画の作者が3億円寄付」という誤情報よりも「イソジンを飲むと放射線予防になる」という誤情報が拡散した時のほうが，社会混乱になるのは明らかである．本研究では，誤情報の拡散収束過程を分析し，ツイートの量や収束するまでの時間を見ることで，社会混乱になりやすい誤情報を同定する手法について提案する．

2.3 誤情報と訂正情報を分類する研究

誤情報の拡散を防ぐ方法としていずれの関連研究でも，誤情報と訂正情報を自動的に検出する技術が必要不可欠となる．宮部ら [2] は訂正情報を提示するために，訂正情報と訂正情報でない情報の二値分類を，SVM を用いた機械学習で行なっている．学習するための素性としてツイート文中の「デマ」の周辺の単語，形態素数，URL の有無，引用 (RT@) の有無を素性として用い，高い精度で訂正情報の検出に成功している．また，平常時と災害時のツイートでの分類性能の比較も行なっており，平常時のツイートで訓練した分類器だと，災害時，通常時のどちらでも高い分類性能が得られると明らかにした．

Carlos ら [8] も機械学習を用いて情報の自動分類を行ったが，誤情報と訂正情報の分類ではく，信頼出来る情報か信頼出来ない情報に分類するという違いがある．Carlos らは機械学習に，ユーザの特徴，ツイート本文から得られる特徴，Twitter 固有の特徴の３つの特徴から得られる素性を用いている．

いずれの研究においても，訂正情報，もしくは信頼できる情報の検出には，根拠となる URL の有無や，誤情報に対して訂正している表現，もしくは否定的な表現の有無が有効な素性であるということを明らかにしている．しかし関連研究では，誤情報を訂正情報でない情報として仮定しているので，「URL が無いと誤情報の可能性が高い」や，「誤情報を訂正している表現が無いと誤情報の可能性が高い」のような特徴では，誤情報の本質的な特徴を捉えているとは言えない．そこで，本研究では誤情報に関するドメインが予め与えられていると仮定し，誤情報を説明する記述とツイート本文との類似度を求めることにより，誤情報の本質的な特徴を捉える素性を入れて機械学習を行ない，分類器の精度の向上を試みる．

第3章　誤情報の拡散収束過程を提示する為に

この章では，誤情報の拡散収束過程の分析や，誤情報と訂正情報の自動分類をするための準備として，前提条件や，必要になる分析・実験データの収集，技術的な面での問題について述べる．

3.1　コーパスの準備

本研究では実際に起こった現象を分析するため，東日本大震災当時の全ツイートを準備する必要がある．よって，東日本大震災ビッグデータワークショップで提供された震災一週間分 (2011/3/11 9:00 ～201/3/18 9:00) の全ツイート (約 1 億 8 千万ツイート) を用いて分析・実験データを作成する．

3.2　誤情報に関するツイートの収集

震災時の情報の流布を再現できたとしても，人手で誤情報に関連したツイートを収集，分類することは非常に困難である．よって，誤情報の拡散収束過程を分析・提示するためには

- 誤情報に関するツイートの自動収集
- 誤情報と訂正情報の自動分類

が必要となる．

ここで言う「誤情報」とは事実とは反する情報のことを言い,「訂正情報」は誤情報を訂正する立場にある情報のことを言う．本研究では誤情報と訂正情報の自動分類について取り扱うが，誤情報に関するツイートの自動収集については鍋島ら [5] を用いて，自動的に誤情報についての記述を収集する．鍋島らは「○○というのはデマです」のような，誤情報を訂正するような表現に合致するツイートを収集してきて，○○に当たる部分から誤情報らしいものを収集するという研究をしている．具体的には，まずコーパス全体に「○○というのはデマです」のような訂正フレーズを用いて○○の部分に当たる被訂正フレーズの抽出を行う．これによって「コスモ石油の爆発で有害な雨が降る」等の誤情報の候補を多数収集できる．

収集した被訂正フレーズの中には,「コスモ石油の爆発は有害だ」「コスモ石油の火災により有害な雨が降る」のように異なる表現だが，内容が同一ものがある．次に，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, 検索クエリとして「コスモ石油」や「有害」などのキーワードを抽出する．キーワードは式 2.1 で定義される，ある語がツイートで言及されるとき，その語が被訂正フレーズに含まれる確率 (条件付き確率) が高いものとする．

表 3.1: 分析・実験データの内訳

誤情報の内容	ツイート数	誤情報	訂正情報	その他
サーバーラックが倒れて動けない	1155	742	401	12
コスモ石油の爆発で有害な雨が降る	979	382	499	98
放射線対策にイソジン (うがい薬) が利く	925	162	700	63
阪神大震災では三時間後に一番大きい揺れが来る	610	506	84	20
ONEPEACE の作者尾田栄一郎が 15 億円寄付	311	170	134	7
東大が合格者の入学取り消し	249	140	81	28
天皇陛下が京都御所へ避難	171	25	129	17
支援物資の空中投下が認められていない	165	38	58	69
トルコが 100 億円支援	164	100	47	17
フジテレビの募金は日本ユニセフに行く	153	82	64	7
阪神大震災でレイプが多発した	152	69	82	1
福島第一原発が核爆発の恐れ	74	16	45	13
辻本補佐官が米軍の救助活動に抗議	46	28	16	2
ポケモンクリエーターの田尻智が死去	41	2	36	3
合計	5195	2462	2376	357

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 2.1 によって計算する．ここで，D は訂正フレーズ集合を表す．

$$P(w \in D|w) = \frac{P(w \in D)}{P(w)} = \frac{w \text{が訂正パターンを伴って出現するツイート数}}{w \text{を含むツイート数}} \tag{3.1}$$

