メーリングリストを利用した質問応答システムのための知識獲得

渡辺　靖彦　　園　　和也　　岡田　至弘

龍谷大学 理工学部 電子情報学科
〒 520-21 大津市瀬田大江町横谷 1-5

E-mail: {watanabe,okada}@rins.ryukoku.ac.jp, sono@mail433.elec.ryukoku.ac.jp

あらまし　本研究では、方法や対処法を問う質問 (how 型の質問) に質問応答システムが答えるための知識を、メーリングリストに投稿されたメールから獲得する方法について述べる。方法や対処法を問う質問に答えるための知識 (「こんな場合にはこうする」など) は、メーリングリストに投稿されたメールから質問や説明の中心になる文 (重要文) を取り出すことによって獲得する。Vine Users ML に投稿されたメールを対象にした実験を行い、メーリングリストに投稿されたメールから方法や対処法を問う質問に答えるための知識を獲得できることを示す。

キーワード　質問応答、メーリングリスト、重要文抽出

Knowledge Extraction for a Question Answer System Using Emails Posted to a Mailing List

Yasuhiko WATANABE, Kazuya SONO, and Yoshihiro OKADA

Dept. of Electronics and Informatics, Ryukoku University
Seta, Otsu, Shiga, Japan

E-mail: {watanabe,okada}@rins.ryukoku.ac.jp, sono@mail433.elec.ryukoku.ac.jp

Abstract　In this paper, we propose a method of extracting knowledge for a question answer system using emails posted to a mailing list. We first report a method of extracting important sentences from emails which were posted to a mailing list. Then, we show that important sentences extracted from emails posted to a mailing list (Vine Users ML) can be used as knowledge for answering how type questions.

Key words　question answer system, mailing list, sentence extraction

1. はじめに

自由に閲覧することができる電子化文書の数が膨大になるにつれ、その中からユーザが必要とする情報を効率的に獲得することが困難になってきている。このため、ユーザからの質問に対して明確な回答を自動的に提示する質問応答 (QA) 技術が注目されている。

質問応答に用いる知識を人工言語で記述した UC [1] などの質問応答システムでは、十分な記述力をもつ人工言語の設計のむずかしさ、知識ベースの高い作成コストといった問題があった。そこで、大量の電子化文書が利用可能になった 1990 年代からは、自然言語で記述された文書を質問応答システムの知識として利用しようとする研究が行われている [2]。近年では、TREC [3] や NTCIR [4] といった評価型ワークショップも行われ、新聞記事や WWW 文書などを知識として用いる質問応答システムの研究もさかんである。しかし、これらの研究の多くは事実を問う質問 (what 型の質問) を対象としていて、方法や対処法を問う質問 (how 型の質問) を扱うものは [5] [6] などまだ少ない。これは、事実を問う質問に答えるための知識に比べ、方法や対処法を問う質問に答えるための知識 (「こんな場合にはこうする」など) を獲得することがむずかしいからである。日笠らや清田らは、方法や対処法を問う質問に答えるための知識として FAQ 文書やサポート文書が利用できることを示した [5] [6]。しかしこれらの研究では、FAQ 文書やサポート文書がもつ文書構造を利用することを前提としていた。FAQ 文書やサポート文書以外の、より多くの文書を知識として利用するためには、文書構造以外の手がかりを利用する方法について研究しなければならない。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から方法や対処法を問う質問に答えるための知識を獲得する方法について述べる。

2. 方法や対処法を問う質問に答える質問応答システムで用いる知識

清田らは、パーソナルコンピュータの利用者を対象にした質問応答システムを作成していて [6]、そこで入力される質問を以下の 3 種類に分類している。

（1） what 型 (事実を問うもの)

（2） how 型 (方法を問うもの)

（3） symptom 型 (症状を示し、その対処法を問うもの)

