースを統合し、新たなウェブニュースコーパスを作成した。

提案手法は、任意の時間幅 Δt を設定し、 Δt 時間経過することによって増加する $C_{\Delta t}$ 中に含まれる単語の特異性を評価する。この研究では、すべてのウェブニュースコーパスに対して、 Δt を 24 時間一定としてキーワード自動抽出手法の適用を試みた。表 1 に示した多くのサイトでは、ウェブニュース記事に配信された日付と時間の情報が明記されている。それに対して、平成 18 年豪雪について配信された共同通信、信濃毎日新聞、新潟日報のウェブニュース記事には配信時間が示されておらず、ニュース記事がもつ時間情報は配信された日付の精度に止まっている。すべてのウェブニュースコーパスから同じ条件で抽出されたキーワード群をもとに提案手法の検証を進めることを方針とし、すべてのウェブニュース記事がもっている更新日の精度でキーワード自動抽出手法を適用した。

4.災害・危機について報じたウェブニュースからのキーワード自動抽出の妥当性の検証

3章に挙げた手法の妥当性を検証する観点のうち、ハザードの種類やコーパスの大きさを問わずに、イベント

の中で意味上重要な単語をキーワードとして自動的に抽出できる可能性の有無について議論する.

(1)地震とは異なるハザードに起因する災害・危機事例への適用の妥当性の検証

先行研究でキーワード自動抽出手法の妥当性を確認したウェブニュースコーパスは、新潟県中越地震災害を取り上げたものであった。地震というハザードとそれに起因して発生した災害現象がもつ特性そのものを反映して、適切なキーワードが提案手法によって抽出された可能性は否定できない。この節では、提案されたキーワード自動抽出手法は、地震災害を取り上げたウェブニュースコーパスからのみ的確なキーワードを抽出するという命題の真偽を問う。

地震災害とは異なる災害・危機のうち、自然災害の例の中から平成 18 年 7 月豪雨と、人為的な原因による危機として北朝鮮の核実験について報じたウェブニュースコーパスを用いて、提案手法のもつハザードへの影響を考察する。図 2 に平成 18 年 7 月豪雨、図 3 には北朝鮮の核実験にキーワード自動抽出手法を適用した結果を示した。それぞれの災害・危機を取り上げた初出のニュース記事が配信されたそれぞれ 7 月 18 日、10 月 9 日から 1 ヶ月間のウェブニュース記事を用いている。

平成 18 年 7 月豪雨のウェブニュースからのキーワードの抽出結果(図 2)では、7 月 23 日から「鹿児島」が出現し、それから急激に残差値が増加していき、7/25 に 7 月豪雨の解析結果中で最も大きい残差値 56.7 を示した。7/24 には災害対策関係省庁局長会議を開催され、政府調査団が鹿児島県に派遣されることが決定し、翌日 7/25 に同団が派遣されているという事実⁶⁾との対応関係が見ら

