

LB/DON/25/2012

LIBRARY  
UNIVERSITY OF MORATUWA, SRI LANKA  
MORATUWA

Chat Bot



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
www.lib.moratuwa.lk  
K.M.C.P. Gunasekera

MS IT 05/10005

University of Moratuwa



102504

004"11"  
004(043)  
TH

Dissertation submitted to the Faculty of Information Technology, University of  
Moratuwa, Sri Lanka for the partial fulfilment of the requirements of the Degree of  
MSc in Information Technology

March 2011

102504

102504


## Declaration

I declare that this dissertation does not incorporate, without acknowledgment, any material previously submitted for a Degree or a Diploma in any University and to the best of my knowledge and belief, it does not contain any material previously published or written by another person or myself except where due reference is made in the text. I also hereby give consent for my dissertation, if accepted, to be made available for photocopying and for interlibrary loans, and for the title and summary to be made available to outside organization.



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)


Name of the Student: K.M.C.P. Gunasekera

  
Signature of Student:

Date: 15/11/2011

Supervised by

Name of the Supervisor: Dr. Gamini Wijayarathna

  
Signature of Supervisor:

Date: 14/11/2011

## Dedication

This thesis is dedicated to Dr. Gamini Wijayarathna with a heart full of gratitude. Without his encouragement, guidance and support from the initiation to the conclusion as the supervisor this thesis would not have been possible.



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

## Acknowledgments

I am heartily thankful to my supervisor, Dr. Gamini Wijayarathna, for his encouragement, supervision and support from the preliminary to the concluding stages of this thesis.

I would also like to thank my parents and friends who have helped me in various ways to make this project a successful one.

Last but not least, I offer my regards to all of the academic and non academic staff members of the University of Moratuwa for their support provided.



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

## Abstract

With the rapid popularity and increased usage of the Internet, online marketing has become a way of selling and purchasing goods through the Internet. This lead to development of various websites dedicated to online marketing and new techniques to support customers and information sharing became a major area of this.

The main objective of this thesis is to propose an efficient way of providing information to clients than the existing systems like shopping charts, operators, etc.

There are well known AI techniques and successful Chat bots and during the initial phase a research was carried out to get an understanding on those techniques and how they are used in the well know Chat bots.

The aim of this Chat bot is to analyze the user queries and to provide successful answers. Analysis is done by matching customer queries with sample queries and then extracting the information from those queries, which will be used to create the answers. The information extracted will be coupled with the predetermined answers and will be presented to the customers. This is also designed in such a way to capture certain information to increase its scope and efficiency.

The proposed Chat bot is in the primary stage and it is limited only for a simple bookshop scenario. Still this can be used for any kind of online shopping with a proper knowledge base attached to it. The proposed chat bot works as a solution provider for single customer queries, and can be improved with some advanced matching mechanisms to maintain a dialog with the customer, which will be much more user friendly and efficient.

## Table of Contents

<b>Chapter 1 - Introduction .....</b>	<b>1</b>
<b>1.1 Introduction to the Project</b>	<b>1</b>
<b>1.1.1 Why Online Shopping</b>	<b>1</b>
<b>1.1.2 Why Online Shopping is Gaining Popularity.</b>	<b>2</b>
<b>1.1.3 Marketing Strategies Used to Provide Information for Online Customers</b>	<b>4</b>
<b>1.1.4 Shortcomings of Providing Proper Information to Customers</b>	<b>4</b>
<b>1.1.5 Solution</b>	<b>5</b>
<b>1.2 Aim</b>	<b>6</b>
<b>1.3 Objectives</b>	<b>6</b>
<b>1.4 Solution</b>	<b>7</b>
<b>1.5 Expected Outcome</b>	<b>7</b>
<b>1.6 Summary</b>	<b>8</b>
<b>Chapter 2 - Literature Review .....</b>	<b>9</b>
<b>2.1 Introduction</b>	<b>9</b>
<b>2.2 Approaches</b>	<b>9</b>
<b>2.2.1 Symbolic</b>	<b>10</b>
<b>2.2.2 Statistical (Stochastic)</b>	<b>16</b>
<b>2.2.3 Connectionist</b>	<b>19</b>
<b>2.3 Existing Systems</b>	<b>19</b>
<b>2.3.1 ELIZA</b>	<b>19</b>
<b>2.3.2 Dr. Sbaitso</b>	<b>20</b>
<b>2.3.3 PARRY</b>	<b>20</b>
<b>2.3.4 Racter</b>	<b>20</b>
<b>2.3.5 MegaHAL</b>	<b>20</b>
<b>2.3.6 Ultra Hal Assistant</b>	<b>21</b>
<b>2.3.7 Elbot</b>	<b>21</b>
<b>2.4 Summary</b>	<b>22</b>
<b>Chapter 3 - Approach to the Solution .....</b>	<b>23</b>
<b>3.1 Introduction</b>	<b>23</b>
<b>3.2 Design</b>	<b>23</b>
<b>3.2.1 Identify the Question.</b>	<b>23</b>

