

平成 12 年度

学士学位論文

自動応答システムにおける
最適検索アルゴリズムに関する研究

A Study on Optimal Retrieval Algorithms
in an Automatic Answering System

1010453 森枝 友之

指導教員 Ruck Thawonmas 助教授

2000 年 2 月 5 日

高知工科大学 情報システム工学科

要旨

自動応答システムにおける 最適アルゴリズムに関する研究

森枝 友之本

論文では、ラック研究室で開発された高知工科大学ヘルプシステムの検索の効率を上げるための実験と、実験結果から最適な検索アルゴリズムに関する研究について述べる。本研究の目的は、検索時にデータベースにある現在未使用の回答文を利用し検索の効率を上げる事にある。実際に従来のものと比較すると、データベースの回答文の使用は有効である事がわかっている。

キーワード 自然言語, 情報検索, ベクトル空間法, 形態素解析

Abstract

A Study on Optimal Retrieval Algorithms in an Automatic Answering System

Tomoyuki MORIEDA

This thesis describes experiments for increasing the retrieval efficiency of the KUT Help System developed at Ruck Laboratory, and discusses optimal retrieval algorithms based on these experimental results. The object of this study is to exploit answer data in the database, previously unused, for increasing the retrieval efficiency. It will be clear from given experimental results that exploitation of answer data is effective.

key words Natural Language, Information Retrieval, Vector-Space model, Morphological Analysis

目次

第1章	はじめに	1
第2章	自動応答システム	
2.1	システムの概要	3
2.2	質問パターンの分	7
2.3	知識ベースの内容	9
2.4	重み付け	9
2.5	類似度演算	10
2.6	質問応答例	12
第3章	実験方法とその評価	15
3.1	9種の計算法	15
3.2	実験結果	16
3.3	現システムとの融合	18
3.4	失敗した具体例と考察	21
第4章	おわりに	26

謝辞

参考文献

目次

2.1	自動応答システムの概要	4
2.2	高知工科大学ヘルプシステム	5
2.3	高知工科大学ホームページ	6
2.4	類似度演算までの流れ	11
3.1	ログ解析システム画面	17
3.2	実験結果の比較	17
3.3	実験結果の比較2	20

表目次

3.1 ログ解析の結果

19

はじめに

現在，人間が入力した質問に対して計算機が回答を返す情報検索システムが多く研究されている．しかし，現在のキーワード型の情報検索システムは，利用者にとって知りたい情報を，素早く適格に検索・閲覧する手段は，まだ十分とはいえない状況である．

キーワード型の情報検索システムの問題として，

1. 自然言語を使用することができず，利用者の使い勝手を考慮していない
2. 言い回しの違いや，類似語による質問に対処できない
3. 知識を，容易に拡張できる形で計算機に与えることが難しい

などが挙げられる．

この問題を解決するものとして，(株)エス・エス・アールと本学情報システム工学科の坂本研究室と同ラック研究室との共同で，自動応答システムが構築された．

この自動応答システムは，質問と回答をデータベースとして蓄え，質問を受け付け自動的に回答する検索システム[1]の一種である．

実際に 2000 年 9 月から自動応答システムの応用として，高知工科大学ヘルプシステムを試験運用している．同ヘルプシステムは高知工科大学への入学を目指す方を対象とし，おもに受験に対する疑問・質問に答えるものである．上記の自動応答システムは，自然言語で書かれた質問に対しても同様に回答できる事の特徴としている．また同ヘルプシステムはブラウザを介して質問を受け付け，大学のホームページのトップからリンクする事で利用度を高めている．そして，同ヘルプシステムは，2000 年 9 月からこれまでに約 1800 の質問を受け付け，このうち約 33.2%の質問に対して適切な回答を返している．

本研究では，このシステムの検索の効率を上げるための実験を行い，この実験結果から効率が高いアルゴリズムを探る．

データベースには，質問とそのペアの回答などが格納されている．そしてその大部分が回答である．現在のシステムでは，実際に質問に対する回答を検索するときには，入力質問文とデータベースの質問との類似度をベクトル空間法により求めている．し

かしデータベースの大部分を占める回答文は、活用されていない。本研究の目的は、検索時にある現在未使用の回答を利用し、検索の効率を上げることにある。

実験方法は、まず、これまで高知工科大学ヘルプシステムが得た、約 1800 件の質問をデータベースの回答を利用して検索しなおした。その際、ベクトル空間における重みの計算法を 9 通り用意し、9 通りのシステムが出した回答の全てのログをログ解析システムによって解析した。そして良い結果の出るアルゴリズムを探る。

実験結果として、高知工科大学ヘルプシステムが返す回答には、回答・基本情報・回答なしがあるが、表 1 は 9 通りの手法のうち一番高い、正しい回答と正しい基本情報の返答率を示す。この表 1 で正しい回答と正しい基本情報の返答率は 44.7%である。このことから、データベースの回答文の活用は、現在の方法と融合して、回答率を上げることが期待できる。

