メーリングリストに投稿されたメールを利用して あいまいな質問に問い返す質問応答システムの作成

西村 涼　渡辺 靖彦　岡田 至弘
龍谷大学 理工学部 情報メディア学科
r_nishimura@afc.ryukoku.ac.jp, {watanabe,okada}@rins.ryukoku.ac.jp

1　はじめに

質問応答システムの多くは、ユーザが適切な質問をすることを前提としている。しかし、どんな情報をどこまで詳しくどのように表現して質問するのかを決めるのは、実はかなりむずかしい。このため、ユーザが適切に質問するのを支援することは重要である。

適切な質問をするのがむずかしいのには、以下の 2 つのような原因が考えられる。

1. 直面している問題にかかわる重要な概念は知っているが、それをどこまで詳しく説明していいのかわからない。
   (例文 1) は、PC を利用していて「音がでない」という問題に直面したユーザが解決する方法を問う how 型の質問である。
   (例文 1a)　vine 2.0 をインストールしたのですが、音が出ません
   (例文 1b)　音が出ません
   (例文 1c)　音が鳴りません
   (例文 1b) や (例文 1c) のような質問に対して、「Vine Linux をつかっているのですか」とか「バージョンは 2.0 ですか」などとシステムが問い返したとする。このような問い返しは OS やバージョンの種類を知っているユーザにとって有効である。システムがそのように問い返せば、ユーザは最初に (例文 1b) や (例文 1c) のような簡単な質問をして、システムの問い返しを見てから、より詳しく適切に質問しなおすことができる。
2. 質問をさまざまに表現できる場合、どの表現が適切なのかわからない。
   (例文 1b) の「音が出ません」という質問に対して、「鳴らないということですか」とシステムが問い返せば、システムの知識では「音が鳴らない」という表現が用いられているようだとユーザは気づくことができる。

そこで、本研究では、メーリングリストに投稿されたメールを利用してあいまいな質問に問い返す質問応答システムについて述べる。メーリングリストには Vine Users ML [*1]に投稿されたメールを用いた。

以下、2 章では同義語関係にある用言の取り扱いについて述べ、3 章では問い返しに利用できる表現 (問い返し表現) をメーリングリストに投稿された質問メールから取り出す方法について述べる。4 章では作成した質問応答システムについて述べ、5 章で問い返しについての検討を行う。

2　質問文中の用言の同義語候補の抽出

2.1　用言のさまざまな同義語関係

自然な文による質問では、さまざまな表現を用いることができる。例えば同義語関係にある用言のうち、どれを用いて質問するかは、ユーザにとってむずかしい問題である。このため、同義語関係の可能性がある用言について問い返すことは重要である。ユーザの質問に問い返すためには、少なくとも以下の同義語関係を取り扱わなくてはならない。

1. 特定の場面や状況だけではなく、一般に成り立つ用言の同義語関係
   (例文 2a)　音が鳴る
   (例文 2b)　音が出る
   大辞林 (第二版) の語義説明文では「鳴る」は「音が出る」と説明されている。この例のような同義語関係は、シソーラスや辞書の語義説明文を利用して取り扱うことができる。
2. 特定の場面や状況で生じる用言の同義語関係
   (例文 3a)　apache で CGI が使用できない
   (例文 3b)　apache で CGI が実行できない
   「使用する」ことと「実行する」ことは一般に同義ではない。しかし、この場面 (コンピュータを働かせる場面) では、これらの用言は同じ意味を表わしている。
3. 場面や状況、対象に対する認識のずれによって生じる用言の同義語関係
   (例文 4a)　IP アドレスが割り当てられません
   (例文 4b)　IP アドレスが取得できません
   この例では、対象 (IP アドレス) に対する認識が (例文 4a) と (例文 4b) では異なっている。(例文 4a) の発話者は「IP アドレスは与えられるものである」と考えている。一方、(例文 4b) の発話者は「IP アドレスは手に入れるものである」と考えている。この認識のずれによって、同じ内容を表わしているにもかかわらず、一般に同義語ではない用言「取得する」と「割り当てる」が用いられている。

