社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2006－ＮＬ－174（10）
2006／7／28

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の内容確認

渡辺 靖彦† 西村 涼† 岡田 至弘†

† 龍谷大学 理工学部 情報メディア学科
〒520-2194 大津市瀬田大江町横谷 1-5
E-mail: †{watanabe,okada}@rins.ryukoku.ac.jp, †r_nishimura@afc.ryukoku.ac.jp

あらまし インターネットで公開されている大量の電子化文書には、誤った情報や矛盾した内容がふくまれている。したがって、それらの文書を質問応答システムの知識として利用する場合、文書中の誤りを検出する方法について検討することは重要である。われわれはこれまでに、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ことができることを明らかにした。本研究では、メーリングリストに投稿されたメールの中に含まれる誤った情報を検出する方法について述べる。
キーワード 質問応答、メーリングリスト、誤った情報

Confirmed Knowledge Acquisition Using Mails Posted to a Mailing List

Yasuhiko WATANABE†, Ryo NISHIMURA†, and Yoshihiro OKADA†

† Dept. of Media Informatics, Ryukoku University
Seta, Otsu, Shiga, Japan
E-mail: †{watanabe,okada}@rins.ryukoku.ac.jp, †r_nishimura@afc.ryukoku.ac.jp

Abstract In this paper, we first discuss a problem of developing a knowledge base by using natural language documents: wrong information in natural language documents. It is almost inevitable that natural language documents, especially web documents, contain wrong information. As a result, it is important to investigate a method of detecting and correcting wrong information in natural language documents when we develop a knowledge base by using them. In this paper, we report a method of detecting wrong information in mails posted to a mailing list and developing a knowledge base by using these mails.
Key words question answer system, mailing list, wrong information

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、知識ベースの作成コストの高さ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、こうした文書を質問応答システムの知識として利用する場合、その中に含まれる誤りが問題になる。以下の例は、メーリングリストに投稿された質問とその回答のメールの抜粋である。

(質問 1) WheelMouse を Netscape で使えるようにするには?
└ (直接回答 1) SD 誌 12 月号 (PlamoLinux 特集) に設定があります。
└ (質問者返信 1) 記事のとおりに設定したのですが、うまくいきません。

(質問 1) と (直接回答 1) から作成された「こんな場合 (条件) にはこうする (説明)」という知識を用いて質問応答システムが応

答すると、ユーザは (質問者返信 1) と同じ失敗をするおそれがある。このため、WWW 文書などから抽出した知識の内容を肯定/否定する情報がないか調べ、その知識の内容が正しいかどうかを確認することが重要になる。

自然言語で記述された文書に含まれる誤りが質問応答システムで特に問題になるのは、(質問 1) のような質問、すなわち、方法や対処法を問う質問 (how 型の質問) に答えるときである。what 型の質問の答えはふつう一つであるので、誤りを含む複数の答えが WWW 文書などから取り出されたとしても、それらの記述回数などを手がかりにして答えを一つに絞りこみ、誤りを含む答えを取り除くことはむずかしくない。一方、how 型の質問の答えは二つ以上あることが多いので、「こんな場合 (条件) にはこうする (説明)」という答えが二つ以上取り出されても、what 型の場合と同様の方法で誤りを含む答えを取り除くことはむずかしい。そこで、how 型の質問に答えるための知識を WWW 文書などから抽出する場合、その内容を肯定/否定する情報がないか調べ、知識の内容を確認することを考えた。これまでにわれわれは、方法や対処法を問う質問 (how 型の質問) に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取り出すことができることを示した [5]。本研究では、メーリングリストに投稿されたメールから獲得した知識の内容を確認する方法について述べる。

2. メーリングリストに投稿されるメール

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 [5]。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問い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。

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 DA1、DA2)。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、質問内容を問い返す場合がある。

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、

(注1)：http://vinelinux.org/ml.html

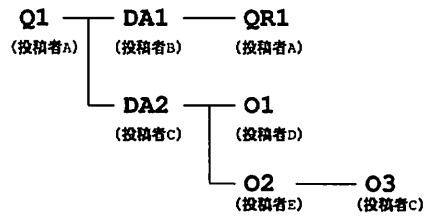

図 1　Vine Users ML に投稿されたメール間の参照関係の例

直接回答メールの内容にしたがって行った作業の報告や問い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)

方法や対処法を問う how 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。日笠らや清田らは、こうした知識を自然言語で記述された FAQ 文書やサポート文書から取り出すことができることを示した [6] [7]。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して知識を抽出していた。FAQ 文書やサポート文書以外の、もっと多くの文書から「こんな場合にはこうする」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。例えば、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。このような文書から「こんな場合にはこうする」知識を獲得するには、文書構造以外の手がかりを使って、質問・説明の中心になる文を取り出さなければならない。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)　質問メールおよび直接回答メールの重要文は、そのメールに直接返信しているメール (直接回答メールの場合は質問者返信メール) 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)　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

