

メーリングリストに投稿されたメールを用いた問い合わせ文の生成

磯貝 直毅 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

{n_isogai,r_nishimura}@afc.ryukoku.ac.jp, {watanabe,okada}@rins.ryukoku.ac.jp

1 はじめに

情報検索システムや質問応答システムがどのような知識をもっているのかわからないと、ユーザが適切な質問をするのはむずかしい。このため、システムがユーザに問い返して、検索要求をより適切・簡単に表現できるように支援することは重要である。例えば、「パソコンから音が出ません」とユーザが質問した場合、システムが「VineLinux を使っていますか?」「スピーカーからは出ますか?」など検索要求で不足している情報を自然な疑問文で問い返せば、ユーザは情報の絞り込みや検索要求の訂正などを簡単に行なえる。

これまでにわれわれは、メーリングリストに投稿されたメールを用いて方法や対処法を問う質問 (how 型の質問) に答える質問応答システムを作成した。また、メーリングリストに投稿されたメールから質問・説明の中心になる文 (重要文) を表層表現を手がかりにして取り出せること、そして重要文とその前後の文に問い合わせ文を生成するのに重要な情報が表現されていることを明らかにした [1]。そこで本研究では、ユーザの検索要求で不足している情報を問い返すことを目的に、メーリングリストに投稿されたメールから自然な疑問形の問い合わせ文を生成する方法について報告する。問い合わせ文は、メーリングリストに投稿されたメールの重要文とその前後の文を利用して生成する。

2 提案手法

提案する手法の概要を図 1 に示す。メーリングリストに投稿されたメールから取り出した重要文とその前後の文 (以下では原文とよぶ) を以下の処理によって問い合わせ文に変換する。

1. 形態素解析および係り受け解析
2. 一人称から二人称への変換 (2.1 節)
3. 単文と複文の判定 (2.2 節)
4. 単文を用いた問い合わせ文の生成 (2.3 節)
5. 複文を用いた問い合わせ文の生成 (2.4 節)

なお、形態素解析には JUMAN [3]、構文解析には KNP [4] を用いた。

2.1 一人称から二人称への変換

原文 (重要文とその前後の文) に含まれる一人称「わたし」「私」「僕」「ぼく」を「あなた」に変換する。

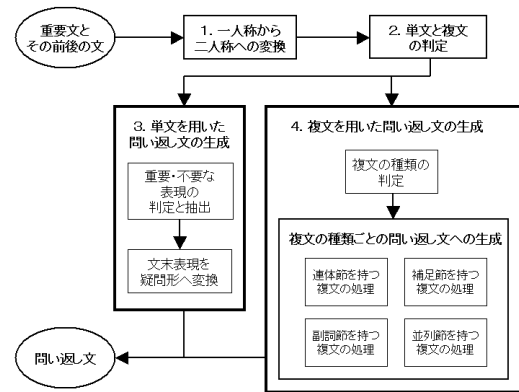


図 1 問い合わせ文の生成手法の概要

2.2 単文と複文の判定

原文 (重要文とその前後の文) に含まれる文を、以下の基準にしたがって単文と複文に区別する。

- 単文** 文中の用言の数が 1 つである文
複文 文中の用言の数が 2 つ以上である文

2.3 単文を用いた問い合わせ文の生成

2.3.1 問い合わせ文にとって重要・不要な表現の判定と抽出

原文 (重要文とその前後の文) から取り出した単文には、問い合わせ文で用いるには不要な表現がある。

(例 1) しかし vedit 上で xwnmo が作動しません。

(例 1) の「しかし」は、元々のメールでは文をなめらかにつなぐために重要な表現であったが、問い合わせ文で用いるには不要な表現である。一方、(例 1) の「vedit」や「xwnmo」などの専門用語や略語などは、問い合わせにおいて重要な役割をはたすことが多い。こうした専門用語が形態素解析用の辞書に登録されていないことが多く、それらは未定義語として扱われている。(例 1) の「vedit」や「xwnmo」も形態素解析結果では未定義語として扱われていた。

そこで、以下の手順で問い合わせ文にとって重要・不要な表現を判定し、重要な表現は取り出す。

step 1 文中にある用言を問い返し文にとって重要な表現として取り出す。

step 2 「ガ格」「ハ格(時相節を除く)」「ヲ格」「ニ格(時相節を除く)」を問い返し文にとって重要な表現として取り出す。

step 3 未定義語を格要素にとる「デ格」「カラ格」「ヘ格」「マデ格」「ヨリ格」がある場合、それらの中で用言に最も近いものを1つ問い返し文にとって重要な表現として取り出す。それ以外の格は、問い返し文にとって不要な表現と判定する。

step 4 取り出すと判定された格を修飾する文節は問い返し文にとって重要な表現として取り出す。また、取り出すと判定された格のうち、用言にもっとも近いものと用言の間に位置する文節も取り出す。

