ANPI_NLP

― NLP 技術を応用した災害時安否情報確認支援―

応用

萩原 正人[1]　村上 浩司[1]　Graham Neubig[2]　松林 優一郎[3]
[1] 楽天技術研究所　[2] 京都大学　[3] 国立情報学研究所

東日本大震災および安否情報の氾濫

2011 年 3 月 11 日，東北宮城県沖を震源とした M9.0 の東日本大震災が発生し，多くの死傷者，大津波による破壊など，甚大な被害をもたらした．警察発表の地震・津波による死者・行方不明者数（図 -1）は 3 月 25 日にピークを示しており，犠牲者の最大数を把握するまでに地震発生後 2 週間を要していることが分かる．被害規模の全容を把握するのにこれだけ時間がかかったのは，地方自治体の庁舎等が被害を受けたこと，および，通信機器の損傷，通話規制等により，必要な情報が断絶されたことが理由として挙げられる☆1．

一方で，震災による影響を比較的受けにくいインターネット上では，被災者の安否情報，避難所・必要物資など震災関連の情報が活発にやりとりされ，被災地からの情報伝達に重要な役割を果たした．特に Twitter 等のマイクロブログサービスおよび mixi などのソーシャルネットワークサービス（SNS）上には，震災直後から大量の震災関連情報が溢れた．

しかしながら，自然言語で記述された未整理の情報が短期間の間に大量に溢れたため，個人にとってこれだけ多くの情報を適切に処理するのは難しく，安否情報をいかにマイニング・整理・構造化し，それらを必要としている人に届けるかが特に重要な課題となった．

☆1 被災地，進まぬ安否確認　役場も交番も水没・情報途絶 http://www.asahi.com/special/10005/TKY201103120549.html

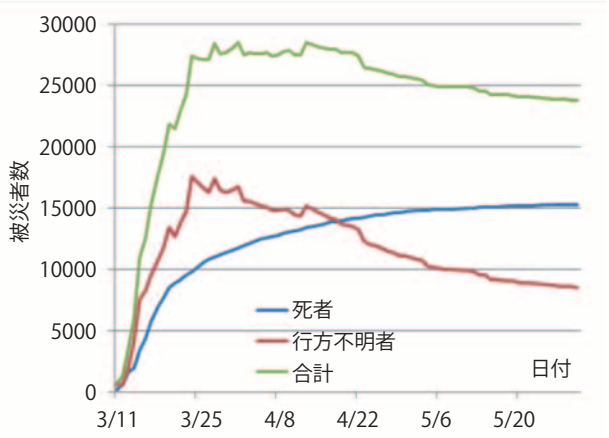

図 -1　東日本大震災の死者・行方不明者の推移

ANPI_NLP プロジェクトの発足

こうした背景の中，震災発生後ほどなくして「震災時に自然言語処理研究者として何ができるか考えてみよう」という趣旨のツイート（Twitter 上の短い投稿メッセージ）の呼びかけに応じて，筆者らを含む日本全国の自然言語処理研究者・技術者・学生らが中心となり「東日本大震災のためのデータマイニング・言語処理 ANPI_NLP」プロジェクトが自発的に発足した．本プロジェクトの目的は，自然言語処理（NLP）の技術を活用して震災関連情報を整理し，情報を必要としている人々に適切な形で提供することである．この呼びかけに応じた参加者によって，最初のツイートから 1 時間も経たないうちに，言語資源等の開発・提供が始まった．本プロジェクトの参加者らは，インターネット上のコミュニティの特徴である，ゆるやかなつながり，中央集権的な管理の不在，Twitter 等のネット上でのオープンな

コミュニケーション形態を最大限に活用しながら，所属する組織の垣根を越えて協力し，各自の専門性・知識・資源を動員して共同で震災関連情報の整理に取り組んだ．本プロジェクト，特に後述の安否情報抽出タスクは，震災後から 1 週間後の 3 月 18 日にかけて集中的に活動が行われた．

