災害対応質問応答システム構築に向けた質問・回答コーパスの構築

川田 拓也　大竹 清敬　後藤 淳　鳥澤 健太郎
独立行政法人 情報通信研究機構

1　はじめに

2011 年 3 月 11 日に日本を襲った東日本大震災とそれに伴う原子力発電所事故においては、情報の発信と伝播そして受容に関してこれまでにない変化が見られた。TV や新聞といった既存のメディアが情報を発信し、大衆がそれを受容する従来の構図に代わり、今回の震災では、ソーシャルメディアを通じて個人が発する情報が災害情報の伝達において重要な役割を果たした。特に、ツイッターや SNS といったソーシャルメディアは誰もが即時的に情報を発信でき、しかも他のユーザによって即座にその情報が拡散し、共有されてゆく様はこれまでに見られなかった情景であろう。また、個人が有用な情報を整理、可視化し公開したり、支援サイトを公開するなど、ネットを通じた支援活動も見られた [4, 6]。

一方でその問題点も浮き彫りとなった。必要な人に必要な情報が効率的に届かないという、情報のミスマッチの問題である。大量の情報が拡散してもそれを必要としている被災者、支援者に対して正確に伝わり、救援、救助が行われた例は限られていた。例えば津波被害にあったある被災地では善意の人々から防寒着が届けられたが、本来必要であったものは防寒ズボンの方であったという事例が我々の行ったヒアリングで得られている。もう 1 つは、想定外の要望への対応である。災害時には「透析器具」や「手話通訳」など平時には想定し得ない要望も数多く出る。このような状況の中で個人がソーシャルメディア上で発する情報を如何に整理し、必要な人に届けられるかが今後の災害に向けての喫緊の課題であると考えられる。

我々はこのような背景から、多様な質問に対して、ツイート情報から回答リストを提示し、想定外の情報も容易に発見可能とし、また、被災の全体像を素早く把握して効率的な救援活動を行えるよう、状況の俯瞰的提示を可能にするシステムの構築を進めている。本稿は主に災害対応質問応答システムの性能評価のために必要な質問とその正解データの構築について報告するものである。

2　アンケート調査

我々は災害対応質問応答システムの構築および、質問作成の準備段階として、ボランティア活動等で震災に関わった人々が現地で感じたことや、要望をすくい上げるために、アンケート調査を行った。アンケートは日本栄養士会に所属する栄養士 103 名に対してメールを通じて行われた。回答者の震災との関わりについては図 1 に示した。なお、図 1 は複数回答が認められているため合計で 103 名を超えているが、7 割以上の人々が実際にボランティア活動に携わっており、3 割近くの人々が自身もしくは身近な人が被災している。

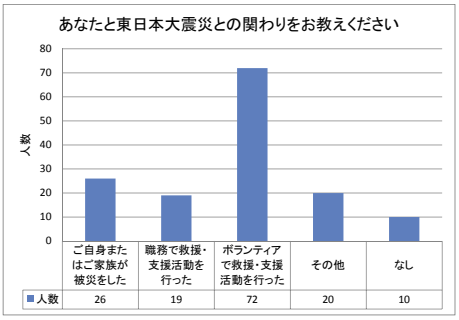

図 1　回答者の震災との関わり

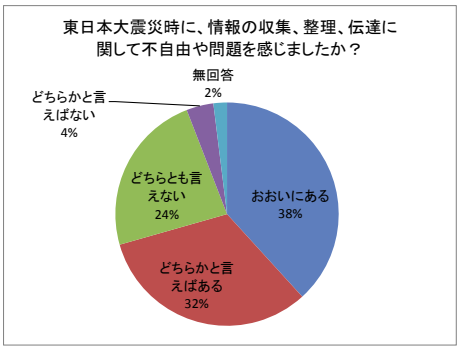

図 2　情報伝達に関する意識

質問は、東日本大震災時に、情報の収集、整理、伝達に関して不自由や問題を感じたかどうか、情報収集の際に、インターネットをどのように利用したか、災害時に質問応答システムがあればどのような質問をどのタイミングでするかといった点について問うたものである。

