質問応答集における質問文の標準形への自動変換

杉水流 英樹

北陸先端科学技術大学院大学 情報科学研究科

1999 年 2 月 15 日

キーワード： FAQ, 標準形, テキストの編集, 情報抽出, 要約.

現在、ネットワーク上では代表的な質問とその解答を集めた FAQ(Frequently Asked Question) と呼ばれる質問応答集を数多く見ることが出来る。通常、このような FAQ は人間の手によってまとめられているのだが、本研究室では自動編集プロジェクトの一環として、この質問応答集をネットニュースグループ fj.sys.sun から自動的に作り出す研究を行なっており、その成果を質問応答パッケージ (Sun QA-Pack) として公開している。

Sun QA-Pack では、fj.sys.sun の質問記事を要約したサマリー文を見出しとして表示している。しかし、サマリー文は元の質問記事から質問に関する重要文を抽出したものであるため、その文の表現は元の記事を書いた人間に依存しており、サマリーが全体として統一の取れた文章になっていないという問題がある。

本研究では Sun QA-Pack のサマリー文を対象として、質問文を標準形へと変換する手法を研究した。標準形とは質問文をその内容ごとに統一して表したもので、異なる表現で書かれた同一内容の質問文は、標準形に変換することで全く同じ形で表現される。質問文を標準形に変換することによりテキストの表示に統一性が生まれ、より高いレベルでのテキスト編集が達成できる。

本研究ではサマリー文の特徴を調査し、サマリー文をその内容から「したい」「できない」「教えて下さい」「状況説明」の 4 種類のタイプに分類した。これらのタイプを元に本研究で設定した標準形の基本型は次の 2 つの形である。

1. （名詞句）を（動詞）したい
2. （名詞句）が（動詞）できない

「したい」型のサマリー文は 1. の標準形に、「できない」型のサマリー文は 2. の標準形にそれぞれ変換する。「教えて下さい」型のサマリー文は 1. の標準形の動詞部分を「知

る」として、「（名詞句） を 知りたい」という形で標準形に変換する。状況説明型のサマリー文には明確な特徴がないため、本研究での標準化の対象外とした。

本研究で作成した質問文の標準化システムは、入力文整形モジュール、標準化モジュール、出力文選択モジュールの 3 つのモジュールから構成されている。

入力文整形モジュールは、サマリー文に対して専門用語タグの削除と文分割を行ない、サマリー文を整形して次の標準化モジュールに渡す。専門用語タグとは QA-Pack の自動分類に用いられている専門用語を示すタグであるが、本研究では必要ないため削除する。また、サマリー文には複数の文から構成されているものがあるので、句点での文分割を行なう。

標準化モジュールでは、まず入力文の文末表現からその文のタイプを判定し、次にそのタイプから適用する標準化ルールを決定する。標準化ルールは、質問文中に含まれる特定の表現と、質問文の標準形との関係をルール化したもので、本システムでは 8 種類の標準化ルールを用いている。なお、標準化ルールを適用する際には、日本語形態素解析システム Juman を用いて入力文を形態素解析し、その情報を利用した。

標準形への変換は、重要な動詞とその動詞に係る目的語を抽出した後、標準化ルールに従って整形することで実現する。文中における重要動詞は、質問文に含まれる意志や否定などの表現から位置を推測し、抽出する。重要動詞に係る目的語は、動詞の前に「を」「について」「が」などの目的語となる助詞を発見し、品詞・品詞細分類などの文法上の特徴を用いて抽出する。

標準化モジュールはサマリー文 1 文に対して標準形を 1 つ出力するため、サマリーが複数の文から構成されている場合は複数の標準形が出力される。出力文整形モジュールは、複数の標準形が出力された場合、その中で最も適切な標準形を選択して出力する。

本研究で作成したシステムについて、ネットニュースグループ fj.sys.sun に投稿された質問記事から抽出したサマリー文を対象として、評価実験を行なった。実験は、本研究での標準化の対象とはしなかった状況説明型のサマリーと、内容が不明瞭なサマリーを除外した 215 件のサマリーに対して行なった。本研究の中心モジュールである標準化モジュールのカバレジは 67%となり、そのうち正しい標準形に変換できたものが 124 文 (70%)、誤って変換したものが 53 文 (30%) であった。この正解率はニュース記事のクオリティを考慮すると悪くない結果と言える。