ブログの投稿時間を利用した地図付きトレンド型質問応答システムの提案

木村 泰知† 荒木 健治‡
† 小樽商科大学 ‡ 北海道大学大学院情報科学研究科

1. はじめに

近年，ブログおよびソーシャルネットワークサービス（SNS）の利用者が増加し，膨大なテキストが WWW 上に日々追加されている．これらのテキスト情報から，ユーザが必要とする情報を抽出する技術が望まれており，情報検索，あるいは，質問応答の技術が期待されている[1][2].

質問応答では，自然言語による質問に対して，情報検索技術を利用することにより，正解の応答が含まれる文書候補を選択し，固有表現抽出を用いて，人名，組織名，日付などを応答する．従来の質問応答では，事実に基づいた factoid 型質問が多く，質問に対して，正解が決まっている応答が多い．しかしながら，流行を尋ねる質問では，応答は時間とともに変化する．このような質問に答えるためには，ブログの投稿時間を考慮した応答が有効と考えられる．

また，質問応答技術は，テキストによる応答が主な研究であるが，応答の表現方法として，図，表，地図などのテキスト以外の形式も考慮する必要がある．最近では，テキストからの可視化に関する研究が注目されており，質問応答に利用可能と考えられる．最近では，電子化された地図を簡単に利用可能となり，質問応答における地図の利用方法は、「場所」に関連した質問以外にも，「人名（出生地）」，「組織名（所在地）」，「理由（場所を含んだ理由）」，「定義（地図を利用した説明）」などがあり，他の質問タイプとも関連性がある．今回，質問に対して，地図とともに応答する．

本研究では地域情報（小樽）に焦点を当て，ブログを 1 時間おきに収集している．収集データからブログの更新時間を抽出し，応答に利用する．たとえば，次のような質問を想定している。

「最近，何が流行っていますか？」

この質問に対する応答は，新聞記事や全ての WWW の情報を利用するよりも，ブログから応答候補を選択することで，適切な応答が可能になる．つまり，ブログの場合，正確な投稿時間を得られるため，流行を分析することに適している．

以前，小樽といえば，海産物が有名で「お寿司」に関して質問されることが予想されたが，ブログを検索した結果，お寿司よりも，お菓子に関する記事が多いことが確認された．その抽出方法について述べる．また，流行の店名と地図情報を統合し，ユーザの満足度が高い応答表現も検討する．

2. 関連研究

WWW の情報を利用した，トレンドに関する研究は 1990 年代後半から行われており，大久保ら[3]は，WWW 検索ログに基づくトレンド情報の抽出に関する研究を行っている．情報検索のキーワードのログを解析することにより，トレンドの分析をしているが，ブログを利用したものではなく，投稿時間を考慮しているものではない．

最近では，キーワードだけではなく，インターネット上に存在するテキストを対象に分析する研究が増えている[4] ．たとえば，要望，要求，印象，賛否の表明などのテキスト評価分析が行われ，各文から意見かそれ以外の 2 値分類を行う研究，テキストを 3 つ組（対象，属性，評価値）に変換し数値へ変換することにより，文の評価を行う研究などが提案され，ブログに注目した研究が多く存在しているが，質問応答に適用している研究は少ない[5].

また，最近では，電子化された地図を利用することが可能になり，地図を利用した質問応答も提案されている[6]．しかし，情

報検索を利用した質問応答であり，ブログの情報および流行の質問を扱ってはいない．

そこで，本稿では流行に関する質問に焦点を当てる．

3. トレンド型

本稿では，流行を尋ねる質問をトレンド型質問と呼ぶ．トレンド型質問のパターンを調べるために，147 名の学生から流行に関する質問を 433 文収集した．その質問を分析した結果次のようなトレンド型の質問パターンが存在した．

0. 曖昧な流行を尋ねる質問
   - 「今，旬なものって何ですか」
0. 年齢，性別など人を限定した質問
   - 「20 代には何が流行っていますか」
0. 場所を限定した質問
   - 「堺町では何が流行っていますか」
0. 時期を限定した質問
   - 「今年の冬は何色が流行っていますか」
0. 条件が複数（2 つ以上）存在する質問
   - 「小樽で今が旬の食べ物は」
0. 比較する必要がある質問
   - 「今年，小樽の雪は多いですか」
0. クイズ形式の質問
   - 「大物女優とお笑い芸人のカップルは誰と誰」
0. 擬人化した質問
   - 「あなたは最近何に注目していますか」
0. その他
   - 「最近石原真理子が有名になった訳は」