自然言語文書を知識として利用する質問応答システムでは what 型の質問を取り扱うものが多く、how 型と symptom 型の質問を扱うものは少ない。方法や対処法を問う how 型や symptom 型の質問に答えるためには、「こんな場合 (条件) にはこうする (説明)」といった、条件と説明を組み合わせた知識が必要である。しかし、こうした知識を自然言語で記述された文書から取り出すのは、事実を問う what 型の質問に答えるための知識を取り出すのに比べてむずかしい。日笠らや清田らは、方法や対処法を問う質問に答えるための知識として、FAQ 文書やサポート文書が利用できることを示した [5] [6]。しかしこれらの研究では、FAQ 文書やサポート文書がもつ文書構造を利用することが前提としていた。FAQ 文書やサポート文書以外の、もっと多くの文書から「こんな場合 (条件) にはこうする (説明)」という知識を獲得するためには、文書構造以外の手がかりを用いる方法を検討する必要がある。

大量の電子化文書から知識を獲得する場合、抽出した知識が正しいかどうかという問題もある。質問応答システムの知識として利用することを前提に作成した文書であるなら、あるいは FAQ 文書やサポート文書のようなものならば、誤った情報がふくまれるおそれは少ない。しかし、インターネットで公開されている大量の電子化文書を質問応答システムの知識として利用する

場合、それらの中に誤った情報や矛盾した内容がふくまれるおそれは十分にある。したがって、それらの文書から取り出した知識が正しいかどうかについての情報も重要である。

質問に直接答えるための知識 (例えば、how 型の質問に対する「こんな場合にはこうする」という知識) 以外にも、質問応答システムにとって重要な知識がある。例えば質問応答システムでは、ユーザの質問の不明確さやあいまいさが問題になる。こうした質問には、システムがユーザに聞き返しを行うことが有効である [5] [6]。このため、どのような聞き返しを行うのかについての知識を用意することは重要である。

3. メーリングリストに投稿されたメールからの重要文の抽出

3.1 メーリングリストに投稿されたメール

メーリングリストには質問と回答のメールが繰り返し投稿されるものがある。たとえば、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 (Vine Users ML (注1)) では質問と回答のメールがさかんに投稿されている。われわれはこうしたメーリングリストに投稿されたメールから質問応答システムで用いる知識を獲得することを考えた。その有利さを以下に示す。

- 特定のドメインについての質問と回答の例を集めやすい
- あいまいな質問に対する聞き返しの例も集めやすい
- 情報のすばやい更新が期待できる
- 回答内容の確認が行われる
- 回答内容に誤りがあると、その誤りが指摘されることが多い

Vine Users ML に投稿されるメールを調査すると、以下の 4 種類に分けることができた。

(注1) : http://vinelinux.org/ml.html

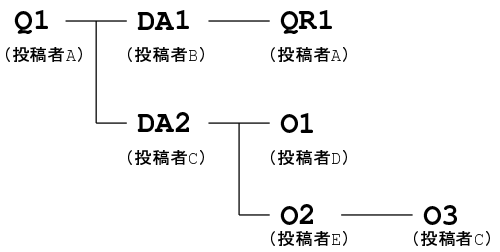

図 1　Vine Users ML に投稿されたメール間の参照関係の例

質問メール　ある問題について、最初に投稿される質問のメール (例:図 1 の Q1)。質問メールでの質問は、質問応答システムにおけるユーザの質問と同様に、その内容が不明確だったりあいまいな場合もある。

直接回答メール　質問メールに直接回答するメール (例 : 図 1 の DR1、DR2)。直接回答メールは、質問メールの質問にそのまま答える場合と、質問内容を聞き返す場合がある。

質問者返信メール　直接回答メールに質問メールの投稿者が直接返信するメール (例 : 図 1 の QR1)。質問者返信メールでは、直接回答メールでよせられた回答にしたがって行った作業の報告や聞き返しに対する回答が述べられている。

その他　(例 : 図 1 の O1、O2、O3)

Vine Users ML などのメーリングリストに投稿されたメールではさまざまな形式で質問や回答が表現されていて、FAQ 文書やサポート文書のような一定の文書構造がない。しかし、質問・説明の中心になる文があった。図 2 に示すメールの例では破線で囲まれた文が質問・説明の中心になる文である。こうした文を重要文とよぶことにする。質問メール、直接回答メール、質問者返信メールの重要文には次のような特徴がある。