まずアンケート調査の中で「東日本大震災時に情報の収集や、整理、伝達に関して不自由や問題を感じたかどうか」についての調査結果を図 2 に示す。図 2 が示すとおり、実に 70% が情報収集、整理、伝達において何らかの不満を感じていたことが明らかとなり、情報技術が発達した現在においても円滑な情報伝達が難しいことを物語っている。

例えば「被災地外から向かう現地事前情報の不明点が多かった」、「現地がどのような状態になっているか、ある程度の情報は得られたが、メディア情報との差があり不安を感じた」、「活動の場ができても実際に被災者の方に、情報を発信するにはチラシの配布等しかできなかったため、当日仮設住宅に出向いてからの訪問勧誘となる。情報の発信によい方法があったらと感じた」、「情報伝達が十分でなかったので、現地に行き、無駄な時間を使った。(中略) 現地で 3 日、4 日の活動と決められていたので、コーディネータ役の存在がしっかりしていない現実

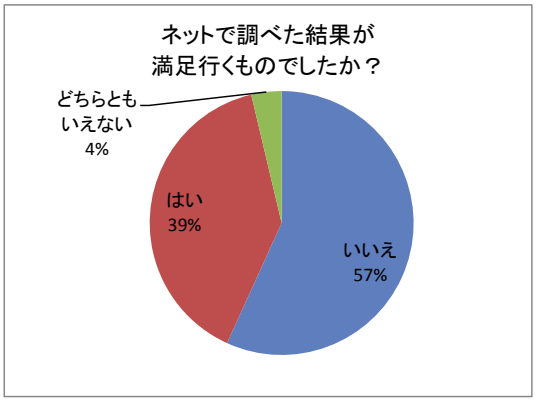

図 3　災害時のネット活用の満足度

があり現地を把握するのに大変だった」等のコメントがあり、情報の収集、発信、共有全てに関して課題が残されていることが浮き彫りとなった。

アンケートではさらに、災害時においてインターネットをどのように活用したかについても質問し、それが実際に役に立ったかどうかも尋ねた。その結果、図 3 のように半数以上が、インターネットでの情報収集に満足のいく結果を得られていなかった。例えば既存の検索エンジンを利用したものの「情報がなかった」というものや放射能の影響について調べたものの「いろんなサイトに分散しており信頼できる情報がわからなかった」という意見が見られた。逆に、「被災地に行くための交通手段を調べる手立てとしては役に立った」という意見や、「TV だけでは得られない情報を得られた」という肯定的意見も挙げられた。

以上のように、実際に被災地と関わりを持った人々からのアンケート結果から見えてくることは、インターネットは支援活動において強力な手段となり得る可能性はあり、部分的には確かに役に立った部分もあるものの、その効果を十分に発揮するためにはまだ多くの課題が残されているということである。特に、情報自体はツイッターを始めとする即時的な情報発信手段により、大量にかつ、迅速に発信されたものの、必要な情報が必要な人々に届かないというミスマッチが生じたといえる。

そのような背景の中で我々は災害対応質問応答システムを構築しているのだが、上記のアンケートの結果で明らかなとおり、被災者および救援者（支援者）の要望に応えられるものでなければ意味がない。そこで我々は今回のアンケートや、被災地の方へのヒアリング調査、震災後に出版された書籍等を参考にし、災害に関連する多様なトピックの質問を災害対応質問応答システムの評価用データとして構築した。以下の節では、我々が構築した質問・回答コーパスについて述べる。