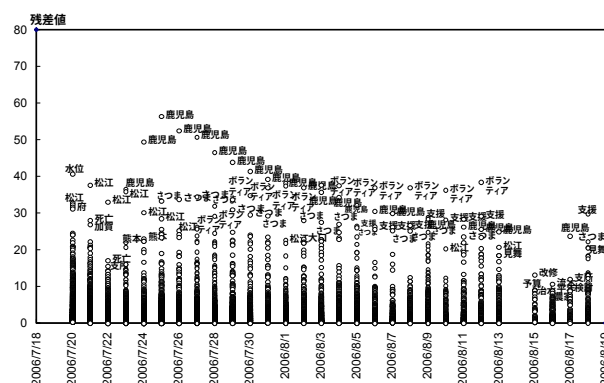


図2 平成18年7月豪雨のウェブニュース
コーパスからのキーワード自動抽出結果

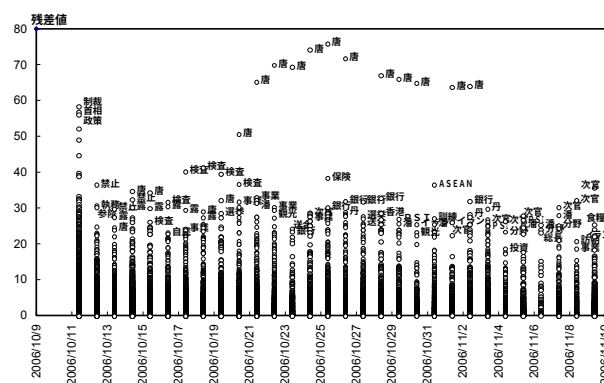


図3 北朝鮮の核実験に関するウェブニュースコーパスからのキーワード自動抽出結果

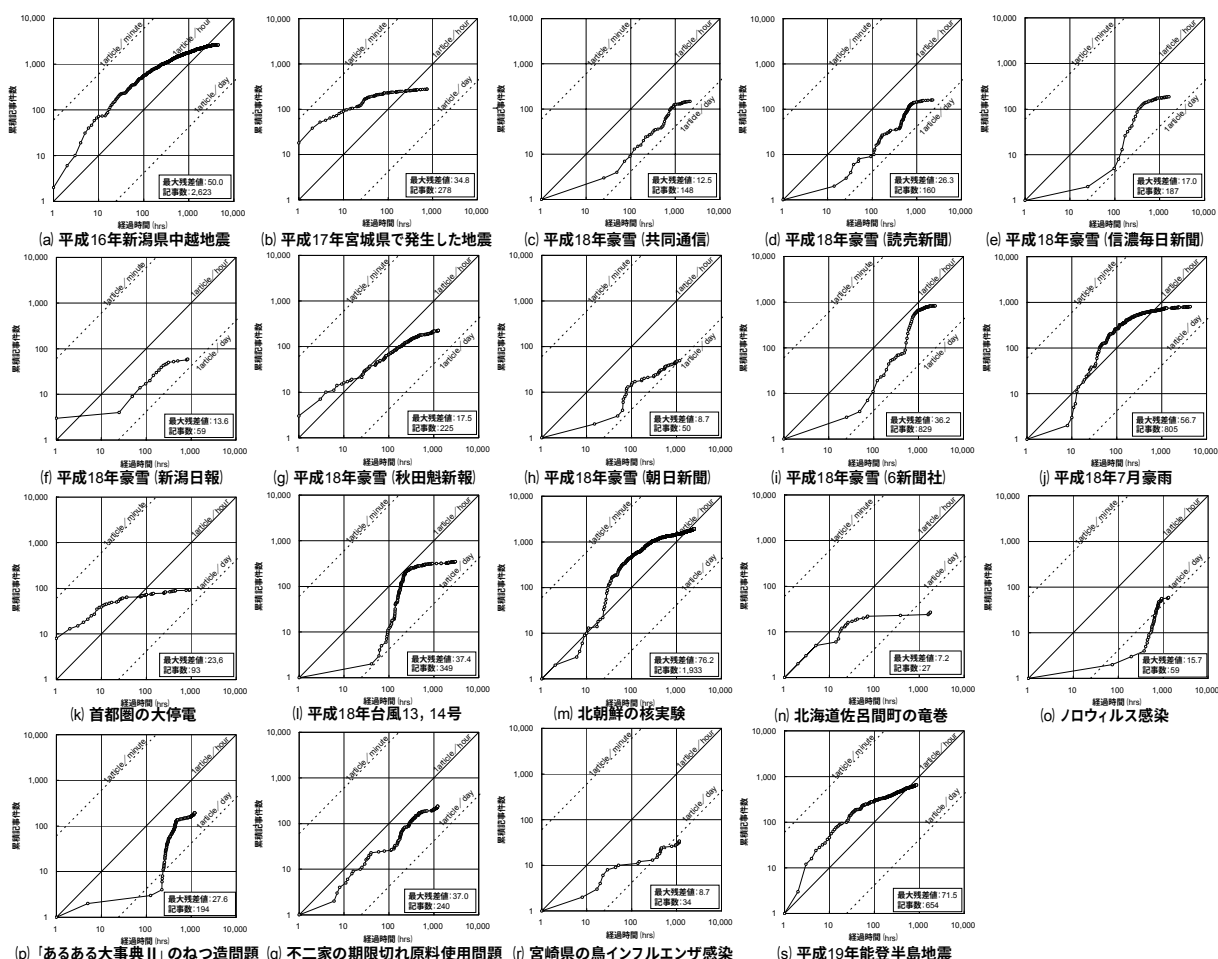


図4 各ウェブニュースコーパスの時間経過にともなう累積記事数の変化

れる。これ以降も「鹿児島」は残差値が低下しながら（特異性が低下しながら）出現しつづけ、それと同時的に「さつま」も出現しつづけている。7/27 から「ボランティア」、8/5 から「支援」が出現しはじめ、それぞれ8/4, 8/9 以降に「鹿児島」よりも高い残差値を示した。キーワード自動検出結果から、7月豪雨全体の被災地として、鹿児島県さつま町が顕在化し、7月末から8月始めにかけてフェーズが移行する様子を読みとれた。

同年に北朝鮮が実施した核実験とその後の対応を取り上げたウェブニュースからの解析結果では、「禁止」や「検査」の残差値が30～40に推移している。これらは、「北朝鮮からの船舶の入港や輸入の全面禁止など政府の追加制裁措置決定（毎日新聞、10/13 9:35）」、「北朝鮮に出入りする貨物の積み荷検査の強化にも着手した（毎日新聞、10/18 10:03）」などの制裁措置について報じたものである。10/20 から「唐」が大きな残差値を示し、10/25 に最も大きい 76.2 を示した。これは、中国の外交担当である唐國務委員が北朝鮮を訪問し、金正日総書記と会談したことと、その後に対応したことが集中的に取り上げられた結果を反映していると考えられる（「唐氏は金総書記との会談で、核実験再実施の中止と6カ国協議への復帰を要請（時事通信、10/20 17:00）」、「潘基文外交通商相が27日から2日間の日程で北京を訪問し、胡錦濤国家主席、唐家セン國務委員、李肇星外相らと会談すると発表した（毎日新聞、10/24 18:56）」）。

7月豪雨や北朝鮮の核実験を取り上げたウェブニュー

スコーパスからも、対応フェーズの変化や、ある時間断面で特異的な位置に関する情報、人物に関する情報を検出することができた。何が社会の関心事なのか、どこがイベントの主な現場なのか、誰が深く関わっているのかを把握することができた。以上の結果から、提案手法は地震災害を報道したウェブニュースコーパスからのみキーワードを的確に抽出するという命題は偽となり、キーワード自動抽出手法に適用できるウェブニュースコーパスはハザードを問わないことが確認された。

(2)多様な大きさもつコーパスへの適用の妥当性の検証

表1に示したように、様々なハザードを原因とする災害・危機に関するウェブニュースを収集したほか、様々な大きさ（記事の総数や収集期間）のコーパスを構築した。これらを用いて、キーワード自動抽出手法がもつコーパスの大きさに関する影響について考察する。

各イベントを最初に取り上げたウェブニュース記事が発信された時間を1時間目として、それから経過した時間と記事の累積件数の関係を図4に示した。図4では、災害・危機が発生した時系列の順に左から右、次に下段の左から右の順で示し、残差値の最大値と記事の総数も併記した。また、記事の増加傾向を比較する目安として、1時間に1件配信されつづけた場合の記事の推移を表す実践を図示した（経過時間 X, 累積記事数 Y, $Y=X$ の直線）。このほか、1分に1件（1時間に60件）、1日に1件（24時間で1件）の記事が配信されつづけた場合の破

線で示している。これら 3 つの直線をそれぞれ、1 件/時間直線、1 件/分直線、1 件/日直線と呼ぶことにする。

構築したコーパスのほとんどは、1 件/日直線から 1 件/分直線の間に推移している。そのうち、宮城県沖の地震を除いて、最初の記事の配信から 100 時間以降もなお、1 件/時間直線を超えている新潟県中越地震、7 月豪雨、北朝鮮の核実験、能登半島地震の最大残差値は 50 以上を示している。一方で、残差値の最大が 10 を下回った平成 18 年豪雪（朝日新聞）、北海道佐呂間町の竜巻、宮崎県の鳥インフルエンザ感染では、収集の最終時点で記事の総数が 1 日/1 件直線上、もしくはそれを下回っていた。残差値の最大値が低い例として、北海道佐呂間町の竜巻のウェブニュースコーパスからのキーワード抽出結果を示す（図 5）。災害の発生から一貫して高い残差値を示さず、「被災」「分かる」「よる」のように、竜巻やそれによって起きた被害や対応との関連性が低い単語が出現していた。