(1)　質問メールの重要文は subject に

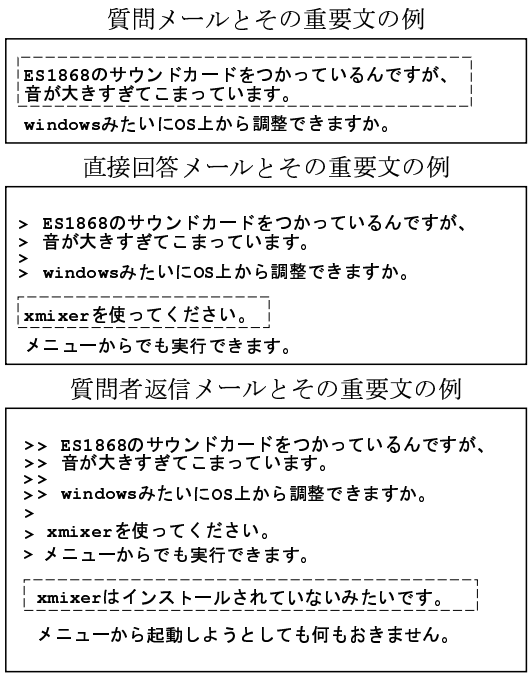

図 2　Vine Users ML に投稿されたメールと重要文の例 (破線で囲まれた文が重要文)

含まれる名詞および未定義語を含むことが多い。これは、質問メールの重要文も subject も、そのメールの質問内容のよい要約になっていることが多いからである。

（2）　それぞれのメールの重要文は、そのメールに直接返信しているメールで引用されることが多い。図 2 では、質問メールと直接回答メールの重要文がそれぞれ直接回答メールと質問者返信メールで引用されている。

（3）　それぞれのメールの重要文には典型的な表現がある。例えば質問メールの重要文には以下に示すような典型的な表現があった。

- 文末に「ません」「しょうか」「います」「ました」がある。

(例) Bluefish で日本語フォントの表示ができ ません。

- 文中に「困って」「トラブって」「ご指導」「？」がある。

(例) 数日前から一般ユーザログインで xstart できなくて 困って います。

- 行頭に＃がない。行頭の＃は、その行の記述については無視することを要請する記号である。

(例) ＃ とても初歩な質問でスミマセン

（4）　それぞれのメールの重要文は、本文のはじめに近い位置にあらわれることが多い。ただし、直接回答メールや質問者返信メールの重要文は、それらのメールが返信しているメールの重要文を引用している場合には、その引用している重要文の後にあらわれることが多い。図 2 の直接回答メールの例では、先頭の 4 行が引用文で、そこでは質問メールの重要文が引用されている。この引用のあとに、直接回答メールの重要文 (破線で囲まれた文) がある。

そこで、これらの特徴を手がかりにして以下の知識を獲得する。

- 質問メールと直接回答メールから取り出した重要文を用いて「この場合にはこうする」という知識を獲得する。
- 質問者返信メールから取り出した重要文を用いて「この場合にはこうする」という知識の正しさについての情報を獲得する。
- 質問メールと直接回答メールから取り出した重要文を用いて、あいまいな質問とそれに対する聞き返しの例を獲得する。

3.2　メーリングリストに投稿された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メールから重要文を抽出する処理の概要を図 3 に示す。前処理によってメールから取り出された文に対し 4 つの規則を適用して重要度を計算する。最も重要度が高い文を重要文として各メールから 1 文ずつ取り出す。

