B2IM2031

修士論文

ウェブにおける誤情報の抽出と集約

鍋島 啓太

2014年2月10日

東北大学 大学院
情報科学研究科 システム情報科学専攻

本論文は東北大学 大学院情報科学研究科 システム情報科学専攻に
修士 (工学) 授与の要件として提出した修士論文である。

鍋島 啓太

審査委員：

乾 健太郎 教授　　（主指導教員）
徳山 豪 教授
伊藤 彰則 教授
岡崎 直観 准教授

ウェブにおける誤情報の抽出と集約*

鍋島 啓太

内容梗概

東日本大震災では，Twitter などのソーシャルメディアが情報源として活躍した一方,「コスモ石油の爆発により，有害な雨が降る」というツイートに代表される誤情報の拡散が問題となった．誤情報の中には人間の健康の安否に関わる情報も存在し，情報の信憑性の確保が急務となっている．本研究の目的は，誤情報の拡散が特に問題となっている Twitter を対象として，誤情報の網羅的な収集を行い，誤情報に対する注意喚起を低コストで実現する仕組みを実現することである．本稿では，誤情報を訂正する表現（以下，訂正パターン）に着目し，誤情報を認識する手法を提案する．具体的には，まず訂正パターンを人手で整備し，訂正パターンにマッチするツイートを抽出し．次に収集したツイートを内容の類似性に基づいてクラスタリングし，最後に，その中から誤情報を過不足なく説明する 1 文を選択する．評価実験では，人手で誤情報をまとめたウェブサイトを正解データとして評価を行い，誤情報の抽出性能の評価を行ったところ，既存のまとめサイトに収録されている 60 件の誤情報の約半数を再現でき，さらにまとめサイトに収録されていない 23 件の誤情報を獲得することができた．また，誤情報の拡散による問題は災害発生時だけではなく，通常時においても発生している．そこで，前述の提案手法が通常時においても有効であることを示すため，災害時以外のデータでも実験を行い，災害時と通常時の抽出結果の比較を行う．

キーワード

自然言語処理, 誤情報, 情報抽出, 訂正, テキストマイニング

*東北大学 大学院情報科学研究科 システム情報科学専攻 修士論文, B2IM2031, 2014 年 2 月 10 日.

Extracting and Aggregating False Information from the Web*

Keita Nabeshima

Abstract

During the 2011 East Japan Earthquake and Tsunami Disaster, we had found a number of false information spread on Twitter, e.g., *Harmful substance will come down with the rain after the Cosmo Oil fire.* This paper extracts pieces of false information exhaustively from all the tweets within one week after the earthquake, and analyzes the processes of diffusions of the false information and its correction information. Desining a set of linguistic patterns that correct false information, this paper proposes a method for detecting false information. More specifically, the method extracts text passages that match to the correction patterns, clusters the passages into topics of false information, and selects, for each topic, a passage explaining the false information the most suitably. In the experiment, we report the performance of the proposed method on the data set extracted manually from Web sites that are specialized in collecting false information.

Keywords:

Natural Language Processing, False Information, Information Extraction, Correction, Text Mining

*Master's Thesis, System Information Sciences, Graduate School of Information Sciences, Tohoku University, B2IM2031, February 10, 2014.

目 次

1 はじめに **1**
1.1 本研究の背景 1
1.2 本研究の目的 2
1.3 本論文の構成 2

2 関連研究 **3**
2.1 情報信憑性に関する研究 3
2.2 Twitter からの誤情報抽出に関する研究 4
2.3 矛盾認識に関する研究 5

3 提案手法 **7**
3.1 ステップ 1：訂正パターンを用いた訂正フレーズの抽出 7
3.2 ステップ 2：キーワードの抽出 10
3.3 ステップ 3：キーワードのクラスタリング 10
3.4 ステップ 4：代表フレーズの選択 11

4 予備実験：訂正パターンの評価 **13**
4.1 データセット 13
4.2 正解データ 13
4.3 評価尺度 14
4.4 結果と分析 14

5 本実験：誤情報の集約の評価 **17**
5.1 実験設定 17
5.2 評価尺度 17
5.3 実験結果 18
5.4 精度に関するエラー分析 20
5.5 再現率に関するエラー分析 23

6 一般ツイートからの誤情報抽出 **26**
6.1 実験設定 26
6.2 実験結果 26

7 Web テキストからの誤情報抽出 **28**
7.1 実験設定 28
7.2 実験結果 28

8 応用：誤情報監視システム **30**

9 おわりに **32**

謝辞 **33**

付録 **37**

A 正解データとして用いた誤情報一覧 **37**

図 目 次

1 Dispute Finder 3
2 誤情報抽出の流れ 8
3 被訂正フレーズを含むツイートの構造 8
4 被訂正フレーズの抽出 9
5 リアルタイム誤情報収集システム 31

表 目 次

1 使用した訂正パターン 9
2 訂正パターンの適合率と再現率 14
3 抽出された被訂正フレーズの内訳 15
4 抽出できなかった誤情報の内訳 15
5 誤情報の抽出結果 18
6 抽出された誤情報のうち，まとめサイトに掲載されていた事例 . . 19
7 抽出された誤情報のうち，まとめサイトに掲載されていなかった事例 19
8 精度に対する誤り分析 20
9 再現率に対する誤り分析 24
10 一般ツイートから抽出されたフレーズの種類 26
11 通常時のツイートから抽出された誤情報 27
12 Web テキストから抽出されたフレーズの種類 28
13 Web テキストから抽出された事例 29
14 正解データとして用いた誤情報一覧 37

1　はじめに

1.1　本研究の背景

2011 年 3 月に発生した東日本大震災では，ソーシャルメディアは有益な情報源として活躍した．野村総合研究所の調査 [1] によると，震災に関する情報源として，ソーシャルメディアを挙げたネットユーザーは 18.3%で，インターネットの新聞社（18.6%），インターネットの政府・自治体のサイト（23.1%）と同程度である．ニールセン社の調査 [2] によると，2011 年 3 月の mixi の利用者は前月比 124%，ツイッターは同 137%，Facebook 同 127%であり，利用者の大幅な伸びを示した．

東日本大震災後のツイッターの利用動向，交換された情報の内容，情報の伝搬・拡散状況などの分析・研究も進められている [3, 4, 5, 6]．Doan ら [4] は，大震災後のツイートの中で地震，津波，放射能，心配に関するキーワードが多くつぶやかれたと報告している．宮部ら [6] は，震災発生後の地域別のツイッターの利用動向，情報の伝搬・拡散状況を分析した．Sakaki ら [5] は，地震や計画停電などの緊急事態が発生したときの地域別のツイッターの利用状況を分析・報告している．Acar と Muraki は [3]，震災後にツイッターで交換された情報の内容を，警告，救助要請，状況の報告，自身の安否情報，周りの状況，心配の 6 つに分類している．

ソーシャルメディアが活躍した一方で，3 月 11 日の「コスモ石油のコンビナート火災に伴う有害物質の雨」に代表されるように，インターネットやソーシャルメディアがいわゆるデマ情報の流通を加速させたという指摘がある．東日本大震災とそれに関連する福島第一原子力発電所の事故では，多くの国民の生命が脅かされる事態となったため，人間の安全・危険に関する誤情報（例えば「放射性物質から甲状腺を守るにはイソジンを飲め」）が拡散した．ネット上のデマをまとめたツイート[1]では，2013 年 12 月時点でも月に二十数件のペースでデマ情報が掲載されている．このように，ツイッター上の情報の信憑性の確保は，災害発生時だけではなく，平時においても急務である．

我々は，誤情報（例えば「放射性物質から甲状腺を守るためにイソジンを飲め」）に対してその訂正情報（例えば「放射性物質から甲状腺を守るためにイソジンを飲めというのはデマ」）を提示することで，人間に対してある種のアラートを与え，情報の信憑性判断を支援できると考えている．

[1] https://twitter.com/#!/jishin_dema

1.2 本研究の目的

訂正情報に基づく信憑性判断支援に向けて，本稿では東日本大震災時に拡散した誤情報の網羅的な収集に取り組む．具体的には「○○というのはデマ」「○○という事実は無い」など，誤情報を訂正する表現（以下，訂正パターン）に着目し，誤情報を自動的に収集する手法を提案する．震災時に拡散した誤情報を人手でまとめたウェブサイトはいくつか存在するが，東日本大震災発生後の大量のツイートデータから誤情報を自動的，かつ網羅的に掘り起こすのは，今回が初めての試みである．評価実験では，まとめサイトから取り出した誤情報のリストを正解データと見なし，提案手法の精度や網羅性に関して議論する．なお，ツイートのデータとしては，東日本大震災ワークショップ[2]において Twitter Japan 株式会社から提供されていた震災後１週間の全ツイートデータ（179,286,297 ツイート）を用いる．

また，誤情報の拡散による問題は災害発生時だけではなく，平時においても急務である．そこで，前述の提案手法が平時においても有効であることを示すため，災害時以外のデータでも実験を行い，災害時と平時の抽出結果の比較を行う．

1.3 本論文の構成

本論文の構成は以下の通りである．まず，第２章では誤情報の検出に関する関連研究を概観し，本研究との差異を述べる．第３章では誤情報を網羅的に収集する手法を提案する．第４章では誤情報抽出に重要となる訂正パターンの評価を行う．第５章では提案手法の評価実験，結果，及びその考察を行う．第６章では提案手法を通常時のツイートに適応し，評価，考察を行う．第７章では本研究の応用として，誤情報をリアルタイムに抽出するシステムを紹介する．最後に，第８章で全体のまとめと今後の課題を述べる．