本論文では、実験により、最適な検索アルゴリズムに関する研究について述べる。本研究によりデータベースの回答文の使用が有効であるか実験データを比較・検証する。現在の方法をどうすれば、検索の効率が上がり、回答率が上がるのかを探る。

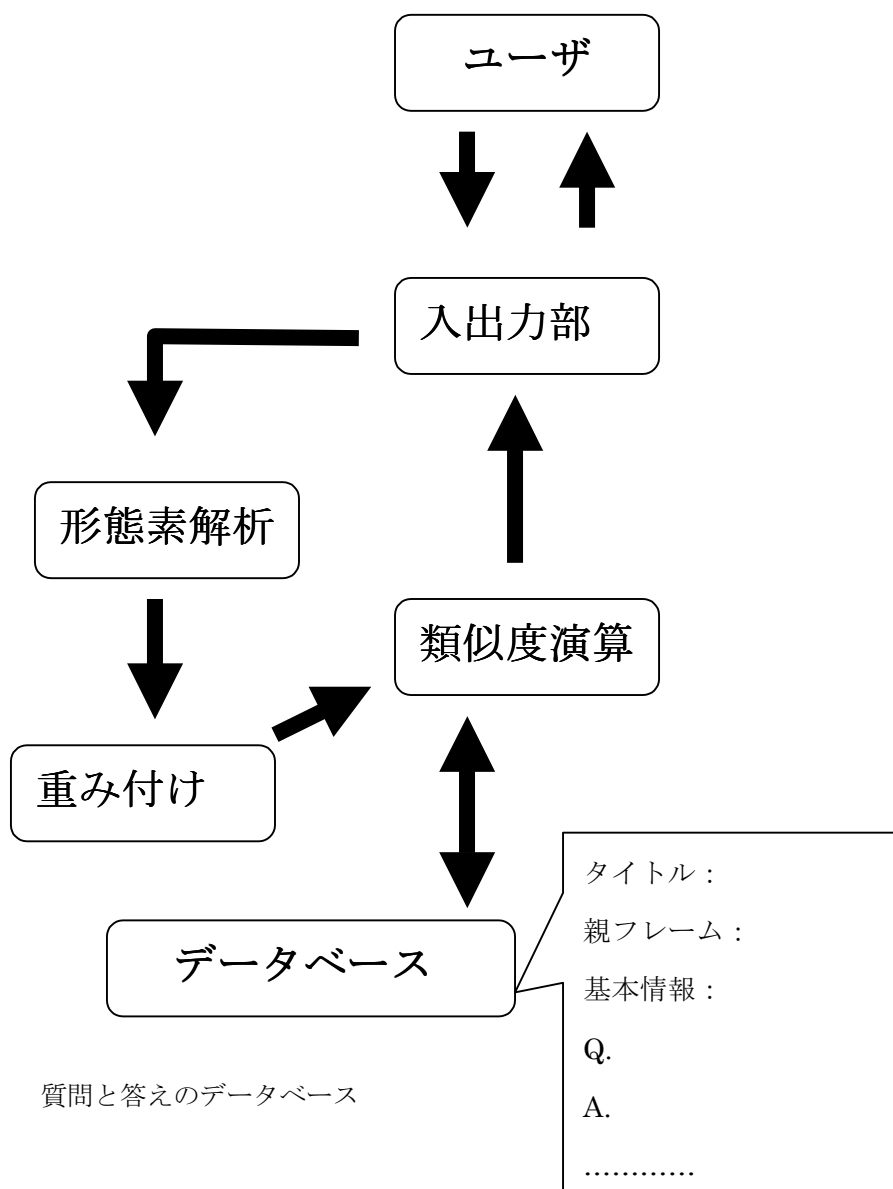
第 1 章

自動応答システム

2.1 システムの概要

自動応答システムとは，計算機と人間との間で行なわれる情報検索システムの一つとして，システムの構築を行なったものである．実際に 2000 年 9 月から自動応答システムの応用として，(株)エス・エス・アールと本学情報システム工学科坂本研究室と同ラック研究室との共同で，高知工科大学ヘルプシステムを試験運用している．同ヘルプシステムは，高知工科大学への入学を目指す方を対象とし，おもに受験に対する疑問・質問に答えるシステムである．

同システムの特徴は，「普通の日本語（自然言語）での問い合わせに適切な答えを返すシステム」のモデルケースとして研究・開発したものだ．同システムの概略を図 2.1 に示す．また同システムの各要素は以下に示す．



質問と答えのデータベース

データベースのフレーム型

- ① ユーザが入出力部に質問を入力する.
- ② 入力質問文が形態素解析される.
- ③ 形態素解析された単語一つ一つに重みがつけられる.
- ④ ⑤ 入力された質問文は、データベースに格納されている質問をベクトル空間法により計算を行い類似度が求められる.
- ⑥ ⑦ 閾値を設定しており、閾値以上の類似度で、類似度の高い回答を出力する.