3.2.2 Extract the Details	23
3.2.3 Database	24
3.2.4 Prepare the Answer	24
3.2.5 Knowledge Capture	24
3.2.6 Information Capture	24
3.2.7 Knowledge Engineer Interface	24
3.3 Scope of the project	25
3.4 Limitations of the scope	25
3.5 Technology Used	25
3.6 Software Licensing Issues	27
3.7 Summary	27
<b>Chapter 4 - Analysis and Design.....</b>	<b>28</b>
4.1 Introduction	28
4.2 Identified Functional Requirements	28
4.3 Identified Non Functional Requirements	28
4.4 Proposed System	29
4.4.1 Main areas of the Chatbot	29
4.4.2 Architectural Diagram	31
4.4.3 Use Case Diagram	32
4.4.4 Data Base Design	32
4.5 How Chatbot Works	33
4.5.1 Identify the Question.	33
4.5.2 Prepare the Answer	34
4.5.3 Update Probability Based on Frequency	35
4.6 Main Issues Faced	35
4.7 Summary	36
<b>Chapter 5 - Conclusion and Future work.....</b>	<b>37</b>
5.1 Conclusion	37
5.2 Future Work	37
5.3 Summary	38
<b>References .....</b>	<b>39</b>
<b>Appendix A - Algorithms Used in the Research .....</b>	<b>41</b>

## List of Figures

Figure 1.1 - US Online Retail Sales [1].....	1
Figure 1.2 - e-Commerce Retail Sales as a Percent of Total Retail Sales [2].....	2
Figure 1.3 - Shoppers' Main Reasons for Buying Online [4].....	4
Figure 2.1 - Tree of Porphyry Drawn by Peter of Spain. [6] .....	11
Figure 2.2 - Concept Map [18] .....	12
Figure 2.3 - An Implicational Network for Reasoning About Wet Grass. [6].....	13
Figure 2.4 - A Data Flow Graph .....	14
Figure 2.5 - A Learning Network .....	15
Figure 2.6 - Markov Model .....	17
Figure 2.7 - Hidden Markov Model.....	17
Figure 2.8 - Parse Tree 1 [19].....	18
Figure 2.9 - Parse Tree 2 [19].....	19
Figure 4.1 - Proposed System.....	29
Figure 4.2 - Chat Window .....	30
Figure 4.3 - Knowledge Engineer Interface.....	31
Figure 4.4 - Architectural Diagram.....	31
Figure 4.5 - Use Case Diagram of Proposed System.....	32
Figure 4.6 - ER Diagram .....	32



## List of Tables

Table 3.1 - Technology Stack .....	25
------------------------------------	----



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

**Introduction**

**1.1 Introduction to the Project**

**1.1.1 Why Online Shopping**

Online shopping is a process where customers directly buy goods and services over the Internet. These shops come in various forms such as, online shop, e-store, Internet shop, eShop, web store, web shop, online store or virtual store. With the increase of computer usage and Internet access, online shopping is becoming a more popular way of doing shopping. Figure 1.1 and Figure 1.2 indicates this trend and predicts this to grow to a much higher percentage of shopping in the coming years.

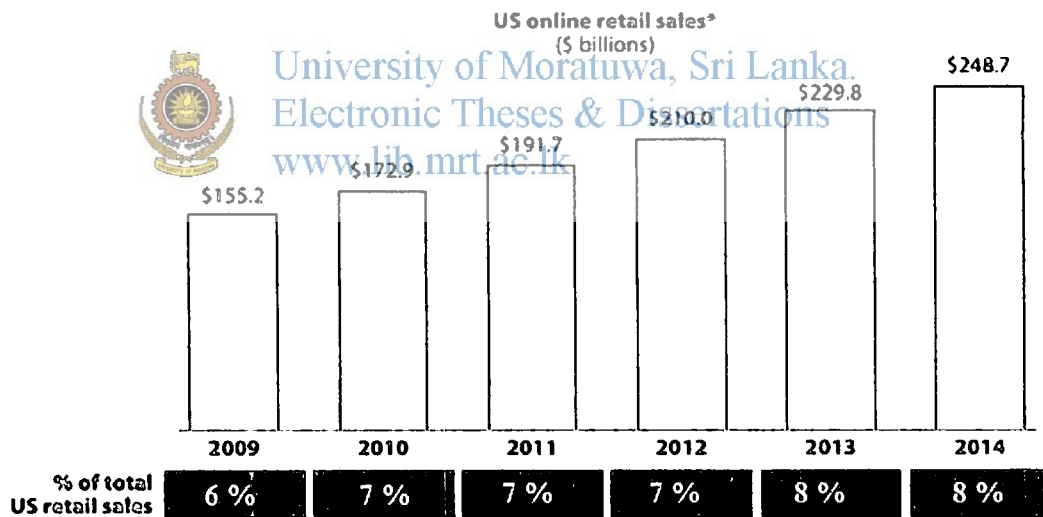


Figure 1.1 - US Online Retail Sales [1]

Estimated Quarterly U.S. Retail E-commerce Sales as a Percent of Total Quarterly Retail Sales:  
4<sup>th</sup> Quarter 1999 - 3<sup>rd</sup> Quarter 2009