3.2.1　前　処　理

メールの各文の重要度を評価する前に、以下の前処理を行う。

step 1:　メーリングリストに投稿されたメールを対象に、メール間の参照関係および投稿者のメールアドレスを利用して、

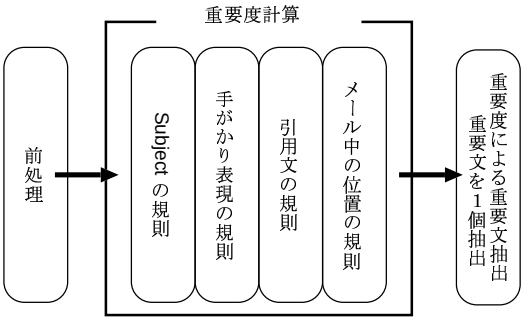

図 3　メーリングリストに投稿されたメールから重要文を取り出す処理

- 質問メール
- 直接回答メール
- 質問者返信メール

を取り出す。

step 2:　取り出したメールの本文を形態素解析する。ただし、以下のものは形態素解析を行う前に取り除く。

- # ではじまる行
- 引用記号 (例:>) ではじまる行
- (　) で囲まれている文字列

図 2 の直接回答メールの例では、先頭の 4 行を引用部分として取り除き、残りの 2 文について形態素解析を行う。また、「実行すると Segmentation fault (core dumped) してしまいます」という文の場合は、「(core dumped)」の部分をとりのぞいてから形態素解析を行う。形態素解析には JUMAN [7] を用いる。

step 3:　形態素解析を行った文が、そのメールに直接返信しているメールで何回引用されているか記述する。

step 4:　質問メールの subject を形態素解析し、その結果から名詞と未定義語を取り出す。

3.2.2　重要度の計算

質問メール、直接回答メール、および質問者返信メールから取り出した文に対し、以下の 4 つの規則を順に適用して重要度を計算する。そして、それぞれのメールで最も重要度が高い文を重要文として取り出す。

規則 1:　[subject の規則] この規則は、質問メールの本文から取り出した文にのみ適用する。subject に含まれている名詞・未定義語を含む文には 1 点を加える。

規則 2:　[手がかり表現の規則]

表 1 に示す手がかり表現を N 個含む文には N 点を加える。

規則 3:　[引用文の規則]

それぞれのメールに直接返信しているメールで引用回数が最多の文に 1 点を加える。

規則 4:　[位置の規則]

規則 1〜3 を適用した時点で最高の重要度が与えられている文が 2 つ以上ある場合、最も先頭に近い文に 1 点を加える。ただし、直接回答メールあるいは質問者返信メールで、それが返信しているメールの重要文を引用している場合は、その引用している重要文の後で最も先頭に近い文に 1 点を加える。

4.　重要文抽出の実験結果と検討

本研究では、Vine Users ML に投稿されたメールを対象に実験を行った。Vine Users ML に投稿されたメール 50846 通から

- 質問メール (8964 通)
- 直接回答メール (13094 通)
- 質問者返信メール (4276 通)

を取り出し、重要文抽出を行った。返信がある質問メールから無作為に 127 通を取り出し、それらの直接回答メール (184 通) と質問者回答メール (75 通) も取り出した。それらに対する重要文抽出の結果を表 2 に示す。重要文抽出に失敗した理由を以下に示す。

- 表 1 に示した手がかり表現を含まない重要文があった。
- 重要文ではない文で表 1 に示した手がかり表現を含む文があった。

表 1　メーリングリストに投稿されたメールからの重要文抽出に用いる手がかり表現

1. 質問メールからの重要文抽出に用いる手がかり表現

（1）「ません」「しょうか」「います」「ました」「?」で終わる文
（2）「困って」「トラブって」「ご指導」を含む文
（3）接続詞「が」「しかし」を含み、「ません」「しょうか」「います」「ました」で終わる文

2. 直接回答メールからの重要文抽出に用いる手がかり表現

（1）以下の表現で終わる文
- 「ますか」(していますか、どうなっていますか、など)
- 「ませんか」(ありませんか、いませんか、など)
- 「ですか」(いかがですか、ってことですか、ないですか、など)
- 「でしょうか」(どうでしょうか、いかがでしょうか、など)
- 「よね」(ますよね、ですよね、など)
- 「できます」「できません」「できています」「ないようです」「簡単です」「可能です」
- 「しました」「いません」「ます」(してます、います、あります、など)
- 「ください」
- 「いかがでしょう」
- 「すればよい」
- 「です」「はず」「と思う」「とか」