図 2.1 : 自動応答システムの概要

入出力部

ユーザとシステムの対話は、WWW ブラウザを介して行なう。ユーザが質問を入力し、システムが回答を返し出力する。

現在試験運用中の自動応答システム（高知工科大ヘルプシステム）は、図 2.2 のようにブラウザのフォームを介して質問を受け付けている。そして図 2.3 のように、高知工科大のホームページのトップに設けていて、すぐ後ろにリンクすることで利用度を高めている。



図 2.2 : 高知工科大学ヘルプシステム



図 2.3：高知工科大学ホームページ

入力解析部

ユーザの自然言語入力を形態素解析し重みを付け、入力された質問とデータベースの質問をベクトルによる類似度の計算を行なう所である。閾値以上の類似度で 1 番高い質問の回答を出力する。

データベース

知識データベースと類似語データベースがある。知識データベースには、質問と答えがある。類似語データベースには、質問されるであろう言葉をあらかじめ、類似語としてまとめてあるデータベースがある。

メール送信部

回答が見つからなかった場合、管理者が直接メールを送り、後日回答する。

2.2 質問パターンの分類

従来のキーワード型の検索システムでは、用意している情報の中から利用者が指定したキーワードを含んでいる情報を探して回答する。これに対して本自動応答システムでは、利用者に自然言語で書かれたテキストで質問してもらい、利用者が何を知らなければならないかを判断する。

これを実現するために、本自動応答システムには、想定される質問の言い回しパターンを用意されている。質問パターンを利用することで、システムは、言い回しによる質問の違いを理解している。

高知工科大学ヘルプシステムを構築するにあたって、まず学生が大学に関してどのような質問をもっているのかを調査した。この調査は、1999年度セミナー3及び2000年度セミナー1

にて、情報システム工学科の3回生、1回生の合わせて約160名の学生に質問を書いてもらい集計したものである。その結果、以下のように質問パターンを分類できることがわかった。

What is 型：内容を問うもの

高知工科大学の誇れる所を教えてください

試験科目はなんですか

Can I 型：可能・不可能を問うもの

ドミトリーや大学構内は携帯電話の電波は入りますか

留学はできますか

Is there 型：存在を問うもの

近くにアルバイトをする場所がありますか

奨学金の制度はあるのか

When 型：時を問うもの

入試はいつですか

夏休みはいつからですか

Where 型：場所を問うもの

インターシップの受け入れ先はどこですか

授業で使う教科書はどこで買うのですか

How much 型：お金の価値を問うもの

授業料はいくらですか

入学時に費用はいくら必要ですか

How many 型：数量を問うもの

就職率はどれくらいですか

学食の座席はいくつありますか

高知工科大学ヘルプシステムでは、これらの What is 型, Can I 型, Is there 型, When 型, Where 型, How much 型, How many 型を質問パターンとして用意している。この質問パターンを利用する事で、質問の言い回しによる違いの解消をしている。

2.3 知識ベースの内容

多量のデータを整理整頓し管理することは、検索効率を上げるばかりでなく、システムの保守・拡張を容易にする点で重要な課題である。また、システムの保守・拡張を容易にするという点においては、データは人間が見てわかる形であるのが望ましい。

そこで、本自動応答システムでは自然言語で記述されたデータをフレーム型で管理する方式が採用された。

知識ベースは、フレームが親子の関係を持ったフレームで構成されている。各フレームは、それぞれ一つの知識ドメインで構成されている。

- ・フレームは、タイトル、親フレーム、基本情報、質問回答のペアの集まりで構成されている。
- ・タイトルは、各フレームの名称であり知識ドメインを代表する言葉が当てはまる。
- ・親フレームは、フレームの親子関係を示す情報である。
- ・基本情報は、タイトルの説明であり自動応答機能の性能向上のために使用される。
- ・質問回答ペアは、よく聞かれる、または、聞かれるであろう質問とその回答のペアである。

このように、自然言語で書かれたデータをフレームという概念で構造化することにより、計算機と人間の相互で利用可能な柔軟な知識ベースを構築することができる。

2.4 重み付け

ユーザから受け付けた入力文は、まず、形態素解析を行い形態素に分解される。形態素解析のツールとして、計算機による日本語の解析の研究を目指す多くの研究者に共通に使える形態素解析ツールを提供するために開発された 茶筌を使用している。

次に分解された形態素に重み付けがなされる。たとえば、「高知工科大学の特徴は何ですか」というユーザの入力の中に含まれている索引語に「1」を、含まれていない索引語に対して「0」を割り当てる。この操作により、この入力に対して、「1 1 1 1 1 0 0 0 0 0 1 0」などのベクトルが作成される。

なお、索引語の作成法は、様々な方法が考えられる。本研究では、この索引語の作成を様々な方法で実験することにより適切な情報検索アルゴリズムを探る。

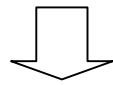
2.5 類似度演算

本節では、ユーザからの入力文の意図に対して知識ベースを検索する際システムが、演算をして回答を行なう基本的な枠組みについて述べる。

実際のユーザの質問は、様々な入力のタイプがあるが、すべての質問にマッチするように知識ベースに記述するのは、無意味である。

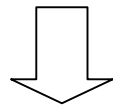
ユーザの入力文は、形態素解析され、各単語に重みを付け知識ベースとの類似度演算を行なう。以上の流れを下図によって示す。

