安否情報ツイートコーパスの詳細分析とアノテーションに関する一考察

村上 浩司　　萩原 正人

楽天株式会社 楽天技術研究所
{koji.murakami, masato.hagiwara}@mail.rakuten.com

1 ANPI_NLP プロジェクト

2011年3月11日午後2時46分，東北地方を中心としてM9の大きな地震が襲い，多くの命が失われた．電気，通信施設や機器などの被害も甚大であった中，インターネットの通信網は比較的被害が少なく，PCおよびモバイル端末からのインターネットへの情報提供が活性化した．その結果SNSやTwitterなどを介して，被災者の安否，被害規模や避難所の状況，必要な物資についてなど，極めて重要な情報の伝達が行われた．しかしながらこれらの媒体を通して流された災害に関する情報量は膨大であり，ある特定の人の安否情報を確認するといった，個人レベルで必要な情報を得ることは非常に難しかった．このような状況の元，自然言語処理技術を使って自分らが今現時点で何ができるかを考え，行動したのがANPI_NLPプロジェクト[1]の始まりである．我々は被災地から発信される個人の安否情報ツイートからの情報の集約と整理に焦点を当てて，ツイートを対象とした“トピック分類”および“人名，組織名に特化した固有表現抽出”タスクに取り組んだ[3]．ツイートには安否情報以外にも重要な情報が多く含まれていたが，ANPI_NLPではあまり注目してこなかった．

そこで本論文では，ANPI_NLPプロジェクトで構築したトピックタグ付きのコーパス（安否情報ツイートコーパス）に改めて着目し，コーパスの詳細分析を行う．またコーパス中の有用な情報を抽出，提供するために，安否情報抽出以外に自然言語処理が取り組むことのできるタスクについて考察する．

2 安否情報ツイートコーパス

このコーパスはTwitterのAPIを用いて収集された，震災に関すると思われるハッシュタグ“#anpi”“#hinan”，“#j_j_helpme”，“#save_[地名]”を含む61,375ツイートから構成される．ツイートが収集されたのは3月11日未明から3月14日未明までの期間である．ANPI_NLPではこのコーパスを100分割し，タグ付けを行った．作業は最終的に65名以上の有志により行われ，最終的に33,242ツイートに対してタグが付与された．

[1]http://trans-aid.jp/ANPI_NLP

コーパスは前節で述べたように，ツイートに記載される安否情報の種類を特定する目的で作成された．表1で示す9種類のタグを付与した．災害時に即時性を重視して自然言語処理技術が応用された研究にはLewisらのハイチ地震での翻訳[2]などがある．また，災害時のツイートへの固有表現のタグ付けを行った研究[1]もある．

3 時間軸からの分析

まず，地震発生直後から数日間の間にTwitterにおいて何が起こったのかを正確に把握するために，ツイートを時間軸の観点から分析した．

なお，当時収集した安否情報ツイートコーパスの各ツイートにはタイムスタンプは含まれていないため，今回時間軸分析のためにそれらを復元した．なお，Twitterのユーザーもしくはツイートそのものが削除された等の理由により，アクセスできないツイートに関しては，ツイートID（全てのツイートに対する通し番号であり，ほぼ時間順に番号がついている）を基に時間的に近いツイートから線形補間した．

まず，安否情報ツイートコーパス中に安否情報タグが付けられているツイートのみを対象とし，震災発生直後から3日の間に，1時間ごとのツイート数がどのように推移したかを調べた．それを図1(上)に示す．ツイート数は震災直後から増え始め，3月13日21時にピークを迎えている．本コーパスはハッシュタグを用いて収集しているため，必ずしもツイートの総数を表しているとは言えないものの，この時間帯に震災関連情報が活発にやりとりされていることが分かる．

ツイートの内容をさらに詳細に調査するため，各安否情報タグの付けられたツイート数の全体に占める割合がどのように推移したかを図1(下)に示した．まず分かるのは，地震発生直後数時間内における「O-その他」の多さであり，この時点での典型的なツイートは「停電中.揺れがまだ続いてる.」「釜石の状況を教えてください」など，情報の錯綜する中，震災もしくは地域の総合的な情報を提供・要請するものである．この時点でツイートに出現する固有名詞も地名がほとんどである（図1上）．

