自然言語処理 117－22
（1997．1．21）

電子メールを用いた日本語文による質問応答システムにおける類似質問の抽出について

堤　豊
（九州帝京短期大学，九州大学）

牛島和夫
（九州大学）

本稿では，コンピュータによる自然言語質問文の自動回答への一つのアプローチとして，質問文を文の形から類似検索する方式について述べ，その有効性について論じる．従来，質問文を構文解析および意味解析し，対応する回答を生成する方法が提案されている．しかし，これらの方法では，膨大な解析の手間と，知識データベース作成の手間がかかる．また，限られた分野でしか利用できない．そこで，質問文をデータベース中に蓄積し，新規の質問文をデータベースと照合し，過去の質問とその答えのペアを返信する質問文の類似検索方式を提案する．類似検索方式では，形態素解析と簡単な意味抽出を行っているが，処理は軽い．実験では，短大の情報処理実習科目で質問メールを対象にシステムを稼働し，検索のヒット率を求めた．その結果，類似検索方式を有効に使うことにより，学生および教師の双方での負担が減ることが期待される．

A Method for Extracting a Similar Query in a QA System allowing Japanese Sentences based on E-mail

Yutaka Tsutsumi
(Kyushu Teikyo Junior College, Kyushu University)

Kazuo Ushijima
(Kyushu University)

This paper describes an automatic answering method to allow students to send questions by email in Japanese, and discusses its effectivity. There are some conventional methods which analyse input query sentences syntactically and semantically, then generate appropriate answers. These methods require much cost to make knowledge-bases of target areas and to analyze sentences. We propose a similarity based retrieval method for natural language questions. This method uses morphological analysis and simple semantic processing, but it is not heavy task. We made an experiment in a subject of information processing training, collecting email questions, and calculated a hit ratio of the retrieval. As the result, it is expected that both of students and teachers load will be reduced by using the similarly question retrieval method.

1. はじめに

コンピュータ通信網の発達とともに，遠隔授業への取り組みが活発に行なわれている．これは，高速通信回線を介して，サーバから授業を送信し，学生が端末で学習するというものである．この利点として，授業をサーバに保存しておき，いつでも自由な時間に学生が学習でき，分からないときは何度でも授業を受けなおすことができることなどがあげられる．学生から教師への質問手段としては，音声を使って即時に双方向で会話ができるものや，電子メールを使ったものなどが考えられているが，即時に会話ができるものは，本来自由な時間に学習できるという利点を生かせないことになるので，電子メールを使った質問という方式が有効であると考えられる．

しかし，電子メールを用いた質問システムでは，教師が同じ質問に対して，何度も回答を書かなければならない．また，教師が不在の場合には，回答が遅くなるなどの問題がある．そこで，本稿では，電子メールによる質問と回答のペアをデータベースに蓄えておき，学生が発した質問をデータベースから検索し，類似した質問があれば，教師を介さずに直接回答する方式を提案する．これにより，教師の手間は削減される．また学生側から見ても，回答されるまでの時間が短縮されるというメリットがある．

2. 関連研究

学生から来た質問をキーとして，回答とともにデータベースに蓄えておき，それを学生が検索するという方式では，一般的には，アクセス言語としてはSQL[1]などの形式的な言語が使われる．しかし，一般人が使うには，複雑で覚えにくい．そこで，自然言語を使ってデータベースにアクセスする自然言語インタフェース方式が提案されている[2][3]．複雑な検索の絞りこみや，誤った入力の訂正などのために，対話的に検索を進めるシステムも研究されている[4]．これらの方式は，いずれも形式的な言語を自然言語に置き換えているので，学生は，自由に質問文を書くことはできない．