[2] https://sites.google.com/site/prj311/

2　関連研究

本研究の目的は，ツイート集合から誤情報を自動的かつ網羅的に抽出，集約を行い提示することで，誤情報に対する注意喚起を低コストで実現することである．誤情報を自動的に特定し集約を行う技術に関連する，情報信憑性，Twitter からの誤情報抽出，矛盾認識の 3 つに関連する研究をそれぞれ述べ，本研究において解くべき課題について説明する．

2.1　情報信憑性に関する研究

Web 上にある情報の信憑性を判断する研究は，これまでにいくつか研究されてきた．Fact-Finder はその中でも有名なアルゴリズムで、情報信憑性の判断に、文書に書かれている内容と、文書間のリンク関係の 2 つを用いた [7]．Pasternack ら [8] はさらに Fact-Finder の拡張を行い，関連知識や文脈情報を手がかりとして組み入れた．Lex ら [9] は OpenIE によって Web 上から得られた事実が，どれだけ文中に含まれているかを計測することにより，Web 文書の信憑性と重要性を評価した．

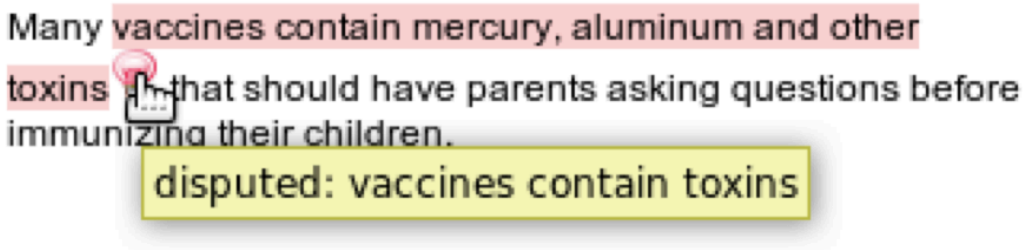

図 1: Dispute Finder

Ennals ら [10] は情報信憑性判断のために，Dispute Finder というシステムを作成した．Dispute Finder は，議論が存在する内容を含む Web ページを観覧しているユーザーに対し、既知の議論を提示するシステムである。図 1 にイメージ図を示す．議論を提示するために，Dispute Finder は論点のデータベースを構築している．このデータベースはユーザーにより作られ，Web 上で議論されている論点と，その論点に関連するサイトで，信頼のある情報源を参照できるリンクを記録してある． Dispute Finder の目的は我々の目指すところは近いが，Dispute

Finder のデータベースの構築は人に頼っており，低コストで誤情報の自動構築を目指している我々の研究の目的とは異なる．

2.2 Twitter からの誤情報抽出に関する研究

近年，ツイッターは自然言語処理の分野において研究対象として注目を浴びている．言語処理学会の年次大会では「Twitter と言語処理」というテーマセッションが 2011，2012 年に企画された．また，国際会議のセッションや併設ワークショップにおいても，ソーシャルメディアに特化した情報交換の場が設けられることが珍しくない．このような状況が映し出すように，ツイッターを対象とした研究は数多くあるが，本節ではツイートで発信される情報の真偽性や信憑性に関連する研究を紹介する．

Ratkiewicz ら [11] は，米国の選挙に関連して，アストロターフィング[3]や誹謗中傷，誤情報の意図的な流布を行っているツイートを検出するシステムを提案した．Castillo ら [12] は、Twitter 上で拡散したニュースの信憑性を分析した．彼らはニュースに関連したツイートを対象とし，そのツイートが信頼できるかどうかの二値分類器を構築した．信憑性を判断するのには，いくつかの要素があると仮定し，その仮定を元にツイートの内容、投稿者、ツイートのトピック、伝搬傾向の 4 つを分類器の素性とした。実験の結果，ツイートに URL が含まれているものやリツイートの伝搬木が深いニュースは、信憑性が高いと述べている。Qazvinian ら [13] は，誤情報に関連するツイート群（例えば「バラク・オバマ」と「ムスリム」を含むツイート群）から，誤情報に関して言及しているツイート（例えば「バラク・オバマはムスリムである」）と，誤情報に関して言及していないツイート（例えば「バラク・オバマがムスリムのリーダーと面会した」）を分類し，さらに誤情報に関して言及しているツイート群を，誤情報を支持するツイートと否定するツイートに分類する手法を提案した．Qazvinian らの研究は，誤情報に関連するツイート群（もしくはクエリ）が与えられることを想定しており，本研究のように大規模なツイートデータから誤情報をマイニングすることは，研究対象の範囲外である．

日本では，東日本大震災時にツイッター上で誤情報が拡散したという問題意識から，関連する研究が多く発表されている．白井ら [14] は，デマ情報とその訂正情報を「病気」とみなし，感染症疾患の伝染モデルを拡張することで，デマ情報・

[3]団体や組織が自発的な草の根運動に見せかけて行う意見主張のこと．一般市民を装って，特定の候補者を支持したり，否定する意見をツイートで発信し，複数のユーザアカウントを使って多勢を装ったり，一般市民のリツイートを誘発させるなどして，選挙活動を行う．

デマ訂正情報の拡散をモデル化した．藤川ら [15] は，ツイートに対して疑っているユーザがどの程度いるのか，根拠付きで流言であると反論されているか等，情報に対するユーザの反応を分類することで，情報の真偽判断を支援する手法を提案した．鳥海ら [16] は，あるツイートの内容がデマかどうかを判別するため，ツイートの内容語と「デマ」「嘘」「誤報」などの反論を表す語の共起度合いを調べる手法を提案した．大和田ら [17] は，情報信憑性や重要性を評価するために，ツイートの返信および非公式リツイートといった返信ツイートを認識する手法を提案した．具体的には，返信ツイートを「同意」「反論」「疑問」の 3 つの態度を推定する分類器を構築した．これにより，多くのツイートに「反論」や「疑問」を持たれているツイートの信憑性は怪しいと判断することができる．

梅島ら [18] は，東日本大震災時のツイッターにおけるデマと，デマ訂正の拡散の傾向を分析することを目標とし,「URL を含むリツイートはデマである可能性が低い」「デマは行動を促す内容，ネガティブな内容，不安を煽る内容が多い」「この 3 つのいずれかの特徴を持つツイートはリツイートされやすい」等の仮説を検証した．彼女らのグループはその後の研究 [19, 20] で，誤情報のデータベースを構築するために,「デマ」や「間違い」といった訂正を明示する表現を用いることで，訂正ツイートの認識に有用であることを示した．さらに彼女らは，訂正を明示する表現を含むツイートを収集し，各ツイートが特定の情報を訂正しているか，訂正していないのか[4]を識別する二値分類器を構築した．

これらの先行研究は，ツイートが誤情報を含むかどうか，もしくはツイートが特定の情報を訂正しているかどうかを認識することに注力しており，ツイート中で言及されている誤情報の箇所を同定することは研究対象の範囲外となっている．したがって，大規模なツイートデータから誤情報を網羅的に収集する研究は，我々の知る限り本研究が最初の試みである．

2.3 矛盾認識に関する研究

あるツイートの内容が別のツイートの内容と矛盾していれば，そのどちらかのツイート内容は間違った情報である可能性がある．そこで矛盾認識を行うことで，誤情報の同定を行うことが可能である．しかし，矛盾認識というタスクは，自然言語処理の中でも難しいタスクであると知られている [21]．

RTE-3 で行わた矛盾認識のタスクにおいて，De Marneffe ら [22] の研究では，適合率と再現率がそれぞれ 23%, 19%であったと報告している [23]．しかしなが

[4]例えば「ツイート上には様々なデマが流れているので注意を！」というツイートには「デマ」という表現を含んでいるが，特定の情報を訂正しているわけではない

ら，RTE-3 の矛盾関係のデータセットは人手によって作成されたものであり，実際の文中で起きている矛盾関係とは必ずしも一致しない．そこで彼らは矛盾関係を現実的なデータセットから収集し，実験を行った．収集した矛盾関係のデータセットで実験し評価を行ったところ，性能は非常に限定的であったと述べている．

De Marneffe ら [24] も矛盾関係の認識に取り組んでいる．彼らは矛盾関係の問題を， 反義語，否定， 数量，事実性，文構造，語彙，世界知識の 7 つのグループに分類しており，この分類をもとに素性を作成した．RTE-3 のデータセットに対しての実験結果に比べ，他のデータセットへ適応した場合性能の低下が見られ，矛盾認識を他のデータセットへ適応するのは困難であると指摘している．

Ritter ら [25] は関係の一意性が矛盾関係認識を解く上で有用であると指摘した．一意性がある関係とは，例えば，[$arg1$ の出身地は $arg2$] という関係について，$arg1$ にある語が代入されたときに，$arg2$ に代入できる語が唯一に決まる関係である．この例の場合，$arg1$ の値がモーツァルトの時，$arg2$ はザルツブルクただひとつに決まり，他の文に [モーツァルトの出身地はウィーン] と書かれていれば，この 2 つが矛盾していると分かる．逆に [$arg1$ と国境を接する $arg2$] などの関係は，$arg1$ が決まっても $arg2$ がただひとつに決まらないので，一意性はない．彼らは関係の一意性をスコア付けする手法を提案し，矛盾関係認識に用いた．