（2）以下の語を含む文
- 「あれば」「すれば」「ならば」「ときは」「したら」
- 「では」

3. 質問者返信メールからの重要文抽出に用いる手がかり表現

（1）以下の表現で終わる文
- 「です」「ました」(できました、いきました、なりました、など)
- 「ません」「だめでした」
- 「ありがとう」「ありがとうございました」
- 「ますか」「ます？」

表 2　重要文抽出の結果

メールの種類	正	誤	合計
質問メール	96	31	127
直接回答メール	153	31	184
質問者返信メール	45	30	75

- 質問あるいは回答の中心になる文が複数の文で構成されていて、それらのうち 1 文しか取り出せなかった。
- 重要文中に誤字・脱字があった。

つぎに、重要文抽出の結果が「こんな場合にはこうする」という条件と説明の知識として適切であるかどうか、

- 文のつながりが正しいかどうか
- その知識が問題解決に有効かどうか

という点に注意して検討を行った。例えば、以下の例では質問メール (質問 A) の重要文と直接回答メール (直接回答 A–1) の重要文とでは正しく文がつながっている。一方、(質問 A) と (直接回答 A–2) の重要文の間では文のつながりがない。しかし、(質問 A) と (直接回答 A–1) の知識は問題解決に役立つとして、この質問メールと回答メールからは有効な知識が獲得できたと判定した。

```
(質問 A) vedit は、存在しないファイルを
  |   ひらこうとするとコアはきますか
  ├(直接回答 A-1) はい、コアダンプしま
  |   す
  └(直接回答 A-2) 将来、GNOME はインス
      トール後すぐつかえるのですか？
```

127 個の質問メールとそれらの直接回答メールを調べると、92 例で有効な知識の獲得に成功し、35 例で失敗した。知識の獲得に失敗した原因を以下に示す。

- 質問メールからの重要文抽出に失敗

(質問 1) サウンドの設定でこまっています。
 ├(直接回答 1-1) まずは、sndconfig を実行してみてください。
 │ └(回答者返信 1-1) これでうまくいきました
 └(直接回答 1-2) sndconfig で、しあわせになりました。

(質問 2) パーティション設定時に SCSI ディスクが表示されないので、インストール
 │ できません。
 ├(直接回答 2-1) えーと、「パーティション設定時に SCSI ディスクが表示されない」
 │ というのは diskdruid での話でしょうか
 └(直接回答 2-2) typical problems に書いてある問題じゃないでしょうか

(質問 3) 1.0.6 のパッチはありますか。
 └(直接回答 3-1) gtk+-1.0.4 を利用するほうがいいでしょう。

(質問 4) ES1868 のサウンドカードをつかっていますが、音が大きすぎてこまってい
 │ ます。
 └(直接回答 4-1) xmixer を使って下さい。
  └(質問者返信 4-1) xmixer も xplaycd もインストールされていないみたいです。

(質問 5) いくつか問題がありますが、この件のレポートはどこに送ればいいのですか。
 └(直接回答 5-1) この ML で構いません。

(質問 6) これはどういう意味ですか。
 └(直接回答 6-1) ちゃんと質問しないと、だれも答えられません。

図 4　Vine Users ML に投稿されたメールからの重要文抽出によって獲得した、方法や対処法を問う質問に答えるための知識の例

した (21 例)

- 直接回答メールからの重要文抽出に失敗した (14 例)

質問メールからの重要文抽出に失敗したことが原因で知識の獲得に失敗した例はそれほど深刻ではない。誤って抽出した文の多くは質問文ではなく、質問応答システムでユーザの質問とマッチする可能性は低いからである。一方、直接回答メールからの重要文抽出に失敗したことが原因で知識の獲得に失敗した例はより深刻である。質問メールから取り出した文は質問文として適切で、質問応答システムでユーザの質問とマッチする可能性が高いからである。その場合、直接回答メールから誤って抽出した、回答や聞き返しとして不適切な文がユーザに示されるおそれがある。