質問：学生食堂のメニューを教えてください



類似語データベースに問い合わせる。

学食のメニューを教えてください



知識データベースに問い合わせ類似度を求める

※類似度を計算する (0.1~1.0)

図 2.4: 類似度演算までの流れ

類似度演算の結果、3パターンの結果を出力する。この際、閾値を使う。閾値の値によって回答を出力する方法が変わる。その3パターンを以下に示す。

- ① 閾値以上のとき
 - ・ 知識ベースの回答を返す。
- ② 閾値未満であり基本情報が存在（該当するフレームがある）とき
 - ・ 基本情報を返す。
- ③ 閾値未満であり基本情報が存在しないとき
 - ・ 回答できませんを回答する（回答不可能）

本研究の自動応答システムの特徴として、間違った答えを返さないことを目的としている。そのためにこの閾値を使用する。本システムでは、適格な回答でないと返事を返さないところが、大きな特徴といえる。

質問応答例

高知工科大学ヘルプシステムと一般ユーザとの実際の対話例を以下に示す。

現在のバージョンのシステムが対応した種々の現象が見られる。質問者(U)とヘルプシステム(S)とする。

正しい回答を返した例

(09/15 14:01)

U : 奨学金は利用できますか

S : (奨学金制度を利用できますか) とペアになっている回答を返しました

入力文が類似度変換されて、閾値が高いので回答文を返した。類似度変換された入力文は、データベースに用意されている質問文と比較し、最も高い類似度の質問の回答を返す。この例は、高知工科大学ヘルプシステムの返した回答文が、正しい回答を返した成功例である。

誤った回答を返した例

(09/18 17:49)

U : 高知工科大学の図書館の特徴を教えてください

S : (高知工科大学の特徴は何ですか) とペアになっている回答を返しました

この例も上記と同様にして、回答文を返した。しかし、質問者は「図書館の特徴」を質問しているのであって「工科大学の特徴」を質問しているのではない。よってこの例は、同ヘルプシステムの返した回答文が、誤った回答を返した失敗例となる。

正しい基本情報を返した例

(10/02 15:36)

U : 就職内定状況はどうですか

S : 卒業見込みの学生が現在就職活動中です。主な就職先及び就職率などの詳細につ

いては、こちらのホームページを 御覧ください

この例は類似度変換された際、閾値が低い為に回答を返すことができない。そこで、データベースに格納されている、基本情報を返す。この基本情報とは、入力質問文に関する近い情報の事である。この例は、同ヘルプシステムの返した基本情報が、正しい基本情報を返した例である。

誤った基本情報を返した例

(09/27 13:23)

U : 工科大にはどのような学科がありますか

S : 高知工科大学は、高知県香美郡土佐山田町宮ノ口 185 にあります。郵便物は、〒782-8502 高知県土佐山田町で届きます

この例も、上記の正しい基本情報を返した例と同様に、基本情報を返した。しかしこの基本情報は、「工科大」についての基本情報である。質問者は「学科」について質問しているので、誤った基本情報と判断できる。よって、返した基本情報が、誤った基本情報を返した例である。

回答なしを返した例

(09/27 13:09)

U : 生徒の数は

S : 何も回答できませんでした

同ヘルプシステムは、閾値未満であり、その質問入力文に関する基本情報がないときには、回答なしを返す。

以上のように、実際に同ヘルプシステムは3種(回答・基本情報・回答なし)の情報を返す。そして、そのログを解析する過程で5種になる。そして、正しい回答と正しい基本情報をより多く返すアルゴリズムを探ることが、本研究の目的となる。

第2章

実験方法とその評価

本システムは、2000年9月15日から、高知工科大学のホームページ上で公開し、現在も試験運用中である。この間のログはすべて保持している。

システムの問題点は、知識ベースの大部分を回答が格納されている。しかし、いままではユーザからの質問を受け、回答を検索するときには、質問どおしだけで検索され使用されていない。本実験では、この回答も利用しつつ検索を行なう。結果としては、未使用の回答を利用し検索の効率を上げることができると考えられる。検索の効率を上げるということは、自動応答システムの効率が上がることになり、本研究の目的である。

本章では、2000年9月15日から11月30日までに、高知工科大学ヘルプシステムが受け付けた約1800個あたりの質問を利用して検索しなおした。その際、ベクトル空間における重みの計算法を9通り用意し、9通りのシステムが出した回答のすべてのログを、ログ解析システム[2]を用いて解析した。そして良いアルゴリズムを探る事が、本実験内容である。その評価を以下に述べる。また本システムと、ユーザとのやり取りにおける失敗例から、今後のシステムについて考察を行なう。