コーパスの大きさが残差値の増減に与える影響を調べるために、平成 18 年豪雪を報じた 6 新聞社それぞれを情報源とする 6 つのコーパスと、これらを統合したコーパスから、キーワードの自動抽出をおこなった結果をもとに、7 つのコーパスの最大残差値を比較した（図 6）。ウェブニュースの場合、情報源があるひとつの事象を複数回取り上げることは希であるから、関心の高い事象が発生することを前提とし、それに対して複数のメディアが一斉に報道することが、単位時間当たり（1 時間あた

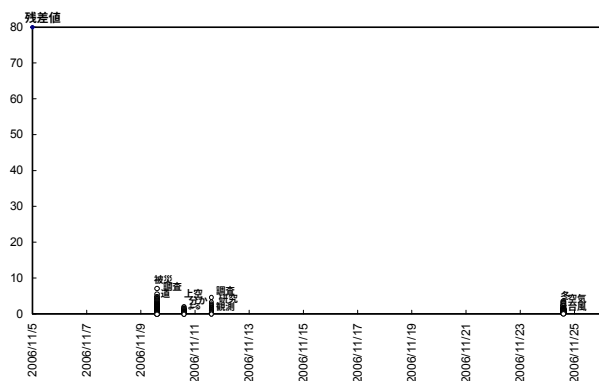


図 5 北海道佐呂間町の竜巻のウェブニュースコーパスからのキーワード自動抽出結果

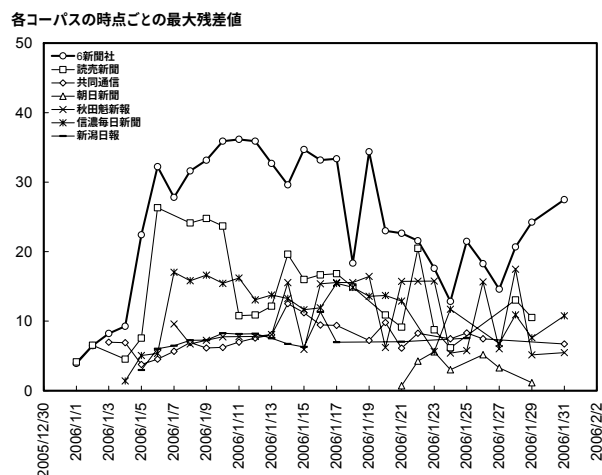


図 6 平成 18 年豪雪に関するウェブニュースコーパスから自動抽出したキーワードの最大残差値（2006 年 1 月分）

り、1 日当たり）に配信される記事が増加させる条件になる。6 つの新聞社を統合したコーパスは、このような状況を仮想的に表現したものである。統合されたコーパスの解析から得られた最大残差値は、1/1 と 1/2 を除いて、6 つの新聞社個々の解析から得られた最大残差値を上回っている。このような結果は、キーワード自動抽出手法は、複数のメディアが注目するような災害・危機に対して、キーワードに高い重みを付けることを示している。

大きな残差値を示すのに必要なウェブニュースコーパスの大きさについての分析を進めるために、コーパスを構成した記事の総数と残差値の最大値の関係の考察を試みた（図 7）。記事の総数が 300 件以内のコーパスでは、記事の総数が増加するとともに、残差値の最大値も徐々に大きくなるような傾向にある。全体的に記事の総数と最大残差値との相関が見られるものの、記事の総数が 500 件を超えると、残差値がそのまま線形的に増加するようなコーパス、残差値が増加するものの増加傾向が緩やかになるコーパスが存在しており、300 件以内での直線的な傾向を逸脱する関係が見られた。記事の総数が 300～500 件の間には、残差値の最大値の変化に影響するような変曲点が存在すると考えられる。これは、著しく高い残差値を示す条件として、キーワード抽出手法を適用するさいに、500 件を超えるウェブニュースが必要になることを示している。

以上の結果から、提案されたキーワード自動抽出手法は、1 日に 1 件以上配信されつづけ、総記事数が 500 件を上回るような社会の関心や注目度の高いイベントに対して、イベントのウェブニュース記事からなるコーパスに対して、特異性の高い単語に大きな重みを与えることが明らかになった。

ここまでの解析・適用を通して、以下のような課題が明らかになっている。先行研究では、提案手法によって得られた残差値の高いものから上位 3 つの単語をキーワードとして抽出してきた。しかし、図 2 の 7 月豪雨や図 3 の北朝鮮の核実験のウェブニュースコーパスからのキーワード抽出結果をみれば、上位 4 つ以下のプロットでも高い残差値を示し、キーワードとして採用すべき意味

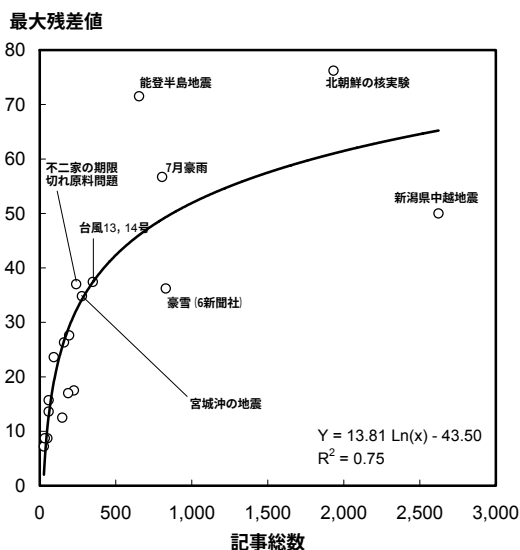


図 7 各ウェブニュースコーパスの記事総数と残差値の最大値の関係

のある単語である可能性があり、こうした単語をキーワードとして抽出できないというオMISSIONエラー（Omission Error）が発生していることが危惧される。また、図 5 の北海道佐呂間町の竜巻のキーワード抽出結果を見れば、単純に上位 3 つ以上をキーワードとするという判定基準を採用すれば、「分かる」「する」などをキーワードとして抽出してしまうことから、本来キーワードとして採用すべきでない単語を抽出してしまっているというコミッションエラー（Comission Error）が存在することは明らかである。キーワード自動抽出手法を複数のコーパスに適用したことによって明らかになった現状を受けて、次章にて上記の問題を解決する検討を行う。