Watanabe ら [26] は 2 文間の各項のアライメントをとり，Natural Logic [27] で定義された意味関係を個別に付与する手法を提案している．このアライメント毎に付与された意味関係を用いて，文間関係を論理的に導くことができる．NTCIR-10 で開催された RITE-2 [28] のタスクの一つである矛盾関係認識において，彼らの手法は一番高いスコアをマークしたが，その際の性能は F 値で 28.57%であり，我々の目的を実現するにはまだ性能不足である．

矛盾認識では性能面だけではなく，計算量の側面から見ても困難である．これは全 N 件のツイートに対し，ツイート間の矛盾関係を求めるのにかかる計算量は，$O(N^2)$ となるためであり，ツイッター上のあらゆるツイート間の矛盾関係を求めることは困難である．さらに新しいツイートが投稿される度に，N 回の矛盾関係認識が必要となってしまう．それに対し，我々の手法はツイート単体で，誤情報かどうか判断するため，計算量は $O(N)$ で十分である．

3　提案手法

本研究では，ツイッター上で拡散している誤情報に対して，別の情報発信者がその情報を訂正すると仮定し，誤情報の抽出を行う．例えば,「コスモ石油の爆発により有害な雨が降る」という誤情報に対して，ツイッター上で以下のような訂正情報を含むツイート（以下，訂正ツイート）が発信された．

ex1 コスモ石油の爆発により、有害な雨が降るという事実はない。

ex2 コスモ石油の科学物質を含んだ雨が降るというデマが Twitter 以外にも出回ってるので注意を

訂正ツイートは，訂正表現（下線部）と，その訂正対象である誤情報から構成される．そこで，ツイート中の訂正表現を発見することで，誤情報を抽出できると期待できる．本節で提案する手法の目標は，訂正表現を手がかりとして，ツイート本文から誤情報を説明する箇所を推定する抽出器を構築することである．さらに，構築した抽出器によって，ツイート集合から誤情報を過不足なく収集したい．

図 2 に提案手法の流れを示す．手順は大きく 4 つに分けられる．まず，ツイート本文に訂正パターン（後述）を適用し，訂正対象となる部分（被訂正フレーズ）を抽出する（ステップ 1）．次に,「昨日のあれ」のように具体的な情報を含まないフレーズを取り除くために，ステップ 2 において被訂正フレーズに含まれやすいキーワードを選択する．同一の被訂正情報を言及しているが，表現や情報量の異なるフレーズをまとめるために，フレーズに含まれるキーワードをクラスタリングする（ステップ 3）．その結果,「コスモ石油」や「イソジン」といった，誤情報の代表的なキーワードを含むクラスタが構築される．図 2 左上の表は，被訂正フレーズに含まれやすいキーワードが上位に来るよう，クラスタをステップ 2 の条件付き確率 (式 1，後述) で並べ替えたものである．最後に，ステップ 4 で，各クラスタごとに誤情報を最もよく説明しているフレーズを選択する．図 2 右上はステップ 3 で並べ替えたクラスタからフレーズを抽出し，出力された誤情報のリストである．以降では，各ステップについて詳細に説明する．

3.1　ステップ 1：訂正パターンを用いた訂正フレーズの抽出

ステップ 1 では，ツイート本文から被訂正フレーズを見つけ出す．被訂正フレーズは,「デマ」や「間違い」といった表現で，訂正や打ち消されている箇所のこと

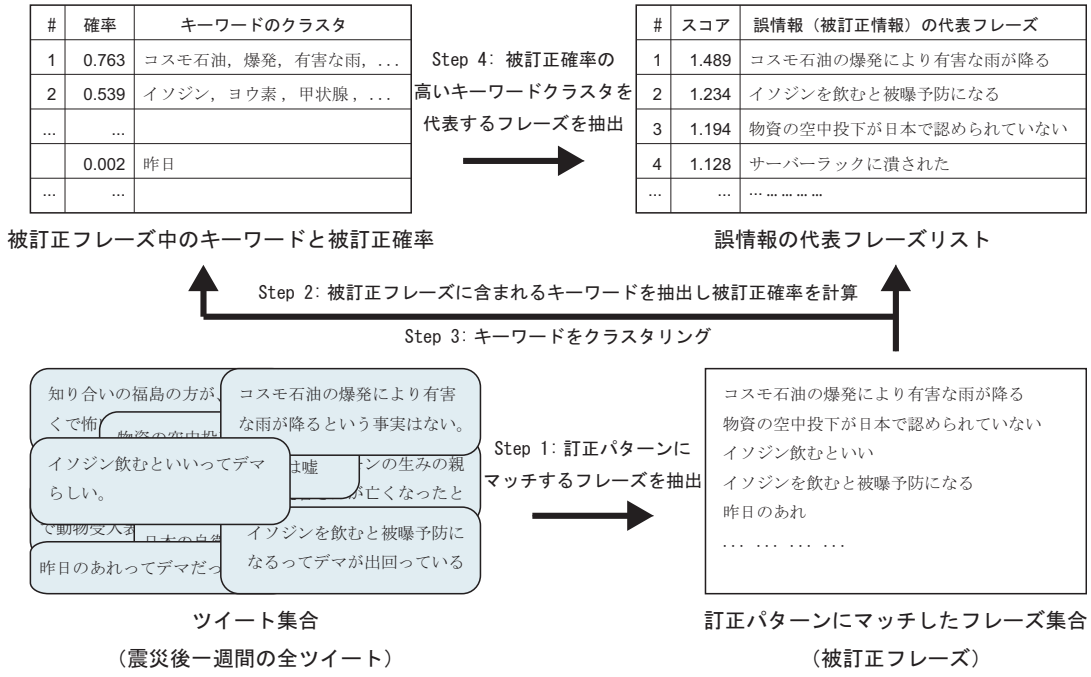

図 2: 誤情報抽出の流れ

である．被訂正フレーズは,「イソジンは被曝を防ぐ」といった単文や,「コスモ石油の火災により有害な雨が降る」といった複文,「うがい薬の件」といった名詞句もある．被訂正フレーズと訂正表現は,「という」や「のような」といった連体助詞型機能表現で繋がれ，図 3 に示す構造をとる．

イソジンは被曝を防ぐ（被訂正フレーズ） ＋ という（連体助詞型機能表現） ＋ デマ（訂正表現） ＋ が流れています

コスモ石油の火災により有害な雨が降る（被訂正フレーズ） ＋ といった（連体助詞型機能表現） ＋ 事実はありません（訂正表現）

図 3: 被訂正フレーズを含むツイートの構造

被訂正フレーズに続く表現を，すなわち連体助詞型機能表現と訂正表現の組み合わせを,「訂正パターン」と呼ぶ．例えば，図 3 において,「というデマ」,「といった事実はありません」が訂正パターンである．

表 1: 使用した訂正パターン

連体助詞型機能表現	など，なんて，とか，とかいう, との，って，といった，という, というのは，の様な，のような
訂正表現	デマ, 嘘, ガセ, 不確定, ソース (が | は | の)(ない | ありません), チェーンメール, チェンメ, 事実は (ありません | ない), 今のところない, 否定, 必要はない, 事実では (ありません | ない)，意味が (ない | 無い), 虚偽, 誤 (り |った | る | 報 | 情報 | 解), 関知しない, 意味が (ない | 無い), 未確認, 訂正, 虚報, 流言, 風説, 出 (どころ | 任せ | 処), でまかせ, 真贋, 真偽, 根拠 (の | が) ない, 効果がない, そんなことはない, ということは (ない | ありません), まずない, 訳ではない,

全ツイートを形態素解析し，訂正パターンに対して形態素レベルでのパターン照合を行う．マッチしたツイートに対して，文頭から訂正パターンの直前までを被訂正フレーズとして抽出する．被訂正フレーズを漏れなく抽出するには，質のよい訂正パターンを整備することが重要である．そこで，どのような表現が訂正パターンになり得るのかを調べた．具体的には，既知の誤情報１５件を含むツイートを検索するようなクエリを考え，そのツイートの内容を確認することにより，訂正パターンを収集・整理した．このようにして得られた訂正パターンの一覧を表 1 に示した．表 1 の訂正パターンのいずれかを含むツイートに対して，文頭から訂正パターンの直前までを被訂正フレーズとして抽出した例を図 4 に示した．図 4 の下線部が訂正パターンである．

イソジンを飲むと被曝予防になるってデマが出回っている
⇒ イソジンを飲むと被曝予防になる
コスモ石油の爆発により有害な雨が降るという事実はない
⇒ コスモ石油の爆発により有害な雨が降る

図 4: 被訂正フレーズの抽出

3.2 ステップ 2：キーワードの抽出

前節で抽出された被訂正フレーズには,「昨日のあれ」のように具体的な情報が提示されていないフレーズも含まれている．これらは誤情報としては不適切であるため，取り除く必要がある．そこで，被訂正フレーズ中の名詞句が訂正情報中に偏って出現しているかどうかを調べる．ここで分析の対象とする名詞句は，単名詞および名詞連続に限定する．具体的には，ある名詞句がツイートで言及されるとき，その名詞句が被訂正フレーズに含まれる確率（条件付き確率）を算出する．被訂正フレーズ中には頻出し，その他のツイート中では出現頻度の低い名詞句は，被訂正時にのみ頻出することから，誤情報のキーワードとなる名詞句である可能性が高い．逆に，被訂正フレーズ以外でも頻出する名詞句は，一般的な名詞句であり，誤情報のキーワードとなる可能性は低い.「昨日のあれ」の「昨日」や「あれ」は，被訂正フレーズ以外でも頻出するため，一般的な名詞句であると判断できる．

フレーズ中の名詞句 w が誤情報のキーワードらしいかどうかを，式 1 によって計算する．ここで，D は訂正フレーズ集合を表す.

$$P(w \in D|w) = \frac{P(w \in D)}{P(w)} = \frac{w\text{が訂正パターンを伴って出現するツイート数}}{w\text{を含むツイート数}} \tag{1}$$

このように求めた条件付き確率が高い上位 500 個を，キーワードとして選択する．ただし，コーパス中での出現頻度が極端に低い名詞句を除くため，コーパス全体での出現回数が 10 回以上かつ，被訂正フレーズ集合での出現回数が 2 回以上の名詞句のみをキーワードとして認定する．また，ひらがなや記号が半数以上の名詞句 (例えば「○○町」) はキーワードとして不適切と考え，キーワードから取り除いた.