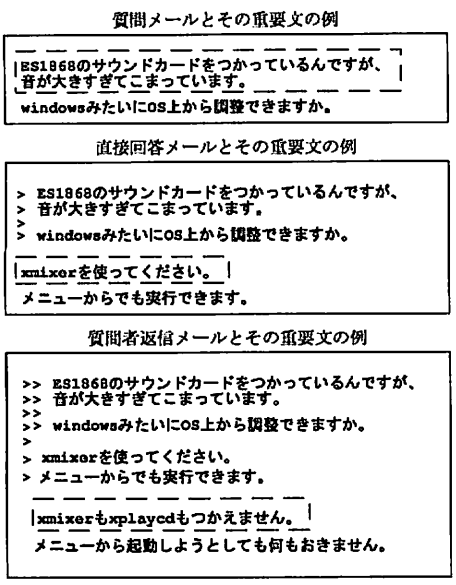

図 2 Vine Users ML に投稿されたメールと重要文の例 (破線で囲まれた文が重要文)

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

質問者応答メールの重要文には、以下のような典型的な表現があった。

- 文中に「できた」「できない」「使える」「使えません」「うまくいく」など、質問メールと直接回答メールの内容を肯定あるいは否定する表現がある。

(例) xmixer も xplaycd も つかえません。

- 文中に「ありがとう」「すみません」などがある。

(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では、それらのメールが返信しているメール (直接回答メールの場合では質問メール) の重要文を引用している場合、引用している重要文の後にそのメールの重要文が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。

われわれは、これらの特徴を手がかりにして質問メールと直接回答メールから重要文を取り出し、それらが how 型の質問に答えるための「こんな場合にはこうする」知識として利用できることを示した [5]。本研究では、how 型の質問に答えるために獲得した知識の内容を、質問者返信メールから取り出した重要文を用いて確認する方法について述べる。

3. メーリングリストに投稿されたメールの内容確認

メーリングリストに投稿されたメールを利用して how 型の質問に答えるための知識を獲得し、その内容を確認する方法を以下に示す。how 型の質問に答えるための知識は、質問メールとその直接回答メールから重要文を取り出し、それらを組み合わせて獲得する。獲得した知識の内容は、質問者返信メールから取り出した重要文を用いて確認する。

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、(1) 質問メール、(2) 直接回答メール、(3) 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、引用記号を手がかりにして先頭の 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 [8] を用いる。

step 3 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。

step 4 形態素解析を行った文について、その文が含まれるメールに直接返信しているメールの中で引用されている回数が最も多いものを調べる。

step 5 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、それぞれのメールから重要度が最も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適用する。subject に含まれている名詞・未定義語を含む文には、その重要度に 1 点を加える。

規則 2: [手がかり表現の規則]

表 1 および表 2 に示す手がかり表現を N 個含む文には、その重要度に N 点を加える。

規則 3: [引用文の規則]

この規則は、質問メールと直接回答メールの本文から取り出した文に適用する。メールの本文から取り出した文で、その文を含むメールに直接返信しているメールの中で引用されている回数が最も多いものには、その重要度に 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、先頭により近い文に 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 (直接回答メールの場合は質問メール) の重要文を引用している場合は、引用している重要文の後で先頭により近い文に 1 点を加える。

規則 1、2、4 は、新聞記事などを対象にして用いられている重要文抽出手法をメールに適用したものである [9]。一方、規則 3 は、引用が多用されるメールから重要文を抽出するための規則

表 1 質問メールおよび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. 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」

表 2 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

type 1 の表現

- できる、できた
- 解決する、解決した
- 使える、使えた
- うまくいく、うまくいった、うまくいきました

type 2 の表現

- できない、できなかった
- 解決しない、解決しなかった
- 使えない、使えなかった、使えません
- うまくいかない、うまくいかなかった
- だめです、だめでした

である。

step 6 質問および直接回答メールから取り出した重要文をくみあわせて「こんな場合 (条件) にはこうする (説明)」という知識とする。獲得した知識に対し、質問者返信メールから取り出した重要文に含まれる手がかり表現を利用して、以下の内容確認ラベルを与える。

positive ラベル 知識の内容を肯定する情報があることを示すラベル。質問者返信メールから取り出した重要文が表 2 に示す type 1 の手がかり表現を含む場合に、質問およびその直接回答メールから取り出した重要文を用いて作成した「こんな場合にはこうする」知識にこの内容確認ラベルを与える。

negative ラベル 知識の内容を否定する情報があることを示すラベル。質問者返信メールから取り出した重要文が表 2 に示す type 2 の手がかり表現を含む場合に、質問およびその直接回答メールから取り出した知識にこの内容確認ラベルを与える。

表 3 質問者返信メールから取り出した重要文を利用した内容確認の結果

判定結果	正	誤	合計
positive	35	18	53
negative	10	4	14
other	48	6	54

表 4 内容確認に失敗した例のタイプと件数

誤って判定したタイプ	正解のタイプ			
	positive	negative	other	total
positive	–	4	14	18
negative	2	–	2	4
other	4	2	–	6