しかし，このままでは被訂正フレーズが「コスモ石油のガソリンスタンド」などの関係のない情報が混ざり，誤情報らしいかがわからないので，収集した被訂正フレーズから代表的なフレーズの選択を行なう．選択には式 2.2 で定義されるスコアで誤情報らしさをランキングする．

$$Score(s,t) = \left(\sum_{w \in C_s} PMI(t,w) \right) * hist(len_s|t) \tag{3.2}$$

複数の似た表現の中から，最も誤情報らしいフレーズを選ぶために，第 1 項は式 2.1 で求めたキーワード w と被訂正フレーズ t との相関を PMI(自己相互情報量) を測る．s は被訂正フレーズ，t は各クラスタの代表キーワード，Cs 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．PMI により，キーワードとよく共起し，内容的に重要な文を選ぶことができるが，長い文ほど高いスコアになってしまう．よって第 2 項ではそれを補正している．len_s は被訂正フレーズ s の単語数を示し，$hist(l, t)$ は，s の中で最重要キーワード t を含み，かつ単語数が l である文の出現頻度を表す．

鍋島らの手法を用いることで「コスモ石油の爆発で有害な雨が降る」のような誤情報を説明する記述を多数得ることができる．本研究では，鍋島らが収集した誤情報の中の 14 件 (表 3.1) に対し，各誤情報を説明する記述 (例えば「コスモ石油の爆発で有害な雨が降る」) に対し，適切な検索クエリ (例えば「コスモ石油 AND 雨」) を選び，誤情報を拡散するツイート，訂正するツイートの両方を区別せずに収集した．なお，影響力の大きいツイートを重点的に調べるため被リツイー

ト数の多いツイートを優先的に採用した[1].

3.3 分析・実験データの作成

収集したツイートに対し，誤情報 (誤情報を支持する立場にある情報)，訂正情報 (誤情報を訂正・阻止する立場にある情報)，その他 (誤情報について言及していない情報，誤情報か訂正情報か判断に悩む情報) のいずれかのラベルを手作業で付与した．ラベル付与の基準はツイートをしたユーザの立場を優先する．例えば,「コスモ石油の公式サイト．チェーンメールに関して記載されています http://...」というツイートは「チェーンメール」に関して記載されているという事実に関してしか記載していない．しかし，このツイートはコスモ石油の公式の訂正発表の Web ページの URL を参照しているため,「コスモ石油の爆発で有害な雨が降る」という誤情報に対して訂正の立場である．よって，訂正情報のラベルを付与する．このように文章だけでは判断できないが，ツイートしたユーザの立場がわかる場合は誤情報または訂正情報のラベルを付与するようにした．

また，収集したツイートの中には「○○にあるコスモ石油のガソリンスタンが開いているそうです」のような，誤情報と直接関係のないものはその他のラベルを付与する．

手作業での分類はコストが大きいので，本研究ではクラスタリングを用いて，効率的にラベリングした．似た表現を用いたツイート群は，同一の主張である場合が多いので，まずツイート群を類似した文字列でクラスタリングした (この時点で誤情報・訂正情報・その他クラスタが多数生成される)．次に各クラスタ内に別の主張が混ざっていないかをチェックした (例えば誤情報クラスタ内に訂正情報のツイートが混ざっていたらクラスタを分割する)．最後に，各クラスタを誤情報・訂正情報・その他の 3 クラスタにマージした．全部で 5195 件のツイートを対象とし，2462 件の誤情報ツイート，2376 件の訂正情報ツイート，357 件のその他のツイートを同定した (表 3.1)．

3.4 技術的問題

前述したように，本研究では**誤情報と訂正情報の自動分類**の手法について考える．実際に

> 千葉市近辺に在住の方！　コスモ石油の爆発により有害物質が雲などに付着し、雨などといっしょに降るので外出の際は傘かカッパなどを持ち歩き、身体が雨に接触しないようにして下さい！

のような誤情報が拡散されたが，誤情報を検出するためには,「有害な雨が降る」という情報から「有害物質が雲に付着し，雨などといっしょに降る」という情報が同じであることを，コンピュータが自動的に理解する必要がある．このように，コンピュータが文章内の情報から誤情報だと同定するのは非常に困難なタスクである．しかし

> コスモ石油の有害物質の雨は **全くのデマ** です!! どんどん回してください！ まだこの事実を知らない人がいます

[1]実際には，被リツイート数が x 件以上のツイートのみを採用した．誤情報によって関連するツイート数が異なるため，閾値 x は誤情報毎に調整した．

のような訂正情報は「全くのデマ」などの誤情報を訂正している手がかりが存在するため，誤情報の同定よりは比較的容易に訂正情報を同定できる．以後，誤情報を訂正している表現のことを**訂正表現**と呼ぶ．関連研究 [1, 2, 4, 5, 6, 7] においても誤情報と訂正情報の分類に訂正表現を用いていたため，これら訂正表現を用いる方法は非常に有効だと期待できる．

第4章　誤情報の拡散収束過程の分析

前章で誤情報の収集，誤情報と訂正情報の分類には訂正表現を用いる方法が有効だと述べた．しかし，訂正表現を用いることはすなわち，訂正情報の出現が無ければ誤情報の自動収集，自動分類ができないことを意味する．訂正情報が出現してから誤情報を検出しても，後手になってしまい社会混乱を防ぐ役割が果たせない可能性がある．よってこの章では，訂正情報が出現した後に誤情報の検出しても実用性があるかどうかを調査する．