5. キーワードの判定基準の妥当性に関する検証

提案手法で得られる残差値の上位 3 つを単純にキーワードとして採用することの妥当性を検証しながら、オMISSIONエラーとコミッションエラーを除去する方法について検討し、先行研究で開発したキーワードの候補への重み付け結果から、キーワードとなる単語を自動的に判定するアルゴリズムの開発を目指す。

(1) 評価の基準となる手動抽出キーワード群

提案手法による解析結果では、本来キーワードとして抽出すべき単語がどれだけ抽出できていないかということ（オMISSIONエラーとなる単語の数）と、キーワードとして適さない単語をどれだけ抽出しているかということ（コミッションエラーとなる単語の数）を評価する基準として、人間がキーワードにすべきと判断したキーワードのデータセットの作成を試みた。

著者のうち 1 人が判定者 A として、各コーパスのキーワード自動抽出結果にもとづき、キーワードとして適する単語を手動で抽出した。キーワードを手動で抽出する手続きは以下の通りである：1) 1 つの記事を 1 レコードとしたウェブニュースコーパスを 24 時間ごとに区分し、それぞれを 1 つのレコードとするウェブニュースコーパスを再編したデータベースを作成する（24 時間分のニュース記事を 1 ケースとする）。2) 分析時間間隔 Δt を 24 時間一定とし、キーワード自動抽出手法によって、キーワード候補の重みづけをおこなう。3) 分析単位ごとに、残差値が高い上位 15 語のリストを作成する。4) 「府」「する」など、明らかにキーワードに適さないようなものを除く。5) 上位から単語を見ていき、ウェブニュースコーパスの文脈の中で、意味があると予想されたものについては、ウェブニュースコーパス中の元の記事を参照し、単語の重要性や代表性を判断して、キーワードにすべきか否かを判定する。

以上のような手続きで抽出したキーワードは、全部で 1,525 語となった（延べ数）。キーワードが 1 語以上抽出されたケースは、全部で 453 ケースだった。判定者 A によって抽出された語の信頼性を評価するために、著者ではない判定者 B に同様の手続きをおこなわせ、新たなキーワードのデータセットを作成した。判定者 B には、全 453 ケースに対して等間隔サンプリングをおこなった 45 ケースに（抽出率 10%）、上記の 1) ～5) の手続きにもとづいて、キーワードの手動抽出をおこなわせた。評価対象の単語のうち、判定者 A と判定者 B が判定した結果は、一致率で 81.8%、評価者間の信頼性を評価する

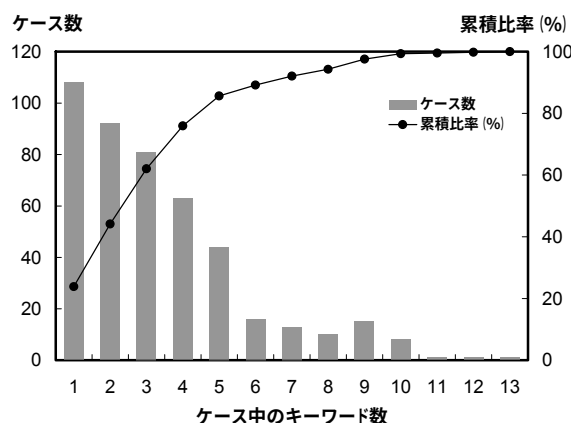


図 8 キーワードが手動抽出されたケースのキーワード数別の出現分布

Cohen の κ 係数は 0.60 であった。ここで採用したキーワードを手動で抽出する手法の適切性についての議論は残るものの、判定者の間である程度の一致性が得られたことから、以降の評価に判定者 A が抽出した手動抽出キーワード群（1,525 語）をもとに分析を進めていく。なお、一般に κ 係数は 0.4 以上の値であれば、信頼にたるとされている。

(2) オMISSIONエラーの除去

残差値の高い上位 3 つをキーワードとして採用することによって発生するオMISSIONエラーの評価を試みる。

図 8 には、各ケースに含まれる手動抽出キーワードの数ごとのケースの出現頻度分布を表した。

上位 3 つをキーワードとして採用した場合、手動でキーワードが抽出できたケースのうち 62.0% が自動的に抽出されることから、約 4 割のオMISSIONエラーが発生していることが分かる。各ケースからキーワードとして抽出する単語の上限数を 5 つまでとすれば、全体のうちの 85.7% のケースが手動でキーワードを抽出した結果と一致する。上限数を 7 つまでにすれば、全ケースのうち 92.1% まで増加する。

キーワードの上限数を 5 つから 9 つに増やしても、手動でキーワードを抽出した結果との一致率は約 6% しか増加しないため、ここでは Principle of Parsimony⁽⁴⁾ に従って、キーワードの上限数を最大 5 つとし、オMISSIONエラーを低減することにする。

(3) コミッションエラーの除去

ここでは、上位の最大 5 つをキーワードとして抽出したさいに生じる、キーワードとして適さない単語を自動的に抽出してしまうというコミッションエラーの除去について検討する。コミッションエラーを除去する手法としては、以下の 2 つが考えられる：1) 残差値に下限値（閾値）を設ける、2) あるケースの中で、残差値のプロットの間を評価する。前者は、1 つのケースの中で残差値が上位 5 つのうち、ある一定の残差値を示さないものはキーワードとして採用しないという方式である。北海道佐呂間町の例をみれば、「分かる」「被災」のようなキーワードとして適さない語は、残差値そのものの値が低い傾向にある。後者は、1 ケースの中で、残差値が著しく高い値を示したものをキーワードとして採用する方式である。これの詳細については、b) の冒頭で述べる。