一夜明けた3月12日の朝から，行方不明者や地域の

表 1: ツイートに付与した安否情報タグ

ラベル	定義	例	数
I-本人	私が本人である	私は大丈夫です.	405
L-生存	誰かの生存情報を入手	探していた日本太郎さんの無事が確認されました.	1,154
P-死亡	誰かが死亡情報を入手	—	93
M-不明	誰かの安否情報を探索	釜石市にいる日本太郎さんと連絡がつきません.	4,438
H-救助	救助要請	**の農協の隣の民家に 10 人ほど取り残されています. 救助を！	280
S-要請	不特定の個人の安否情報又は地域の情報を探索	岩手県陸前高田市に住む大学時代の友人の安否が全く分からない. 是非とも大槌町，避難中の 4,600 名の氏名が，早く知りたいです.	1,903
O-その他	非安否情報	横手市山内地域の停電がほぼ全域復旧.	24,035
R-リンク	安否情報等の外部リンク	南三陸町の情報はこちら！ http://...	773
U-外国語	判断が困難，又は外国語	Información oficial:la explosión #nuclear ocurrir muy pronto..	1, 235

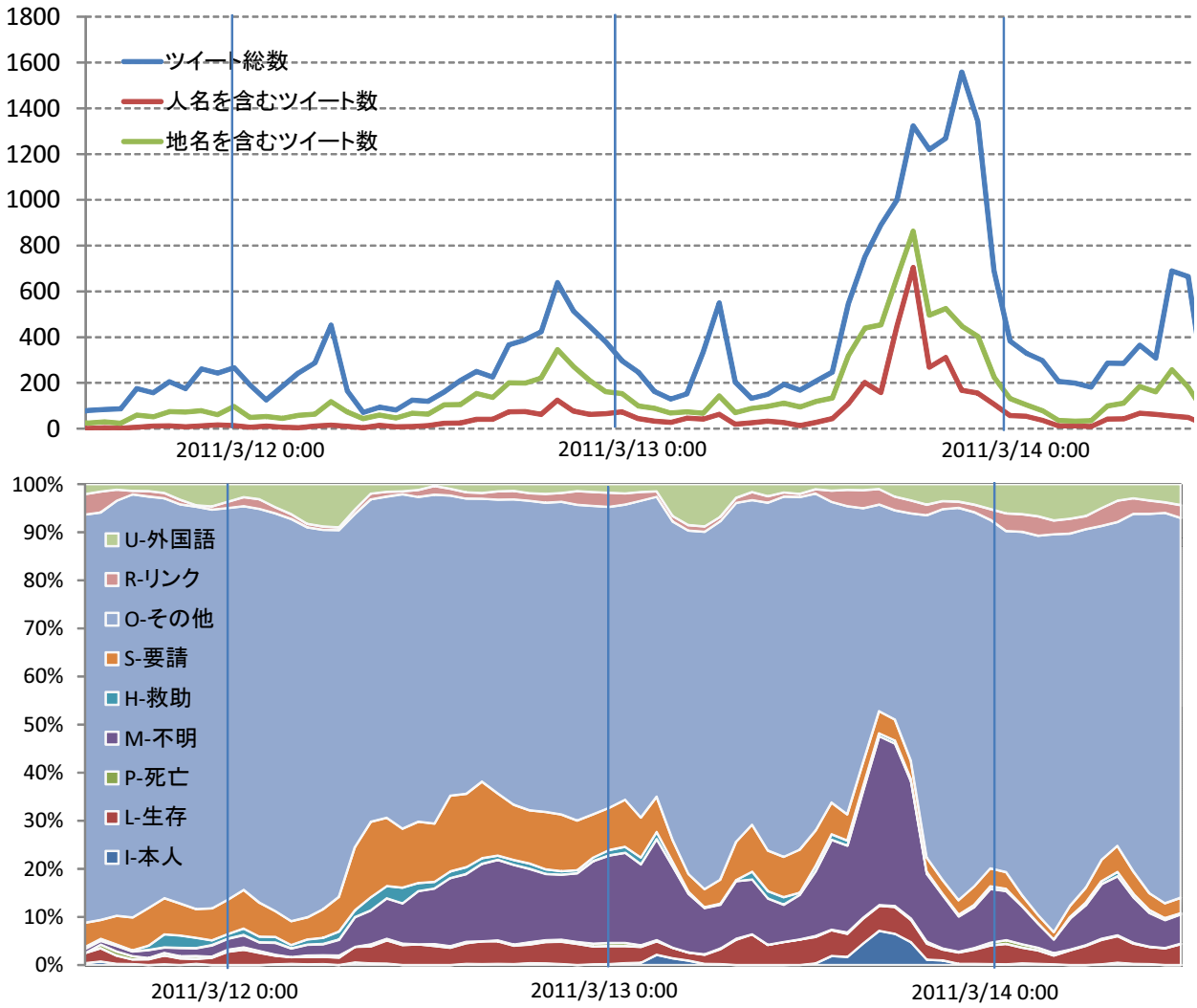

図 1: ツイート数およびトピックの時間遷移

安否情報を求めるツイートの割合が急増している.「気仙沼，市街地の状況なんでもいいので教えてください.」というような広い範囲の情報を求めるツイートに加え,「多くの親類，友人と連絡がついていません.」など被災者の安否に関する情報要請が増えている．このような個人名を特定しない安否情報要請ツイートは，次第に個人名・住所を明記したツイートに取って代わられており，そのことは図 1(上) の人名を含むツイート数がじわじわと増加していることからも分かる．なお，ほとんどの「H-救助要請」のツイートが地震発生翌日の朝までに集中している．このようなツイートをいかに拾えるかが人命救助には非常に重要であり，震災発生直後から数時間という初動の速さを達成するためにも平常時からの準備が不可欠であると考えられる．