3　質問と回答の構築

評価用の質問とその回答コーパスの構築にあたっては概ね次のような手順で行われた。

1. 現地のヒアリングやアンケート結果の分析および被災者、支援者双方のニーズの把握
2. 人手による質問作成。ヒアリングやアンケートで得られた質問だけでなく、それを補完する形で各種資料 ([1] など) を参照して質問作成 (309 質問)
3. 質問中からキーワードとなる語を選定し、我々が開発したツイートの全文検索エンジンを用いてその語を含むツイートを収集
4. 収集したツイートから人手で質問の回答を抽出
5. 収集する回答を増やすため作成した質問と同義の質問をさらに人手で作成し、その質問に含まれるキーワードで再検索。3. に戻る

以下ではそれぞれの段階について具体的に説明する。

3.1　質問の作成

質問作成に関しては、前述したアンケートの他に、被災地等で 20 団体以上に対して我々が行ったヒアリングに基づいて作成されている。

まずアンケート調査の中で、災害対応質問応答システムに対して具体的に質問したい内容を尋ねたところ、「避難所で不足している食品」、「避難所の収容人数」、「高齢者や嚥下障害者の状況」「避難所の食事でのアレルギー対応状況」、「被災者の栄養状態」に関する質問が挙げられていた。いずれも、栄養士ならではの視点からの質問が挙げられていたが、共通するのは現地の状況把握に関する情報が重要視されていたことである。

我々が行った被災地でのヒアリング調査からは、被災者側からの切実な要望が得られた。まず被災直後は生命を守るための情報と、さらには長期的なスパンを見越したニーズの把握といった、それぞれの時期に応じた対応が必要であるというコメントを頂いた。震災直後は当然のことながら、当座の食糧や毛布、燃料の支援に関する質問が想定される。日常を取り戻しつつある状況においては、例えば、教育関係であれば「学校を再開するにはどうしたらよいか、再開したあとで何が必要か」といった事に関する情報が重要になる。また、支援者と被災者の間のコーディネートの重要性も指摘された。被災地には様々なボランティア団体が支援に駆けつけるが、現地が必要としている支援内容とのマッチングがうまくいかないという問題が生じていたことが報告されている。これも現地の状況把握を共有する難しさを物語っている事例と言える。

以上の知見を考慮しつつ我々は、現地の状況把握や、震災後、長期的に必要となる情報も含めて質問を作成することにした。さらにそれを補完する形で、東日本大震災に関する資料 ([1] など) や、Web 情報などを参考に、計 6 名の作業者が被災者、支援者の立場になり質問を 300 問弱作成した。その際、質問はできる限り「なに／どんな」「どこ」「いつ」を問う質問とし、「どのように」「なぜ」を問うものは今回は対象外とした。また、質問は原則として名詞（もしくはサ変名詞）を含むものとした。質問自体は災害に関連する限りにおいて制限は設けなかったが、トピックが偏らないよう適宜取捨選択しながら 300 弱の質問を作成した。

表 1 に質問の例を示す。「カテゴリ」については後述する。なお、災害対応質問応答システムでは「どこに給水車が来ますか」という質問は実際には「**宮城県の**どこに給水車が来ますか」というように場所を限定した質問が入力されることを想定している。