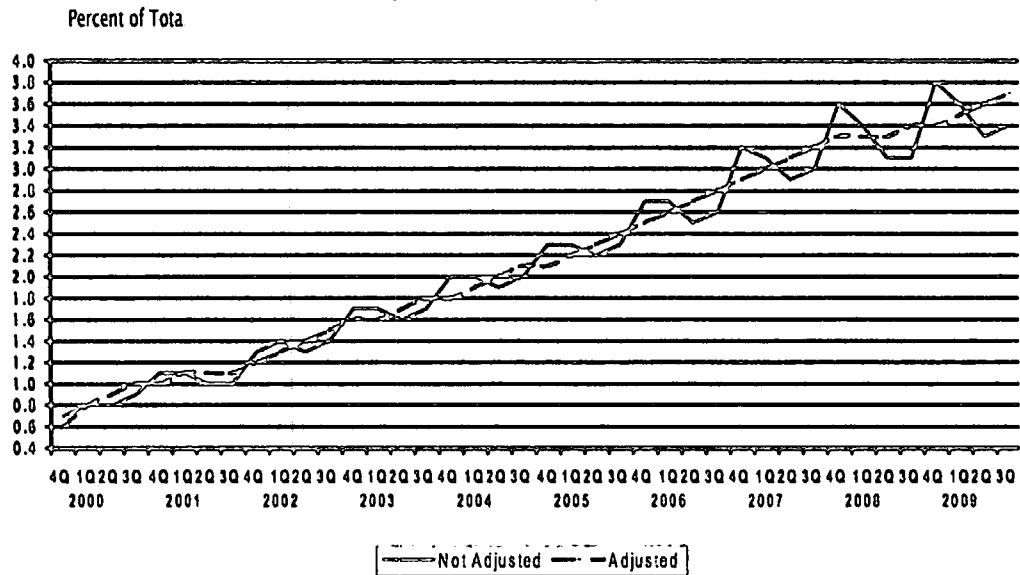


Figure 1.2 - e-Commerce Retail Sales as a Percent of Total Retail Sales [2]

1.1.2 Why Online Shopping is Gaining Popularity.



University of Moratuwa, Sri Lanka.

Following are some of the main reasons for online shopping to become popular [3].

- Total shopping convenience:

These shops are available 24 hours a day and 7 days a week which means customers can shop day or night while relaxing at his/ her home.

- Browse the web instead of driving:

Travelling to various shops at various places is a costly thing and time consuming. With online shopping, customers can sit at their homes and visit any number of shops, even some located in other countries with ease.

- Price comparisons:

Online shops do not require buildings with lots of facilities and other customer attractions which amount for huge sums. Since, with online shopping those expenses do not exist, companies can sell their goods at a much lower price than normal shops.

- Unlimited selection:

Online shops usually provide much more selection of items than a normal shop. They could even sell some items which they do not have in stock.

- No hurry:

Unless the website is down there is no close time for online shops, and customers can take any amount of time to do their shopping.

- Information and reviews:

Some online shops provide facilities to rate or add comments about the goods they bought and this feature will be helpful to new customers to make decision about purchasing goods and services. This an experience which traditional shops never provide to their customers.

- Collectible and hard to find items:

Online shopping is helpful to find collectibles and hard to find items, due to the vast range of selection for goods online shops provide.



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
www.lib.mrt.ac.lk

Figure 1.3 illustrates the main reasons why shoppers are attracted to online shopping and their percentages.



Figure 1.3 - Shoppers' Main Reasons for Buying Online [4]

### 1.1.3 Marketing Strategies Used to Provide Information for Online Customers

All forms of online systems have to make one important thing. That is making the customer buy their product without physically contacting them. To achieve this, websites must present proper information to customers and most web sites use various methods such as, graphics, animations, and special features like shopping charts, providing statistics and security. [5]

### 1.1.4 Shortcomings of Providing Proper Information to Customers

One of the most important aspects of online shopping is to provide correct information easily to the customer. To achieve this some websites allocate large areas of space to display information or make them available through a large set of links. This results in the customer going through a large amount of information which is not relevant to him/her to find some specific information. This can easily result in the customer leaving such websites. Additionally, they cause problems such as, maintaining a large web site, band-width problems, etc.

Another area of focus to winning customers is to provide satisfactory customer support as soon as inquiries are made. The majority of web sites provide facilities to send e-mails whilst some have support staff dedicated to the customer inquiries.

The drawback with the e-mail option is that it takes a lot of time to reply even for a small information request. Further, having a large number of customer support staff working 24 hours a day, 7 days a week comes with a high price tag. In addition, those resources could be idle for most of time, which results in huge losses to the companies.

#### 1.1.5 Solution

Therefore, the best way to provide information through a small web site will be to have a chat window (chat bot) attached to it. This chat bot should have facilities similar to ordinary chat systems with an automated system replacing the operator. The automated system should identify the customer queries by matching key words and provide answers through a predefined set of answers with the help of some stored knowledge.