3.3 ステップ 3：キーワードのクラスタリング

被訂正フレーズには,「コスモ石油の火災により有害物質を含む雨が降る」と「コスモ石油の爆発は有害だ」のように，同一の被訂正情報を言及しているが，表現や情報量の異なるフレーズが含まれている．誤情報を過不足なく抽出するために，これらをまとめる必要がある．そこで，ステップ 2 で抽出されたキーワードを，同一の被訂正情報を説明するキーワードがまとまるようにクラスタリングする．

クラスタリングにおけるキーワード間の類似度計算では，キーワードと文内で共起する内容語（名詞，動詞，形容詞）を特徴量とした文脈ベクトルを用いた.

これは，周囲に同じ単語が表れていれば，２つのキーワードは類似しているという考えに基づく．文脈ベクトルの特徴量には，各単語との共起度合いを表す尺度である自己相互情報量 (PMI) を用いた．この値が０以上の内容語を文脈ベクトルの特徴量に加えた．各文脈ベクトルの類似度はコサイン類似度によって計算した．クラスタリング手法は，階層クラスタリングの一種である最長距離法を用いた．今回のデータでは，類似度の閾値を 0.2 に固定してクラスタリングを行ったところ，500 個のキーワードから 189 個のクラスタが得られた．

得られた各クラスタに対し，式 1 の示す確率が最も高いキーワードを代表キーワードとする．代表キーワードは，クラスタの誤情報を説明するために最も重要なキーワードであると考える．

3.4 ステップ 4：代表フレーズの選択

クラスタごとに被訂正フレーズを抽出し，誤情報として出力する．誤情報に相応しい被訂正フレーズは，誤情報を過不足なく説明できるような一文である．例えば，以下の例では，b は説明が不足しており，c は冗長な情報が含まれているため，a を誤情報として出力したい．

a コスモ石油の火災により，有害物質を含む雨が降る

b コスモ石油の件で，有害な雨が降る

c コスモ石油が爆発したというのは本当で，有害な雨が降るから傘やカッパが必須らしい

このような選択を可能にするため，内容語の種類と含有率に着目する．

まず，代表キーワードを含む被訂正フレーズを誤情報の候補として抽出する．次に，この候補の中から誤情報の内容を過不足なく説明するものを抽出する．文書自動要約における重要文抽出の考えから，前段で用いたキーワードとよく共起する内容語を多く含むものは，より重要な文であると考えられる．そこで，共起度合いを自己相互情報量 (PMI) で計る．

$$\mathrm{Score}_{\mathrm{p}}(s,t) = \sum_{w \in C_s} \mathrm{PMI}(t,w) \tag{2}$$

s は被訂正フレーズ，t は各クラスタの代表キーワード，C_s は s 中の内容語の集合を表す．ここで，内容語とは被訂正フレーズに含まれる名詞，動詞，形容詞とする．この式により，誤情報クラスタを代表するキーワードと共起性の強い内容語を多く含むフレーズに対して，高いスコアが付与される．

しかし，この式では，被訂正フレーズに含まれる内容語の数が多い，長い文ほど高いスコアが付与されてしまう．そこで，代表キーワードを含む文の中でも，典型的な長さの文に高いスコアを付与し，短い文および長い文に対して低いスコアを与える補正項を用いる．

$$\mathrm{Score_n}(s,t) = \mathrm{hist}(\mathrm{len}_s, t) \tag{3}$$

len_s は被訂正フレーズ s の単語数を示す．$\mathrm{hist}(l,t)$ は，代表キーワード t を含み，かつ単語数が l である文の出現頻度を表す．

最終的なスコアは，式 2 と式 3 を乗算したものとする（下式）．

$$\mathrm{Score}(s,t) = \mathrm{Score_p} * \mathrm{Score_n} \tag{4}$$

最後に，各クラスタから式 4 のスコアが最も高いフレーズを一つずつ選択し，誤情報として出力する．

4　予備実験：訂正パターンの評価

提案手法は，訂正パターンで表明されない誤情報を獲得することができず，誤情報の抽出性能に大きく影響する．そこで本章では，ステップ 1 で用いた，人手で整備した訂正パターンの性能を評価する．

4.1　データセット

誤情報の抽出元となるコーパスには，東日本大震災ビックデータワークショップ[5]で Twitter Japan から提供された，2011 年 3 月 11 日 9 時から 2011 年 3 月 18 日 9 時までに発信された日本語のツイートデータ全 179,286,297 ツイートを利用した．このデータのうち，リツイート（自分の知り合いへのツイートの転送）は単純に同じ文が重複しているだけであるため，取り除いた．

4.2　正解データ

今までに，東日本大震災の際に発信された誤情報を網羅的にまとめたコーパスは存在しない．そこで正解データを作成するため，誤情報を人手でまとめた以下の４つのウェブサイトに掲載されている事例を利用した．

1. 絵文録ことのは「震災後のデマ 80 件を分類整理して見えてきたパニック時の社会心理」[6]
2. 荻上式 BLOG「東北地方太平洋沖地震, ネット上でのデマまとめ」[7]
3. 原宿・表参道.jp　地震のデマ・チェーンメール[8]
4. NAVER まとめ 注意！地震に関するデマ・チェーンメールまとめ[9]

以上の 4 サイトに掲載されているすべての事例のうち，Twitter データの投稿期間内 (2011 3/11 09:00 から 2011 3/18 09:00 まで) に発信されたと判断できる事例は全部で 60 件存在した．この 60 件の誤情報を正解データとした．作成した正解データの一部を以下に列挙する．全 60 件は後述の付録に記述した．

[5] https://sites.google.com/site/prj311/
[6] http://www.kotono8.com/2011/04/08dema.html
[7] http://d.hatena.ne.jp/seijotcp/20110312/p1
[8] http://hara19.jp/archives/4905
[9] http://matome.naver.jp/odai/2130024145949727601

- 関西以西でも大規模節電の必要性
- ワンピースの尾田栄一郎さん 15 億円寄付
- 天皇陛下が京都に避難された
- ホウ酸を食べると放射能を防げる
- 双葉病院で病院関係者が患者を置き去りにして逃げた
- いわき市田人で食料も水も来ていなく餓死寸前
- 宮城県花山村が孤立
- 韓国が震災記念 T シャツを作成
- 民主党がカップ麺を買い占め

4.3 評価尺度

訂正パターンは，適合率と再現率で評価した．収集した被訂正フレーズ集合約 2 万件からランダムに 150 件サンプリングし，その中で発信者が訂正パターンで情報を否定・訂正していると判断できる割合を適合率とした．再現率は，収集した被訂正フレーズ集合約 2 万件によって正解データの誤情報 60 件をカバーできた割合とした．

4.4 結果と分析

表 2: 訂正パターンの適合率と再現率

適合率	再現率
0.79 (118/150)	0.83(50/60)

表 2 に訂正パターンの適合率と再現率を示す．約 8 割の適合率，再現率で誤情報を抽出することができた．表 3 に抽出された被訂正フレーズの内訳を示す．

(あ) と (い) は表 2 の評価で正解と判断した事例である．そのうち，(い) は「昨日のあれはデマだ」の「昨日のあれ」のように，具体的な情報に言及していない

表 3: 抽出された被訂正フレーズの内訳

被訂正フレーズの種類	件数
(あ) 情報を訂正していると判断できる被訂正フレーズのうち，内容が十分なもの	76
(い) 情報を訂正していると判断できる被訂正フレーズのうち，内容が不十分なもの	42
(う) 誤抽出のうち，パターンが曖昧な事例	24
(え) 誤抽出のうち，著者の態度が不明な事例	8
合計	150

フレーズや,「イソジンの件ってデマだったのか。」の「イソジンの件」のように説明が不足している事例である．ステップ 2 の条件付き確率によるランキグや，ステップ 4 の代表フレーズの選定を行うことで，(い) のような訂正フレーズを取り除くことができると考えられる．

(う) と (え) はどちらも誤って抽出された事例である．そのうち，(う) は「こういう災害の時ってデマがよく流れる」のように，訂正パターンの用法の違いにより訂正されていないフレーズを抽出した事例である．(え) は「募金するとモテるってデマを流せばいい」のように，訂正パターンに続く表現により，著者の訂正に対する態度が曖昧になっている事例である．