3.1 9種の計算法

本実験では、従来の自動応答システムが回答を検索する際に、使用していない、データベースにある回答も利用し、類似度計算して回答を検索する。その計算法を9パターン用意し、より良い結果を探る。その9パターンを以下に説明する。

1. 日本語は、長い文章のとき1番最初の文を見るとその段落のことがほぼ判るようになっている。このことを利用して実験する。知識ベースの回答を使用する際、3パターン用意する。

回答の先頭から1文 (k01)

回答の先頭から 2 文 (k02)

回答の先頭から 3 文 (k03)

これ以上やると、回答文が 1 文しか無いデータが多いので、かえって不平等になり、効率が上がらないと考えられる。

2. 索引語の作成法の違いによる重み付けを変えたものを 3 パターン用意する。

重み付けを変えた (a) , (b) , (c)

1 と 2 をそれぞれの条件を合わせると、

{a-k01, a-k02, a-k03, b-k01, b-k02, b-k03, c-k01, c-k02, c-k03}

となる。この 9 種の計算法を用いて、約 1800 個の質問をしなおし、より良い結果の出したアルゴリズムを探る。

3.2 実験結果

約 1800 個の質問をしなおし、自動応答システムが回答を出した物が、9 パターン分ある。その 1 パターン分、約 1800 個の回答をログ解析システムによってログ解析を行い、評価する。評価する際の回答の種類は以下に示す。

- ・ 正回答 : 質問にあった正しい答えをかえす
- ・ 誤回答 : 質問にあってない誤った答えをかえす
- ・ 正しい基本情報 : 質問に近い情報を返す
- ・ 誤った基本情報 : 質問にあってない誤った答えをかえす
- ・ 回答無し : 知識不足や現在の機能での返答不可能