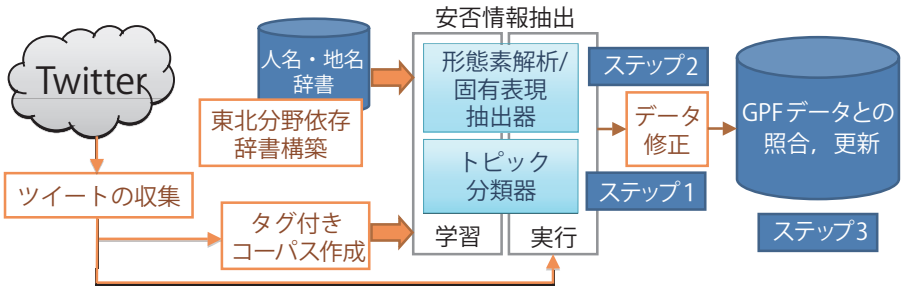

図 -2　ツイートからの安否情報抽出とデータ照合の流れ

初期の参加者らに加え，言語資源・分野適応など各分野の専門家らが主に先頭に立ち，プロジェクトを推進していった．たとえば，各ツイートに対するタグ付けは，経験者がタグ仕様等をまず Twitter 上で共有し，参加者からのフィードバックにより素早く改善することにより進行した．発足後まもなく，Wiki ページ☆2 も有志によって準備され，プロジェクト関連情報の集約・共有の重要なハブとなった．

ツイートからの安否情報抽出

本プロジェクト内では，後述するように複数のサブプロジェクトが自発的に立ち上がった．本稿では，その中でも特に重要性・緊急性の高い，被災者の安否情報を Twitter からマイニングするタスクに焦点を当て，安否情報ツイートの分類，分野適用，形態素解析，固有表現の処理などの自然言語処理分野から見た技術的課題およびその解決について述べる☆3．また最後に，有事の際の教訓として，質を確保しつついかに迅速に共同開発を進めるか，専門家が協調できる体制をいかにして構築するか等，本プロジェクトを通じて得られた知見についても述べる．

ここで紹介する安否情報マイニングタスクは，

1. 安否情報を含むツイートとその他大量の無関係のツイートを分類すること
2. ツイートなど特殊なテキストから人名・地名など重要な情報を抽出すること
3. 抽出された情報を検証し，それを必要とする人々に届けること

のステップから主に構成される（図 -2）．この時点で，Google によって提供されている安否情報データベース Google Person Finder（GPF）☆4 が，マスメディア等の助けもあり，広く認知されていた．そこで，我々は，Twitter から抽出した安否情報を最終的に GPF に登録することによってそのような情報を必要としている人に届けることを目標とした．情報の正確さはもちろんのこと，タスクを完了するスピートも重要である．また，インターネット上の言語現象という「くだけた」表現に対応できる言語資源の構築，NLP ツールの構築，GPF への照合および更新についても技術的課題がある．これらについて次章以降解説する．

言語資源およびツイートコーパス

プロジェクトの初期段階の主な活動は，有志の間での言語資源の共有であった．有志により共有された言語資源には，ツール等の開発者・管理者によって共有された既存の辞書や，人手によって集められた分野依存の辞書等がある．主なものを次に挙げる：

☆2 http://trans-aid.jp/ANPI_NLP/
☆3 ここでは普及度という点から Twitter に限ったが，本タスクの知見は SNS 等で共有されているあらゆる安否情報に適用可能である．
☆4 http://japan.person-finder.appspot.com/

トピック	定義	例	ツイート数
I	私が本人である	○○市の××です．無事です．	405
L	誰かが生きているという情報を入手した	○○市の××さんは△△避難所にいらっしゃるようです．無事です．	1,154
P	誰かが亡くなっているという情報を入手した		93
M	誰かの安否情報を探している	大船渡町 XXに住んでいた XXXXさんの安否が確認できずにおります．	4,438
H	救助要請	誰々がどこどこに取り残されています	280
S	特定できない個人の安否情報（探している，判明している）	いわき市南部に下宿していた親戚の安否が分からない	1,903
	もしくは，地域の広い意味での情報を探している	○○小学校に避難されている方々に物資は足りていますか？	
O	安否情報ではない	被災者の安否について情報提供できる方はこちらのサイトで！	24,035
R	安否情報・人名リスト等の外部リンクがある	○○市の安否確認リストです http://～	773
U	安否情報かどうか判断がつかない／日本語以外の言語		1,235