また，関連研究で提案している誤情報自動収集システムでは，誤情報を網羅的に集めるようになっているため．ユーザが大量に有る誤情報の中から社会混乱を引き起こしやすい誤情報を調べて探すという手間ができる．よって本章では，どのような条件が誤情報の収束に寄与しているのかを調査することで，社会混乱を引き起こしやすい誤情報を自動的に収集する手がかりについて述べる．

4.1　拡散収束過程の可視化

前章で用意した14件の分析データを用いて，

- 各誤情報の発生から訂正情報が出始めるまでの時間 (TTC: Time To Correction)
- 誤情報の数が減り始めるまでの時間 (TTS: Time To Suppress)
- 誤情報が発生してから収束するまでの時間 (TTE: Time To End)

を計測し，各計測時間を表4.1に示す．表4.1のコスモ石油やイソジンの例のように，長い時間拡散し続ける誤情報は社会に大きな損害をもたらす可能性がある．しかし，どのような誤情報についても誤情報の発生から最初の訂正情報が出始めるまでの時間は，概ね数時間である．よって，訂正情報をうまく捉えることで，誤情報の検出と注意喚起を行うことは十分可能であるといえる．

更に，誤情報の個別のケースを詳しく調べるため，誤情報の「拡散」ツイートと「訂正」ツイートの数を，それぞれ一定時間おきに折れ線グラフにプロットし，誤情報の拡散状況を可視化するシステムを開発した．このシステムでは，各時点でどのようなツイートが拡散していたのか，ツイート本文を閲覧できるようになっている．なお，グラフにプロットするツイートの数はリツイート数も考慮し，ツイート空間上での情報の拡散状況を表している．

4.2　4種類の拡散収束過程

14件の誤情報に対して，本システムで可視化したグラフを観察すると，誤情報の拡散状況は，主に訂正ツイートの量と収束までの時間で特徴づけられることが分かった．これらの2つの要素

表 4.1: 誤情報が拡散してから収束するまでの経緯

誤情報の内容	TTC(時間)	TTS(時間)	TTE(時間)
福島第一原発が核爆発の恐れ	0.0	0.5	1.5
サーバーラックが倒れて動けない	0.0	1.0	10.0
フジテレビの募金は日本ユニセフに行く	0.0	4.0	33.0
ポケモンクリエーターの田尻智が死去	0.5	0.0	2.5
東大が合格者の入学取り消し	0.5	1.5	9.5
阪神大震災では三時間後に最大の揺れが来る	0.5	2.0	13.5
支援物資の空中投下が認められていない	0.5	32.5	-
トルコが 100 億円支援	1.5	3.5	40
辻本補佐官が米軍の救助活動に抗議	2.0	0.0	5.0
ONEPEACE 作者尾田栄一郎が 15 億円寄付	2.0	1.0	35.0
阪神大震災でレイプが多発した	2.0	1.5	13.5
コスモ石油の爆発で有害な雨が降る	2.0	16.5	34.5
放射線対策にイソジン (うがい薬) が利く	9.5	32.5	74.0
天皇陛下が京都御所へ避難	12.0	0.0	12.5
平均	2.3	6.9	21.9

の組み合わせにより，大きく 4 種類の拡散収束状況に分類でき，14 件の誤情報の一部を分類したものを図 4.1 に示した．また，図 4.1 には，縦軸を訂正情報の量，横軸を誤情報が収束するまでの時間とし，各マスには各誤情報の拡散収束過程をグラフにプロットしたものを示した．この図の右上は訂正情報の量が少なく，長時間拡散したことになり，逆に図の左下は訂正情報の量が多く，短時間で収束することになる．以降では，図 4.1 を用いて，訂正情報の数や収束までの時間を決定づける要因について考察する．

4.2.1 訂正情報の量を決める要因

誤情報より訂正情報の量が少ない場合，訂正情報の信憑性・影響力が小さいことが考えられる．例えば「支援物資の空中投下は法律で認められていない」という誤情報において，「許可があれば可能」という訂正情報が流れたが，決定的な証拠や公式発表がなかった．このため，危機意識に駆られた人々が誤情報をどんどん拡散し，訂正情報が浸透しなかった可能性が高い．

逆に訂正情報の量が誤情報より多い場合，訂正情報の信憑性・影響力が強いことが考えられる．例えば「被災者の合格者が期限までに書類を提出できないと東大の入学が取り消される」という誤情報に対し，東大がウェブサイト上で「合格者本人の意志を確認せずに入学の資格を取り消すようなことはありません」と発表した．人々の不安を取り除くに十分な訂正情報だったため，誤情報よりも訂正情報の量が多くなり，誤情報を効果的に抑制することができた可能性が高い．