そのログ解析画面、実験結果を以下に示す。

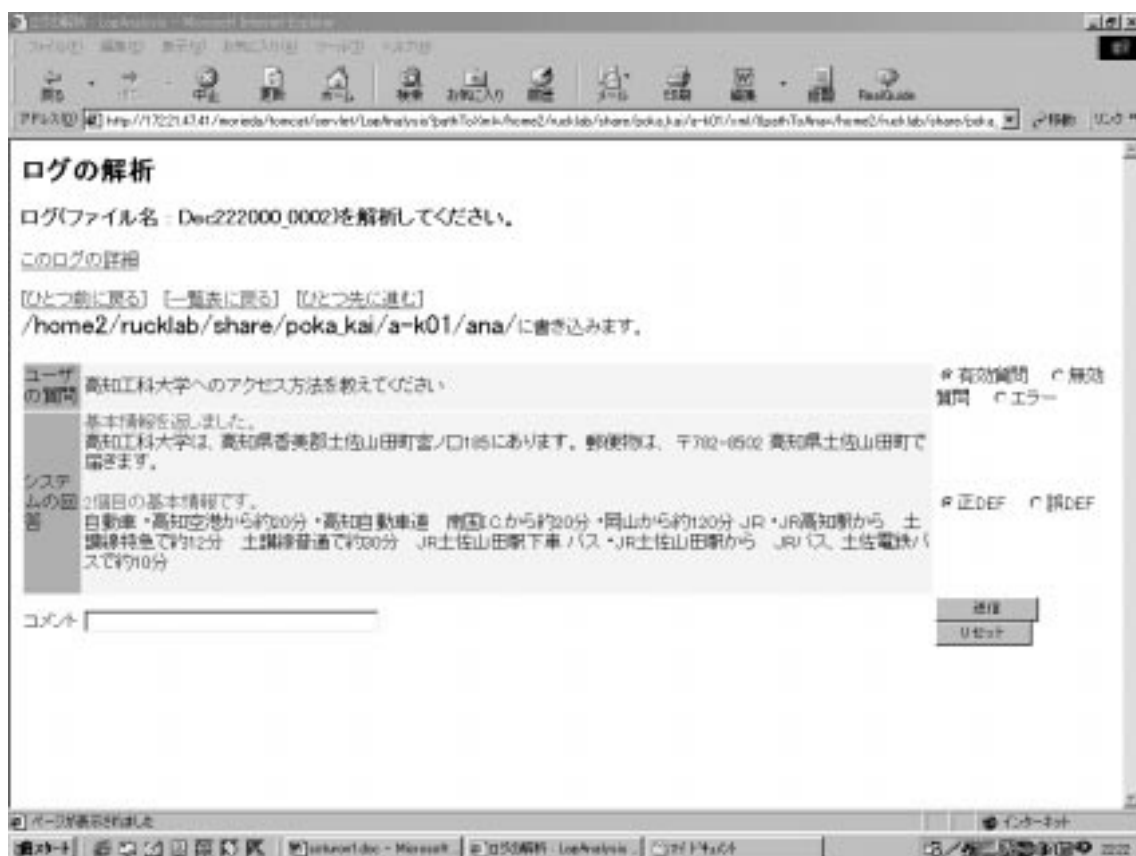


図 3.1: ログ解析システム画面

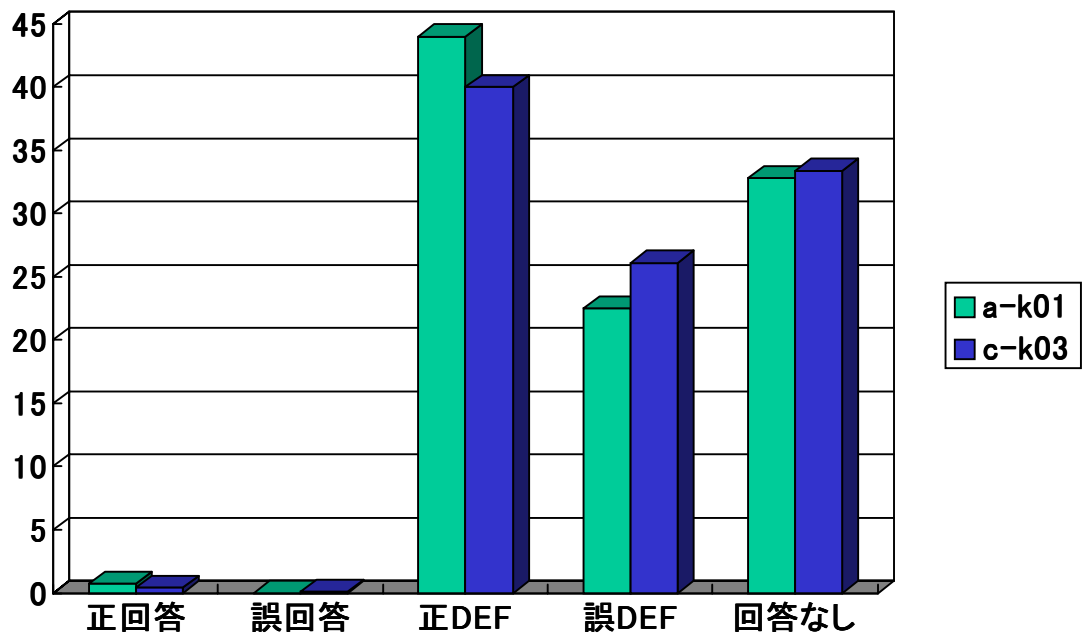


図 3.2 : 実験結果の比較

ログ解析の結果から、a (回答の先頭から 1 文) と c (回答の先頭から 3 文) とを比較すると、a の方は正しい基本情報を返した割合が少し高い。また、k01 (重みを変えたパターン) と k03 を比較すると、k01 の方が若干、正しい基本情報を返した割合が高い。以上のことを踏まえると、9 パターンの中で最も良い結果をだしたものが、a-k01 であることがわかる。このパターンの正しい答えの返答率は 44.7% となる。最も低いものと比較すると、低いものは 40.4% で 4.3% の開きがある。また現在の方法と融合して、回答率を上げることが期待できる。

3.3 現システムとの融合

そこで、今までの実験では、類似度計算をする際、入力質問文とデータベースの回答を利用し閾値を求めていた。今から述べる実験は、従来の高知工科大学ヘルプシステムとデータベースの回答を利用し回答を求めた方法とを、融合させたものである。融合した手順を示すと、

1. 従来の方法で類似度計算する。検索して回答が見つければそのまま回答を返す。

2. 基本情報を返す・回答なし の場合にデータベースの回答を利用してもう 1 度検索
しなおす

ここで、2 種類の閾値を 0.55 と 0.4 の 2 種類を用意し、約 1800 個の質問を上記と同様に質問しなおす。従来のものと比較し回答率が上がったかを比較・検証する。

	正回答	誤回答	正基本情報	誤基本情報	回答なし
a-k01 0.55	542 (35.5%)	24 (1.6%)	283 (18.5%)	228 (14.9%)	450 (29.5%)
a-k01 0.4	578 (33.7%)	26 (1.7%)	264 (17.2%)	222 (14.5%)	445 (29.0%)
normal	505 (33.2%)	21 (1.4%)	297 (19.5%)	252 (16.5%)	448 (29.4%)