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

The following additional facilities can increase the efficiency of the chat bot:

Providing a facility in the chat bot with the ability to forward customer queries to a human operator if it fails to recognize the query will increase the success rates of the system.

- Having a good knowledge about what the customers wants, what sort of information they like to have, their buying patterns, and market trends will be helpful to have a successful selling mechanism. This sort of information is normally collected by various research groups' through surveys and sold to vendors. If the system has a facility to capture customer queries against the time of the queries, will be an added advantage.

## 1.2 Aim

The aim of this project is to design a tool to provide support/ information to online customers for their requests and helping to vendors in decision making and to achieve higher customer satisfaction.

## 1.3 Objectives

1. This system will help reduce the amount of data displayed on web pages.
2. Provides an easy way for customers to get the required information quickly without going through various links.
3. Provide vendors with details about customer needs, requirements, customer views about the items on the site to sell, and market trends.
4. The system will generate requests to customer support staff when required.
5. Ability to function 24 hours and 7 days a week ideally suited to websites dealing global customers.
6. Reduce the requirement of keeping a large number of customer support staff.



University of Moratuwa, Sri Lanka.

Electronic Theses & Dissertations

[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

#### 1.4 Solution

The solution proposed is an automated chat system which can handle customer queries up to a certain level, removing customer care officer's work load.

The main feature of the system is the chat window, which is used by the customer to submit queries and to get answers. The system will generate the answers and display it in the chat window. When required support staff can answer customer queries.

In the back-end of the system, there is an engine to analyse customer queries. This should decide whether the engine can answer the query. If it is answerable it will generate an answer using a pre-defined set of answers or using available data. If it fails to generate an answer it will direct the question to an operator.

To support the engine there is a database. It stores a pre-defined set of answers to support the engine. Further, the database should store other relevant information (prices, types of goods, etc...) to support the engine. Moreover, it should store the information (customer queries, etc...) for forecasting purposes.

The main technologies required for this project are Java as the development language, MySQL as the database and eclipse as the development environment. All of the above technologies are open source technologies and therefore, freely available to use.

#### 1.5 Expected Outcome

The expected outcome of the project is to produce a tool which will assist online shopping reducing the workload, time and cost for assigning and maintaining a large force of operators.



## 1.6 Summary

In this chapter we have discussed how and why online shopping is popular and why we need a new system to provide information to customers.



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

## Chapter 2

### Literature Review

#### 2.1 Introduction

Natural language processing (NLP) is a computerized way to analyze text and speech based on a set of methods.

It has various definitions and few of them are listed below.

It's a field of computer science and linguistics concerned with the interactions between computers and human languages [20].

Instead of using Boolean logic, the user simply can type in a question as a query. The simplest processing just removes stop words and uses statistical approaches. Natural language processing is the process of using linguistic analysis to infer meaning from human-written text that could not be extracted using the individual word meanings [21].

It's a simulation of human language processing on the computer by programming the knowledge of human cognitive mechanisms [22].

- It's a branch of artificial intelligence that deals with analyzing, understanding and generating the languages that humans use naturally in order to interface with computers in both written and spoken contexts using natural human languages instead of computer languages [8].

#### 2.2 Approaches

There are three main approaches for natural language processing, they are, Symbolic, Statistical and Connectionist. [9]

### 2.2.1 Symbolic

This method originated in the late 1950's and this tries to model intelligent behaviour using physical symbol systems. Under this approach knowledge about the language is included in to rules or other ways of representation (algorithms, data, etc.), [7]

#### 2.2.1.1 Logic based systems

These systems use mathematical logic, and symbols are taken as logic propositions. Logic programs use a declarative style of programming and computation is done through logic deduction. One of the famous logic based programming languages is Prolog.

#### 2.2.1.2 Rule based systems

These systems use a set of rules, inference engine and working memory. A rule consists of 2 parts, condition (left hand side) and conclusion (right hand side) and it is used to represent knowledge. The condition should be true to fire the conclusion.

If (Condition) then (conclusion).

If ( $A > B$ ) then ( $B + 1$ )

The inference engine matches the facts against the rules to select a rule and execute it [9].

#### 2.2.1.3 Semantic networks [9]

This is a way of representing knowledge through interconnected nodes and arcs. There are various types of semantic networks, and graphic representation is the most common feature of them. The most famous classification for semantic networks is John F. Sowa's classification. They are as follows:

- Definitional networks
- Assertional networks
- Implicational networks
- Executable networks
- Learning networks
- Hybrid networks

### Definitional Networks

This highlights sub type or is-a relation between a concept and its sub type. The oldest known network of this type is the ‘Tree of Porphyry’ drawn by Porphyry a Greek philosopher. Figure 2.1 shows the Tree of Porphyry.