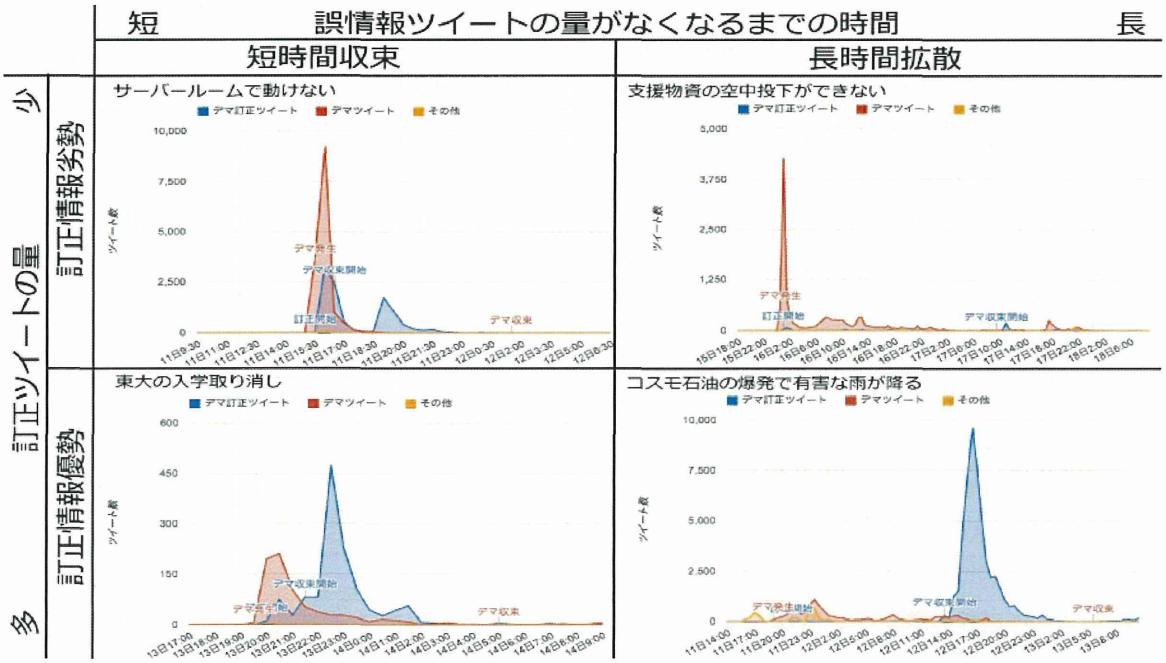

図 4.1: 4 種類に分けられる拡散状況

表 4.2: 分析データを 4 種類の拡散収束過程に分類した内訳

拡散収束過程の種類	誤情報の内容
訂正情報劣勢・長時間拡散	支援物資の空中投下が認められていない | 放射線予防にイソジン (うがい薬) が効く
訂正情報優勢・長時間収束	3時間後に最大の揺れが来る | サーバーラックが倒れて動けない | 辻本補佐官が米軍の救助活動に抗議
訂正情報劣勢・短時間収束	トルコが 100 億円支援 | ONEPEACE 作者尾田栄一郎が 15 億円寄付 | コスモ石油の爆発で有害な雨が降る | フジテレビの募金は日本ユニセフに行く
訂正情報優勢・短時間収束	阪神大震災でレイプ多発 | ポケモンクリエイターの田尻智が死去 | 東大の合格発表の入学取り消し | 天皇陛下が京都御所へ避難 | 福島第一原発が核爆発の恐れ

4.2.2 収束までの時間を決める要因

誤情報が収束するまでの時間が短いということは，有効な訂正情報が早期に現れたということである．例えば「サーバールームで身動きが取れない」という誤情報では，この情報の発信者がジョークとしてつぶやいていたことが早期に発覚したため，短期間で収束した．

逆に誤情報が収束するまでの時間が長い場合，有効な訂正情報の出現が遅いことが考えられる．例えば「コスモ石油の爆発で有害な雨が降る」という誤情報は，コスモ石油本社からの「そのような事実はない」という発表が遅れたため，長時間にわたり拡散した．

4.3 即座に拡散を止める必要がある誤情報の同定

政府や自治体，公式からの発表が，訂正情報の量や誤情報の収束までの時間を決めている要因であるとわかった．この事実から，「支援物資の空中投下が法律で認められていない」のような，訂

正情報の量が少なく，長時間拡散している誤情報は人々が誤情報を信じ続け，社会混乱を引き起こしやすく，直ちに拡散を止めるように促す必要があることがわかる．他にもどのような誤情報の拡散を止めるべきか分析するため，作成した 14 件の分析データを 4 種類に分類した結果を表 4.2 に示した．

表 4.2 より，訂正情報劣勢・長時間拡散の誤情報に「放射線予防にイソジンが効く」が存在した．この誤情報も，即座に拡散を止めるように促さなかったため，非常に大きな社会混乱になった．つまり，訂正情報の量や誤情報が拡散し続けている時間を監視することで，優先的に拡散を止めなければならない誤情報を提示することも十分可能である．

第5章　誤情報と訂正情報の自動分類

前章で，訂正情報の出現後に誤情報の拡散収束過程を提示しても，社会混乱を防ぐ支援になりうることがわかった．よって自動的に誤情報の拡散収束過程を提示することは大きな貢献となるはずである．これを達成するには

- 誤情報に関連するツイートの自動収集
- 誤情報と訂正情報の自動分類

が必要になる．前者については3.2節で述べた鍋島ら[5]の手法を用いることで可能となる．よって本研究では予め誤情報に関するツイートが収集されていると仮定し，後者の誤情報と訂正情報の自動分類の手法を提案・実験・評価する．

5.1　分類器の構築

訂正情報には「○○という情報はデマです」のように「デマ」や「風説」のような誤情報を訂正している**訂正表現**が含まれている可能性が高い．幾つかの関連研究[2, 5, 6, 7]は訂正表現の有無で訂正情報と誤情報の判別を行なっている．しかし，この方法では「誤情報」と「訂正情報」の分類にしか対応しておらず，誤情報とは無関係な「その他」のツイートを分類することができない．そこで，本研究では2.3節で構築した実験データを訓練事例として，最大エントロピー法を用いた3クラス分類モデルを学習した．

本研究では以下の素性を設計した．

- 訂正表現の有無 (T):