擬人化した質問以外の質問は，ブログから収集したデータを利用することで応答可能と考えられる．

4. 処理過程

本手法の処理過程を図 1 に示す．

4.1. 入力文解析

入力文解析では質問タイプの判定およびキーワードの抽出を行う．質問タイプの判定は，入力文に含まれる単語に基づいて，予め与えた規則により質問タイプを判定する．質問タイプの種類は「人名，地名，組織名，固有物名，日付，時間，金額，割合，Why 型，Definition 型，How 型」に加え，TREND（流行）型を付け加え，12 種類とした．本システムは 12 種類の質問タイプを識別できるが，本稿では TREND 型に焦点を当て，説明および実験の説明をする．

キーワードは，ChaSen により形態素解析を利用し，名詞を抽出した．

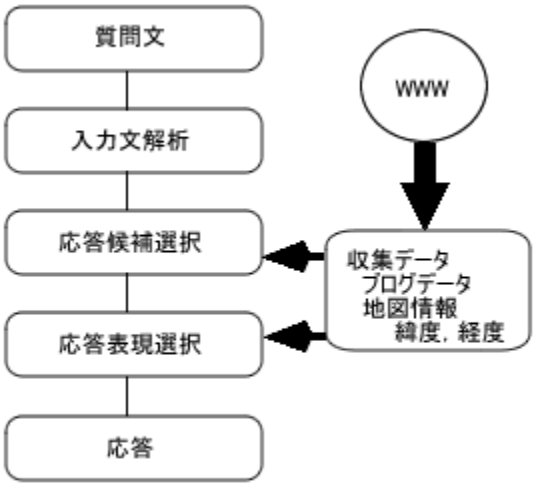

図 1　処理過程

4.2. 応答候補選択に利用する収集データ

今回，質問応答の対象を「小樽」とした．そのため，「小樽」を含むブログだけを収集している．そのブログの収集方法について説明する．

4.2.1. ブログデータ収集方法

2006 年 9 月から「小樽」および「おたる」を含むブログ情報を 1 時間おきに収集している．2007 年 1 月の時点で「小樽」に関連するブログは 12,000 記事以上収集した．「小樽」および「おたる」を 1 日に収集するブログ数は約 106 件であった．参考として，「札幌」，「さっぽろ」を含むブログの場合，同じ収集方法で 1 日約 460 件であった．

本稿で処理するブログ情報は RSS 形式で処理できる状態のものである．

RSS は，サイトの更新情報をまとめたフォーマットであり，XML によって記述されている．フォーマット形式により，RSS は，RSS1.0, RSS0.91, RSS0.92, RSS2.0 などが存

在し，他にも Google が採用している atom などのフォーマットも存在する．ここでは，CPAN の XML::RSS を利用し，RSS1.0 のフォーマットを対象としている．

RSS などのフォーマットがあるため，容易に投稿時間を抽出できるようになり，トレンド型の質問応答にも応用できる．

4.2.2. 応答候補選択

本稿では，応答方法を下記の 2 つとする．

- 曖昧な質問
- 条件が含まれる質問

「曖昧なトレンド型質問」に対しては，最近収集された 1,000 件のブログを対象とする．収集した 1,000 件の記事から，句点（“.”，“。”）で文単位に分割し，同じ文を含むブログは営業目的である可能性が高いため，対象外とした．収集されたブログから，2 回以上出現する名詞を抽出する．その単語が 2000 年と 2001 年の読売新聞と日本経済新聞に出現しているかを比較し，新聞には出現しないが，ブログには出現する回数が多い単語をトレンドの単語とする．

「条件（キーワード）が含まれる質問」に対しては，キーワードを含むブログだけを対象とし，そのキーワードの付近に応答候補が存在すると仮定する．キーワード付近にある単語に対して，キーワードからの距離，出現頻度，単語の長さを考慮する．

4.3. 応答表現の選択

応答にはテキストに加えて地図を利用する．Google MAPS API を利用し，質問に関連する地図を表示する．関連する店は経緯度情報を用いてプロットする．現在，Google MAPS API が公開されており，自由にマーカーを追加することが可能であり，そのマーカーをクリックすることで関連情報の表示を行う．最近では，緯度経度の情報を直接入力する必要のない，ジオコーディングが可能となっているが，本システムのデータベースは約 500 件のお菓子店，お寿司，観光などに関する位置情報を予めデータベースに登録してある．データベースには，位置情報の他に簡単なコメント，写真情報がある．