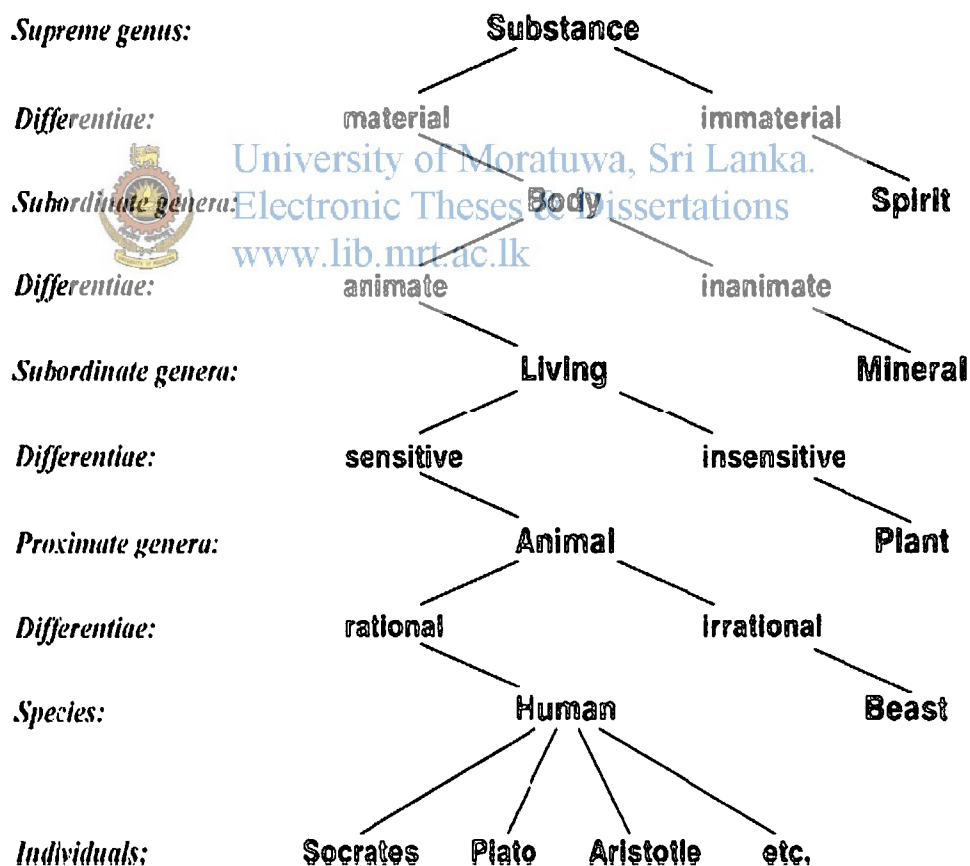


Figure 2.1 - Tree of Porphyry Drawn by Peter of Spain. [6]

## Assertional Networks

Assertional networks are a type of a semantic graph, which is used to make assertions about the world and are believed to be true. In these graphs the nodes represent concepts and the edges represent statements, beliefs or assumptions about the concepts. Figure 2.2 shows a sample concept map.

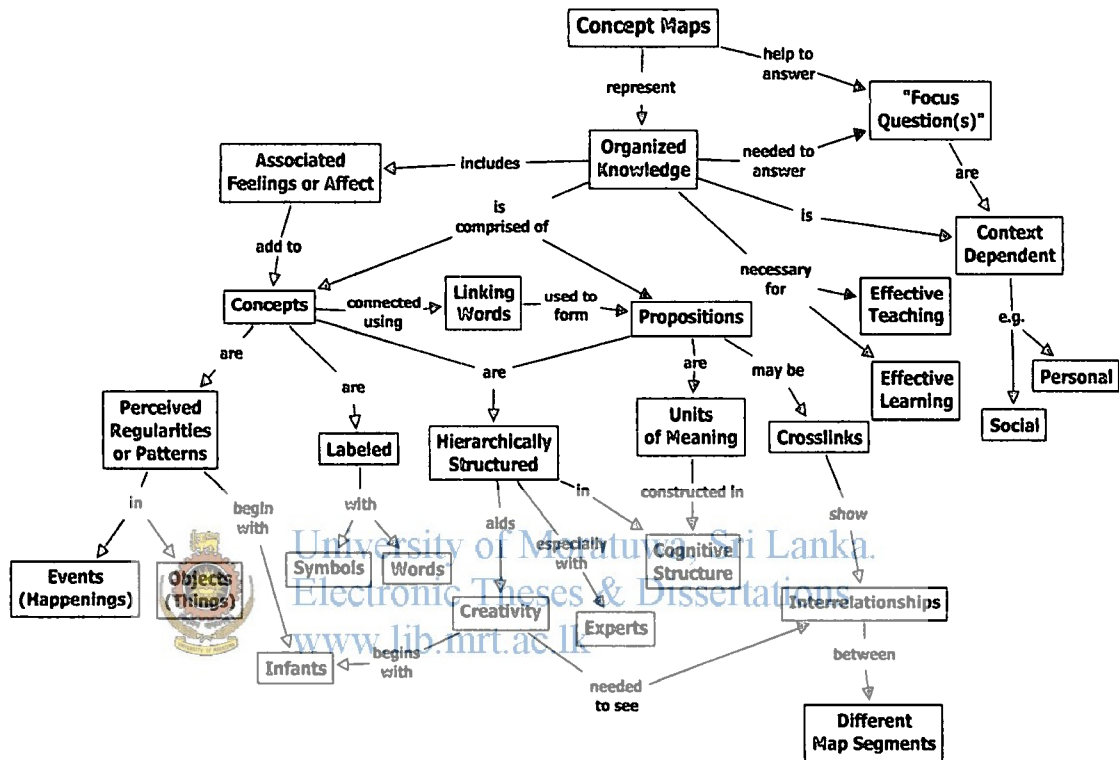


Figure 2.2 - Concept Map [18]

## Implicational Networks

This type of semantic networks has propositional nodes and a relationship called implication with an arrow head indicating the next proposition. If the proposition is true then it will imply to the next proposition. If it fails there is no indication about the second proposition. Figure 2.3 shows a sample implicational network.