  本文中に「デマ」や「事実はありません」のような訂正表現が含まれていれば，訂正情報である可能性が大きい．本研究では，震災時のツイートから121個の訂正表現を手作業で収集したもの (付録A) を使用する．

- Bag of words (B):

  ツイートを形態素解析し，単語の表層系を素性とした．拡散したい情報がある場合，ユーザは情報をそのままコピー&ペーストする可能性が高い．よって拡散される情報内には，特定の単語 (「拡散希望」「コピペ」等) が用いられる傾向にあるため，誤情報や訂正情報に良く用いられる単語を素性にすることができると考えた．

- URLの有無 (U):

  訂正情報の中にはしばしば誤情報であるという根拠を提示するためにソースや理由となるwebページのURLを記載している場合がある．よってURLがツイート本文中にあれば訂正情報の可能性が高い考えられる．

- 拡散 (RT @) の有無 (R):

  「RT @」が文字列が含まれている場合，ツイートを拡散させようとしている，すなわち，広く知られるべき情報と考えられるため，誤情報か訂正情報である可能性が高い．

- 訂正表現周辺の単語 (TW):

  単に訂正表現の有無のみでは,「デマではありません」などの訂正表現を否定しているツイートのように，実際は誤情報であるツイートを訂正情報にしてしまう可能性がある．よって訂正表現の周辺単語を調べることにより，それらのツイートを正しく分類できることが期待できる．本研究では訂正表現の前後 5 単語を素性として加える．

- 訂正表現から誤情報キーワードまでの距離 (D):

  ある誤情報を訂正したい時は「(誤情報キーワード) についてはデマです」のように，定型的な言い回しが多い．よって誤情報に関するキーワードから訂正表現までの距離 (文字数) が小さければ，訂正情報である可能性が高い．ここで，誤情報に関するキーワードは，2.1 節でコーパスを作成した際に用いた検索クエリ (例えば「イソジン」と「うがい薬」) とする．

- 誤情報とツイートの類似度 (SU，SB):

  誤情報を説明する記述とツイート本文の類似度を素性にすることで，誤情報を支持するツイート認識をできると考えられる．本研究では，誤情報を説明する記述とツイート本文の単語ユニグラムと単語バイグラムのコサイン距離をもとに類似度を算出し，素性として用いる．(それぞれ SU，SB)

5.2 実験

5.2.1 実験設定

実験の機械学習では，訓練データとして 3.3 節で作成したデータ (表 3.1) を使用する．提案手法を評価するため，訓練データに含まれる 14 件の誤情報ごとに，学習データを 14 グループに分割し，交差検定を行う．つまり,「コスモ石油の爆発で有害な雨が降る」などの誤情報を評価データとして,「イソジンは放射線予防になる」などのその他の誤情報を学習データとして評価する．なお，評価のためにまず誤情報・訂正情報・その他の各クラスの True Positive，True Negative, False Positive，False Negative を計算する (表 5.1)．これらの値を用いて，各クラスにおいての分類性能を Precision, Recall, F1 スコアを以下の式 5.1, 式 5.2, 式 5.3 で算出する．

$$Precision_{class} = \frac{TruePositive_{class}}{TruePositive_{class} + FalsePositive_{class}} \tag{5.1}$$

$$Recall_{class} = \frac{TruePositive_{class}}{TruePositive_{class} + FalseNegative_{class}} \tag{5.2}$$

$$F1_{class} = \frac{2 * Precision_{class} * Recall_{class}}{Precision_{class} + Recall_{class}} \tag{5.3}$$

表 5.1: 分類したデータの分割表

	各クラスに属する	属さない
各クラスで有ると予測	True Positive	False Positive
そうでないと予測	False Negative	True Negative

これらの値を用いて三値分類器の性能を，Accuracy のマイクロ平均 (式5.4) で算出し，Precision，Recall，F1 のマクロ平均 (式 5.5, 式 5.6, 式 5.7) で算出する．

$$Accuracy = \frac{(TP_{誤情報} + TN_{誤情報}) + (TP_{訂正情報} + TN_{訂正情報}) + (TP_{その他} + TN_{その他})}{3 * AllTweet} \tag{5.4}$$

$$Precision = \frac{Precision_{誤情報} + Precision_{訂正情報} + Precision_{その他}}{3} \tag{5.5}$$

$$Recall = \frac{Recall_{誤情報} + Recall_{訂正情報} + Recall_{その他}}{3} \tag{5.6}$$

$$F1 = \frac{F1_{誤情報} + F1_{訂正情報} + F1_{その他}}{3} \tag{5.7}$$

誤情報と訂正情報を自動分類する最も単純な方法は，ツイート本文中に訂正表現が存在するかどうかで分類する方法である．よって，実験でのベースラインは，素性「訂正表現の有無 (T)」のみを使用した分類器の精度とする．

5.2.2 実験結果 1：素性セットによる性能の違い

提案した素性のうち，どの素性が有効に作用したかを調べるため，表 5.2 に素性の組み合わせ別の精度，適合率，再現率，F1 スコアを示した．提案手法の全ての素性を用いた時 (全 8 素性) の精度は 0.6562，マクロ F1 スコアは 0.5266 であった．訂正表現のみを素性に用いた場合 (ベースライン) の精度は 0.7578 で，全素性を用いた提案手法の性能の方が悪くなってしまった．この現象を調べたところ，Bag of words 素性が性能低下の原因となっていることが判り，これを除いた提案手法 (7 素性) の精度は 0.8125，マクロ F1 スコアは 0.5606 であった．Bag of words 素性を用いた時に性能が低下するのは，誤情報と関連が深い単語を分類器が丸暗記してしまうためだと考えられる．例えば,「コスモ石油の爆発で有害な雨が降る」という誤情報のデータで機械学習すると，「コスモ石油」や「有害」などの単語に重みづけてしまうので，この分類器で「イソジンを飲むと放射線予防に効く」に関連したツイートを分類することができないということである．