表 3.1 : ログ解析の結果

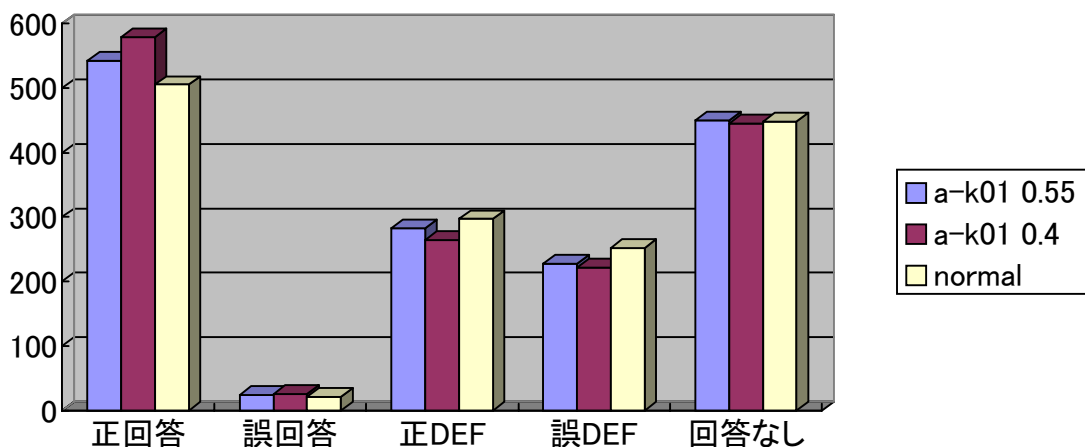


図 3.3: 実験結果の比較 2

実験結果から

- a-k01 閾値 0.55 の正しい返答率は 54.0%
- a-k01 閾値 0.4 の正しい返答率は 50.9%
- 従来通りのものは 52.7%

このことから、データベースの回答文の活用は、現在の方法と融合して実際に回答率が上がったことを示す。また、閾値の設定法により、回答率に影響を与えることがわかる。

3.4 失敗した具体例と考察

本節では、ログ解析をして、これまでの失敗（誤回答・誤った基本情報・回答無し）の原因は知識不足にある。知識ベースにない質問をユーザが使用すると、たいてい回答できないか、違う回答、または基本情報を返す。しかし、知識を増やしたところで単純に成功率が上がるとは思えない。本システムのログ解析から、失敗した具体例を参考にし、今後の自動応答システムの向上ため以下に考察していく。

単語の重みに関する問題

(2000/12/22 21:45)

U：大学へのアクセスを教えてください

S：高知工科大学は、高知県香美郡土佐山田町宮ノ口 185 にあります。郵便物は、〒782-8502 高知県土佐山田町で届きます。

質問のキーワードとなる言葉、つまり重みの大きい単語は 大学・アクセス となる。そして、ユーザが聞きたい事は、アクセスについてである。減システムでは、文章の左側にある方に重みが大きくなるので、この文章では、大学がキーワードとなり、大学について質問を返している。ユーザはこのことを知らず、また工科大学ヘルプシステムなので、どうしても質問するとき頭に、大学という言葉がきがちだ。将来的には、文章のキーワードを正しく選択し対処する必要がある。

(2000/12/22 21:45)

U：高知工科大学の特徴を教えてください

S：高知工科大学は、高知県香美郡土佐山田町宮ノ口 185 にあります。郵便物は、〒782-8502 高知県土佐山田町で届きます。

この質問も上記と同様のことがいえるが、管理側から見れば、この回答は明らかに間違っているが、ユーザも必ずしも同様だとはかぎらない。こういう事も考慮する必要がある。

知識ベースの拡張について

(2000/12/22 21:48)

U：何人の学生がいますか？

S：何も回答できませんでした。

(2000/12/22 21:46)

U：何階建て

U：何も回答できませんでした。

このような例は、重み付けで一番重たい語、つまりキーワードとなる言葉に関するデータが知識ベースに無いときに起こる現象だ。当然回答は見つからないわけで、「回答できません」を返すことになる。こういう事はログを解析しなければわからない。知識ベースに無いデータは、随時加えていって対処する予定だ。

疑問形の質問に対する要求

(2000/12/22 21:57)

U：どうすれば高知工科大に入学出来ますか。

S：高知工科大学は、高知県香美郡土佐山田町宮ノ口 185 にあります。郵便物は、〒782-8502 高知県土佐山田町で届きます。

(2000/12/22 21:57)

U：どんな学校ですか。

S：何も回答できませんでした。

質問するときに文頭に、どんな・どこに・どうすれば などがくると、現システムでは回答を返さないもしくは、間違った答えを返すことが多い。しかし、このシステムはヘルプシステムなのでユーザはこういう形式で質問するのはあたりまえである。今後のシステムとして、このような疑問形にも対応できるように、知識ベースの知識量を増やしていくことを今後の課題とする。

長い質問文の対処法

(2000/12/22 21:46)

U： 高知大学はどのような学科あり尚且つどのような授業を行っているのでしょうか。また、今回のシステムはどのような形で造られたのか教えてください。

S： 卒業するためには、共通基礎科目 48 単位以上、専門科目 76 単位以上の修得が必要ですが、必修科目はありません。各学科で用意した履修モデルを参考に、自分の興味のある科目を選択できるシステムになっています。また所属学科の専門科目を 60 単位以上修得すれば、他学科の科目も履修可能。放送大学の講義も自由に選択でき、しかも単位の修得も認定されます。そして、どうしても自分のやりたいことが、他の学科にあったという場合、2 年次での転学科制度も設けています。

(2000/12/22 22:57)

U： 親戚の者が工科大学を推薦で受験します。(情報システム工学科を専願)