より自由な自然言語文を取り扱う試みとして，質問文を意味解析し，あらかじめ用意された知識ベースから回答を生成するものがある[5]．しかし，この方式では，矛盾のない知識ベースの作成が不可欠であり，また，分野毎に知識ベースを作成しておかなければならないため，現実的には，限られた分野にのみ適用が可能である．また意味解析では，曖昧さの問題が避けて通れないため，処理時間がかかるのも欠点である．

本稿で提案する類似質問検索方式は，質問と回答をデータベースに蓄えておき，かつ，自然言語で質問が可能という点で従来の方式とは異なっている．著者らは，日英翻訳支援という観点から類似用例の検索方法について研究してきた[6]．ここでは，構文が類似している用例を探すということを目的としていた．本稿での類似質問検索方式では，質問の主題に着目して，類似質問を検索する．

3. 類似質問検索方式

3.1 質問メールの特徴

本稿で対象としている，授業で学生の発する質問メールについて，どのような特徴があるかを調べるために，予備調査を行った．対象は，短期大学学生68名で，電子メールの授業に対する質問を電子メールで送らせた．サンプルを図1に示す．

[仮定] 質問メールは，状況説明と，質問の主題および，回答依頼の3つの部分から成り立っている．

例えば，次の質問メール，「今日の授業で行った電子メールについてですが，受信や送信の方法について詳細を教えて下さい.」という文では，「今日の授業で行った電子メールについて」が状況説明であり，「受信や送信の方法」が質問の主題，「教えて下さい」が回答依頼となっている．

3.2 システムの概要

図2にシステムの概要を示す．

命令文の一覧表みたいなものはありますか?
入力方法はローマ字入力しかできないのですか?
今日送られてきたメールは印刷することができますか。
メールを書く時，はてなマークはどうやって使うのですか。

図 1: 質問メールのサンプル

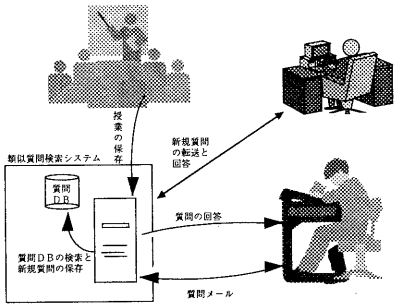

図 2: システム概要

表 1: 類義語表

方法，やりかた，仕方
こと，事
とき，時
なぜ，どうして，なんで
わけ，訳，理由

学生は授業の中で質問があれば，いつでも電子メールで質問をすることができる．この質問メールは，いったん類似質問検索システムに送られ，そこでデータベースに蓄えられた質問との間で類似検索を行う．もし，類似する質問があれば，その質問と回答のペアが学生に送られる．もし類似質問が検索できなければ，質問メールは教師に転送される．学生に送られた類似質問と回答のペアが学生が意図したもので無い場合には，学生は，新たに質問の表現を変更してもう一度質問をしなおすか，あるいは直接教師に質問することができる．教師が回答した質問は，その回答とともに新たに質問データベースに加えられる．

図3に類似検索システムの構成を示す．図で示すように，データベースへの登録時と検索時では，標準形式への変換までの部分を共有している．

3.3 形態素解析

入力として与えられる文を形態素解析する．現在は，形態素解析のために，京都大学で開発されたJUMAN[7]を使っている．また形態素解析用の辞書としてJUMAN用に開発された辞書を利用している．

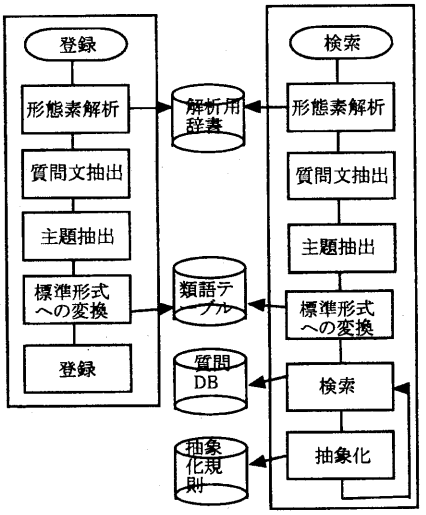

図 3: 類似検索システム

3.4 質問文抽出

一般には，質問文を抽出するためには，終助詞の「か」を検出することや，疑問の意図がある動詞のチェックなどを行い対象となる文が質問文であるかどうかを判定している．しかし，我々が対象としている質問メールに関しては，単純な処理で質問文の抽出が可能と考えた．そこで，前節で行った予備調査のデータについて文末で分類した結果を表2 に示す．