表 1　質問の例

カテゴリ	例		
インフラ	給水所はどこですか	宿泊できる場所はどこですか	どこでガスが復旧していますか
防災	どこで避難訓練は行われていますか	防災袋には何を入れると良いですか	地震で避難するときの注意点は何ですか
法律・制度	どこで被災者ローンが借りれますか	失業給付を受けるために何が必要ですか	復興給付金は何に使われますか
交通	どこで道路が寸断していますか	機能している空港はどこですか	被災した港はどこですか
救援・避難施設	避難所はどこですか	避難所での娯楽は何ですか	仮設住宅には何がありますか
災害状況	津波の高さはどのくらいですか	震災による経済的損失はいくらですか	どこで土砂崩れが起きていますか
病気・負傷	はやっている病気は何ですか	クラッシュ症候群の注意点はなんですか	不安解消に効くものは何ですか
原発・汚染	放射能が高いのはどこですか	なにが汚染されていますか	線量はどのくらいですか
救援情報	どこで避難指示が出ていますか	被災者支援対策は何がありますか	どこでペットを預けられますか
安全・信憑性	風評被害の原因は何ですか	デマ拡散を防ぐには何をすればいいですか	野菜の放射線量の基準値は何ですか
政府・行政	どこに自衛隊がいますか	自治体の就労支援には何がありますか	緊急車両が通れる道はどこですか
支援物資	必要な家電は何ですか	スポーツドリンク代わりになる物は何ですか	支援物資の受付窓口はどこですか
安否確認	どこで安否確認ができますか	どこで身元の確認ができますか	救援を求めているのはどこですか
ボランティア	ボランティアに適した服装は何ですか	ボランティアの作業は何になりますか	どこで復旧作業が行われていますか
支援活動	どこで募金ができますか	支援が必要なのはどこですか	どこで炊き出しをしていますか
介護・育児	どこで粉ミルクを買えますか	オムツの代わりになるものは何ですか	高齢者に対するボランティアは何ですか
教育	学校はいつ再開しますか	どこに仮設校舎が建てられますか	普段の給食の代わりに何がありますか
医療・医薬品	出産ができる病院はどこですか	救急病院はどこですか	どこで献血できますか

表 2　質問とツイートから抽出された回答例

質問	抽出回答数	回答例
どこで孤立していますか	304	宮城県柴田町仙南中央病院, 宮城, 仙南中央病院, 宮城県県南地域, 茨城, 気仙沼, いわき市, 福島,...
ボランティアはどこでできますか	104	石巻, 宮城, いわき, 福島, 浦安, 栄村, 仙台, スーパーアリーナ, 浦安市, 気仙沼, 東京ＡＲＫ, 旭市,...
どこで火災が起きていますか	309	気仙沼, 4 号機, 千葉, お台場, 製鉄所, コスモ石油, 福島第一原発 4 号機, 千葉の製油所, 気仙沼市,...
どこに支援物資が届いていますか	90	福島県いわき市, 石巻, 東北, いわき市, 茨城, 仙台, 福島, 岩手県, 宮城県, いわき, 福島県, 気仙沼,...
必要な家電は何ですか	18	懐中電灯, 冷蔵庫, ラジオ, ＴＶ, 風呂用ラジオ, 電池充電, 電子レンジ, 蓄電式ラジオ, 発電機器,...
長持ちする食品は何ですか	40	お菓子, 冷凍食品, インスタント食品, 納豆, 米, 発酵食品, 加工食品, パン, カロリーメイト,...
宅急便はどこまで届きますか	55	東京, 東北, 宮城, 福島, 岩手, 仙台, 石巻, 宮城県, ひたちなか市, 青森県, 茨城県, 秋田県, 山形県, ...
どの専門の医師がいますか	40	放射線治療, 被ばく医療, 放射線医学, 放射線, 放射線科, 原子力関連の事故, 内科, 精神科,...

3.2　回答抽出

次に、回答の抽出について述べる。ここで構築する回答データは、災害対応質問応答システムの評価用データとして使用するため、精度、被覆率ともに十分でなければならない。「孤立している場所はどこか」という質問に対して、多くのツイートで言及されている場所を回答として提示するだけでは十分ではない。支援者にとっては、1 つのツイートでしか言及されていない場所でも等しく支援対象とするべき場所であり、被災者の家族や知人にとっては、自分の家族知人の状況を知ることができる貴重な情報なのである。

回答は人手で抽出するため、精度は（人為的なミスを除けば）問題無いと仮定できるが、大量のツイートから人手で回答を抽出する以上、被覆率を保証するデータ作成は困難である。そこで、以下で述べるように、できる限り網羅的に回答を収集できるよう工夫した。

ツイートから回答を抽出するにあたり、風間ら [2] が構築した、2011 年 3 月 10 から同年 4 月 4 日までに投稿されたツイート約 2 億文（データ提供元：(株）ホットリンク）を元にした全文検索エンジンを利用した。