図 4 に、Vine Users ML に投稿されたメールからの重要文抽出によって獲得した「この場合にはこうする」という知識の例を示す。

図 4 の質問メール (質問 1) には、2 つの直接回答メール (直接回答 1–1) と (直接回答 1–2) があった。どちらのメールでも質問者に sndconfig を使うことをすすめているが、(直接回答 1–1) はその内容が質問者返信メール (質問者返信 1–1) によって保証されている。方法や対処法を問う質問に答える時に回答候補が複数ある場合、(質問 1) の (直接回答 1–1) のように情報内容の保証

がある情報から優先して回答することができる。

図 4 の質問メール (質問 2) と (質問 3) からは、質問としてはあいまいで不完全な文が重要文として取り出されている。

(質問 2) のメールでは、ハードディスクのパーティションの設定についての質問が行われていた。しかし、質問者の質問そのものがあいまいであったため、そこから取り出した重要文もまたあいまいな内容になっていた。具体的には、質問者がどんなプログラムを利用してハードディスクのパーティションの設定したのかについての情報が欠ていた。これに対して、(直接回答 2–1) の回答者は、質問者が利用したプログラムが diskdruid であるかどうか聞き返している。この例を知識として用いれば、(質問 2) に類似するあいまいな質問に対して質問応答システムは、ユーザに diskdruid を利用したのかどうか聞き返すことができる。実験では、このようなあいまいな質問に対する聞き返しの例が 15 例あった。

(質問 3) では、gtk+についての質問が行われていた。この質問メールでの質問にはあいまいさはなかったが、質問の中心になる文が複数あった。そのうち 1 文だけを重要文として取り出したため、何について質問しているのかという情報 (この場合は、gtk+) が失われていた。しかし、(直接回答 3–1) から取り出した重要文がこの失われた情報を補っている。そこで、この例では (質問 3) からの重要文抽出には失敗と判定したが、(質問 3) と (直接回答 3–1) から抽出した重要文を組み合わせた知識については正しいと判定した。実験では、このような例が 10 例あった。

(質問 4) の質問に対する (直接回答 4–1) の回答は (質問 4) の質問者にとっては適切な内容ではなかった。(質問 4) の質問者は (直接回答 4–1) の回答内容を試し、問題が解決しなかったことを (質問者返信 4–1) で報告している。実験では、このように回答の誤り・不適切さを指摘する例が 4 例あった。

(質問 5) と (質問 6) では、分野に依存しない質問が行われている。したがって、これらの例はわれわれの方法が分野に依存したものではないことを示している。ただし、(質問 6) に対する (直接回答 6–1) の例はあまり丁寧な文ではない。このような用例を利用してシステムが回答すると、ユーザにそのシステムを利用しようとする意欲を失わせるおそれがある。

5. おわりに

本研究では、メーリングリストに投稿されたメールから方法や対処法を問う質問に答えるための知識を抽出する方法について述べた。すでにわれわれは、質問メールから取り出した重要文とユーザの質問のマッチングを行い、ユーザの質問に類似する質問メールの重要文を探し出す方法について検討している [8]。今後は今回獲得した知識を用いて聞き返しを行う方法について検討を行い、質問応答システムを作成する予定である。

参考文献

[1] Wilensky, Arens, Chin: “Talking to UNIX in English: An Overview of UC”, Communications of the ACM, 27(6), (1984)

[2] Hammond, Burke, Martin, Lytinen: “FAQ Finder: A Case-Based Approach to Knowledge Navigation”, 11th Conference on Artificial Intelligence for Application, (1995)

[3] TREC: http://trec.nist.gov/

[4] NTCIR: http://www.nlp.cs.ritsumei.ac.jp/qac/

[5] 日笠, 古河, 黒橋: 大学における計算機環境下での対話的ヘルプシステムの作成, 言語処理学会第 5 回年次大会,(1999)

[6] 清田, 黒橋, 木戸: 大規模テキスト知識ベースに基づく自動質問応答–話し言葉ナビ–, 言語処理学会 第 8 回年次大会, (2002)

[7] 黒橋, 長尾: 日本語形態素解析システム JUMAN version 3.61 使用説明書, 京都大学, (1998)

[8] 横溝, 渡辺, 岡田: メーリングリストを用いた質問応答システムの作成, 「情報アクセスのためのテキスト処理」シンポジウム, (2003)