表 -1　ツイートに対するトピック分類

- オープンソースの仮名漢字変換システム Mozc の辞書（26,514 の姓，50,848 の名を収録．開発者により提供）
- Web から収集した東北地方に特有な姓の辞書
- 郵便番号辞書をダウンロードし整形した岩手・宮城・福島・茨城の各県の地名辞書

ほかにも有志の手によって，日本のすべての駅名，Wikipedia から抽出した関東地方・東北地方のランドマーク名などの辞書が提供されたが，本タスクにおいて主に使用されたのはプロジェクト初期に提供された上記 3 つの資源であった．将来的にも，なるべく数多くの資源が，災害関連プロジェクトに精通した人々の手に素早く渡るようにする体制作りが必要不可欠であろう．

次の段階として，本タスクは Twitter からの情報抽出を対象としているため，震災に関係するツイートを収集するプロセスに進んだ．具体的には，「#anpi」「#hinan」「#j_j_helpme」「#save_［地名］」などのハッシュタグを含んだ 61,376 ツイートを Twitter から収集した☆5．安否情報に関連する典型的なツイートと，その具体的な解析例を示す：

☆5 RT（リツイート，再ツイート）や QT（引用ツイート）は削除した．

石巻市合戦谷の田中太郎，次郎，花子さんについての情報を！

本タスクではまず，ツイートのトピックすなわち記述されている安否情報の種類を認識する必要がある．上の例の場合，「不明者の情報を求めている」ツイートであると分類できる．ツイートに関する 9 種類のトピックを表 -1 のように定義した．

さらに，ツイート中に現れる固有表現（Named Entity ; NE）の識別も必要不可欠である．固有表現とは，人名・地名などの固有名詞，数値・時間表現の総称であり，上記のツイートの場合，対象となっている人名が「田中太郎」「次郎」「花子」であり，住んでいる地名が「石巻市合戦谷」という情報を認識する必要がある．

言語資源に対し抽出したい情報の正解を付与する作業を「タグ付け」と呼ぶ．上記のツイート例に対してタグ付けをすると，以下のようになる：

<location> 石 巻 市 </location><location> 合 戦 谷 </location> の <person type="M"> 田 中 太 郎 </person>，<person type="M"> 次郎 </person>，<person type="M"> 花子 </person> さんについての情報を！

人名タグには，前記のように安否情報の対象であることを示すタグ（type）を付けられる．後述する解析器および分類器の教師情報として用いるため，コーパスの収集が終わり次第，すぐに有志によりコーパスのタグ付けが始まった．タグ付け作業の話題は Twitter を媒介として瞬く間に広がり，65 人以上の有志がそれぞれコーパスを 100 等分して分担し，結果として，33,242 ツイートのトピックおよび NE タグ付けが 2 日もかからずに完了した．

今回タグ付けに参加した有志は，Amazon Mechanical Turk のようにクラウドソーシングによってタグ付けする場合に比べて総じて知識・経験ともに高いといえるが，プロジェクトの時間制約上，コーパス中のタグ付け基準の一貫性を保つ上で課題が残った．これを回避するために，まずテストラン的にタグの定義を公開した後改善する，少数の専門家が Twitter 上で質問を受け付け，Wiki 上のタグ付けガイドラインを更新するなどの方法により統制を図った．

教訓として，タグ付けの経験の少ない参加者が，難しいケースを「パス」できるような仕組みづくりがあれば，より時間の無駄やタグの曖昧性に伴う困難さを避けられたと考える．さらに，予想以上の有志が短期間のうちに殺到したため，プロジェクトを少人数で統制していくことが難しく，これがタグ付けなど一部のタスクへ人的リソースが集中してしまった原因となったことも，1 つの反省点として挙げられる．