質問の正解となる回答をまず正確に得るために、作成した質問から任意に選んだキーワード（名詞か動詞）を全文検索エンジンのクエリとして入力し、そのキーワードを含むツイートを 1 つの質問ごとにランダムで 1,000 ツイート収集した上で、質問の回答となる部分を抽出した。回答の抽出は人手で行った。回答は原則として名詞単位で抽出した。

表 2 に質問と得られた回答数、および実際に抽出された回答例を示している。309 の質問のうち、上記のキーワード検索を利用した回答の抽出を行った結果、約 200 の質問で、実際にツイートから 1 つ以上の回答が得られ、平均すると、1 つの質問につき 113 個の回答が得られた。

3.3　同義質問の作成と回答抽出

しかし、これでは質問文に含まれるキーワードを持つツイートからしか回答を抽出することができない。そのため、被覆率を考えると不十分である。また、ユーザは同じような意味を持つ質問を様々な表現で質問する可能性もある。例えば支援者は「孤立している場所」を問う意図で「物資が来ないのはどこか」と問うことも可能である。そこで、各質問につき、それと同義の質問を作成した。以降では同義質問と呼ぶこととする。ここでいう同義とはオリジナルと同じ回答が得られるという意味での同義であり、語彙的な同義性だけでなく、文全体としての同義性も捉える必要がある。そのため、同義質問は機械的に生成するのではなく、オリジナルの質問と同様に人手で作成した。同義質問の作成にあたっては、最低限オリジナルとは異なる名詞もしくは動詞を含むように作成するという制約を設け、その制約の下で様々な表現の質問を作成した。オリジナルの質問 1 つにつき平均約 5 つの同義質問を作成した。表 3 に例を示す。

表 3　同義質問の例

オリジナル質問	同義質問
孤立しているのはどこですか	物資が来ないのはどこですか
孤立しているのはどこですか	支援を待っているのはどこですか
孤立しているのはどこですか	連絡が取れないのはどこですか
どこに避難していますか	どこで被災者を受け入れていますか
どこに避難していますか	どこに逃げていますか
どこに避難していますか	どこに身を寄せていますか
営業しているお店はどこですか	再開しているお店はどこですか
営業しているお店はどこですか	やっているお店はどこですか
営業しているお店はどこですか	開いているお店はどこですか
どこで床上浸水の被害がありましたか	どこで家が水浸しになりましたか
どこで床上浸水の被害がありましたか	どこで床が水につかりましたか
どこで床上浸水の被害がありましたか	どこで家が水没しましたか

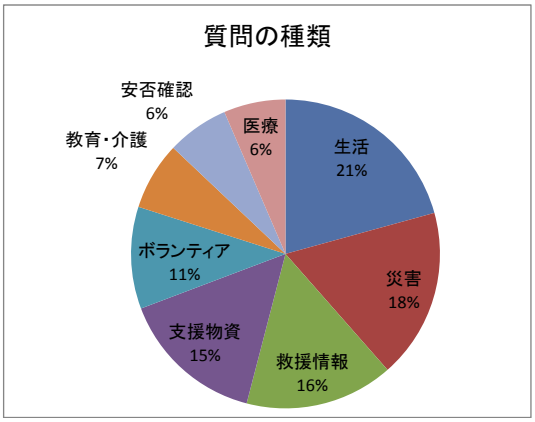

図 4　質問の分類（大分類）

回答の抽出方法はオリジナルの質問と同様で、同義質問に含まれるキーワードを 1 つ以上選定し、それらをツイート全文検索エンジンで AND 検索することでツイートを 500 件収集する。そこから、同義質問とオリジナルの質問の両方に当てはまる回答を収集する。このような手法により、質問の回答数を増やし、回答をできる限り網羅的に収集するよう努めた。2013 年 1 月現在も作業中であり、同義質問による回答収集により、最初のサイクルでは回答を得られなかった質問の回答も得ることができるようになった。回答が得られた質問は計 264 となり、1 つの質問につき平均回答数も 195 となった。