以上の処理の結果、(例 1) は (例 2a) のように変換される。

(例 2a) vedit 上で xwnmo が作動しません。

2.3.2 文末表現の疑問形への変換

2.3.1 節で述べた手法によって問い返し文にとって重要と判定されて取り出された表現を、その時制と否定の有無によって以下の4つに分類する。

- (1) 現在形肯定文
- (2) 過去形肯定文
- (3) 現在形否定文
- (4) 過去形否定文

それぞれの文における文末表現の疑問形への変換方法を以下に述べる。

(1) 現在形肯定文

文末に一番近い動詞、形容詞、判定詞、動詞性接尾辞、形容詞性述語接尾辞の活用を連用形に変え、その後に続く単語をとりのぞき、文末に「ますか?」をつける。以下に例を示す。

(例 3a) エラーメッセージが出て いる のですが、

(例 3b) エラーメッセージが出て いますか?

(例 3a) の場合、下線部の動詞性接尾辞「いる」を連用形「い」に変え、その後に続く「のですが、」をとりのぞき、文末に「ますか?」を加えることで (例 3b) にする。

(2) 過去形肯定文

文末に一番近い動詞、形容詞、判定詞、動詞性接尾辞、形容詞性述語接尾辞の活用を連用形に変え、その後に続く単語をとりのぞき、文末に「ましたか?」をつける。

(例 4a) ASUS の TX97-X では鳴って いた んですが、

(例 4b) ASUS の TX97-X では鳴って いましたか?

(例 4a) の場合、下線部の動詞性接尾辞「いた」を連用形「い」に変え、その後に続く「んですが、」を取り除き、文末に「ましたか?」を加えることで、(例 4b) にする。

(3) 現在形否定文

表 1 現在形否定文での変換処理

形態素の原形	変換内容
ない	なくなっているのですか?
ます + ぬ	ないのですか?
する + ます + ぬ	しないのですか?
ぬ	ないのですか?
いる、おる	いるのですか?
ある	あるのですか?
んだ	のですか?

表 2 過去形否定文での変換処理

形態素の原形	変換内容
ない	なくなっていたのですか?
ます + ぬ	なかったのですか?
する + ます + ぬ	しなかったのですか?
ぬ	なかったのですか?
いる、おる	いたのですか?
ある	あったのですか?
んだ	のですか?

用言を含む文節に含まれる形態素について、表 1 に示した変換を行う。変換した形態素の後に続く形態素はとりのぞく。

(例 5a) ppxp だけが使え ません が、

(例 5b) ppxp だけが使え ない ののですか?

(例 5a) の場合、下線部の「ません」が表 1 に示された表現で、これを「ないのですか?」に変換し、その後に続く「が、」を取り除くことで (例 5b) にする。また、(例 2a) は現在形否定文と判定され、(例 2b) のように変換される。

(例 2b) vedit 上で xwnmo が作動 しない ののですか?

(4) 過去形否定文

用言を含む文節に含まれる形態素について、表 2 に示した変換を行う。変換した形態素の後に続く形態素はとりのぞく。

(例 6a) メッセージが配信され ません でした

(例 6b) メッセージが配信され なかった ののですか?

(例 6a) の場合、下線部の「ません」が表 2 に示された表現で、これを「なかったのですか?」に変換し、その後に続く「でした」を取り除くことで (例 6b) にする。

2.4 複文を用いた問い返し文の生成

複文は用言を2つ以上含む文で、主節と接続節から構成されている。接続節は、その役割から、(1) 連体節、(2) 補足節、(3) 並列節、(4) 副詞節に分けられる [2]。以下では、それぞれの接続節をもつ複文について問い返し文への変換方法を述べる。