4.4. 本システムの実装例

本手法の実装例を図 2 に示す．

図 2 では，質問文に含まれる単語を利用して関連するブログを収集し，「関連ブログ」として，投稿時間が最新のものから表示する．「トレンドブログ」は，質問に関係なく，流行しているキーワードを含むブログを表示する．地図は Google Map を利用し，予め与えられた位置情報を持つ店や場所であれば，ターゲットの場所を中央に表示する．

図 2 実装例

5. 評価実験

本実験では，流行に関するトレンド型の質問の評価を行う．トレンド型は一定の正解が存在しないため，根拠記事を示した応答を求めることにする．その根拠記事に質問に対して，正解を含む文書があれば正解とする．

評価は，「流行の質問に関する評価」とする．実験データはトレンド型質問の分析に利用した 433 文中 54 文選択した．分類された各質問群から，ランダムに質問 10 文を抽出し，流行とは関係ない文を削除した．各質問群からランダムに質問を抽出することにより，質問の偏りをなくした．

流行に関しての評価方法は，佐藤ら[6]の評価方法を利用する．上位 3 つのシステム応答に対して評価する．応答は 3 段階で評価する．

- 正解
  - 根拠記事があり，適切な応答.
- 準正解
  - 根拠記事の文章から判断し，質問に対して関連はあるが，不十分である応答
- 不正解
  - 正解，準正解に当てはまらない応答

精度は，次のように評価する.

$$精度 = \frac{正解数 + 0.5 * 準正解数}{正解数 + 準正解数 + 不正解数}$$

今回は，トレンド型応答に下記の 2 つの手法で応答した.

【手法 1】出現頻度，距離，文字の長さに基づいて応答を選択する.
【手法 2】手法 1 に，新聞記事の出現頻度と比較した基準を追加

表 1　実験結果

	正解数	準正解数	不正解数
手法 1	15	4	143
手法 2	46	33	83

表 1 に質問に対する正解数，準正解数，不正解数を示す．手法 1 の精度は 0.104，手法 2 は 0.385 となり，一般的に出現するデータと比較することが有効であった.

5.1. 考察

下記に正解応答，準正解応答の例を示す.

正解例
質問「今何がはやっていますか」
応答「チョコラングドシャ」,「ルタオ」,「ボジョレーヌーボー」
準正解例
質問「どんなダイエット方法が流行っていますか」
応答「寒天」

このとき，「寒天」に関する記述は「寒天以上のダイエット･･･」との記述があり，「寒天」を指していないため，準応答とした.

次に不正解の例を示す．手法 1 において「小樽は何が流行ってますか」に対して，「北海道」,「運河」などが上位に出現し，流行とは関係なく，頻度情報の高い単語を選択する傾向があった.

他には，「小学生の間では何が流行していますか」のように年代を限定した質問には，ブログの内容から判断することが困難な場合がある．「札幌へ越して来た小学生からの友人･･･」,「図書館なんて、小学生ぶりかな･･･」のように，小学生の回想が含まれる内容が多く，小学生が書いているブログは少ないため抽出が困難である．また，地図については，予め保持している地図の位置情報が限られているため，表示できない応答も存在した．今後はジオコーディングと WWW 上に存在する住所情報を利用することで，解決する必要がある.

6. まとめ

本稿では，時間とともに質問の正解が変化するトレンド型の質問に焦点を当て，分析した結果に基づいて，システムを作成し，実験を行った．出現頻度，距離，文字の長さのみで判断した場合，出現頻度の高い単語が優先された．一方，過去と現在の出現頻度の違いを測る尺度を導入することによって，流行に関連した応答の精度が向上し，有効性が確認された.

文　献

[0] 加藤恒昭，福本淳一，桝井文人，神門典子，"質問応答技術は情報アクセス対話を実現できるか，" 情報処理学会研究報告，2004-NL-162，pp.145-150，2004.
[0] NTCIR http://research.nii.ac.jp/ntcir/
[0] 大久保雅且，杉崎正之，井上孝史，田中一男，"WWW 検索ログに基づくトレンド情報の抽出について" デジタル・ドキュメント研究会，Vol.1997 No.49 pp.23-30, 1997.
[0] 小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一．意見抽出のための評価表現の収集．自然言語処理，Vol.12, No.2, pp.203-222, 2005.
[0] 乾孝司，奥村学．テキストを対象とした評価情報の分析に関する研究動向．自然言語処理　Vol.13，No.3，pp.201-241.2006.
[1] 佐藤　充，森　辰則（横浜国大），画像や地図を用いて回答できる質問応答システム，2006-NL-176，pp.113-120，2006.