S： 高知工科大学は、高知県香美郡土佐山田町宮ノ口 185 にあります。郵便物は、〒782-8502 高知県土佐山田町で届きます。

ひとつめの例は、一見システムが適切な回答を返したように見られる。しかし、ユーザが後方(文章の終わりぐらい)に質問したい事を入力していたら、適切な答えを返していない。要するに長い文章を入力すると、文章の始めの方にあるキーワードに関する答えをかえしてしまう。2 つめの例で、質問者は、情報システムについて、もしくは推薦について聞いているが、システムは工科大についての答えを返している。長い文章や、2 つの並列な文章のどれを優先するか、システムが判断するのは、現時点ではできない。

基本的には、キーワードがたくさん含まれる文章が質問者の質問したい文章である事が多い。解決方法としては、長い文章だとシステムはそれらをすぐに判別してしまえばいい。これらの問題はすぐに対処しなければならない。

2 つ以上の返答ができるタイプ

(2000/12/22 22:21)

U： 高知工科大学の学部・学科を教えてください

S : 高知工科大学は，高知県香美郡土佐山田町宮ノ口 185 にあります．郵便物は，〒782-8502 高知県土佐山田町で届きます．

(2000/12/22 22:33)

U : 高知工科大学の学科は，どのような学科，学部がありますか

S : 高知工科大学は，高知県香美郡土佐山田町宮ノ口 185 にあります．郵便物は，〒782-8502 高知県土佐山田町で届きます．

上記のような質問の場合，現段階の自動応答システムでは，2つの同じ重みのキーワードならば，先に出た方について返答を返す．ある意味システムは正しい返答をしたと解釈する．しかし質問者側はAとBについて質問したいときにAだけの質問の返答しかなければ，実際の質問者は納得しない．対処法としては，ペアで質問されるパターンをあらかじめ知識ベースに記述していればいい．しかし実際そうになると，莫大なコストがかかり，効率も上がらない．この問題は形態素解析などの言語処理技術が向上しないと，解決しない問題である．

質問の追加要求に対する問題

(2000/12/22 22:01)

U : 推薦入試の募集定員よりも多く合格者を出してくれますか．それは何人位ですか．

S : 推薦入試についての詳細は，入試ガイドに掲載されています．詳しくは，こちらのホームページを御覧ください．

(2000/12/22 22:31)

U : 高知工科大学の入試科目の変更はありますか？

S : 高知工科大学は，高知県香美郡土佐山田町宮ノ口 185 にあります．郵便物は，〒782-8502 高知県土佐山田町で届きます．

ひとつめの例は，システムは質問に対して適切な答えを返している．しかし，さらに続いて質問者は「それは何人位ですか」とより多くの情報を求めている．現段階のシステムでは2つの文脈や，キーワードの追加情報を考慮に回答することはできない．上述の問題を解決するとなると，今まで以上の豊富な知識量と，意味理解などの処理技

術が向上しないといけない。これは今後自動応答システムが向上する為の課題である。

質問文が単語だけのタイプ

(2000/12/22 21:53)

U : 英語

S : 何も回答できませんでした

(2000/12/22 21:52)

U : 教授

S : 何も回答できませんでした

このような、工科大に関する事であろう質問には、文章になっていなくとも答えた
い語句である。システムでは、語句だけでは回答を返さない場合が多く目立つ。しか
し質問者にすれば、そのようなシステムの構造を知っているわけではないので、この
ような状況にも対処しなければならない。

第4章

おわりに

本論文では、実験により、最適な検索アルゴリズムに関する研究について述べた。

9月15日から11月30日までの、本システムの試験運用中に約1800件の質問を受け付けた。その期間の本システムにおける正しい回答と正しい基本情報の返答率(工科大学に関する事でなく、無意味な質問を除いた上での成功の割合)は、52.7%である。現在の方法と融合してデータベースの回答文の活用も利用した返答率は、一番高いもので54.0%であり、従来のもものと比較する1.3%の開きがあり、実際に回答率が上がったことを示す。このことにより、データベースの回答文の利用も有効である事がわかった。

しかし、従来のもものと、データベースの回答の活用法とを融合するときに、閾値を変えなければならない。閾値を下げると回答率が上がる可能性もあるが、それに伴い誤った回答が増える可能性も出てくる。実際どのようにして、現在の方法と融合し回答率も上げ、さらに誤った回答を増やさないアルゴリズムの研究が今後の課題である。

謝辞

本研究を進めるにあたり終始懇切丁寧にご指導下さいましたラック ターウォンマツト助教授に心より御礼申し上げます。

また、共同研究で貴重な御意見と適切なアドバイスをいただきました(株)エス・エス・アール関係者各位にも、心から謝意を述べさせていただきます。

最後に、本研究に関して援助して下さいました人工知能研究室の皆様に、本当にありがとうございました。

参考文献

[1] 長尾誠

“岩波講座 ソフトウェア科学 15 自然言語処理”，岩波書店，
p. 441-p. 417, 1996

[2] 友池貴之，溝渕真司，ラック ターウォンマツト，“データ解析方法およびその装置ならびにコンピュータ・プログラム”，特許出願中(審査請求)，Jan. 30, 2000.

“