このように，文末をチェックすれば，予備調査のデータに関しては，質問文を漏らさず抽出できることが分かった．もちろん，実験で得られたデータ数が多くないため，より大きいデー

表 2: 予備調査の結果

分類	出現数
文末が?で終わる	14
文末が「か」で終わる	20
文末の動詞が「教える」	12
文末が「分かりません」で終わる	13
質問ではないメール	9
合計	68

タで行えば上記に含まれない文例も出現することは容易に想像がつくが，それらについては，相対的に数が少ないので，運用しながら改善していくことにした．

3.5 主題抽出

本システムでは，質問の主題が一致したものについてデータベースから検索する方式を採っている．したがって，質問文から，どこが主題なのかを抽出する必要がある．先に述べたように，質問文を状況説明，主題，回答依頼の3つの部分に分けるとすると，質問文から状況説明と回答依頼を除いたものが主題ということになる．回答依頼については前節で求めた質問文かどうかの判定条件がそのまま利用できる．状況説明については，例えば，「今日は，電子メールを行ったのですが，送信方法が分かりません」については，状況説明は「今日は，電子メールを行った」であり，主題は「送信方法」となる．

主題中の単語が表1の類義語表にあれば，各項目の先頭の単語で置き換えられる．これは，表現を統一しておくことで，検索時に似た単語を探す手間を省くことができるからである．従って，完全に置き換えても副作用がない単語のみを登録している．

3.6 データ格納

質問文をデータベースに登録する．データベースの見出しとして，前節で抽出した主題を用いる．すなわち，形態素解析で単語に分割された主題部分がキーワードとして登録される．また，見出しに関連付けられた内容は，質問文全体と回答のペアである．

3.7 データ検索

データベース中の質問文と，新規の質問文が似ているかどうかは，データベースへの照合と，質問文の変形とを繰返しながら行う．ここでは，これを抽象化と呼ぶ．

最初に，質問の主題について，データベース中の主題で完全に一致するものがあるかを調べる．この作業は，質問の主題中に含まれる各単語がすべて含まれる文を検索する．該当する文があれば，対応する質問文と回答を返送する．該当する文がなければ，質問の主題を図4の規則の順に抽象化しながら，照合を繰り返す．もし，最後の規則まで適用しても照合が成功しなければ，検索失敗である．この抽象化の規則は，質問の主題のうち，質問の内容にそれほど関わりのない要素から順に省いていくように設計している．

(1) **連体詞，指示詞，接尾辞など，文の構造に関与せず，かつ内容語でもない単語の省略**
　(例)この学校では電子メールが使えるのですか→学校では電子メールが使えるのですか
(2) **代名詞の省略**
　(例)私のパソコンでも電子メールができますか→～のパソコンでも電子メールができますか
(3) **連用修飾語の省略**
　(例) 電子メールがとても早く届くのには理由がある→電子メールが早く届くのには理由がある
(4) **連体修飾語の省略** (ただし[名詞]「の」[名詞]という形を除く)
　(例) インターネットの上手な活用方法→インターネットの活用方法
(5) **形式名詞の省略**
　(例) インターネットのこと→インターネット
(6)[名詞1]「の」[名詞2]を[名詞2]にする．ただし，[名詞2]が「方法」のときは，[名詞1]を残す．
　(例) 電子メールの送信→送信
　(例) 印刷の方法→印刷