さらに，7 素性の設定から，残りの素性を削除している場合の結果を 6 素性として表 5.2 に載せた．6 素性の時，URL の有無を削除した分類器以外の分類器の精度は 0.81, F1 スコアは 0.56 と概ね良好であった．削除することにより性能が低下した素性は，分類に貢献したことになり，逆に上昇した素性は過学習を引き起こすなどして，分類の邪魔になっていったと考えられる．以上の事から，貢献していた素性は「URL の有無 (U)」で，貢献していなかった素性は「Bag of words」であることが明らかになった．

表 5.2: 提案手法の性能と素性セットによる性能の違い

素性		スコア			
		Accuracy	Precision	Recall	F1
ベースライン	T(訂正表現の有無)	0.7578	0.5337	0.5413	0.5204
全 8 素性		0.6562	0.5540	0.5333	0.5266
7 素性	除 B(Bag of words)	0.8125	0.5437	0.5816	0.5606
6 素性 (B を除く)	除 SB(バイグラム類似度)	0.8181	0.5480	0.5855	0.5644
	除 SU(ユニグラム類似度)	0.8120	0.5485	0.5808	0.5601
	除 D(誤情報と訂正表現との単語距離)	0.8169	0.5462	0.5848	0.5637
	除 TW(訂正表現の周辺単語)	0.8088	0.5437	0.5787	0.5578
	除 T(訂正表現の有無)	0.8094	0.5415	0.5793	0.5585
	除 R(RT の有無)	0.8092	0.5466	0.5789	0.5582
	除 U(URL の有無)	0.7870	0.5245	0.5634	0.5431

5.2.3 実験結果 2：有効な素性の働き

表 5.3 に，学習により高い重みが与えられたトップ 7 の素性と，低い重みが与えられたトップ 7 の素性を示した．素性セットとしては，Bag of Words を削除した 7 つの素性の分類器を用いた．最も高い重みが与えられた素性は．誤情報とツイートの類似度から誤情報を予測する素性で，重みが 3.88 であった．つまり誤情報に似ている文字列のツイートを，誤情報であると認識できていることになる．他にも得られたモデルからは，直観的に理解できるような重みが与えられていることが判る．例えば，「○○はデマじゃない」といった，ツイートに対しては，訂正表現周辺の単語 (TW) の素性によって，訂正情報を訂正している，つまり誤情報であると判断されやすくなる．また，URL が本文中に存在すると，何らかのソースや理由となる Web ページへのリンクの可能性が高く，訂正情報であり，誤情報ではないと判断されやすくなる．

5.2.4 実験結果 3：各クラスの分類性能

表 5.4 に誤情報・訂正情報・その他の各クラスについての適合率・再現率・F1 スコアを示した．誤情報分類の再現率と，訂正情報分類の適合率が下がっているが，それ以上に誤情報分類の適合率と訂正情報分類の再現率が高くなっているため，F1 スコアはベースライン手法よりも向上していることが分かる．よって提案手法で用いた素性は正しく特徴を捉えろことができ，より正確に誤情報と訂正情報の分類ができたと言える．しかし，その他クラスにおいては全く分類することができ無かった．これは，提案した素性は誤情報，または訂正情報の特徴を捉えた素性のみの為．その他の特徴を捉えることができなかったのが原因だと考えられる．

表 5.3: 分類器のモデル (Bag of Words を除く 7 素性)

ラベル			
誤情報		訂正情報	
重み	素性	重み	素性
2.73	誤情報とツイートの類似度 (SU)	1.25	訂正表現あり (T=True)
2.05	誤情報とツイートの類似度 (SB)	0.79	訂正表現周辺の単語 (TW[0]=デマ)
0.82	訂正表現なし (T=False)	0.69	URL あり (U=True)
0.71	拡散あり (R=Ture)		
-0.91	拡散なし (R=False)	-0.89	訂正表現なし (T=False)
-1.03	訂正表現あり (T=True)	-1.05	訂正表現周辺の単語 (TW[1]=じゃ)
		-1.22	訂正表現周辺の単語 (TW[2]=ない)
		-1.75	誤情報とツイートの類似度 (SU)

表 5.4: 各クラスの分類性能

クラス	素性	Precision	Recall	F1
誤情報	T(訂正情報の有無)	0.68	0.98	0.80
	7 素性 (除 Bag of Words)	0.79	0.93	0.85
訂正情報	T(訂正情報の有無)	0.92	0.64	0.76
	7 素性 (除 Bag of Words)	0.84	0.82	0.83
その他	T(訂正情報の有無)	0.0	0.0	0.0
	7 素性 (除 Bag of Words)	0.0	0.0	0.0

5.2.5 実験結果 4：訓練データ数による性能の違い

本研究では，実験設定でも述べたように,「コスモ石油の爆発で有害な雨が降る」のような，ある誤情報で機械学習を行い,「イソジンを飲むと放射線予防に効く」などの別の誤情報の分類の性能を測定した．誤情報の種類をさらに増やすことで，分類器の性能が向上するかどうか見積もるために，学習曲線を求めた．具体的には，14 種類の誤情報から，ランダムに 1 個をテストデータとして選び，残りの 13 個を訓練データとした．次に 13 個の誤情報から成る訓練データから，1 個ずつ訓練データをランダムに選ぶことで学習データの量を調整し，学習曲線をプロットした．ランダムにデータを選ぶという以上の試行を，学習曲線の形が安定するまで繰り返した．以上の方法で，表 5.2 で高い精度を示した 7 素性 (除 B(Bag of words)) を用いて，学習曲線をプロットした．