地震発生から 2 日が経過した 3 月 13 日は，個人の安否情報が最も活発にやりとりされた日である．このことは半数以上のツイートが人名・地名，もしくはその両方を含んでいることからも分かり，固有表現のツイートからの適切な抽出がいかに重要かを物語っている．また，実際に ANPI_NLP のプロジェクトがスタートしたのはその 1 日後の 3 月 14 日であり，その時点では既に 1，2 日以上遅れてツイートにタグ付けの作業を開始していたことになる．このように，ツイートからの安否情報の抽出タスク自体が初めての試みとはいえ，いかに迅速に情報抽出システムを構築するか，は今後の課題であろう．

なお，震災発生後の話題遷移に関する研究にはソーシャルメディアの大規模コーパスを用いて調査した研究 [6] やニュース・ブログの相関を調査した研究 [11] がある．

4 非安否情報ツイートに焦点を当てた分析

ANPI_NLP プロジェクトでは，トピック分類と人名，地名に特化した固有表現抽出に焦点を当てたため，構築したコーパス中のタグのうち，一部のタグ（I/L/P）を主に利用した．しかしながらコーパス中のツイートには，救助要請（“M”）や被災地の状況，道路等の交通機関，生活インフラの復旧状況，避難所での炊き出しや給水などの情報も多く含まれており，救助活動や救援物資の配分などの検討に極めて重要な役割を果たす．そこで，タグが付与されたツイートの多くが属する非安否情報（“O”）に着目し，どのような情報を含んだツイートがあるかかを調査した．まず，前節で述べた時間情報補完を行った非安否情報ツイートを 1 時間毎にグループ化した．次に各時間毎のツイート数の比率を保持するよう各グループからランダムに，合計 1,000 ツイートを抽出した後，述べられている情報の種類に着目して分類した．

分類結果を表 2 に示す．本来ならば “U” タグが付く非日本語ツイート，“P” タグとなる安否情報なども多く存在する．これはタグ付け作業は短時間，かつ大人数で同時並行で行ったこともあり，タグの定義が作業者の間で統一できなかった事が一つの要因と考えられる．Twitter