[*1] http://vinelinux.org/ml.html(Vine linux に関心のある人たちが情報を交換しているメーリングリスト)

2.2 用言の同義語候補の推定

類似した内容を表現する文では、同義語関係にある用言はよく似た格要素をもつことがある。そこで、メーリングリストに投稿された質問メールとユーザの質問文で用いられる用言の格要素を比較して、よく似た格要素をもつ用言を同義語の候補として質問メールから取り出した [1]。

本研究では、同義関係にある用言を推定するために、ユーザの質問中にふくまれる用言 V_{user}(N 個の格要素 C_i をもつ) と検索対象の文に含まれる用言 V_{target} の類似度、すなわち同義語らしさ $Sim(V_{user}, V_{target})$ を以下のように定義した。

$$Sim(V_{user}, V_{target}) = \sum_{i=1}^{N} f_{dp}(C_i, V_{target}) \cdot isf(C_i)$$

ただし $f_{dp}(C_i, V_{target})$ は、体言 C_i およびその同義語を格要素としてもつ用言 V_{target} の数である。$isf(C_i)$ は体言 C_i の格要素としての重要度をあらわす値で、以下の式で表わす。

$$isf(C_i) = log\left(\frac{N}{sf(C_i)}\right)$$

N は検索対象の文の総数で、その中で体言 C_i を含む文の数が $sf(C_i)$ である。この同義語らしさ $Sim(V_{user}, V_{target})$ を用いて、ユーザが入力した質問文中の用言と同義語である可能性が高い用言を質問メールから取り出す。

3 ユーザの質問に問い返すのに有効な表現の抽出

方法や対処法について問う how 型の質問に応答するためには、問い返しが重要である。そこで、われわれは、メーリングリストに投稿された質問のメールから how 型の質問に応答するのに役立つ問い返し表現を取り出す研究を行った [3]。

メーリングリストに投稿された質問のメールを調査すると、問い返しに役立つ情報は以下の特徴を持つことがわかった。

- 問い返しに役立つ 6 種類の情報 (前提、症状、時間、目的、予想、例示) が表現されていた。また、それらの情報の表現には典型的な表現があった。図 1 に前提と症状の情報の典型的な表現を示す。
- 前提と症状の情報は、以下に示すように同じ表現で表現されることがある。
  - ··· + ました
  - ··· + のです/ なのです/ んです + が/けど
- 前提と目的の情報は、重要文とその前の文で表現されることが多い。一方、症状の情報は、重要文の後の文で表現されることが多い。

そこで、メーリングリストに投稿された質問メールの重要文とその前後の文から、ユーザの質問に問い返すための知識を以下の手順で取り出した。

1. 前提 (conditions)
   - ··· + ました/しました + (が/けど)
     (例) このたび、vine2.1.5 をインストール しました
   - ··· + います/ してあります + (が / けど)
     (例) vine2.1.5 を使用して います。
   - ··· + のです/なのです/んです + が/けど
     (例) 現在、ELECOM の LD-BBR4 というルータを使用している のですが
2. 症状 (symptom)
   - ··· + ません
     (例) このマシン、イーサーのカードが 1 枚しか入らなくて、イーサーのコネクターもないので、NIC の 2 枚ざしができ ません
   - ··· + ます/ました
     (例) kinput2 はうごいて ました が、jserver は死んで ました
   - ··· + のです/なのです/んです + が/けど
     (例) 漢字入力には成功している のですが ···

図 1 質問メールの重要文で表現されている問い返しに役立つ情報とその典型的な表現 (前提と症状の場合)

step 1 [質問メールからの重要文抽出]

メーリングリストに投稿された質問メールから重要文とその前後の文を表層表現を手がかりにして取り出す [2]。

step 2 [問い返し表現の抽出と意味ラベルの付与]

step1 で抽出した文に対して、問い返しに役立つ 6 種類の情報を含む表現を表層表現を手がかりにして取り出す。例えば、前提と症状の情報を含む表現は図 1 に示す表層表現を手がかりにして取り出す。さらに、その情報の種類を示す意味ラベル (前提、症状、時間、目的、予想、例示) を与える。ただし、取り出した表現が