a) 残差値の下限値にもとづく方法

残差値の上位 5 つのキーワードとして採用するさいに、元来キーワードとして適さない単語が選定される量を調べるために、上位 5 つの単語の残差値カテゴリ別の出現頻度分布を作成した（図 9）。図 9 は、上位 1 番目の単語のみ、2 番目の単語のみ、3 番目の単語のみ、4 番目の単語のみ、5 番目の単語のみの出現頻度分布を示している。そのうち、手動抽出キーワードに該当する頻度を黒棒で併記してある（積み上げグラフ）。上位 1 番目の単語は、残差値が最も大きいものから、20-25 のカテゴリまでのすべてが手動抽出キーワードになっていた。20-25 のカテゴリのみで、上位 2 番目では全単語のうち 92.0% が、上位 3 番目では 81.7% が手動抽出キーワードになっていた。4 番目では 75.4%、5 番目では 55.1% が残差値 22-25 のカテゴリを手動抽出キーワードが占める。残差値 15-20 のカテゴリでは、どの分布でも手動抽出キーワードが半分を占めることはなかった（11.3~31.0%）。これは、残差値 20 未満をキーワードとして採用する残差値の下限値とした場合、多くのキーワードとして適さない単語が多く採用されてしまうことを示している。手動抽出キーワードについて、残差値が大きいものから累積していったさい、残差値が 18.8 で全体の 95% を超えた。これらの結果を受け、残差値の下限値を 20 としたキーワード判定を試行することとする。残差値に下限値を定めた場合のコミッションエラーの定量的な評価は c) 項に譲る。

b) 1 ケース内の残差値の差にもとづく方法

解析結果の各ケースでは、残差値が 1 番大きいもの（1 番目の単語の残差値）と 2 番目に大きいもの（2 番目の単語の残差値）との残差値の差は、2 番目の残差値と 3 番目の残差値の差より大きく、これはさらに、3 番目と 4 番目の単語の残差値の差より大きいという傾向が見られる。なかには、このような傾向にしたがわないような、となりあう単語と単語どうしの残差値が著しく離れてるような解析結果も見られた。ケースの中で、著しく残差値が高いような単語のみをキーワードとして判定すれば、キーワードとして適さない単語を除去できることを期待した。

以上を評価する方法案について述べる：1) n 番目の単語の残差値と $n+1$ 番目の単語の残差値の差 (D_{Un})、 $n+1$ 番目の単語の残差値と $n+2$ 番目の単語の残差値との差 (D_{Ln}) の大きさを比較する。2) n を 1 ずつ増やしていき、 $n=5$ までの計算をおこなう。3) $D_{Un} < D_{Ln}$ になる次点があれば、最も大きい D_{Un} の箇所における $n+1$ 番目のキーワードの残差値を、キーワードとして採用する閾値とする。このような方法でキーワードを判定する方法をおこなった場合のコミッションエラーの定量的な評価は c) 項に譲る。

c) 各方法によるコミッションエラーの評価

図 10 に、a) と b) に記述した方法によって、キーワードを判別し、手動抽出キーワードとの一致の程度を分析した。図 10 には、本章 (2) 節で提示したオMISSION エラーを除去する方法としてのキーワードの上限数を 5 つとした場合の結果を示した（図 10 の①）。この結果を基準とし、a) の残差値の絶対値が 20 以上であるもの（図 10 の②）、b) に示した残差値の差を評価する方法にもとづいて選定されたもの（図 10 の③）の分布を図示した。図 10 の中央やや右よりの縦線を境界とし、右側に a) b) の方法で抽出した単語のうち、手動抽出キーワードと一致した単語数、左側に a) b) の方法で抽出した単語のうち手動抽出キーワードと一致しない単語数（コミッションエラー）を棒グラフで示している。したがって、

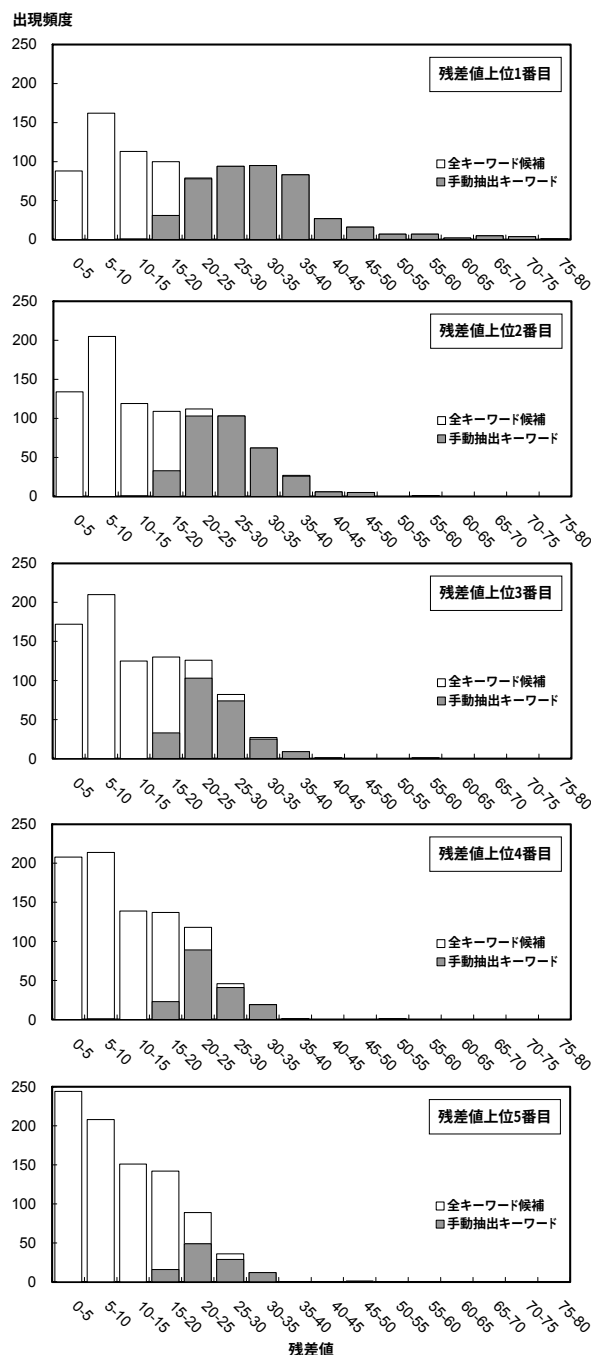


図 9 残差値上位 5 つの出現頻度分布

右側の棒グラフが長く、左側の棒グラフが短い手法の組み合わせが、抽出の精度が高いことを示す。

図 10 の①で抽出することができた 1,319 語のうち、残差値に下限値を設ける方法（図 10 の②）では 1,180 語が抽出された（89.5%）。手動抽出キーワードと一致する語は、若干減少するものの、①で生じるコミッションエラーの 3,095 語を②では 125 語まで低減できる。一方、残差値の差を評価する方法（③）では、手動抽出キーワードと一致するものが 1,054 語（79.9%）と①のうち約 8 割を抽出することができるが、コミッションエラーが 1,847 語という結果になった。図 10 には、②と③の方法を組み合わせ手動抽出キーワードとの一致をみたものも示した（図 10 の④）。④は、コミッションエラーを最も低減することができる（52 語）、手動抽出キーワー