表 5 適切と判断された「こんな場合にはこうする」知識に対する内容確認の結果

内容確認の結果	positive	negative	other	合計
正	29	8	27	64
誤	4	4	15	23

other ラベル 知識の内容を肯定/否定する情報が見つかっていないことを示すラベル。質問者返信メールから取り出した重要文が表 2 に示す type1 および type 2 の手がかり表現を含まない場合に、質問およびその直接回答メールから取り出した知識にこの内容確認ラベルを与える。

4. 内容確認の実験結果と検討

Vine Users ML に投稿されたメールから、直接回答メールをもつ質問メール 100 通を無作為に取り出した。この 100 通の質問メールには、質問者返信メールを少なくとも 1 通もつ直接回答メールが 121 通あった。この 121 組の質問および直接回答メールを対象に、それらの内容を確認する実験を行った。直接回答メールに対する質問者返信メールが複数ある場合は、最初に投稿された質問者返信メールを用いて内容確認を行った。

最初に、質問者返信メールが直接回答メールの内容を肯定しているのか否定しているのかを、質問者返信メールから取り出した重要文を用いて正しく判定できているか調べた。その結果を表 3 に示す。表 4 に内容確認に失敗した例のラベルのタイプと件数を示す。内容確認に失敗した原因は、質問者返信メールからの重要文抽出の失敗によるものであった。質問者返信メールからの重要文抽出の失敗の原因を以下に示す。

- 表 2 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 2 に示した手がかり表現を含む文があった。
- 内容の中心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、質問メールと直接回答メールから抽出した重要文を組み合わせたものが「こんな場合にはこうする」という how 型の質問に答えるための知識として適切であるかどうかを、取り出

(質問 2) 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 2-1) はい、コアダンプします
└ (直接回答 2-2) 将来、GNOME はインストール後すぐつかえるのですか？

(質問 3) サウンドの設定でこまっています。
├ (直接回答 3-1) まずは、sndconfig を実行してみてください。
| └ (質問者返信 3-1) これでうまくいきました
└ (直接回答 3-2) sndconfig で、しあわせになりました。

(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 もつかえません。

図 3 Vine Users ML に投稿されたメールからの重要文抽出の結果の例

した重要文のつながりが妥当かどうかという点に注意して検討した。図 3 の質問メール (質問 2) と 2 つの直接回答メール (直接回答 2-1)(直接回答 2-2) から取り出した重要文を例にして説明する。(質問 2) の重要文と (直接回答 2-1) の重要文は、文が正しくつながっている。そこで、(質問 2) と (直接回答 2-1) から取り出された知識は適切であると判定した。一方、(質問 2) の重要文と (直接回答 2-2) の重要文は、文が正しくつながっていない。そこで、(質問 2) と (直接回答 2-2) から取り出された知識は不適切であると判定した。この実験では、質問メールとその直接回答メールの重要文の組合わせ 121 例のうち 87 例が「こんな場合にはこうする」知識として適切であると判定された。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した (20 例)
- 直接回答メールからの重要文抽出に失敗した (14 例)

本研究で提案した方法は、質問メールと直接回答メールから重要文が正しく取り出されていることを前提としていて、その内容について質問者返信メールで肯定あるいは否定の情報があるかどうかを調べて示すものである。質問メールと直接回答メールからの重要文抽出が正しく行われているかどうかを確認してはいない。このため、図 3 の (質問 2) と (直接回答 2-2) から取り出した知識のように、how 型の質問に答えるのに不適切な知識を内容確認ラベルを利用して取り除くことはむずかしい。しかし、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が低いからである。一方、図 3 の (質問 2) と (直接回答 2-2) の場合のように、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取り出した、回答として不適切な文がユーザに示されるおそれがある。

最後に、質問メールと直接回答メールから取り出した重要文の組み合わせのうち、「こんな場合にはこうする」知識として適切と判断された 87 例について、その内容を正しく確認できたかどうか調査した。その結果、87 例中 64 例について適切な内容確認ラベルが与えられていた。その内訳を表 5 に示す。以下に positive ラベルと negative ラベルが与えられた例を示す。

図 3 の質問メール (質問 3) には、2 つの直接回答メール (直接回答 3-1) と (直接回答 3-2) 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。(直接回答 3-1) に対する質問者返信メール (質問者返信 3-1) から取り出した重要文が表 2 の type 1 の表現 (「うまくいきました」) を含むため、(質問 3) と (直接回答 3-1) の組み合わせに positive ラベルが与えられた。一方、(直接回答 3-2) には質問者返信メールがないので、内容確認ラベルは与えられなかった。how 型の質問に対する回答候補は複数個ある場合が多く、この場合のように内容確認ラベルがあると、回答をしぼりこむのに役立つ。

図 3 の (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。そこから取り出した重要文には表 2 の type 2 の表現 (「つかえません」) が含まれていたため、(質問 4) と (直接回答 4-1) から取り出した重要文の組合わせには negative ラベルが与えられた。

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] 渡辺, 横瀬, 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).
[6] 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)
[7] 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)
[8] 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)
[9] 奥村, 難波: テキスト自動要約に関する研究動向, 自然言語処理, Vol.6, No.6, (1999)