- ··· + ました
- ··· + のです/ なのです/ んです + が/けど

である場合、その意味ラベルは次のように決める。

- その文が重要文かその前の文から取り出された場合、前提の意味ラベルを与える
- その文が重要文の後の文から取り出された場合、症状の意味ラベルを与える

4 問い返しを行う質問応答システム

作成したシステムの概要を図 2 に示す。作成したシステムは、ユーザの質問に対して以下のように問い返す。

1. ユーザの入力した質問文で用いられている用言の同義語を推定し、ユーザに示す。同義語候補の推定には、2 章で述べた手法を用いる。

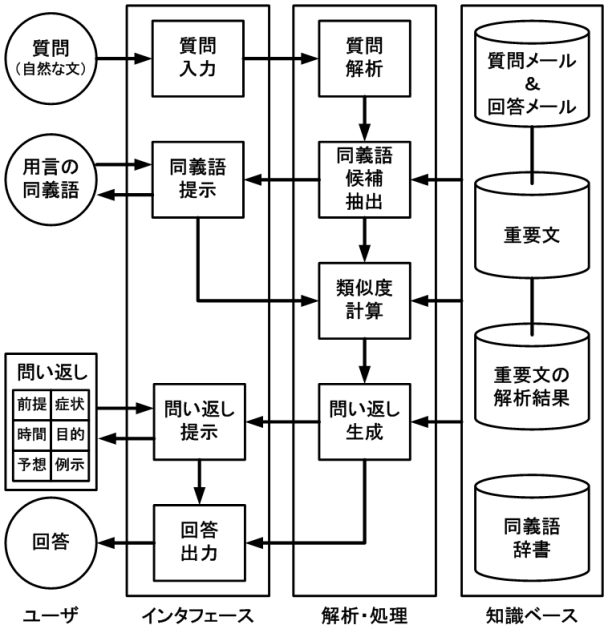

図 2　システムの概要

2. ユーザが選んだ同義語を用いて質問を拡張し、それと類似した内容の質問メールを取り出す。取り出した質問メールはその回答メールとあわせて質問に対する答えとする。
3. 答えと判定した質問メールから条件/症状などの表現を取り出し、問い返しとしてユーザに示す。
4. ユーザが選んだ問い返しに応じた質問メールとその回答メールを質問に対する回答としてユーザに示す。

システムを構成するモジュールの機能と内容や知識ベースの内容について以下に示す。インタフェースには Web ブラウザを用い、システムは Perl で実装されている。

質問入力モジュール　自然な文で表現されているユーザの質問を受けつけ、質問解析モジュールに送る。ただし、質問の内容にかかわらない文末表現、例えば、
- 〜について教えてください
- 〜について知りたい
- 〜方法を教えてください

などが含まれていた場合は、それらを取り除いておく。

質問解析モジュール　ユーザの質問文を対象に形態素解析および係り受け解析を行い、解析結果を同義語候補抽出モジュールに送る。形態素解析には JUMAN[4]、係り受け解析には KNP[5] を用いた。

同義語候補抽出モジュール　2 章で述べた手法で、ユーザの質問文で用いられる用言と質問メールの重要文で用いられている用言の類似度 (同義語らしさ) を計算し、上位 5 つまでの情報を同義語提示モジュールに送る。

同義語提示モジュール　同義語候補抽出モジュールが取り出した同義語候補をユーザに示し、同義語として適切なものを選ばせる。ユーザが指定した用言は類似度計算モジュールに送る。