ツイートの言語解析

■形態素解析と固有表現抽出

ツイートの言語解析の最終的な目標は，ツイートから固有表現（NE）を抽出することである．この最初のステップは，日本語の平文を分析し，各形態素に品詞を付与する形態素解析である．しかし，本タスクにおいては，ツイートという特殊な分野のテキストを扱っており，東北地方の人名・地名など，既存の形態素解析や辞書では高精度で解析できない言語表現を多く扱わなければならない．この問題に対処するため，本タスクではオープンソースの形態素解析器 KyTea を使用した．KyTea は分野外のテキストに対しても比較的頑健であり，さまざまな言語資源を柔軟に素性として組み入れることができるという特徴を持つ[1]．本タスクの時間的制約上，「人名（姓）・人名（名）の連続は人名として 1 つにまとめる」というように，形態素解析によって出力された品詞に基づく簡単なルールを用いて固有表現を抽出した．この 3 つのステップを図 -3 に示す．

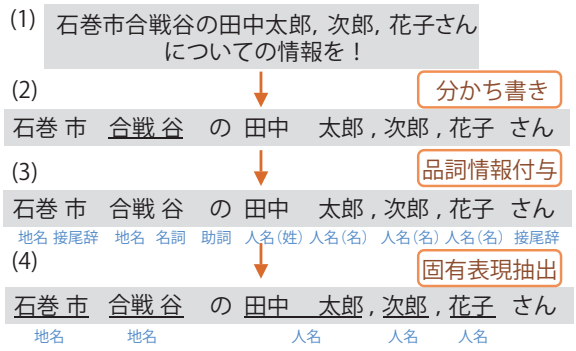

図 -3　ツイートからの固有表現抽出の各ステップ

■分野適応

本タスクでは最初に，日本語の汎用コーパスおよび辞書☆6 を用いて汎用的な形態素解析モデル「ベースライン」を開発した．このモデルは一般的な分野のテキストに対しては有効であるが，本タスクで扱ったツイートなど特殊な分野やスタイルの文書に対しては解析精度が低下することが知られている．そのため，人名・地名の解析精度を高めるために，前章で述べた東北地方の人名および地名の辞書を追加した．

またさらに，たとえば「ツイート」という単語そのものやハッシュタグなど，ツイート特有の言い回しに対応するため，能動学習と部分アノテーションの枠組みを用いて単語分割を再学習した．具体的に

☆6 現代日本語書き言葉均衡コーパス
http://www.kotonoha.gr.jp/shonagon/
および形態素解析辞書 UniDic
http://www.tokuteicorpus.jp/dist/

は，形態素解析の信頼度が低い単語を 100 個選び，その単語境界を人間の評価者が修正し，教師データに加えるというプロセスを 4 回繰り返した．この結果を図 -4 に示す．これは，50 ツイートからなるテストコーパスにおける定量的性能である．