また，抽出出来なった誤情報 10 件を調査したところ，表 4 にある 3 つに分類することができた．

表 4: 抽出できなかった誤情報の内訳

原因	件数
(お) 新しい訂正パターンが存在	3
(か) 訂正ツイート内に手がかりあり	4
(き) 訂正ツイートなし	3
合計	10

(お) は今回整備した訂正パターンでは網羅できなかった事例である．例として「天皇が 24 時間御祈祷に入ってるってのはソースがない」の下線部の訂正パター

ンは，今回整備した訂正パターンには含まれていなかったが，今後パターンを拡充することで抽出できる．

(か) は本研究が対象とする訂正パターンの型によらず，誤情報を訂正した例である．例として,「日本に韓国が借金の申し出。しかも管は快諾」という誤情報に対して以下のような訂正ツイートが存在した．

> これデマなんじゃ？ソースないし。 RT @xxx　RT こんな非常事態の日本に韓国が借金の申し出。しかも管は快諾！

この例のように，元のツイートにコメントする形で，情報を訂正するツイートがいくつか見られた．

(き) の誤情報は今回の実験で用いたツイート内に存在するが，それに対する訂正ツイートが存在しない事例である．本手法は，誤情報には何らかの訂正ツイートが存在することを前提としているため，抽出は困難であるが，その数は少ない．

5　本実験：誤情報の集約の評価

本章では，3 節のステップ 2 から 4 を評価する．前章で抽出された被訂正フレーズを，その代表キーワードの式 1 で並べ替え，上位 100 件を評価対象とした．(い) に含まれる具体的な情報に言及していない被訂正フレーズが取り除けたか，誤情報を過不足なく説明する被訂正フレーズを抽出できたか，という観点で評価をする．考察では，ツイートデータから抽出できなかった事例や，誤って抽出された事例を分類し，今後の対策について述べる．

5.1　実験設定

抽出された誤情報の正否は，同等の内容が 60 件の正解データに含まれるかどうかを一件ずつ人手で判断した．また，正解データに含まれていないが，誤情報であると判断できるものもある．そこで抽出された情報が正解データに含まれなかった場合は，関連情報を検索することで，その正否を検証した．

本研究の目的は，出来るだけ多くの誤情報を抽出し，人に提示することにある．しかし人が一度に見ることのできる情報には限界があり，出来るだけ多くの誤情報を人に提示するには，提示する誤情報の中にある，冗長な誤情報を取り除きたい．この目的のため，抽出した誤情報のうち，同じ内容と判断できるものが複数ある場合は，正解は一つとし，他の重複するものは不正解とした．また，日本語として不自然なものも不正解とした．

5.2　評価尺度

提案手法はスコアの高い順に N 件まで出力可能であるため，N をいくつか変化させたときの精度@N，再現率@N，F 値@N によって評価した．精度には，正解データに含まれるかどうかで判断したもの (精度@N(60 件)) と，人手により検証を行ったもの (精度@N(人手)) を用意した．また，人手による検証に加え，重複を許した場合 (精度@N(重複)) も評価に加えた．この評価を行うことで，目的の一つである「誤情報抽出」がどの程度達成されているかを知ることができる．それぞれは以下の式で表される．

$$\text{精度@}N(60\text{ 件}) = \frac{N\text{ 事例のうち, }60\text{ 件の誤情報に含まれる数 (重複除く)}}{N} \quad (5)$$

$$\text{精度@}N\text{(人手)} = \frac{N\text{ 事例のうち，人手で誤情報と検証された数 (重複除く)}}{N} \quad (6)$$

$$\text{精度@}N\text{(重複)} = \frac{N\text{ 事例のうち，人手で誤情報と検証された数 (重複許す)}}{N} \quad (7)$$

$$\text{再現率@}N = \frac{N\text{ 事例のうち, 60 件の誤情報に含まれる数 (重複除く)}}{\text{正解の誤情報の数（60 件）}} \quad (8)$$

$$F\text{ 値@}N = \frac{2 * \text{精度@}N\text{(60 件)} * \text{再現率@}N}{\text{精度@}N\text{(60 件)} + \text{再現率@}N} \quad (9)$$

5.3 実験結果

表 5: 誤情報の抽出結果

	精度@N(60 件)	精度@N(人手)	精度@N(重複)	再現率@N	F 値
N = 25	0.44(11/25)	0.68(17/25)	1.00(25/25)	0.18(11/60)	0.26
N = 50	0.34(17/50)	0.60(30/50)	0.90(45/50)	0.28(17/60)	0.31
N = 75	0.36(27/75)	0.59(44/75)	0.80(60/75)	0.45(27/60)	0.40
N = 100	0.31(31/100)	0.54(54/100)	0.76(76/100)	0.52(31/60)	0.39
上限 (N=189)	—	—	—	0.63(38/60)	—
上限 (クラスタなし)	—	—	—	0.83(50/60)	—

評価結果を表 5 に示す．N が 100 のとき，提案手法が抽出した情報のうち，60 件の正解データにも含まれる情報は 31 件であった．さらに，正解データには含まれないが，誤情報と判断できる事例が 23 件存在したことから，提案手法は 54%の精度で誤情報を抽出できた．実際に抽出できた誤情報を表 6 に示す．上位を見ると，震災当時デマとして拡散した誤情報が抽出 s されていることが分かる．また，正解データには含まれないが，誤情報と判断できた事例を表 7 に示す.「カラオケ館が便乗値上げした」のように，信じたとしても一見害がない情報も抽出された.

もし表にある「新宿高島屋が無料開放」という情報を信じてしまった場合，緊急時に開放していない避難先に誤って向かい，貴重な時間を失う可能性がある．このようにまとめサイトには掲載されておらず，かつ情報を信じた場合のリスクが高い,「有用な」誤情報も抽出することができた.

表 6: 抽出された誤情報のうち，まとめサイトに掲載されていた事例

順位	キーワード	誤情報
1	田尻智さん	ポケモンの生みの親の田尻智さんが亡くなった
2	尾田栄一郎先生	尾田栄一郎先生が 15 億円を寄付
3	女性暴行	「阪神大震災の際には女性暴行が増えた」
4	コスモ石油千葉製油所	市原市のコスモ石油千葉製油所 LPG タンクの爆発により、千葉県、近隣圏に在住の方に有害物質が雨などと一緒に飛散する
5	有毒物質	コンビナート火災に関し『有毒物質が発生し、雨に混じって降ってくるので肌をさらさないように』

表 7: 抽出された誤情報のうち，まとめサイトに掲載されていなかった事例

順位	キーワード	誤情報
29	新宿高島屋	新宿高島屋が無料開放
96	値上げ	カラオケ館が便乗値上げした

次に，上位 N 件に限定しない場合の再現率について述べる.「上限 (N=189)」は 500 個のキーワードをクラスタリングし得られた 189 個のクラスタから，代表フレーズをすべて出力した時の再現率であり,「上限 (クラスタなし)」は，提案手法ステップ 1 で収集された被訂正フレーズ集合約 2 万件をすべて出力した時の再現率である.「上限 (N=189)」は，キーワードを 189 個に絞った時の，ランキング改善による性能向上限界を表すに対し，後者はキーワードの選択，ランキング，クラスタリング改善による性能向上限界，つまり訂正パターンに基づく抽出手法の

限界を表す．被訂正フレーズ集合の段階でカバーされている 50 件は，キーワードの選択やクラスタリングなど，後段の処理を改善することで抽出できる可能性があるが，残る 10 件は，訂正パターンに基づく抽出手法の改善が必要となる，難解な事例である．

5.4 精度に関するエラー分析

本節では，評価結果の誤りを分析する．抽出された誤情報の上位 100 件のうち，31 件は正解データに含まれていたが，残りの 69 件は正解データに含まれていなかった．そこで，不正解データに対する誤判定の原因を調べたところ，８種類の原因に分類できた．表 8 に理由と件数を示す．

表 8: 精度に対する誤り分析

原因の内容	件数 (件)	割合 (%)
(a) キーワード抽出による誤り	6	8.70
(b) クラスタリングによる誤り (重複)	22	31.9
(c) 内容が不明確な情報	5	7.25
(d) 正しい情報	1	1.45
(e) まとめサイトに掲載されていない誤情報 (過去)	9	13.0
(f) まとめサイトに掲載されていない誤情報 (現在)	14	20.3
(g) 未来予測	6	8.70
(h) 真偽不明	6	8.70
統計	69	100.0

(a) から (d) は，明らかに誤抽出と判断できる事例である．(e) と (f) は，正解データの構築に用いた４つの誤情報まとめサイトに掲載されてはいなかったが，ウェブ上で調べることで，明らかに誤情報であると認められる事例である．(g) と (h) は，人手でも誤情報であるかを判断できない事例である．

以下でそれぞれの詳細と，改善案を述べる．

(a) キーワード抽出による誤り

代表キーワードが誤抽出につながったと考えられる事例である．以下に例を示す．括弧の中は，選定に利用した代表キーワードである．

> 陰謀論とか、「悪意の行動があった」とかいうデマを信じる人って … (悪意)

「善意」や「悪意」といった単語は，元々「デマ」などの訂正表現の周辺文脈に出現しやすい単語であるため，条件付き確率 (1) が高く，キーワードとして選ばれた．しかし，特定の誤情報に関連するキーワードではないため，上記の例のように，具体性に欠ける被訂正フレーズが誤情報として抽出された．このようなキーワードは，誤情報の拡散時に限らず，通常時から訂正表現と共起すると考えられる．そこで対策として，被訂正フレーズに含まれる確率 (式 1) を使用するのではなく，通常時の共起度合いを組み込むことで，改善が望めると考えらる．