図 4: 抽象化の規則

4. 評価実験

類似質問検索部がうまく検索できるか評価するために，実験を行った．実験方法は，授業で，被験者55名に同一内容の質問をさせ，最初の質問をデータベースに登録し，残りの質問を類似質問検索した．類似質問検索した後，質問メールはデータベースに登録する．従って，類似検索質問をするたびにデータベースに含まれる質問文の数は増えていく．

評価は，1つでも類似質問検索できた場合を類似質問検索成功とする．1つも類似質問検索の出力がない場合を類似質問検索失敗とする．

4.1 実験結果

類似質問検索実験の出力例を付録に示す．また，課題番号別に類似質問検索を行った結果を表3に示す．

失敗の原因を分類したのが表4である．複雑な文としては，次の(例1)のようなものがある．

(例1) 図と図が重なっていてその片方の図の中に線が入らないようにするにはどうしたらい

表 3: 実験結果

質問番号	類似質問検索成功	類似質問検索失敗	合計	類似質問検索成功率
1	44	9	53	83%
2	26	28	54	48%
3	32	21	53	60%
4	32	21	53	60%
5	32	21	53	60%
6	18	8	26	69%
7	29	24	53	55%
8	41	13	54	76%
9	28	24	52	54%
10	34	11	45	76%
11	34	14	48	71%
12	35	15	50	70%
13	27	24	51	53%
合計	412	233	645	64%

表 4: 検索失敗の原因

原因	出現数
形態素解析誤り	83
複雑な文	52
データ不足	46
抽象化失敗	28
2文に分割	24
合計	233

いですか。

2文に分割というのは，次の(例2)のように質問文が2文から構成されているもので，現在のシステムでは質問の主題を抽出できないものである．

(例2) かな漢字変換の方式が難しかった．もう一度教えて下さい．

4.2 応答時間

電子メールでの質問においては，質問をタイプする時間や質問が送られるのに費やされる時間などと比べ極端に遅くなければ，問題がないと考えられる．図5に応答時間を示す．

5. 考察

5.1 他方式との比較

オンライン授業での質問システムという面から，次の4種類の間での比較を試みる．

(1) テキスト方式

FAQ(Frequently Asked Question) などのデータをテキスト形式で利用するもの．学生は各自の端末からテキスト処理プログラムなどを利用して質問に対する回答を探す．

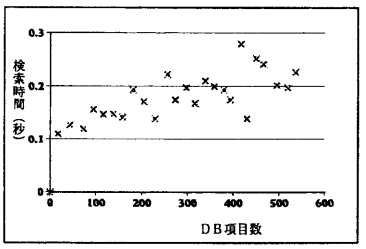

図 5: DB項目数に対する検索時間

(2)DB方式

SQLのようなデータベースに授業に関する内容を入れておくもの．学生は，自分の探す内容に対応するキーワードを入力する．

(3) 意味解析方式

自然言語での質問を意味解析し，知識データベースから回答を生成するもの．学生は，自然言語で質問を発することができる．

(4) 類似質問検索方式

本方式．学生は自然言語で質問を発することができる．

それぞれの長所短所をまとめたものを表5に示す．

事前の手間に関しては，DB方式は，データベースに入れる項目に対して，いちいちキーワードを付けていかなければならない．意味解析方式では，膨大な知識データベースを作っておく必要がある．類似質問検索方式では，初期データは無くても運用は可能であるが，授業で予想される質問を予め登録しておけば，学生の利用という点では，便利である．また，今年蓄積したデータを次年度は初期データとして使うことができる．