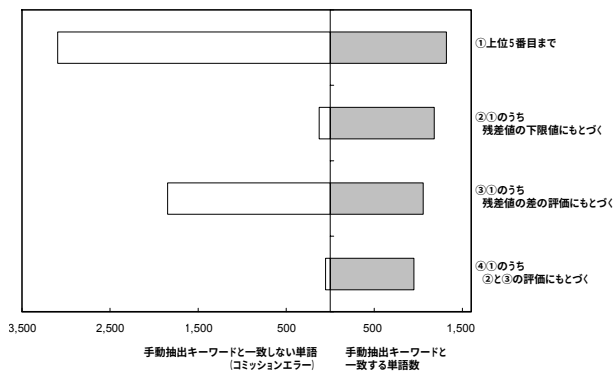


図 10 各方法で得られる単語の手動抽出
キーワードとの一致とコミッションエラー

ドとの一致が最も少ない（949 語，71.9%）。

③の残差値の差を使ってキーワードを判定しようとする方法は，結果としてコミッションエラーを十分に除去できるものではなかった，②と③を組み合わせた方法④では，②に比べてコミッションエラーとなる単語を半分以下に低減することができるが，手動抽出キーワードとの一致率が 8 割を下回ってしまうことから，②をコミッションエラーを除去する方法として採用する。

先行研究で提案した TFIDF を用いたキーワード自動抽出手法によって得られた残差値を求めたあと，1) 各ケースから残差値が大きい 5 つの単語をキーワードの候補とし，2) 残差値が 20 より大きいものをキーワードとする。ここに示したキーワードを抽出・判定する手法は，数値指標と数値基準を用いており，一連のプロセスを自動的に処理することができ，今回の参考となった判定者が抽出したキーワードのうち，77.4%を抽出することができる。

6. キーワード自動判定手法の適用

ここでは，自動抽出されたキーワードの候補から，前章で開発した自動的にキーワードを判定する手法の適用を試みた。本稿で提案するキーワードの上限数と残差値の下限値を定める方法をキーワード自動判定手法と呼び，先行研究で提案した自動抽出手法と呼び方を区別する。

図 11～13 に新潟県中越地震，宮城県沖で発生した地震，能登半島地震のウェブニュースコーパスについて，地震の発生から 1 ヶ月間のキーワード自動判定結果を示した。図中には，各ケースで残差値が上位から 5 つの単語を示し，キーワード判定アルゴリズムによって，キーワードとして判定された単語には下線を付している。新潟県中越地震の例では，「IC（インターチェンジ）」「優太（妙見町の県道からがけ崩れ巻き込まれた少年）」，「ダム（記事の例：山古志村の芋川に大量の土砂が流れ込んでつくられた天然の「ダム湖(天然ダム)」が，1 日夜から 2 日にかけての降雨で満水に近い状態になった(11/2 12:53 毎日新聞)）」，「授業（授業の再開）」，「入居（仮設住居への入居開始）」などがキーワードとして判定されている。宮城県沖の地震の例では「耐震」「構造」「振る」「止め」「つり」，能登半島地震の例では「現」「新」「支持」「選挙」「定数」がキーワードとして多く判定された。宮城県沖の地震では，仙台市泉区の運動施設「スポパーク松森」などの施設で，天井板の

落下事故が 4 施設で発生した。この事故は，国が平成 13 年（2001 年）芸予地震をうけて，非構造物である天井板の落下の防止のために「振れ止め」と呼ばれる耐震用の補強材設置の通達をおこなっていたにもかかわらず，これが設置されていないことに起因した事故であった。この落下事故の原因について多くの報道がなされ，「耐震（耐震用，耐震確認など）」「構造（天井がねじれやすい構造，非構造物など）」「振る（振れ止め）」「止め（振れ止め）」などが出現したと考えられる。能登半島地震では，「現（選挙立候補者が現職であることを示す，略字）」は，他に抽出された「新」「選挙」「定数」などから明らかなように，同時期に行われていた県議会選