(b) クラスタリングによる誤り

抽出された誤情報上位１００件のうち，同じ内容と判断できる誤情報が重複している事例である．例を以下に示す．括弧の中は，選定に利用した代表キーワードである．

> 市原市のコスモ石油千葉製油所 LPG タンクの爆発により，千葉県，近隣圏に在住の方に有害な雨などと一緒に飛散する (コスモ石油千葉製油所)
> 千葉県の石油コンビナート爆発で，空気中に人体に悪影響な物質が空気中に舞い雨が降ると酸性雨になる (石油コンビナート爆発)

これはステップ 3 でクラスタリングを行ったとき，同じクラスタに分類できなかったため，重複として表れた．誤情報検出の目的は達成できているものの，冗長な誤情報を抜き出しているため厳しめに評価して不正解とした．キーワードのクラスタリングには，被訂正フレーズの中で共起する単語を素性としているが，素性に表層の情報を加えることで，誤りを減らすことができると考えられる．

(c) 内容が不正確な情報

抽出された誤情報の内容が，誤情報を説明するのに内容が不足していると思われる事例である．以下に例を示す．

> 餓死者や凍死者が出た．

正解データの中には「いわき市で餓死者や凍死者が出た」というものが存在するが，それと比べると具体性に欠けているため，不正解とした．より的確な候補を抽出するには，候補が多いほど作成したパターンの精度や再現率を考慮した選定が必要である．

(d) 正しい情報

誤情報として抽出されたが，事実を確認したところ，誤情報ではなかった事例である．以下に例を示す．

> 東京タワーの先端が曲がった

この例に関連するツイートを観察したところ，根拠とされる写真を提示されても信じてもらえないほど，突拍子のない情報として扱われていた．そのため，訂正ツイートが多く投稿されたようである．提案手法は訂正の数が多い情報ほど，ランキングが上位になる仕組みになっているため，この事例は誤って抽出された．本研究の目的は「誤情報の抽出」であることを考えると，(a) から (c) の誤りに比べ，深刻な誤りである．しかし，始めは誤情報として疑っていたユーザーの中には，誤情報出なかったことを知り，以下のようなツイートをしている人も存在した．

> 東京タワーが曲がったってデマじゃなかったんだ東京タワー曲がったとかデマだと思ったら本当だった

このように，訂正を訂正しているツイートも存在し，二重否定を判別することが出来れば，この問題の改善につながると考えられる．

(e) まとめサイトに掲載されていない誤情報 (過去)

これは誤情報まとめサイトに掲載されていないが，人手で検証したところ，誤情報と判別された事例である．その中でも今回利用したツイートコーパスの期間より前の事象に関する誤情報である．以下に例を示す．

> 関東大震災の時「朝鮮人が井戸に毒を入れた」というのはデマだったはず
> 阪神淡路大震災は三時間後に最大の揺れが来たというのは誤った情報のようです。
> 明治 43 年（1910 年）にハレー彗星が大接近した時、地球上の空気が 5 分間ほどなくなるというデマが一部で広まり，…

上記の例は訂正ツイートであり，下線部は被訂正フレーズとして抽出された部分である．一度過去に誤情報として認識されたことは間違いないが，人々に悪影響を与える可能性があり，誤情報として抽出し，拡散・訂正の動向を監視する必要がある．

(f) まとめサイトに掲載されていない誤情報 (現在)

これは誤情報まとめサイトに掲載されていないが，人手で検証を行ったところ，誤情報と判別された事例である．その中でも今回利用したツイートコーパスの期間中に発生した誤情報である．以下に例を示す．

VIP で韓国の救助犬 1 匹が逃亡
巷説にある遺体には感染症のリスクがある

(g) 未来予測

(h) の真偽不明の事例のうち，未来に起こりうる事象について述べたものを抽出した事例である．以下に例を示す．

福島で核爆発が起こる
富士山が噴火する

未来に起こりうる事象である以上，現時点での真偽は不明である．抽出されたものの多くは，上記の例のように人々の不安を煽る情報であり，パニックを防ぎたいと思い訂正ツイートを発信した人が多かったため，抽出されたと考えられる．

(h) 真偽不明

複数のウェブサイトを検索して検証を行ったが，誤情報かどうかを判別できなかった事例である．以下に例を示す．

サントリーが自販機無料開放
築地で魚が余っている

5.5 再現率に関するエラー分析

次に，正解データにある誤情報 60 件のうち，抽出されなかった誤情報 29 件についても前節と同様に原因を調査したところ，33 つに分類できることが判明した．3 つの原因の件数と割合を表 9 に示す．

表 9: 再現率に対する誤り分析

原因の内容	件数 (件)	割合 (%)
(i) 訂正パターンで候補を抽出できなかったもの	10	34.5
(j) 訂正パターンで抽出できたが，クラスタリングによる誤り	2	6.9
(k) 訂正パターンで抽出できたが，ランキング外	17	58.6
統計	29	100.0

(i) 訂正パターンで候補を抽出できなかったもの

今回作成した訂正パターンでは，抽出できなかった誤情報である.「仙台市三条中学校が中国人・韓国人が７割の留学生の心ない行動で避難所機能停止」という誤情報に対して，以下のようなツイートが数多く存在した.

> コレ本当? RT @XXXXX 今，祖母と叔母に確認．何と仙台市の三条中学校の避難所，閉鎖！避難所用救援物資を根こそぎ，近隣の外国人留学生（中国韓国で七割強）が運び出してしまい，避難所の機能停止だそうです.

上の例では，明示的に誤情報だと否定している人は少ないが，元のツイートコメントする形で，その情報を疑っている人は多かった．このことから，改善案とし訂正パターンのみではなく，懐疑を表す表現も利用できるのではないかと思われる.

(j) 訂正パターンで抽出できたが，クラスタリングによる誤り

訂正パターンにより候補の抽出はできたが，クラスタリングにより，誤って他の誤情報に含まれた事例である．しかし，全体に比べ，事例数が少ないため，それほど問題ではないと思われる.

(k) 訂正パターンで抽出できたが，ランキング外

訂正パターンにより候補を抽出できたが，条件付き確率が低かったため，キーワードとして抽出できなかった事例である．例えば,「東京電力を装った男が表れた」という誤情報では,「東京電力」というキーワードは誤情報以外の話題でも頻出したため，条件付き確率が低くなった．対策としては,

キーワード単独をスコアリングするのではなく，被訂正フレーズそのものをスコアリングするような手法が必要である．

6　一般ツイートからの誤情報抽出

誤情報は非常時に限らず，通常時でも問題となっている．我々の手法が，震災時を対象とするだけでなく，通常時のツイートに対しても有効であることを示すため，本章では通常時のツイートを用いて誤情報の抽出を行う．

6.1　実験設定

誤情報抽出元となるコーパスとして，Twitter API を用いて独自にクロールした，2013 年 2 月 6 日から 2014 年 1 月 31 日までの日本語の 9,424,868,844 ツイートを用いた．評価方法について，東日本大震災の時に比べ，現在誤情報を収集している Web ページは少ない．よって評価は，抽出されたインスタンスの上位 25 件の正否の検証した．正否の検証は前章と同様に，人手により Web で関連情報を検索することで行った．

6.2　実験結果

表 10: 一般ツイートから抽出されたフレーズの種類

タイプ	# 事例数
(A) 誤情報	18
(B) 真偽不明	3
(C) 重複する誤情報	1
(D) 抽出エラー	3
合計	25

実験結果を表 10 に示す．表 10 の (A) にあるように，上位 25 件のうち，18 件の誤情報を抽出することが出来た．人手による検証により正解を判断した場合，通常時のデータに対する誤情報抽出の精度は 0.72 であり，災害時の精度 0.68(17/25) と比べると，同程度の性能で誤情報を抽出することが出来た．

通常時のツイートから実際に抽出された事例を表 11 に示す．

残り 7 件は，真偽不明（B）と重複する誤情報（C），抽出エラー（D）の 3 つに分類できる．真偽不明（B）は人手で判断しても，誤情報かどうかを判断できなかった事例である．例えば，収集された真偽不明な情報の中には,「同性婚を認めると同性愛者が増える/出生率が下がる」のようなものがあった．真偽の判定に

表 11: 通常時のツイートから抽出された誤情報

事例一覧	タイプ
iPhone を電子レンジで充電すればすぐに充電完了する	誤情報
ホワイトハウスでテロ爆破があり，オバマ大統領が負傷	誤情報
同性婚を認めると同性愛者が増える/出生率が下がる	真偽不明
野間さんが在日特権	抽出エラー

は社会学的調査が必要となり，判定が難しい事例である．このような情報は，真偽を判定するのは難しいが，何かしらの否定がされている以上，間違っている可能性がありユーザーにその情報を提示することで，ユーザーに何かしらの疑いを与えることができ，ユーザーに信憑性判断の機会を与えることが出来ると考えている．

重複する誤情報（C）は，25 件のうち，同じ内容と判断することの出来る誤情報が含まれていた事例である．

エラーとして判断した事例（D）は,「野間さんが在日特権」の様に，具体性の無いものや，文の意味を理解できないものである．誤情報を抽出するという目的からすると，出力するべきでないもである．しかし，誤って抽出した時に特に問題となる真実の情報は含まれておらず，本手法は通常の誤情報に対しても有効である．

7　Webテキストからの誤情報抽出

情報の信憑性の問題は，Twitterに限らず，Web全体で問題となっており，昔から様々な研究がされてきた．提案手法は，ツイート中のテキスト情報を元に誤情報の抽出，集約を行うため，一般のWebテキストに対しても適応可能と考える．そこで本章では，Webテキストからの誤情報の抽出，集約を行い，ツイートデータを対象とした場合との違いについて詳しく述べる．