(1) 連体節を持つ複文と補足節を持つ複文の問い返し文への変換

名詞を修飾する接続節(連体節)および主節の用言を修飾する接続節(補足節)は、重要な表現として問い返し文で利用する。連体節あるいは接続節をもつ複文にはまず2.3.1節で述べた単文から問い返し文に重要な表現を抽出する処理(step 1~4)を適用し、その後、以下の処理を適用する。

step 5 連体節を問い返し文にとって重要な表現として取り出す。さらに、連体節を修飾する文節も取り出す。

step 6 補足節を問い返し文にとって重要な表現として取り出す。さらに、補足節を修飾する文節も取り出す。

取り出した表現に対して2.3.2節で述べた文末表現の疑問形への変換を行い、問い返し文を生成する。例えば、(例7a)には連体節「不要な」が含まれていて、その問い返し文が(例7b)である。

(例7a) プリンタから不要な紙が印刷されます

(例7b) プリンタから不要な紙が印刷されますか？

一方、(例8a)には補足節「管理させたい」が含まれていて、その問い返し文が(例8b)である。

(例8a) なるべくRPMで管理させたいと思っています。

(例8b) RPMで管理させたいと思っていますか？

(2) 並列節を持つ複文と副詞節を持つ複文の問い返し文への変換

(例9)は、主節と対等な関係にある接続節(並列節)を持つ複文である。

(例9) SCSIのホストアダプタは認識されるのに、HDDなどのSCSI機器が認識されません。

並列節は、主節の用言と接続節の用言との並列構造を手がかりに判定している。一方、(例10)は連用修飾の働きをする接続節(副詞節)を持つ複文である。

(例10) ES1868 互換のサウンドカードを使っていて、サウンドカードからスピーカーに直結しています。

本研究では、連体節、補足節、並列節と判定されなかった接続節を副詞節と判定している。

連用節や補足節に比べて、並列節と副詞節は主節、特にその用言との関係が弱い。このため、並列節と副詞節を問い返し文で利用するかどうかは、そこで表現されている情報の量によって決定する。具体的には、接続節(並列節あるいは副詞節)と主節に含まれている格の数でその接続節を問い返し文で利用するかどうかを決定する。

1. 接続節がもつ格要素が0~1個で、主節がもつ格要素が2個以上の場合

その接続節は問い返し文では利用しない。主節には2.3.1節で述べた処理(step 1~4)を適用し、取り出した表現に対して2.3.2節で述べた文末表現の

疑問形への変換を行い、問い返し文を生成する。

2. 接続節と主節それぞれが2個以上の格要素をもつ場合

主節と接続節とに文を分割し、それぞれを問い返し文に変換する。分割された主節と接続節には2.3.1節で述べた処理(step 1~4)を適用し、取り出した表現に対して2.3.2節で述べた文末表現の疑問形への変換を行い、問い返し文を生成する。

3. その他の場合

主節と接続節とに文を分割せず、そのまま問い返し文に変換する。接続節には2.3.1節で述べた処理(step 1~4)を適用し、問い返し文に重要な表現を抽出する。次に、主節に対して2.3.2節で述べた文末表現の疑問形への変換を行い、問い返し文を生成する。

以下に、並列節を持つ複文と副詞節を持つ複文を用いて生成した問い返し文の例を示す。

(例11a) この度ノートにあきたらず、自作デスクトップパソコンにVineをインストールしてみたのですが、

(例11b) 自作デスクトップパソコンにVineをインストールしてみましたか？

(例11a)は接続節に格要素が1つ、主節に格要素が2つある複文である。情報量の少ない接続節「この度ノートにあきたらず、」を捨て、主節だけを用いて問い返し文に変換したのが(例11b)である。

(例12a) liloの設定をwindowsをdefaultにしておいたのですが、update後はlinuxがdefaultで起動してしまいます。

(例12b) liloの設定をwindowsをdefaultにしておきましたか？

(例12c) update後はlinuxがdefaultで起動してしまいますか？

(例12d) liloの設定をwindowsをdefaultにしておいたのですが、update後はlinuxがdefaultで起動してしまいますか？

(例12a)は、接続節、主節ともに格要素が2つ以上ある複文である。この複文を接続節と主節に分割し、問い返し文に変換したのがそれぞれ(例12b)と(例12c)である。一方、(例12d)は、接続節と主節に分割せずに、問い返し文に変換したものである。接続節、主節ともに情報量が多いため、(例12b)と(例12c)に比べ、問い返し文としてはわかりにくい。われわれは、1つの文を2つに分割することで主節と接続節の関係が失われることよりも、問い返し文がユーザにとってわかりやすいことの方が重要であると考えている。