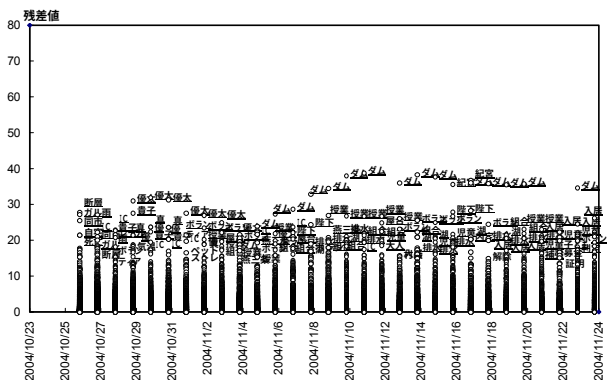


図 11 新潟県中越地震に関するウェブニュース
コーパスからのキーワード自動抽出・判定
結果（発災から 1 ヶ月）

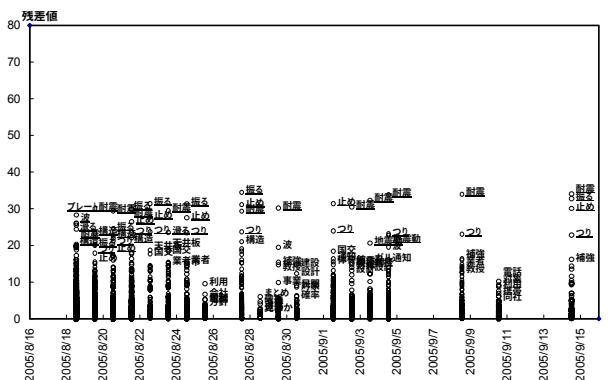


図 12 宮城県沖の地震に関するウェブニュース
コーパスからのキーワード自動抽出・判定
結果（発災から 1 ヶ月）

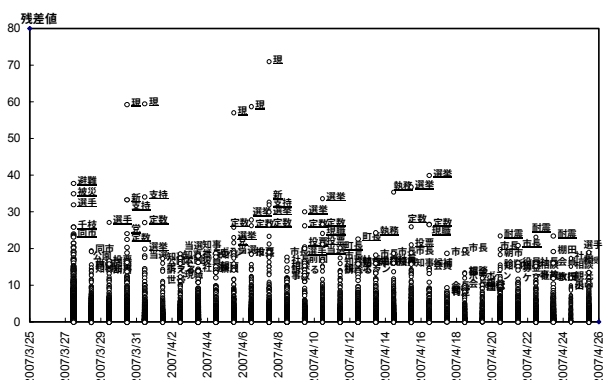


図 13 能登半島地震に関するウェブニュース
コーパスからのキーワード自動抽出・判定
結果（発災から 1 ヶ月）

挙に関するニュースから抽出された単語がキーワードとして判定されている。新潟県中越地震、宮城県沖で発生した地震、能登半島地震の発災から 1 ヶ月の間でキーワードとして判定された単語は、延べ数でそれぞれ 123 語、47 語、51 語であり、異なり数でそれぞれ 28 語、9 語、18 語であった。

災害や危機の中で、様々な事象が発生し、かつ当該イベントの中での固有性が高い事象であったり、過去の事例と比較したときに異例な事象であったりした場合には、社会の関心が高まり、それを取り上げるメディアやニュースの記事が増加することが想像される。4 章 (2) 節の結果からも分かるように、記事の量はキーワード抽出手法によって求められる残差値に影響することから、以上のような場合に多くの単語がキーワードとして判定されることが予想された。地震の発生から 1 ヶ月後の記事数は、新潟県中越地震で 1,616 件、宮城県沖の地震で 275 件、能登半島地震で 609 件となっていた。新潟県中越地震のキーワード判定結果からは、複数の事象を表すキーワードが、それぞれに時間のピークをもち、地震の発生から 1 ヶ月で同時並行的に展開していた様子を読みとれ、宮城県沖の地震では耐震管理の問題、能登半島地震では統一地方選挙の話題に 1 ヶ月間終始していたという結果は、報道すべき事象の存在と社会の関心の度合いを表すことが予想されるニュース記事の数と整合する結果である。このような結果は、キーワード自動判定手法において定義した、選定すべきキーワードの上限数と残差値の下限値に一応の妥当性があることを示している。

先行研究の段階では、本手法で求められた残差値に閾値を設けていなかったために、低い残差値であっても、当該時点で上位になるような単語は、キーワードとして選定されてしまっていた。図 12 の宮城県沖の地震に関するウェブニュースコーパスについての解析結果を例にとれば、8/25 の「利用」や 8/28 の「まとめる」など、当該イベントの関連の低い単語をキーワードとして抽出していたことになる。また、残差値にもとづく単語の上限数を 3 つにしていた先行研究では、上位 4 以降の単語については高い残差値を示し、かつ意味上重要な単語であっても選定されないような状況が発生していた。図 11 の新潟県中越地震のウェブニュースコーパスの解析結果を見ると、高い残差値を示す単語が上位の方で密集しているが、これらの 5 分の 2 がキーワードとして選定されない結果となっていた。ここで開発しているキーワード自動抽出手法は、当該のイベントに対して、時点ごとに特徴的な事象に関連のある言葉を抽出し、膨大な量の災害・危機に関する文書がもつ情報量を縮約して、災害・イベントの時系列的な概略を提示することを目的としている。キーワード自動抽出手法に、残差値の閾値と単語の上限数を設けることによって、災害・危機に関する言語資料を用いた災害の概況の把握を、よりの確におこなえるようになった。

図 14 に不二家の期限切れ原料使用問題について取り上げたウェブニュースからキーワードを自動判定した結果を示す。1/25、26、2/1、7 で「森永（森永製菓）」が、2/7、10 で「山崎（山崎製パン）」がキーワードとして判定された。これら 2 社は、不二家の再建支援を計画、実行した同業社である。森永製菓では不二家の再建支援を検討され、両者のトップが会談したことがニュース記事となった。その後「不二家の桜井康文社長は 24 日の記者会見で、「自主再建を目指す」と述べた（読売新聞、1/27 10:31）」との報道があり、不二家が自主再建する意

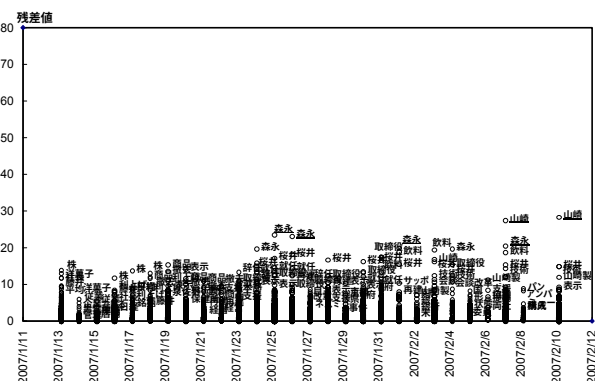


図 14 不二家の期限切れ原料使用問題に関するウェブニュースコーパスからのキーワード自動抽出・判定結果

向が示されたが、不二家離れが進行したことを受けて、山崎製パンが衛生管理技術の支援をおこなうことが報道された（「不二家の桜井康文社長と山崎の飯島延浩社長は同日午後、山崎が採用している衛生管理手法の導入に向けた覚書に調印した（産経新聞、2/6 8:00）」）。キーワード自動判定手法を不二家の不祥事とその対応経過のニュースに適用することによって、不二家を取り巻いた 2 社を表すキーワードを時系列に判定することができた。

7. おわりに

本研究では、先行研究で開発した TFIDF と残差分析手法を用いる災害・危機に関するコーパスからキーワードを自動的に抽出する手法の妥当性を検証を試みた。社会が捉えた災害・危機の像を捉える重要性や災害・危機の事例について基礎的な学習の必要性を踏まえたうえで、デジタルテキストでの収集と加工が容易に行えるウェブニュースに着目し、様々なイベント事例を取り上げたウェブニュースへの適用を通じて、キーワード自動抽出手法の外的妥当性の検証をおこなった：