7.1　実験設定

誤情報抽出元となるコーパスとして，Webからクロールした約150億文を用いた．このデータは予め重複する文が除かれている．訂正パターンについて，3章の表1で述べた訂正パターンを抽出に用いる．このパターンにより抽出された被訂正フレーズのうち，完全に一致するフレーズは，コピーアンドペースト(元あった文章を別の場所にコピーすること）したものと考え，取り除いた．これにより，約10万の被訂正フレーズが収集された．このデータを対象に，提案手法のステップ2からステップ4により集約を行う．

評価方法について，Webにある誤情報を網羅的に収集したページ，データベースは限られている．よって評価は，提案手法により出力された上位25件の正否を検証した．正否の検証は前節と同様に，抽出されたフレーズの関連情報を一件ずつWebで検索を行い調査し、ニュースページや公式サイトから真偽を判断した．

7.2　実験結果

表 12: Webテキストから抽出されたフレーズの種類

タイプ	# 事例数
(A) 誤情報	12
(B) 真偽不明	5
(C) 重複する誤情報	4
(D) 抽出エラー	4
合計	25

実験結果を表12に示す．表12の(A)にあるように，上位25件のうち，10件の誤情報を抽出することが出来た．実際に抽出された事例を表13に示す．抽出

表 13: Web テキストから抽出された事例

事例一覧	タイプ
放射能汚染から体を守るには塩が効く	誤情報
ＣＯ２の増加と地球温暖化の関連性	真偽不明
北朝鮮によるミサイル発射が行われた	抽出エラー

された中には「ＣＯ２の増加と地球温暖化の関連性」のように，突発的に発生した誤情報ではなく，定常的に真偽について議論されている，真偽不明も抽出された．本研究の目的は誤情報の抽出であるが，多くの人が反論していることから真偽不明の中でも重要な事例と考えている．

また誤情報として抽出した事例の中には,「北朝鮮によるミサイル発射が行われた」のように，いつの時点での誤情報なのか分からない事例も存在した．抽出される事例によっては，時間の経過によって真偽がかわるものもあり，データの抽出期間を指定しない場合は，いつの時点での訂正情報なのかを判別する必要がある．

8　応用：誤情報監視システム

前章でも述べたとおり，誤情報は非常時に限らず，通常時でも問題となっている．誤情報の拡散による問題を防ぐには，誤情報をより早く発見し，早期に訂正をする必要がある．そこで本研究の応用課題として今後，通常時から誤情報の収集を行い，リアルタイムでユーザーに提示するシステムの構築を考えている．

図 5 は，現在試作中の誤情報監視システムである．このウェブアプリケーションを利用することで，ユーザーは現在，または今までに発生した誤情報をいち早く知ることができる．

このシステムはまず，Twitter API を用いて 15 分おきに「デマ」や「間違い」といった誤情報と関連するキーワードを用いて，誤情報に関するツイートをウェブからクロールしてくる．さらにこのツイートの中から誤情報を発見，同一の情報の集約を行い，ユーザーに提示する仕組みである．このシステムを運用することで，誤情報に対する注意喚起を容易に出来ると考えている．この誤情報監視システムを構築し，ツイッターユーザーに情報を伝えることのリテラシーを身につけてもらうことが，誤情報拡散を防ぐことに繋がると考えられる．

自動抽出されたものは必ずしも誤情報とは限らないので，今後一般ユーザーが提示された情報に対して訂正，補足できるようシステムを改良していく必要がある．また大量の情報のリアルタイム処理についても，研究していく予定である．

順位	誤情報	訂正数	初出	NEW
1	１８才未満がLINEが使えなくなる	667	2013-07-24	NEW
2	フォロワー以外へのリプライは規約違反だから凍結される	490	2013-07-26	NEW
3	自民党が徴兵制を検討	467	2013-07-17	NEW
4	ドイツの食品基準は日本より厳しい	336	2013-07-24	NEW
5	山本太郎は福島県産は放射性廃棄物と言った	211	2013-07-21	NEW
6	秋から18歳未満のLINE使用が禁止になる年齢制限開始	184	2013-07-24	NEW
7	あと、比例で個人名を書けばワタミの元会長に票が入らない	129	2013-07-20	NEW
8	選挙で議員が選ばれるので、民主主義国家である	103	2013-07-09	
9	田舎の駅では、切符を通す機械がなく、駅員さんが一つ一つ目視で管理している	63	2013-05-18	
10	三宅洋平さんを応援したいと思う人ほど、彼の不正選挙や人工地震	63	2013-07-25	NEW

図 5: リアルタイム誤情報収集システム

9　おわりに

本研究では，誤情報を訂正する表現に着目し，誤情報を自動的に収集する手法を提案した．実験では，誤情報を人手でまとめたウェブサイトから取り出した誤情報のリストを正解データと見なして評価を行ったところ，出力数が 100 件のとき正解データの約半数である 31 件を収集することができた．これは抽出した情報 100 件の約 3 割であるが，残り 69 件の中には，まとめサイトに掲載されていない誤情報も 23 件あり，54%の精度で誤情報を抽出できた．また，収集された誤情報の中に真実の情報が含まれていると深刻な問題であるが，誤って抽出された事例の多くは，内容の重複する誤情報や真偽不明の事例であり，特に問題である真実の情報は 100 件のうち 1 件と非常に少なく，提案手法は誤情報の自動収集に有用であることを示した．

今後は，訂正パターンの拡充や被訂正フレーズのスコアリングの改良を進め，誤情報抽出の性能 を向上させるとともに，リアルタイムでの誤情報獲得に取り組む予定である．

謝辞

本研究を進めるにあたり，ご協力，御助言を頂きました多くの方々に，深く感謝いたします．主指導教員である乾健太郎教授には，お忙しい中，研究活動全般にわたり温かいご指導，御助言を頂きました．心より感謝いたします．審査委員をお引き受け下さいました，徳山豪教授，伊藤彰則教授に深く感謝します．本研究内容に関して，ご指導，御助言を頂きました岡崎直観准教授に深く感謝いたします．本研究内容に関して，いろいろと御助言を頂きました渡邊陽太郎助教授に深く感謝いたします．研究室内での進捗報告の度に，本研究に関して有意義なご指摘を頂きました松林優一郎研究特任助教に深く感謝いたします．本研究内容に関して，数多くの御助言，相談にのっていただきました水野淳太さんに深く感謝いたします．研究生活や学生生活を暖かく見守ってくださいました八巻智子秘書に心から感謝いたします．最後になりましたが，研究生活の様々な場面でお世話になりました研究室の皆様有難うございました．

参考文献

[1] 野村総合研究所. プレスリリース：震災に伴うメディア接触動向に関する調査. http://www.nri.co.jp/news/2011/110329.html, 2011.

[2] ネットレイティングス株式会社. ニュースリリース: 震災の影響により首都圏ライフライン関連サイトの訪問者が大幅増. http://csp.netratings.co.jp/nnr/PDF/Newsrelease03292011_J.pdf, 2011.

[3] Adam Acar and Yuya Muraki. Twitter for crisis communication: lessons learned from japan's tsunami disaster. *International Journal of Web Based Communities*, Vol. 7, No. 3/2011, pp. 392–402, 2011.

[4] Son Doan, Bao-Khanh Ho Vo, and Nigel Collier. An analysis of Twitter messages in the 2011 Tohoku Earthquake. In *4th ICST International Conference on eHelth*, 2011.

[5] Takeshi Sakaki, Fujio Toriumi, and Yutaka Matsuo. Tweet trend analysis in an emergency situation. In *Special Workshop on Internet and Disasters (SWID 2011)*, pp. 3:1–3:8, 2011.

[6] 宮部真衣, 荒牧英治, 三浦麻子. 東日本大震災における twitter の利用傾向の分析. 情報処理学会研究報告, 第 2011-DPS-148/2011-GN-81/2011-EIP-53 巻, 2011.

[7] Jon M Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, Vol. 46, No. 5, pp. 604–632, 1999.

[8] Jeff Pasternack and Dan Roth. Making better informed trust decisions with generalized fact-finding. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three*, pp. 2324–2329. AAAI Press, 2011.

[9] Elisabeth Lex, Michael Voelske, Marcelo Errecalde, Edgardo Ferretti, Leticia Cagnina, Christopher Horn, Benno Stein, and Michael Granitzer. Measuring the quality of web content using factual information. In *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality*, pp. 7–10, 2012.

[10] Rob Ennals, Beth Trushkowsky, and John Mark Agosta. Highlighting Disputed Claims on the Web. In *Proc. of WWW 2010*, pp. 341–350, 2010.

[11] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Goncalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World Wide Web*, WWW '11, pp. 249–252, 2011.

[12] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pp. 675–684. ACM, 2011.

[13] Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. Rumor has it: identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pp. 1589–1599, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[14] 白井嵩士, 榊剛史, 鳥海不二夫, 篠田孝祐, 風間一洋, 野田五十樹, 沼尾正行, 栗原聡. Twitter におけるデマツイートの拡散モデルの構築とデマ拡散防止モデルの推定. 2012.

[15] 藤川智英, 鍛治伸裕, 吉永直樹, 喜連川優. マイクロブログ上の流言に対するユーザの態度の分類. 言語処理学会第 18 回年次大会, 2012.

[16] 鳥海不二夫, 篠田孝祐, 兼山元太. ソーシャルメディアを用いたデマ判定システムの判定精度評価. デジタルプラクティス, Vol. 3, No. 3, pp. 201–208, jul 2012.