利用の容易さという点では，意味解析方式と類似質問検索方式が，自然言語文で質問ができるので有利である．

データ保守性に関しては，類似質問検索方式が優れている．なぜなら，新規データを自動的に蓄積することができるからである．

表 5: 他方式との比較

	事前手間	利用が容易	保守
テキスト方式	△	△	○
DB方式	×	△	×
意味解析方式	×	○	×
類似質問検索方式	△	○	○

5.2 電子メールのための形態素解析

従来開発されて来た形態素解析プログラムのほとんどは，完全にチェックされた，文法に則った文章を対象としているため，電子メールの文章では解析率は良くない．これは質問メールには，口語，タイプミス，ひらがな書きの名詞，などが含まれるからである．本方式では，類似質問検索されるデータベース中の質問文と，入力となる質問文の両方が同じ形態素解析プログラムで解析されているため，文の構造が完全に間違ってしまうような致命的な解析誤りでない場合には，類似質問検索ができる可能性がある．

しかし，検索の精度をより一層あげるためには，形態素解析プログラムの改良が必要である．

6. むすび

以上，類似質問検索方式による質問応答システムについて述べた．この方式を使うことで，学生，教師双方にメリットがあることが分かった．特に教師については，学生が分かっていないところが分かること，同じ質問に何度も答える必要がなくなること，質問データベースに質問が蓄積されるため，今後の授業に利用できること，毎年データベースは充実するため，類似質問検索成功率が上がることなどが期待される．一方学生から見れば，返事が即座に返ってくるため，質問がしやすくなる．

しかし，類似質問検索成功率があまりにも低過ぎると，何度も質問しなおさなければ返事がこないというデメリットも考えられる．我々の実験では，データさえ揃えれば，60%以上の質問に対しては，類似質問が返送できることが確かめられた．また，短い文にするなど，質問の仕方を学生が考えるようになれば，さらに類似質問検索成功率が上がることが期待され，実用化できると考えられる．

謝辞

本研究を進めるにあたり，有意義なコメントをいただいた，九州大学大学院システム情報科学研究科菅沼明助教授に感謝いたします．

参考文献

[1] ISO/IEC 9075, Information Technology Database Languages–SQL, (1992).

[2] 藤崎, 間下ほか:データベース照会システム「ヤチマタ」と名詞句データ模型, 情報処理学会論文誌, Vol. 20, No. 1, pp. 1-24 (1979).

[3] 笠, 小林, 白石, 横田:自然言語問合せ文の意味表現方法とその応用, 情報処理学会論文誌, Vol. 34, No. 5, pp. 925–933 (1993).

[4] 山田, 溝口, 原田:質問応答システムにおけるユーザ発話モデルと協調的応答の生成, 情報処理学会論文誌, Vol. 35, No. 11, pp. 2265–2275 (1994).

[5] 熊本, 伊藤:支援対話の解析に基づく対話処理方式の提案, 電子情報通信学会論文誌, D-II, Vol. J77-D-II, No. 8, pp. 1492–1501 (1994).

[6] 隅田, 堤:翻訳支援のための類似用例の実用的検索法, 電子情報通信学会論文誌, D-II, Vol. J74-D-II, No. 10, pp. 1437–1447 (1991).

[7] 松本ほか:日本語形態素解析システムJUMAN使用説明書version2.0, 京都大学工学部(長尾研究室)奈良先端科学技術大学院大学(松本研究室) (1994).

[類似質問検索の出力例]

例1

入力質問文

印刷の時、紙の位置はどこに設定したらよいのですか。

主題部

印刷（名詞）の（助詞）時（名詞)、(読点）紙（名詞）の（助詞）位置（名詞）は（助詞）どこ（指示詞）に（助詞）設定（名詞）したら（動詞）よい（形容詞）のです（助動詞）

照合パターン

(("紙" 名詞)("位置" 名詞)("する" 動詞))

検索結果

表示している文書を印刷する時に、紙を置く位置を教えて下さい。

例2

入力質問文

送信の方法をもう少し、詳しく教えて下さい。

主題部

送信(名詞) する(動詞) 方法(名詞)

照合パターン

(("送信" 名詞)("方法" 名詞))

検索結果

電子メールの送信の方法を教えて下さい。