図 5.1 によると，精度の上昇傾向は緩やかに続いており，誤情報の種類を増やすことによって，精度のさらなる向上が期待できる．しかし，劇的な精度向上を達成するには，提案手法の問題点を明らかにする必要がある．

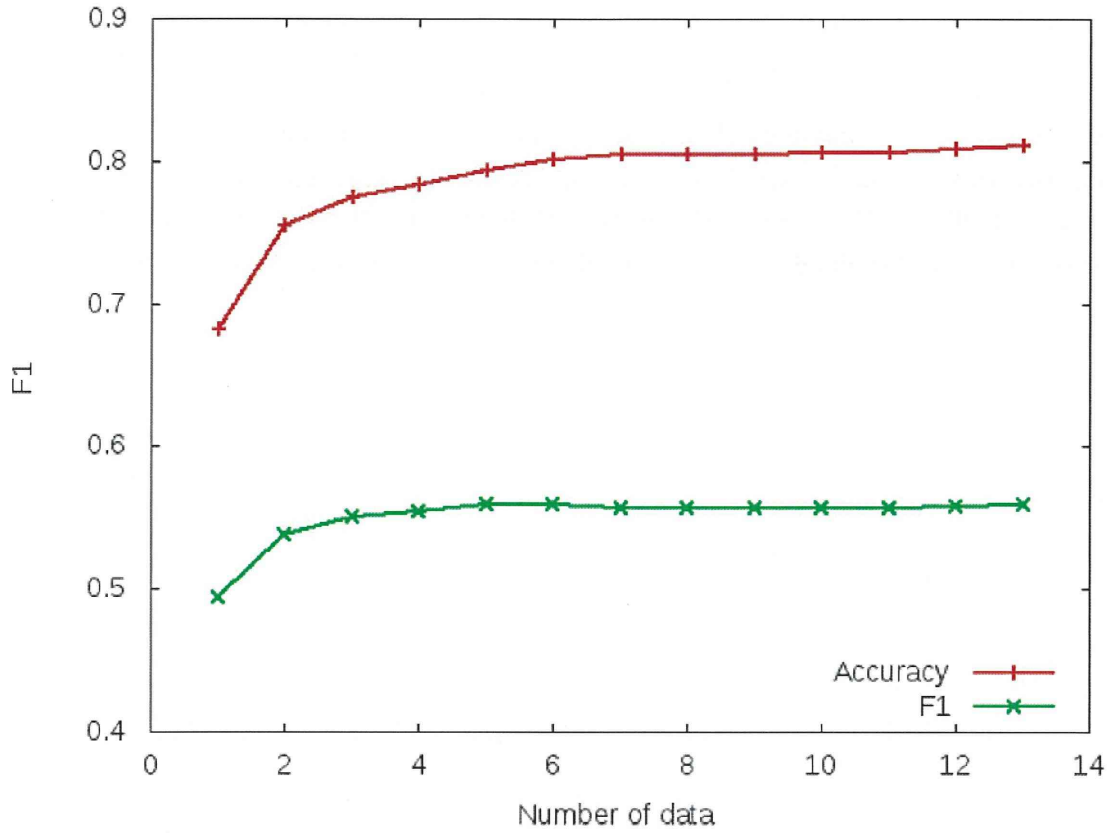

図 5.1: 学習曲線

5.3 考察

実験では，分類器の性能について様々な分析を行い，ベースライン手法である訂正表現の有無による分類よりも，精度の高い分類ができることを確認した．中でも「URL の有無」の素性は有効に働き，下のように訂正表現では分類しにくいツイートを正しく分類できた．

うがい薬「飲まないで」と専門家　買い求め客が急増 http://...

また，訂正表現周辺の単語を素性にすることで，「デマじゃない」のような訂正表現を否定するツイートを正しく認識できることを期待していた．

万が一原発から放射能が漏れ出した際，被爆しない為にイソジンを１５ cc 飲んでおいて下さい！原液です！ガセネタではありません．お医者さんからの情報です．これは RT ではないので信じてください！

しかし，コーパス内でこのような表現を用いたツイートが少ないため，学習がうまく行えなかった．但し，図 5.1 の学習曲線から判るように，訓練データの規模が大きくなると精度の向上が見られるため，学習データの量を増やすことで，有効な素性になると期待できる．

さらに，何の手がかりもないが，誤情報を訂正するツイートも存在する．

厚生労働省です不特定多数の方に送信されている，コスモ石油千葉製油所における火災関連のメールについては，厚生労働省からの発表情報ではありませんのでご留意願います

このツイートでは,「デマ」「嘘」などの訂正表現や，URL や RT は一切使われておらず，また誤情報の内容 (「コスモ石油の火災により有害物質の雨が降る」) も説明していないが，内容から誤情報を訂正するツイートであると判断できる．このようなツイートを訂正ツイートと認識するためには，深い処理 (例えば「火災関連のメール」を「火災により有害物質の雨が降るというチェーンメール」と解釈する) や，ツイートやユーザ間の関係 (例えば，厚生労働省はこの誤情報に関連して別のツイートで訂正表現を用いて打ち消しを行った，等の手がかり) を用いる必要がある．