表 2: 非安否情報ツイートに含まれる情報の分類

ツイートの種類	数	内訳
情報提供	479	（詳細に分析）
非日本語ツイート	254	英語:167，その他の言語:87
要請	28	情報:20，物資:5, その他:3
安否情報	9	個人名など
ハッシュタグの告知	32	
その他	198	

表 3: 情報提供ツイートの分類と例

種類	数	詳細
インフラ関連	116	電気，水道，ガスなど
避難・避難所	56	炊き出し，給水情報など
津波情報	50	注意勧告，津波，火災状況など
原発関連	26	原子炉，内部被曝，対応など
救援活動	22	自衛隊の活動，病院情報
交通機関	20	電車，道路，バス
地震情報	7	注意勧告，予測情報など
サイトへのリンク	120	情報まとめサイトの紹介など
首都圏帰宅困難	9	施設の紹介，食事の情報など
安否確認	17	個人名以外の安否の情報
東京節電	36	計画停電の情報

特有のハッシュタグについてのツイートも一定数存在する．これは，情報の集約が容易にできるよう，様々なハッシュタグが提案されたり，また Twitter 社からオフィシャルの震災に関するハッシュタグが推奨され，それらの情報が広くアナウンスされたものである．分類の結果，最も数が多かった “情報提供” についての 479 ツイートを更に詳細に分類した．結果を表 3 に示す．

災害時には様々な情報が飛び交うことから，有用な情報の分類 [4] だけではなく，情報の種類によって集約，整理，共有することで情報を必要としている側にとって，容易にアクセスすることが可能になる．例えば被災地までの道路状況と生活インフラの情報がそれぞれ構造化されていれば，それらを組み合わせることでインフラ復旧や被災者への救援物資の運搬などに極めて重要な情報となる．また，質問応答システム [8] の情報源として利用することで精度の向上が期待できる．情報の構造化は情報抽出が必要となる．以下に生活インフラ，交通機関，救援活動に関するツイートからの情報抽出について考える．またツイートには伝聞や参照などの情報も多いことからそうした情報を分類する必要がある．情報の信頼性という観点からも課題があると考えられる．

生活インフラ・交通機関の状況情報の抽出

道路状況，電車やバスの運行状況の情報を抽出する．道路情報に関しては石野らも着目 [5] し，危険情報の抽出に利用している．ツイートから例えば道路名と区間，道路状況などが以下のように抽出できれば救援物資の運搬などにも役立つ．

始点	終点	道路名	状況
–	–	45 号線	通れない
花巻空港 IC	東和 IC	釜石自動車道	通行再開

電気，ガス，水道などの生活インフラについての情報

も，以下のように地域名，インフラ名とその状況が次のように集約できると被害・復旧状況が把握できる．

地域名	インフラ	状況
宮城・岩手	電話	不通
横手市内山内	電気	停電復旧
仙台市内	ガス	漏れ報告多数

救援活動ツイートからの病院の診断状況の抽出

救助活動には医療機関の状況が不可欠である．交通機関の情報と組み合わせることで搬入先の検討に役立つ．次の例のような，受け入れだけでなくその対象（重症のみ，透析など）の抽出も必要である．

地域名	病院	受入	対象
仙台市内	私立病院	可能	軽傷者
いわき市	慈愛病院	不可能	新規受け入れなし
宮城野区	宮城野分院	可能	軽傷のみ

情報要請ツイートからの要請内容の抽出

ツイートが，表 1 中の “H” タグのような救助要請だけではなく，表 2 のように何らかの情報や，物資，人材などを求めている場合もある．救助要請の抽出には例えば [9] などがある．情報の要請の場合，“教えてください” や “求む”，“お願いします” といった表現で記述されることが多い．被災地の状況を把握するために以下のように抽出できれば重要な情報となる．