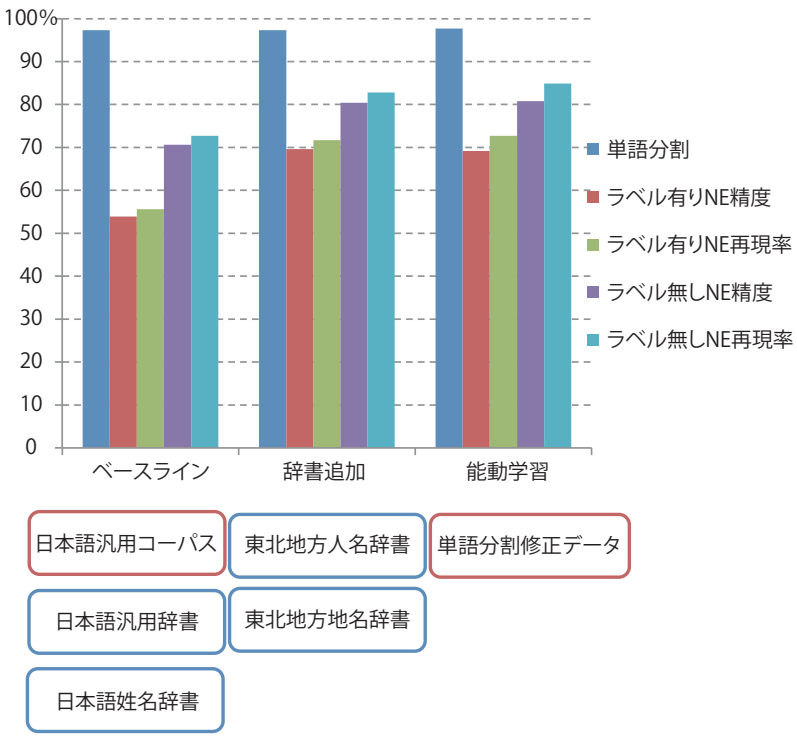

図 -4　各モデルの形態素解析結果

「ベースライン」と「辞書追加」モデルの主な違いは，後者が被災地の地名をより高精度に解析できることである．たとえば，ベースラインでは「気仙沼 _ 市」は「気仙 _ 沼 _ 市」と誤って分割され，「気仙」が人名と認識されたが，辞書追加モデルにより正しく解析された．また，ベースラインでは「平浄 _ 水場」と誤って解析されていたものが，辞書追加モデルでは「平 _ 浄水場」と正しく分割，「平」が地名として正しく認識された．

残る最大の課題は，日本語では曖昧な地名と人名の区別である．住所情報を利用することにより固有表現抽出の結果を改善することができたが，結果として地名に偏ったモデルができてしまった．このようなサンプルの偏りを是正できる機械学習手法（たとえば 文献 4）など）を適用する必要があろう．

ツイートの分類

次に，安否情報を含んでいるツイートを分類する．このために，Support Vector Machine（SVM）アルゴリズムを用い，ツイートから抽出された素性に基づいてツイートのトピックを認識する分類器を学習した．

まず，最初の分類器（分類器 1）は，非常に限られた素性のみを用い，1 つのツイートに単一のタグをつけるものであった．素性として，文中に現れる文字 n グラムおよび固有表現を用いた．加えて，固有表現タグの出現数を素性として利用した．これは，重要なツイートには固有表現が多く出現する傾向があるためである．

その後数日を経ないうちに，分類器 1 の拡張として分類器 2 を開発した．分類器 2 では，単一のツイートに対して当てはまるすべてのタグを付与することが可能である．分類器 2 では，地名と人名の両者が同時に出現したかどうか，人名タグの出現と文字トライグラムの組合せ，さらに，ハッシュタグ “#anpi” や “http(s)” 等が存在するかどうかを，ツイートのトピック分類の精度を高めるため，素性として追加した．

表 -2 に，分類器 2 を用いてツイートのトピックを分類した結果を示す（9,812 ツイートを含む 3 月 17 日のコーパスを使用）．ここで，あるトピック X の精度，再現率，F 値は，それぞれ，

精度＝ X に正しく分類されたツイート数／
　　分類器が X に分類したツイート数

再現率＝ X に正しく分類されたツイート数／
　　コーパス中の X のツイート数

トピック	ツイート数	精度	再現率	F値
O	719	88.4%	98.3%	93.1%
M	134	91.2%	93.3%	92.3%
S	51	72.5%	56.9%	63.7%
L	45	50.0%	11.1%	18.2%
R	32	76.9%	31.3%	44.0%
U	22	0.0%	0.0%	0.0%
I	12	50.0%	58.3%	53.9%
H	6	100.0%	50.0%	66.7%
P	4	0.0%	0.0%	0.0%

表-2　各ツイートトピックに対する分類結果

F値＝精度と再現率の調和平均

である．比較的ツイートの数が多いO，Mについては高い精度を達成できたが，他のトピック，特にLなどでは再現率が低かった．トピックU, S, Oの間で起こる誤判定が最も多かったが，これらはすべて個人の安否確認に関係がないためあまり大きな問題ではない．ただ，注意深くデータを観察してみると，アノテイター間によるタグ付け基準の不一致が非常に多いことが分かった．これは本プロジェクトの時間的制約が厳しかったこと，およびコミュニティによるタグ付けによりアノテイター間の一致を保証するのが難しいといった特性に起因している．

このような状況は，最近の効率的な能動学習手法や公開されているツールを利用し，より少数のアノテイターによって同等の精度をもたらすことができる手法（たとえば，文献5)）を利用すると改善することができると考えられる．

抽出された安否情報の応用

最後のステップは，抽出されたツイートの安否情報が信頼に足るかどうかを検証しつつ，トピックL「誰かが生きているという情報を入手した」というツイートから抽出された人名と地名の固有表現をGPFのデータと照合するところである．被災者の生命という扱いに細心の注意を要する情報であるため，人手によって抽出結果を検証し，照合するという方法をとった．