第6章　おわりに

本稿では東日本大震災時の誤情報の拡散収束過程を可視化・分析することで，誤情報の自動収集・提示の実用性を検証した．分析の結果，誤情報の発生から訂正情報の出現までの時間は数時間程度で有ることがわかり，誤情報が拡散して早期に発見することが期待できるということがわかった．また，訂正情報の量や，誤情報が収束するまでの時間を観察することで，即座に止める必要がある誤情報を同定する方法を提案した．誤情報の拡散収束過程を自動で提示・分析できるようにすれば，誤情報の提示だけでなく様々な観点から捉える事ができ，何らかの重要性を提示することができるので，本研究ではさらに自動化を試みた．

本研究では，収集したツイートを誤情報と訂正情報に分類するためのコーパスを構築した．そのコーパスを用いて，最大エントロピー法を用いた教師あり学習を行なって自動分類する手法を提案した．実験の結果，従来の研究で用いられている訂正表現だけを用いた分類よりも，良い性能を示す分類器を作成できた．誤情報を説明する記述とツイート本文の類似度を計算することで，誤情報の特徴を捉えた素性を入れることができ，これらがうまく働いていることがわかった．また，Bag of words の素性は性能を下げ，ツイートに URL が存在することで，訂正情報の分類性能が向上している事が分かった．教師データの量を増やして分類の精度を計測し，教師データがを増やせば今後精度の向上が見込めることが分かった．

しかし，考察で述べたような訂正表現が存在する誤情報や，訂正表現を使っていない訂正情報の分類は非常に難しいため，文章の意味解析をしたり，URL 先の情報や他のユーザとの関連性などを利用していく必要がある．

今後の課題は分類器のより細かな分析を通じて，自動分類のさらなる性能向上を計ることである．また，別の災害や平常時など様々な環境下での検証も必要である．他にもシステムのリアルタイム化するにあたり，誤情報から適切な検索クエリの自動生成や，「誤情報」や「訂正情報」以外の「懐疑情報」や「検証情報」など，より細かい情報の分類に取り組む予定である．

謝 辞

本研究を進めるにあたり，ご指導を頂いた乾健太郎教授，岡崎直観准教授に感謝いたします．
貴重なデータを提供して頂いたTwitter Japan 株式会社に感謝いたします．
日常の議論を通じて多くの知識や示唆を頂いた乾・岡崎研究室の皆様に感謝致します．

参考文献

[1] 宮部 真衣, 梅島 彩奈, 灘本 明代, 荒牧 英治, マイクロブログにおける流言の特徴分析, 情報処理学会論文誌 Vol54, 2013.

[2] 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 発表論文集, 2012.

[3] 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, 日本データベース学会 第10回年次大会 発表論文集, 2012.

[4] 梅島 彩奈, 宮部 真衣, 灘本 明代, 荒牧 英治, 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, 2011.

[5] 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 発表論文集, 2013.

[6] 白井 崇士, 榊 剛史, 鳥海 不二夫, 篠田 孝祐, 風間 一洋, 野田 五十樹, 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, 人口知能学会 第26回全国大会 発表論文集, 2012.

[7] 鳥海 不二夫, 篠田 孝裕, 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, 情報処理学会デジタルプラクティス Vol3 特集号投稿論文, 2012.

[8] Carlos Castillo, Marcelo Mendoza, Bardara Poblete, Insormation Crebidility on Twitter, International World Wide Web Conference, 2011

[9] 萩上チキ, 検証 東日本大震災の流言・デマ, 光文社, 2011.

[10] 情報支援プロボノ・プラットフォーム (iSPP), 3.11 被災地の証言 東日本大震災 情報行動調査で検証するデジタル大国・日本の盲点, インプレスジャパン, 2012.

付 録A　訂正表現

表 A.1: 訂正表現

訂正表現		
あらず	いたづらツイート	うそらしいよ
かえって危険	ことはない	このような事実はありません
これを覆し	ございません	そのようなことはない
そのような事実はない	そのような事実は無い	そのような心配はありません
そのような心配はありません	そんなことないでしょう	そんなことはありません
ただの噂	ていません	てないそう
でまかせ	とんでもない間違い	なんか変
なんてなかった	ねつ造	はないそう
はなかった	まずない	まちがいのようです
ような事はないとの事	よくない	アカン
ウソ	エライ違い	ガセ
ソースが明確でない	ソースのない	チェーンメール
チェンメ	ツイッターで出回ってるような事はない	ツイッターで出回ってるような事はないそうです
デマ	トバシ記事	ニセ情報
バカ	ホントかいな	ホンマかいな
メール連鎖	悪戯	意味ない
意味なし	違います	嘘
嘘です	噂	何の情報源もない噂
怪しい話	確証はない	間違い
間違った	間違ってる	関知していない
危険かつ悪質	起こりません	偽ツイッター
偽情報	逆に危険	逆効果
虚偽	虚偽のチェーンメール	虚偽情報
虚報であった	誤	誤った情報
誤り	誤解	誤情報
誤報	誤報が拡散	誤報ではないかとの情報があります
公式に否定	効果がない	効果なく
効果なし	効果はありません	効果はない
根拠なし	根拠のない情報	事実じゃない
事実でないと否定	事実ではありません	事実はありません
事実はない	出所不明	出来ない
情報がめちゃめちゃ	情報がめちゃめちゃだ	状況ではない
信じるな	真偽	真逆の情報
正式に否定	全くありません	全くそういう事実はない
全くのデマ	全くの誤報	全くの事実無根
大丈夫なようです	注意喚起	訂正
否定	不確定情報	不思議
不当非難	風説	変なチェーンメール
報道のニュアンスは誤り	報道の悪意	本当??
本当に？	無視	裏付けなし
連鎖メール		