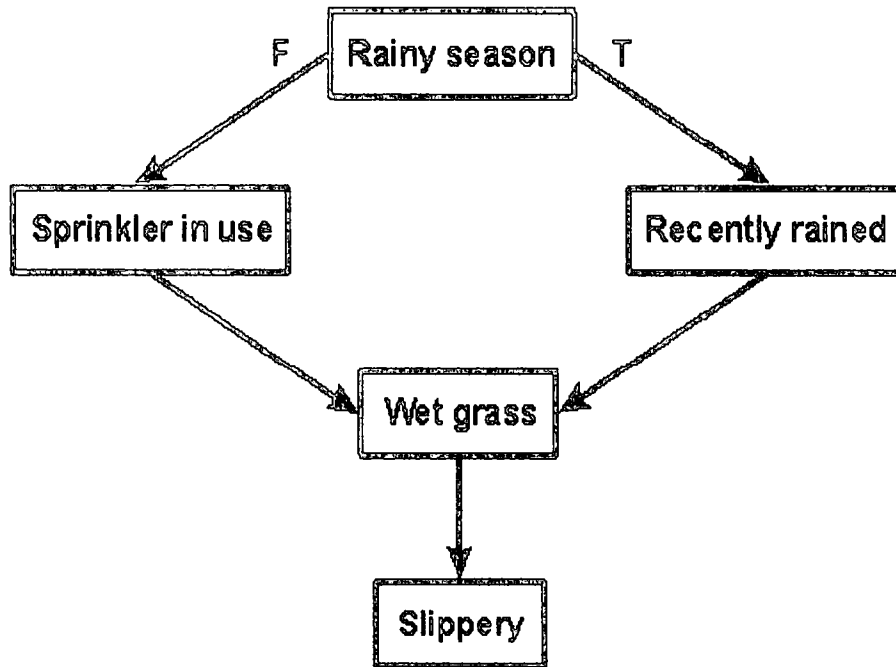


Figure 2.3 - An Implicational Network for Reasoning About Wet Grass. [6]

#### Executable Networks

These networks have mechanism to make some changes itself. The most common kinds are message passing (Pass data from one node to another, where the data have tokens or triggers for other nodes), attached procedures (A program contain/ in a node performs an action/computation on data that node or nearby node has) and graph transformation (Combine, modify or break graphs in to other graphs by programs external to those graphs triggered).

Figure 2.4 shows a data flow graph with rectangles for passive nodes, which contain data and diamonds for active nodes, which contain functions.

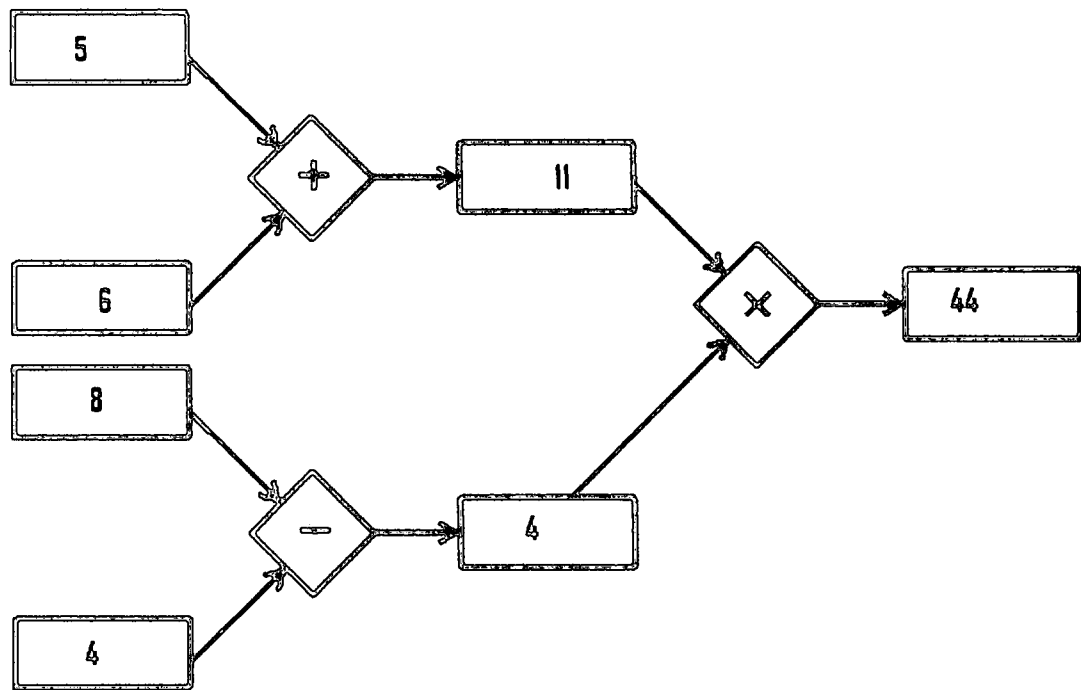


Figure 2.4 - A Data Flow Graph

### Learning Networks



University of Moratuwa, Sri Lanka.

Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

When a learning network encounters a new detail/information it will modify itself to handle the new detail/information. There are 3 main ways to achieve this.

1. Convert the new detail/information and add it as a part of the network.
2. If the nodes of the network have weights/probabilities, the network will update the weight/probability according to the new detail/information.
3. Restructure the network according to the new detail/information. This will be the hardest but the most efficient method.

Figure 2.5 shows a sample learning network.

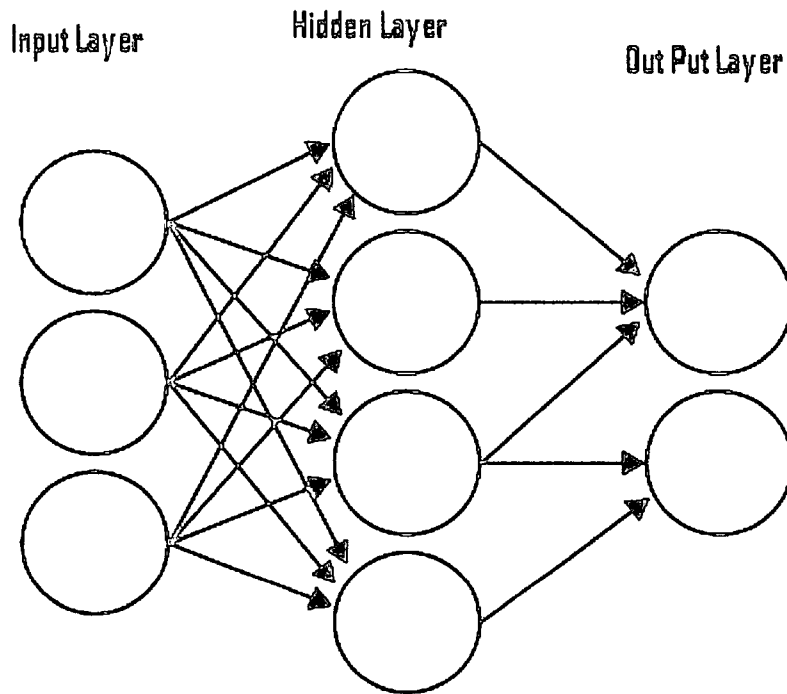


Figure 2.5 - A Learning Network

When a new data/information is available, the weight/probability of the nodes of the input layer and input are combined and will determine the weight/probability of the nodes in the hidden layer. Finally, all the weights/probabilities will decide the weight/probability of the output.

### Hybrid Networks

In this type of network, various problems can be handled by various methods. But when one technique is inadequate, researches start developing hybrid networks which have the strengths of several techniques and are better prepared to handle problems more efficiently.



## 2.2.2 Statistical (Stochastic)

This uses various mathematical techniques with large sets of data (Corpora) to create models. Some of the main areas of statistical techniques are:

- N-Gram Model
- Hidden Markov Models
- Probabilistic Context Free Grammar

### 2.2.2.1 N-Gram Model

The N-gram (When  $n = 1$  called unigram, when  $n = 2$  bigram,  $n = 3$  trigram.) model is a stretch of  $n$  words and can be used to predict the next word based on the available  $n$  words. This uses a corpus to gather prior information (training data).

Unigram model  $P(W_1)P(W_2)...P(W_n)$

Bigram model  $P(W_1)P(W_2|W_1)P(W_3|W_2)...P(W_n|W_{n-1})$

Trigram model  $P(W_1)P(W_2|W_1)P(W_3|W_1W_2)...P(W_n|W_{n-2}W_{n-1})$

N-gram model  $P(W_1)P(W_2|W_1)...P(W_n|W_{n-n-1}...W_{n-1})$

### 2.2.2.2 Markov and Hidden Markov Model

In the Markov model all the states are visible to an observer and the parameters will be the state transitions probabilities. But in the Hidden Markov model, only the outputs which depends on states are visible and not the states. A state has different probabilities for all outputs, and a sequence of outputs will give information about the sequence of states. These models are mainly used for pattern recognition in speech, handwriting, speech tagging, etc.

## Markov Model

Figure 2.6 shows a sample Markov model.