GPFは，被災者の姓名，自宅の住所，市町村，県名などのフィールドから構成される．人手によって更新した固有表現データをGPFのフィールドと照合したところ，100件以上の被災者の生存がツイートによって確認できた．

ただし，ツイートから抽出された多くの情報が以下のような問題を含んでいた．

- **不完全な住所情報**
  これは，住所の市町村区や番地のような重要な情報が欠けている場合である．たとえば，「宮城県の田中太郎と無事に連絡が取れました！」というツイートでは，宮崎県に「田中太郎」という個人が複数存在する可能性があり，個人を同定できない．
- **不完全な個人名**
  ツイートに 家族などが同時に記述されている場合，姓や名が省略される場合が多い．たとえば「仙台市に住む田中太郎，次郎，花子が避難所にいました.」「仙台市に住む田中一家の安否が確認できました.」などが例にあたる．前者は姓の情報を復元できるが，後者の場合個人を特定できない．人名が漢字ではなくひらがなやカタカナの読みのみで書かれるケースもある．

本タスクでは行方不明者の安否情報に主に焦点を当てたが，災害時においてTwitterはさまざまな情報を提供する情報源にもなり得る．たとえば「気仙沼で被災し，市民会館の裏山に50人避難しています」というツイートは，被災者の人数（50人）と具体的な避難場所（市民会館の裏山）を含んでおり，どこに救助資源を集中すればよいかを示す貴重な情報である．ここからの情報抽出は，ツイートの信頼性に関する近年の研究（たとえば文献6）等）の対象となり得るであろう．

災害と情報処理

本稿で紹介した安否情報マイニングタスクだけではなく，ANPI_NLPプロジェクトの枠では，他の活動も活発に行われた．代表的なものを次に挙げる：

- ツイートと避難所との結びつけ：ツイートから抽出された地名から緯度経度を求め（ジオコーディングし），避難所リストと結びつける
- 地図上でのツイートの可視化：検索を容易にするために，ジオタグ（発信位置情報）の付けられたツイートを地図にマッピングする
- 震災関連情報の外国語への翻訳：震災関連用語の多言語辞書の作成・共有，地震関連情報の外国語（特に英語，中国語，韓国語，ポルトガル語の各言語）への自動翻訳および提供

もちろん，災害時の NLP の役割に注目したのは我々が初めてではない．Lewis[2] は，2010 年のハイチ大地震の際に，短期間でハイチ語の機械翻訳システムを開発した．短い開発スパンで製品を出荷する必要がある点で，我々のプロジェクトと共通する点が多い．また，文献 3）では，ツイートを用いて地震の発生をリアルタイムで検出している．

本プロジェクトの教訓と課題

本稿では，限られた時間制限の中で，ツイートから安否情報をマイニングする有志によるプロジェクト ANPI_NLP を紹介した．

総合的には，本プロジェクトは一定の成功を見たと言ってよいであろう．非常に短い時間内に，処理の難しい，ツイートという特殊な情報源から情報を抽出するシステムを設計・実装し，100 人以上の安否情報を更新することが可能であったことを示したからである．このタスクを通じて，大規模災害において情報提供するために必要な枠組みを明らかにし，さらに大規模な状況に対応できるツールを開発することができた．

もちろん，解決すべき課題が数多く残されている．主な問題には，以下のようなものがある：

- **スピードこそがすべて**

  データの収集，解析ツール，分類器の開発，情報の提供といった一連のサブタスクを限られた時間内に，かつ順番にこなす必要があるため，ある時点の遅れはその後の全体の作業に影響してしまう．精度を犠牲にしても次のステップに進み，後に改善するといったアジャイルな開発スタイルが必要である．

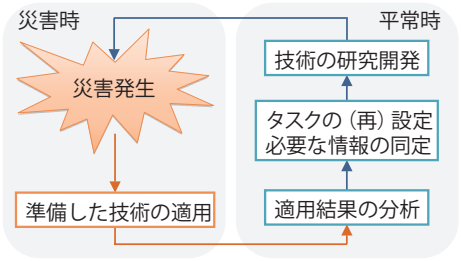

図 -5　災害用自然言語処理技術の開発サイクル

- **タグ付け枠組みの必要性**

  短期間にタグ付けをする際に起きる問題に対処するため，必要な情報のみにフォーカスして明確な基準を作成すること，起こり得る問題に対処するためにタグ付けの試行をまず実施してみること，難しいケースを後回しにする，移譲できるようにするなどの枠組みを作る必要がある．