質問メール & 回答メール　メーリングリストに投稿された質問メールとその回答メールが格納されている。

重要文　質問メールから取り出した問い合わせの中心になる文 (重要文) とその前後の文が格納されている。また、回答メールから取り出した重要文も格納されている。

重要文の解析結果　質問メールとその回答メールから取り出した重要文の形態素解析および係り受け解析の結果が格納されている。重要文の解析結果は同義語候補抽出モジュール、類似度計算モジュール、問い返し生成モジュールで利用される。

同義語辞書　同義語候補抽出モジュール、類似度計算モジュールで用いる同義語の辞書。名詞を中心に 519 語が登録されている。

類似度計算モジュール　ユーザの拡張された質問文と質問メールから取り出した文の類似度を、文の構文的な構造にもとづいて計算する [3]。

問い返し生成モジュール　類似度計算モジュールの計算結果にしたがって、ユーザの質問文に類似すると判定した質問メールを 10 個取り出し、回答出力モジュールに送る。取り出した質問メールから問い返しに利用する表現を 3 章で述べた手法を使って取り出し、問い返し提示モジュールに送る。

問い返し提示モジュール　問い返し生成モジュールによって生成された問い返し表現を意味ラベルごとにまとめてユーザに示し、条件や症状などが似ている問い返し表現を選ばせる。ユーザが選んだ問い返し表現が関連づけられている質問メールとその回答メールを回答出力モジュールに送る。

回答出力モジュール　ユーザの質問に類似すると判定した質問メールとその回答のメールを出力する。

5　問い返しの例と検討

提案手法を評価するために、作成した質問応答システムに「DHCP で IP を再取得できない」という質問を与えた。この質問は、Vine Users ML に類似したメーリングリスト Linux Users ML [*2]に 2001 年 10 月に実際に投稿された質問メールの標題である。このメールの本文には、「DHCP で IP を再取得できない」以外の情報も詳しく述べられている。

「DHCP で IP を再取得できない」という質問を与えると、システムはまず「再取得」の同義語の候補として「割り当てる」「設定」「もらう」「指定」「割り振る」の 5 つをユーザに問い返した。図 3 はシステムが用言の同義語候補をユーザに問い返している画面である。これらの中から「割り当てる」「もらう」「割り振る」の 3 つを同義語であると指示し、質問文を拡張した。

[*2] http://www.linux.or.jp/community/ml/linux-users/(Linux に関するユーザ同士の情報交換を目的としたメーリングリスト)