4　質問の分類

表 1 で示したように 309 の質問はカテゴリ別に整理されている。表 1 では一部カテゴリをマージして例示しているが、合計で 28 カテゴリに分類し、災害に関わる様々なトピックが網羅されるよう、留意して作成されている。質問のカテゴリの全体の分布を図 4 に示す。ここでは全体像の把握を容易にするため、8 カテゴリにマージして示している。この図が示すとおり、概ねインフラや避難所状況などを含む日常生活に関する質問（生活）、地震、津波、原発事故などの被災状況に関する質問（災害）、避難指示や行政などによる救援状況、対策などを問うた質問（救援情報）、支援物資に関する質問が過半数を占めている。さらにボランティア、安否確認や医療、福祉に関する質問があり、災害時に、被災者と支援者の双方が必要とする質問が網羅されていると思われる。

5　関連研究

災害時において ICT 技術が果たす役割が大きくなっている中で、自然言語処理を応用した支援として何ができるかという観点からの議論も盛んになってきている。例えば Neubig ら [3] によって、東日本大震災時のツイートから情報抽出する手法について議論されている。また、災害時には、被災者、支援者双方から様々な要望とそれに対する回答として対応策が発信される。佐野ら [5] は災害関連のツイートから、要望やそれに対応する対応策を蓄積、整理したデータを構築している。この要望／対策データを利用することで、さらに質問の種類を増やすことが可能となる事が考えられる。事前の調査で例えば「(○×町の状況を）メディアに取り上げて欲しい」「物流を（福島へ）解放して欲しい」というような要望が我々の質問では網羅されていなかった。今後は要望／対策データを利用して質問を拡充していく予定である。

6　おわりに

本稿では、災害対応質問応答システムの性能評価に向けた質問と回答のコーパス構築について報告した。これからも自然言語処理技術を応用した災害に対応したシステムとして、質問応答を始めとして様々なものが開発されていくことが考えられる。本稿で述べたような実際のアンケート調査や各種資料に基づき、災害時に人々が知りたいと思う事柄を網羅した質問と回答のセットは、そのようなシステムを構築し、評価する上で重要な基盤データになると考えられる。現在災害対応質問応答システムの評価実験を本コーパスを用いて始めている。今後は質問、回答を増やし、より大規模な評価を進めていく。

また、将来は今回の震災とは大きく異なる要素を含む大災害が発生することを想定しておく必要があるだろう。そのような事態にも対応できるような、分野を限定しない汎用的な質問応答システムの開発も継続し、今後起こるであろう災害に対しては、現在の災害対応質問応答システムと、より汎用的な質問応答システムの両方を活用した災害対応ができる仕組みを構築していく予定である。

謝辞

アンケート調査を行うにあたり、日本栄養士会の清水詳子氏および、国立保健医療科学院奥村貴史氏にご協力頂きました。ツイートデータについては（株）ホットリンク様より頂きました。また、質問の作成と抽出には計 8 名のアノテータの方々にご協力頂きました。この場にて感謝いたします。

参考文献

[1] 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, インプレスジャパン (2012).
[2] 風間淳一, De Saeger, S., 鳥澤健太郎, 後藤淳, István, V.: 災害情報への質問応答システムの適用, 平成 24 年度情報処理学会関西支部 支部大会 (2012).
[3] Neubig, G., Matsubayashi, Y., Hagiwara, M. and Murakami, K.: Safety Information Mining — What Can NLP Do in a Disaster —, *Proceedings of IJCNLP 2011*, pp. 965–973 (2011).
[4] 西條剛央: 人を助けるすんごい仕組み, ダイヤモンド社 (2012).
[5] 佐野大樹, Varga, I., 風間淳一, 橋本力, 鳥澤健太郎: 災害関連ツイート要望・対応策マッチングコーパスの作成, 平成 24 年度情報処理学会関西支部 支部大会 (2012).
[6] 徳田雄洋: 震災と情報 —あのとき何が伝わったか, 岩波書店 (2011).