- **人的リソースの有効活用**

  プロジェクトでは，当初の予想を上回る支援を得ることができたが，人的リソースを有効に活用できたかどうかについては疑問が残った．有志による支援を活用するためには，プロジェクトの総合的なビジョンを参加者が共有し，新しいタスクを素早く割り当てるなどの工夫が必要である．

これらの課題に対しては，あらかじめ準備しておくことができる点が数多く存在する．図 -5 に示すように，災害用自然言語処理の研究開発に平常時から取り組んでおき，災害時にその知見をすぐに利用できるような体制づくりが必要であろう．

地震をはじめ，災害は悲劇であり困難な時間であり，個人が処理できる以上の情報に圧倒されてしまう．本プロジェクトによって，NLP が災害支援活動に貢献でき，人命救助に役立てる可能性があることを示すことができたと自負している．将来に同様のプロジェクトがもしあるとすれば，本プロジェクト経験が糧として活かされることを願って止まない．

参考文献

1）Neubig, G., Nakata, Y. and Mori, S. : Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis, *In Proc. of ACL-HLT 2011* (2011).
2）Lewis, W. D. : Haitian Creole : How to Build and Ship an MT Engine from Scratch in 4 days, 17 hours, & 30 minutes, *In Proc. of EAMT 2010* (2010).
3）Sakaki, T., Okazaki, M. and Matsuo, Y. : Earthquake Shakes Twitter Users : Real-time Event Detection by Social Sensors. In *19th International Conference on World Wide Web*, pp.851-860 (2010).
4）Huang, J., Smola, A. J., Gretton, A., Borgwardt, K. M. and Scholkopf, B. : Correcting Sample Selection Bias by Unlabeled Data, *Advances in Neural Information Processing Systems*, 19:601 (2007).
5）Settles, B. : Closing the Loop : Fast, Interactive Semi-supervised Annotation with Queries on Features and Instances, *In EMNLP 2011*, pp.1467-1478, Association for Computational Linguistics (2011).
6）Qazvinian, V., Rosengren, E., Radev, D. R. and Mei, Q. : Rumor has It : Identifying Misinformation in Microblogs, *In EMNLP 2011*, pp.1589-1599, Association for Computational Linguistics (2011).

（2011 年 11 月 18 日受付）

謝辞　プロジェクト全体を通してご意見をいただいた公立はこだて未来大の藤田篤氏，東京工業大学の橋本泰一氏，GPF データ利用の快諾をいただいた Google 社，および，65 名以上のプロジェクト参加者全員に心より感謝する．

萩原 正人（正会員）masato.hagiwara@mail.rakuten.com

2009 年名古屋大学大学院情報科学研究科博士後期課程修了．同年よりバイドゥ（株）において検索エンジンの研究開発に従事．現在楽天技術研究所に所属．博士（情報科学）．自然言語処理の研究に従事．

村上 浩司 koji.murakami@mail.rakuten.com

2004 年北海道大学大学院工学研究科博士課程単位取得退学．ニューヨーク大学コンピュータサイエンス学科，東京工業大学，奈良先端科学技術大学院大学を経て 2010 年より楽天技術研究所に所属．博士（工学）．自然言語処理の研究に従事．

Graham Neubig（学生会員）neubig@ar.media.kyoto-u.ac.jp

2005 年米国イリノイ大学アーバナ・シャンペーン校工学部コンピュータ・サイエンス専攻卒業．2010 年京都大学大学院情報学研究科修士課程修了．同年同大学院博士後期課程に進学．現在に至る．

松林 優一郎 y-matsu@nii.ac.jp

国立情報学研究所特任研究員．2004 年九大・理・物理卒業．2010 年東大大学院・情報理工・博士課程修了．現在に至る．自然言語処理の研究に従事．言語処理学会，人工知能学会，ACL 各会員．