地域名	対象	条件，補足情報
郡山方面	49 号	通行可能
いわき市	ガソリンスタンド	朝から営業
福島駅周辺	場所	処方箋扱う

ツイートの情報信頼性判定

個人からの情報発信であるツイートは必ずしも報道情報だけではなく，人から伝え聞いた情報や引用，憶測なども含まれる．以下のような事実であるかどうかが不明であるツイートは数多く含まれる．

(1) 私が通っていた仙台市六郷幼稚園，石巻門脇小学校は津波に飲まれたんだろうな．

(2) 鹿折にある東陵高校に避難者がいる みたい です．

(3) 昨日多賀城にいた方の 話だと，産業道路沿いに多数の遺体が放置されていた ようだ とのこと，…

こうした情報は情報の集約・整理という観点から考えると，情報の信頼度が低いと判断するべきである．文のテンス，アスペクト，モダリティなどの解析（例えば [7] など）を利用して情報の信頼性を取り扱う必要がある．根拠や条件などの抽出 [12] も情報の信頼性を扱うために重要であるまた，誤った情報の拡散を防ぐための流言訂正ツイートの抽出 [10] なども有用である．

5 おわりに

本稿では，ANPI_NLP プロジェクトで用いた安否情報ツイートコーパスに改めて焦点を当て，時間軸からの分析と，非安否情報ツイートの分析を行った．災害時に溢れる様々な情報から有用な情報を抽出することは，被害状況の把握や救助活動にとって有用であり渇望されている技術である．平常時から災害時を想定した言語処理課題に取り組んでいくことが必要ではないだろうか．

謝辞

ANPI_NLP プロジェクトではコーパスのタグ付けには 65 名以上の方々に参加頂いた．心より深謝する．

参考文献

[1] William J. Corvey, Sarah Vieweg, Travis Rood, and Martha Palmer. Twitter in mass emergency: What NLP can contribute. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, June 2010.

[2] William D. Lewis. Haitian Creole: how to build and ship an MT engine from scratch in 4 days, 17 hours, & 30 minutes. In *14th Annual Conference of the European Association for Machine Translation*, 2010.

[3] Graham Neubig, Yuichiroh Matsubayashi, Masato Hagiwara, and Koji Murakami. Safety information mining -what can NLP do in a disaster-. In *International Joint Conference on Natural Language Processing 2011*, p. (to appear), November 2011.

[4] Graham Neubig, 森信介. 能動学習による効率的な情報フィルタリング. 言語処理学会第 18 回年次大会, 2012.

[5] 石野亜耶, 小田原周平, 難波英嗣, 竹澤寿幸. Twitter からの被災時の行動経路の自動抽出および可視化. 言語処理学会第 18 回年次大会, 2012.

[6] 榊剛史, 鳥海不二夫, 篠田孝祐, 風間一洋, 栗原聡, 野田五十樹, 丸井淳己, 松尾豊. 大規模災害時におけるソーシャルメディアの変化. 言語処理学会第 18 回年次大会, 2012.

[7] 松吉俊, 江口萌, 佐尾ちとせ, 村上浩司, 乾健太郎, 松本裕治. テキスト情報分析のための判断情報アノテーション. 電子情報通信学会論文誌 D, Vol. J93-D, No. 6, pp. 705–713, 2010.

[8] 風間淳一, Stijn De Saeger, 鳥澤健太郎, 後藤淳, Istvan Varga. 災害時情報への質問応答システムの適用の試み. 言語処理学会第 18 回年次大会, 2012.

[9] 相田慎, 新堂安孝, 内山将夫. 「東日本大震災関連の救助要請情報抽出サイト」構築と救助活動について. 言語処理学会第 18 回年次大会, 2012.

[10] 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.

[11] 小池大地, 横本大輔, 牧田健作, 鈴木浩子, 宇津呂武仁, 河田容英, 吉岡真治, 福原知宏. 震災を題材としたニュース・ブログ間の話題の相関と遷移の分析. 言語処理学会第 18 回年次大会, 2012.

[12] 岡崎直観, 成澤克麻, 乾健太郎. Web 文書からの人の安全・危険に関わる情報の抽出. 言語処理学会第 18 回年次大会, 2012.