3 実験結果と評価

2章で述べた手法により、メーリングリストに投稿されたメールを利用して問い返し文の生成を行い、その結果を

表3 生成した問い返し文の評価結果

評価	単文を利用した 問い返し文の生成	複文を利用した 問い返し文の生成
○	79	406
△	0	4
×	1	37
合計	80	447

評価した。実験には、Vine Users ML に投稿されたメールから取り出した重要文とその前後の文 (単文が 4746 文、複文が 5045 文) を用いた [1]。問い返し文への変換結果は、以下の 3 段階で評価をした。

- 自然な疑問形の問い返し文が生成されている。
- △ 問い返し文に不要な部分が残っている。または、文末表現が自然な疑問形に変換されていない。
- × 問い返しで重要な表現が変換過程で失われている。

3.1 単文を利用した問い返し文の生成

4746 文の単文から 100 文を無作為に抽出した。その中の 20 文は、挨拶文 (19 文) や抽出に失敗した文 (1 文) といった問い返し文の生成には役立たない文だったので、あらかじめ人手で取り除いた。以下に、取り除いた挨拶文や抽出に失敗した文の例を示す。

- (例 13a) 宜しくお願いします。(挨拶)
- (例 13b) いきなり質問で恐縮なのですが、(挨拶)
- (例 13c) がなのですが、(抽出失敗)

残りの 80 文を用いて問い返し文を生成した。その評価結果を表 3 に示す。×と評価した文は、形態素解析の誤りによる失敗であった。

3.2 複文を利用した問い返し文の生成

複文を利用した問い返し文の生成の実験は、文中に用言を 2 つ含む 5045 文を用いた。その内訳は、連体節を持つ複文 669 文、補足節を持つ複文 2846 文、並列節を持つ複文 301 文、副詞節を持つ複文 1229 文であった。これらの複文から 500 文を無作為に取り出し、挨拶文 (43 文) や抽出に失敗した文 (10 文) を取り除き、残りの 447 文を対象に問い返し文に変換する実験を行った。実験結果を表 3 に示す。

表 3 で△と評価した 4 文は、(例 14) のように文末表現の疑問形への変換が不十分と判定されたものだった。

- (例 14a) sndconfig で自動認識して難なく使えておりました。
- (例 14b) sndconfig で自動認識して難なく使えておりましたか？

連体節を持つ複文と補足節を持つ複文を利用した場合は、自然な疑問形の問い返し文が生成されている (○) と

判定されたものが多かった。一方、並列節を持つ複文と副詞節を持つ複文を利用した場合は、

- 接続節がもつ格要素が 0 ~ 1 個で、主節がもつ格要素が 2 個以上の場合
- 接続節と主節それぞれが 2 個以上の格要素をもつ場合

では、問い返しで重要な表現が変換過程で失われている (×) と判定された例が、その他の場合に比べて多かった。本実験では、格の数のみで接続節 (並列節および副詞節) と主節で表現されている情報の重要性を推定している。今後は、格要素の内容、格の出現する位置なども利用してそれぞれの格で表現されている情報の重要性を推定することを検討している。また、取り出す表現の選択に失敗した例もあった。

(例 15a) 変更すると、BS で 1 文字前、DEL でカーソルの文字を削除します。

(例 15b) BS で DEL でカーソルの文字を削除します

(例 15a) では、「1 文字前、」が取り除かれてしまい、(例 15b) のような問い返し文に変換されてしまった。

主節と接続節を分割したことで、時制の表現が問題になる例もあった。

(例 16a) 私は Live! にサードパーティーのドーターボードをつないで光出力でアンプにつなげてるのですが、

(例 16b) あなたは Live! にサードパーティーのドーターボードをつなぎますか？

(例 16c) アンプにつなげてますか？

(例 16a) では、「ドーターボードをつなぐ」ことも「アンプにつなぐ」ことも過去に「つなぐ」ことが行われ、現在もその状態が継続していることが表現されている。(例 16b) と (例 16c) は、接続節と主節を分割して問い返し文へと変換した結果であるが、(例 16c) では、(例 16a) と同様に、「アンプにつなぐ」ことは過去に行われ、現在もその状態が継続していることが表現されている。一方、(例 16b) では、(例 16a) で表現されていた時とは異なり、これから「ドーターボードをつなぐ」こととして表現されている。

参考文献

- [1] 西村 涼, 渡辺 靖彦, 岡田 至弘, “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”, 言語処理学会第 13 回年次大会, E5-2, (2007).
- [2] 益岡 隆志, “24 週日本語文法ツアー”, くろしお出版, (2007)
- [3] 黒橋 禎夫, 河原 大輔, “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)
- [4] 黒橋 禎夫, 河原 大輔, “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).