- 1) 既存のキーワード自動抽出手法は、地震のみならず、その他の自然災害や人為災害について報じたウェブニュースコーパスからも、ニュース記事の時系列的な特徴の変化を表すキーワードを抽出できることが確認された。
- 2) コーパスがもつ大きさは、キーワードとしての特異性を評価する残差値の大きさと相関があり、1 日に 1 件以上ニュース記事が配信されつづけるようなイベントのうち、500 件程度のニュース記事を有するコーパスから残差値 50 を超えるようなキーワードを抽出することができる。

また、キーワード自動抽出手法から得られるキーワード候補群からキーワードとして単語を判定する基準の妥当性を検証し、その結果から得られた残差値の傾向をもとに、キーワードを自動的に判定するアルゴリズムを検討した：

- 3) 残差値の上位 3 つを採用する既存のキーワード判定基準において、大きなオMISSIONエラーが発生することを示し、これを低減する方法としてキーワードの最大数として上位 5 つを採用するという判断基準を提案した。
- 4) 既存のキーワード抽出に見られた、キーワードとし

て適さない単語をキーワードとして抽出するコミッションエラーの存在を指摘し、これを低減するために残差値に下限値を設ける方法を提案した。

- 5) 先行研究で導出したキーワード自動抽出手法の後続に 3) 4) に示した手続きを組み込むキーワード自動判定手法を提案した。この方法によって人間が手動で抽出したキーワードのうち約 8 割を自動的に抽出・判定することができる。

ここに提案した災害・危機に関するウェブニュースコーパスから、時間断面ごとに特異なキーワードを自動的に抽出・判定する手法を用いることによって、災害・危機対応の実務者や研究者に、今起きているイベントに対して状況認識の統一や過去の事例からの学習を支援することができることが期待される。

今回は、大量の文書データからの情報縮約に主眼を置いているために、ある時間の幅やある程度の数をもった文書群からのキーワードの抽出と判定をおこない、可視化することを目的とした手法の観点からの検証を行っている。過去の災害・危機の事例についての学習を想定すれば、コーパスの中からのニュース記事の電子的な検索の場面が想定される、このことを念頭におけば、可視化の素材としてのキーワードではなく、インデックスとしてのキーワードの妥当性を検証する必要がある。ニュース記事個々からのキーワードの自動抽出・判定の適用可能性の議論を今後の課題としたい。

ここで検証した手法は、当該の災害・危機を特徴づけるキーワードの時系列的な変遷を出力するものである。この結果は、当該イベントの概況を把握したり、コーパス中の情報資料を閲覧するさいの検索語の手がかりとして用いられることを想定している。前者の目的を達成しようとするれば、キーワードを検出し可視化された結果を用いて、当該イベントの現象の理解やコーパス中の本文の内容の推測がどこまで可能なかを考察する必要があることから、これを今後の検討課題として位置づける。

先行研究では、新潟県中越地震のウェブニュースから本手法で抽出されたキーワードの時系列的な変遷が 1995 年阪神・淡路大震災のエスノグラフィ調査によって得られた「命を守る活動」「社会のフローを安定させる活動」「社会のストックを再建させる活動」が同時並行的に展開するという災害過程モデルと整合することを確認している。その他の多くの災害・危機に対しても同様な分析を行うことにより、既往の研究で得られている以上のような被災者支援が進行していくモデルの一般性・再現性を確認することが可能になる。災害事象や災害対応課題に関連するキーワードを限定して抽出する手法へと精緻化し、災害過程モデルの妥当性の検証や被災者支援過程をモニタリングする手法の開発を進めたい。

謝辞

本研究は、文部科学省大都市大震災軽減化特別プロジェクトⅢ-3 第 5 課題「新公共経営 (New Public Management) の枠組みにもとづく地震災害対応シミュレーターによる災害対応力向上」(研究代表者：林春男 京都大学) および独立行政法人科

学技術振興機構 社会技術研究開発センター研究開発プログラム「ユビキタス社会のガバナンス」研究開発プロジェクト「危機に強い地域人材を育てる GIS 活用型の問題解決塾」(研究代表者：林春男 京都大学) によるものである。

補注

(1) 時間増加型 TFIDF

佐藤ら²⁾では修正 TFIDF として定義した指標で、これ以降は時間増加型 TFIDF と呼ぶことにする。この指標は TFIDF という単語の重み付け指標に、時間のパラメータを導入したものである。通常の TFIDF では、式[1]中の N_j のように変数のかたちはとらず、解析対象となるコーパスに含まれるすべての文書数 N (一定) を用いる。災害・危機に関する文書資料は時間の経過とともに増加することから、時間経過に耐えうる指標として再構成している。

(2) 単語

国語学上、単語の厳密な定義はない。ここでは便宜的に形態素の単位を単語と呼ぶ。

(3) 特異語

佐藤ら⁴⁾で定義した特徴語にあたり、これ以降は特異語と呼び改める。前述の残差値が正の値を示すキーワードで、各時間断面で特性の高い単語として位置づけている。これに対して、残差値が負の値を示すものを共通語と呼ぶ(旧：一般語)。本研究では、特異語のみを対象とした検証をおこなうため、キーワードという呼び方を特異語と同意で扱うことにする。

(4) Principle of Pasimony

儉約家の原理のことで、オッカムの剃刀とも呼ばれる。同程度の説明能力のある仮定が複数ある場合は、より単純なものを採用するという原理。

参考文献

- 1) 松村 劭：オペレーショナル・インテリジェンス 意思決定のための作戦情報理論，日本経済新聞社，220pp.，2006.
- 2) 佐藤翔輔，林春男，牧紀男，井ノ口宗成：TFIDF/TF 指標を用いた危機管理分野における言語資料体からのキーワード自動検出手法の開発—2004 年新潟県中越地震を取り上げたウェブニュースへの適用事例—，地域安全学会論文集，No. 8，pp. 367-376，2006.
- 3) Yahoo! JAPAN：Yahoo! ニュース，<http://headlines.yahoo.co.jp/hl>
- 4) 林春男：率先市民主義 防災ボランティア論 講義ノート，昇洋書房，p. 52，2001.
- 5) 松本裕治：形態素解析システム「茶釜」，情報処理，Vol.41，No. 11，pp. 1208-1214，2000.
- 6) 内閣府：7 月 4 日からの梅雨前線による大雨の被害状況について(第 9 報)，http://www.bousai.go.jp/kinkyu/060710_baiuzensen/baiuzensen09.pdf

(原稿受付 2007.5.28)

(登載決定 2007.9.15)