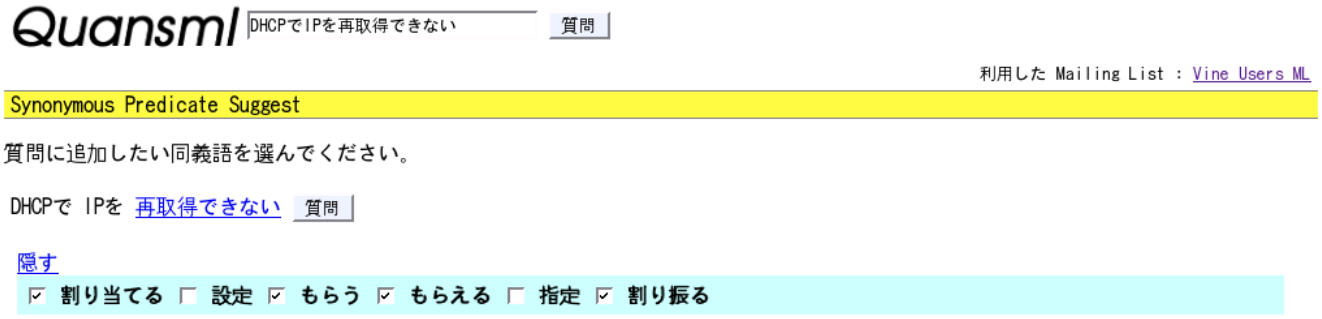

図 3　用言の同義語候補の問い返し

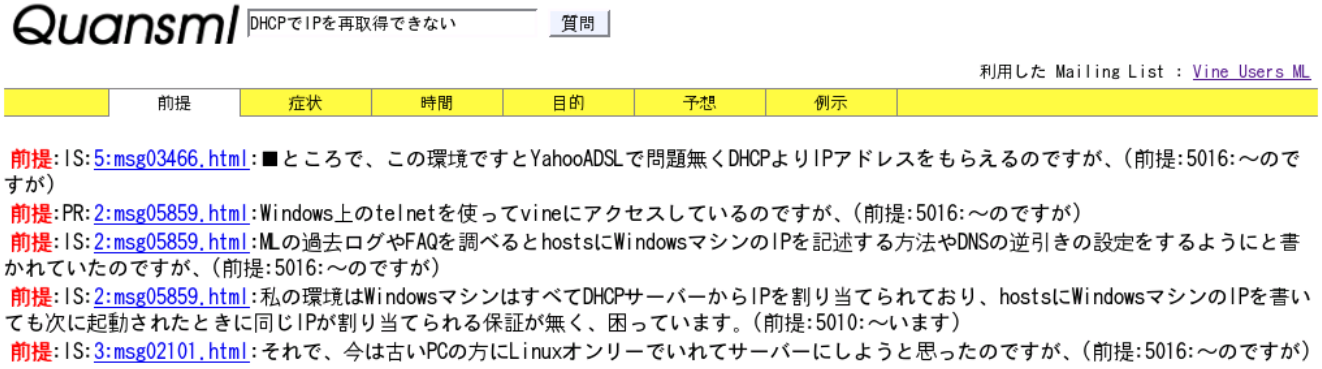

図 4　条件や症状などの問い返し

システムは拡張した質問で検索を行い、10 個のスレッド (質問メールとその回答メールの組) を答えに選び、そこから 39 個の問い返しを生成した。図 4 はシステムがユーザに「前提」を問い返した画面である。システムが生成した 39 個の問い返しを質問を作成するのに利用した質問メールの本文を用いて、以下の 3 種類の評価を与えた。

○　質問を作成するのに利用した質問メールで表現されている情報に一致あるいは関連する内容が問い返されているもの
△　問題の解決に参考になるかもしれない情報が問い返されているもの
×　問題の解決になる情報が問い返されていないもの

その結果、16 個の問い返しが○と評価された。さらに、問い返しと答えの関連を調べると○の評価が与えられた問い返し 16 個のうち適切な答えと関連づけられているものは 8 個であった。図 5 に○、△、×の評価の問い返しの例を示す。図 4 の上から 1 つ目の問い返しは、問い返しとして○の評価が与えられ、さらに適切な答え (ユーザの条件にあう答え) と関連づけられていた。

作成したシステムにさまざまな質問を与えてみたが、答えから取り出した重要文とその前後の文を利用するよりも問い返しを利用した方が早く簡単に適切な答えを見つけられることが多かった。この理由としては、問い返し表現は短くて読みやすく、ユーザの条件にあうかどうかを判断するのに重要な情報が述べられていることが考えられる。

例 1 (評価○)　ある CATV のグローバル契約で Internet を楽しんでいるのですが

例 2 (評価△)　ネットに接続するため NAT を選択したのですが

例 3 (評価×)　いろいろ調べたんですが

図 5　「DHCP で IP を再取得できない」に対する問い返しの評価の例

参考文献

[1] 西村, 渡辺, 岡田: 同義語を用いた質問文の拡張による係り受け関係の柔軟な照合, 情報処理学会研究報告, 2006–NL–176, (2006).
[2] 渡辺, 横溝, 西村, 岡田: 質問応答システムのための知識獲得, 自然言語処理, Vol.12 No.6, (2005).
[3] 西村, 渡辺, 岡田: あいまいな質問に問い返すためのメーリングリストを利用した知識獲得, NLC 言語理解とオントロジーシンポジウム, (2007).
[4] 黒橋, 河原: 日本語形態素解析システム JUMAN version 5.1 使用説明書, 京都大学, (2005)
[5] 黒橋, 河原: “日本語構文解析システム KNP version 2.0 使用説明書.”, 京都大学, (2005).