[17] 大和田裕亮, 水野淳太, 岡崎直観, 乾健太郎, 石塚満. 返信・非公式リツイートに基づくツイート空間の論述構造解析. 自然言語処理= Journal of natural language processing, Vol. 20, No. 3, pp. 423–460, 2013.

[18] 梅島彩奈, 宮部真衣, 荒牧英治, 灘本明代. 災害時 twitter におけるデマとデマ訂正 rt の傾向. 情報処理学会研究報告. データベース・システム研究会報告, Vol. 2011, No. 4, pp. 1–6, jul 2011.

[19] 梅島彩奈, 宮部真衣, 灘本明代, 荒牧英治. マイクロブログにおける流言マーカー自動抽出のための特徴分析. 言語処理学会第 18 回年次大会, 2012.

[20] 宮部真衣, 梅島彩奈, 灘本明代, 荒牧英治. 流言情報クラウド: 人間の発信した訂正情報の抽出による流言収集. 言語処理学会第 18 回年次大会, 2012.

[21] Sanda Harabagiu, Andrew Hickl, and Finley Lacatusu. Negation, contrast and contradiction in text processing. In *Proceedings of the 21st national conference on Artificial intelligence*, pp. 755–762, 2006.

[22] Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. Finding contradictions in text. In *Proceedings of ACL-08: HLT*, pp. 1039–1047, 2008.

[23] Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pp. 1–9, 2007.

[24] Marie-Catherine de Marneffe, Anna R. Rafferty, and Christopher D. Manning. Identifying Conflicting Information in Texts. In *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer, 2011.

[25] Alan Ritter, Stephen Soderland, Doug Downey, and Oren Etzioni. It's a Contradiction – no, it's not: A Case Study using Functional Relations. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 11–20, 2008.

[26] Yotaro Watanabe, Junta Mizuno, and Kentaro Inui. THK's natural logic-based compositional textual entailment model at NTCIR-10 RITE-2. In *Proceedings of the NTCIR-10 Conference*, pp. 531–536, 2013.

[27] Bill MacCartney, Michel Galley, and Christopher D. Manning. A phrase-based alignment model for natural language inference. In *Proceedings of 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP-08)*, pp. 802–811, 2008.

[28] Yotaro Watanabe, Yusuke Miyao, Junta Mizuno, Tomohide Shibata, Hiroshi Kanayama, Cheng-Wei Lee, Chuan-Jie Lin, Shuming Shi, Teruko Mitamura, Noriko Kando, Hideki Shima, and Kohichi Takeda. Overview of the recognizing inference in text (rite-2) at ntcir-10. In *Proceedings of the NTCIR-10 Conference*, pp. 385–404, 2013.

付録

A　正解データとして用いた誤情報一覧

正解データとして用いた誤情報の一覧を以下に載せる

表 14: 正解データとして用いた誤情報一覧

コスモ石油の火災により、有害な雨が降る
関西以西でも大規模節電の必要
筑波大学の連絡で約一時間後には茨城にも放射能が来る
天皇陛下が京都に避難された
農水省に非常事態通達「明日は出勤しなくていい」
今日の夕方、友人に防衛省の夫を持つ人が家族に 「東京から家族を逃がせ」と言われたとのこと。 その後、総務省の友人にも連絡したところ、 総務省はほとんど空になっています。東京にいる方、逃げて !!
東京電力アカウント@OfficialTEPCO は公式を騙った偽物
富士山から煙が出ている
市原火災で有害な雨が降る
放射線対策にイソジン・ワカメが良い
ホウ酸を食べると放射線を防げる
福島・双葉病院で病院関係者が患者を置き去りにして逃げた
仙谷前官房長官が 11 日徳島で地震に関する不謹慎発言
菅首相がこの間も豪華な夕食をとってる
鳩山前首相　「原発から半径２００キロは住めない」　九州から
仙谷由人氏 16 日から訪韓」「仙石氏が韓国を訪問している
小沢一郎被災で安否不明
蓮舫発案のコンビニの深夜営業禁止より、節電の為 全国パチンコ店 1 週間の営業規制を政府に訴えよう」
辻本補佐官が米軍救援活動に抗議（NHK）
さすが阪神大震災の被災地で「自衛隊は違憲です。 彼らから救援物資をもらわないで！」とビラを配った馬鹿者
在日米軍ポンプ車到着も日本側から支援断られる
台湾救助隊が日本政府に拒否され、台湾は民間から派遣

日本では物資の空中投下が認められていない
民主党がカップ麺を買占め
ヤマダ電機大船渡店で単一電池 4 本 2 千円
阪神淡路大震災のとき、地震で、朝鮮人による レイプ多発の事実と、放火説があります
仙台市三条中学校が中国人・韓国人が 7 割の留学生らの 心無い行動で避難所機能停止
歌手の BoA さんが Twitter で不謹慎発言
韓国が震災記念 T シャツを作成
こんな非常事態の日本に韓国が借金の申し出。しかも管は快諾！
中国の救援隊が遺体写真を撮影
アグネスさんの折鶴紛失
日本ユニセフは国連の UNICEF とは無関係の団体だから 募金してはいけない
フジテレビ募金の行き先は日本ユニセフ
宮城県花山村が孤立
いわき市田人で食料も水も来ていなく餓死寸前
北茨城・高萩乳児餓死
石巻の避難所で幼児が餓死
ワンピースの尾田栄一郎さん 15 億円寄付
トルコが 100 億円支援
天皇陛下が地の神を鎮めるために徹夜で 24 時間の祈祷
枝野官房長官 105 時間ぶりに就寝
静岡ガンダム崩壊
オバマ大統領の演説（諸君がまもなく赴く戦いは、 人類史上最強の救出活動となるだろう）
AC の『ぽぽぽぽーん』を歌っているのは矢野顕子
現地の人が SOS を求めている
外国人が犯罪を起こす
日本が地震兵器で攻撃された
著名人（山口裕子さん，田尻智さん,etc）が死亡
埼玉の水道水が危ない
東京電力を装った男がうろついている
自衛隊が県庁で物資を募集

ヨードを含んだものを食べるべき？
東大入学予定者が地震で合格取り消し？
海外セレブたち（ブリトニー，ガガ，etc）の多額の寄付
仙谷由人氏が東日本大震災を「ラッキー」と発言？
日本から放射線が飛んでくる
九州電力も節電の呼びかけ？
茨城県知事が災害派遣要請を出してない
自衛隊が支援物資を受け付けている
食塩を摂取すれば放射線の害が防げる

発表文献一覧

受賞一覧

- 言語処理学会 2013 年度論文賞 (2014)
- 情報処理学会第 75 回全国大会学生奨励賞 (2013)

学術論文誌

- 鍋島啓太, 渡邉研斗, 水野淳太, 岡崎直観, 乾健太郎. 訂正パターンに基づく誤情報の収集と拡散状況の分析. 自然言語処理, Vol. 20, No. 3, June 2013.
- Yotaro Watanabe, Junta Mizuno, Eric Nichols, Katsuma Narisawa, Keita Nabeshima, Naoaki Okazaki and Kentaro Inui. Leveraging Diverse Lexical Resources for Textual Entailment Recognition. ACM Transactions on Asian Language Information Processing (TALIP), Vol. 11, No. 4, pp.39:1-39:21, December 2012.

国際会議論文

- Yotaro Watanabe, Junta Mizuno, Eric Nichols, Katsuma Narisawa, Keita Nabeshima and Kentaro Inui. TU Group at NTCIR9-RITE: Leveraging Diverse Lexical Resources for Recognizing Textual Entailment. The 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-9), pp.418-421, December, 2011.
- Naoaki Okazaki, Keita Nabeshima, Kento Watanabe, Junta Mizuno, Kentaro Inui. Extracting and Aggregating False Information from Microblogs. In Proceedings of the Workshop on Language Processing and Crisis Information 2013, October 2013.

国内会議・研究会論文

- 車 智修, 鍋島啓太, 水野淳太, 岡崎直観, 乾健太郎. 文書構造を用いた Why 型質問応答システム. 第 27 回人工知能学会全国大会

- 渡邉研斗, 鍋島啓太, 岡崎直観, 乾健太郎. Twitter 上での誤情報と訂正情報の自動分類. 言語処理学会第 19 回年次大会, pp178-181 March 2013.
- 鍋島啓太, 水野淳太, 岡崎直観, 乾健太郎. マイクロブログからの誤情報の発見と集約. 言語処理学会第 19 回年次大会, pp182-185 March 2013.
- 渡邉研斗, 鍋島啓太, 水野淳太, 岡崎直観, 乾健太郎. Twitter における誤情報の拡散収束過程の可視化. 情報処理学会 第 75 回全国大会予稿集 pp1-657 ー 1-658, March 2013.(学生奨励賞受賞)
- 鍋島啓太, 水野淳太, 岡崎直観, 乾健太郎. 訂正パターンに基づく誤情報の抽出と集約. 情報処理学会 第 75 回全国大会予稿集 pp2-179 ー 2-180, March 2013.(学生奨励賞受賞)
- 鍋島啓太, 水野淳太, 渡邉陽太郎, 岡崎直観, 乾健太郎. 命題ネット構築にむけて. NLP 若手の会 第 7 回シンポジウム, September 2012.

解説記事

- 岡崎直観, 鍋島啓太, 乾健太郎 言語処理による分析-日本栄養士会活動報告の分析. 日本栄養士会雑誌, Vol.55, No.12, pp.6 ー 8, December 2012.