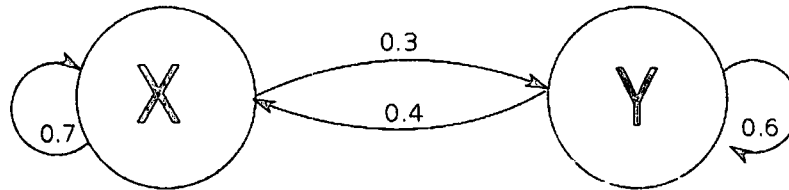


Figure 2.6 - Markov Model

Starting Probability = {'X': 0.6, 'Y': 0.4}

If we want to calculate the probability for the sequence {'Y', 'Y', 'X', 'X'} =  $P\{X | X\} P\{X | Y\} P\{Y | Y\} P\{Y\} = 0.7 * 0.4 * 0.6 * 0.4$



Figure 2.7 shows a sample Hidden Markov model.

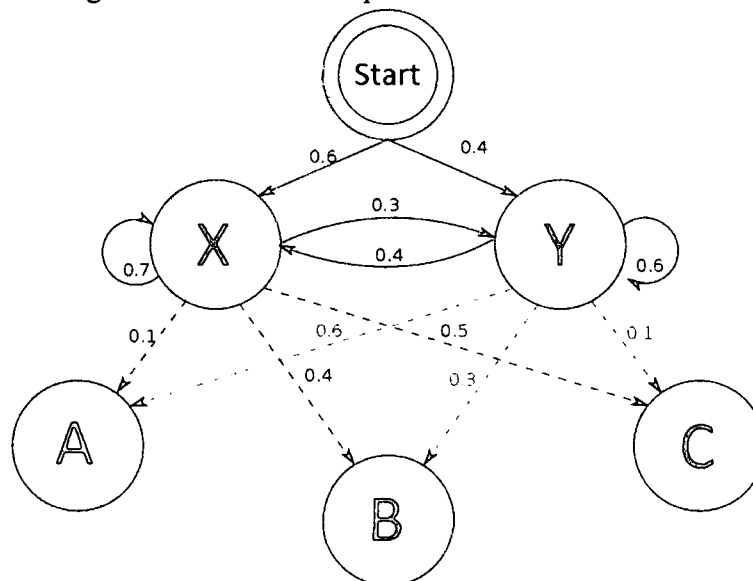


Figure 2.7 - Hidden Markov Model

States = ('X', 'Y')

Observations = ('A', 'B', 'C')

Starting Probability = {'X': 0.6, 'Y': 0.4}

Transition Probability = {'X' : 'X', : 0.7, 'X' : 'Y': 0.3, 'Y' : 'X': 0.4, 'Y' : 'Y': 0.6}

Emission Probability = {'X' : 'A': 0.1, 'X' : 'B': 0.4, 'X' : 'C': 0.5, 'Y' : 'A': 0.6, 'Y' : 'B': 0.3, 'Y' : 'C': 0.1}

If we calculate the probability for {'C', 'B'}

$$P\{\text{'C','B'}\} = P(\{\text{'C','B'}\},\{\text{'X','X'}\}) + P(\{\text{'C','B'}\},\{\text{'X','Y'}\}) + P(\{\text{'C','B'}\},\{\text{'Y','X'}\}) + P(\{\text{'C','B'}\},\{\text{'Y','Y'}\})$$

For first term,

$$P(\{\text{'C','B'}\},\{\text{'X','X'}\}) = P(\{\text{'C','B'}\} | \{\text{'X','X'}\}) P(\{\text{'X','X'}\}) = P(\{\text{'C'} | \text{'X'}\}) P(\{\text{'B'} | \text{'X'}\}) P(\{\text{'X'}\}) P(\{\text{'X'} | \text{'X'}\}) = 0.5*0.4*0.6*0.7$$

### 2.2.2.3 Probabilistic Context Free Grammar

In this method probabilities are assigned to each parse tree to resolve syntactic ambiguity and pick the most probable parse tree.

e.g. Book the flight through Houston

According to the probability we can select first parse tree (D1) from Figure 2.8.

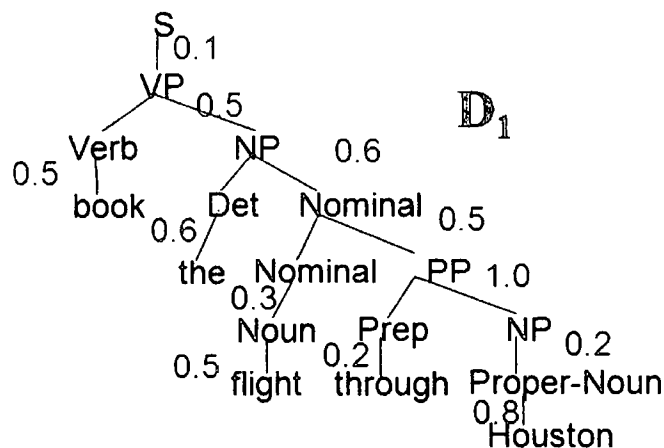


Figure 2.8 - Parse Tree 1 [19]

$$P(D1) = 0.1*0.5*0.5*0.6*0.6*0.5*0.3*1.0*0.2*0.2*0.5*0.8 = 0.0000216$$

According to the probability we can select second parse tree (D2) from Figure 2.9.

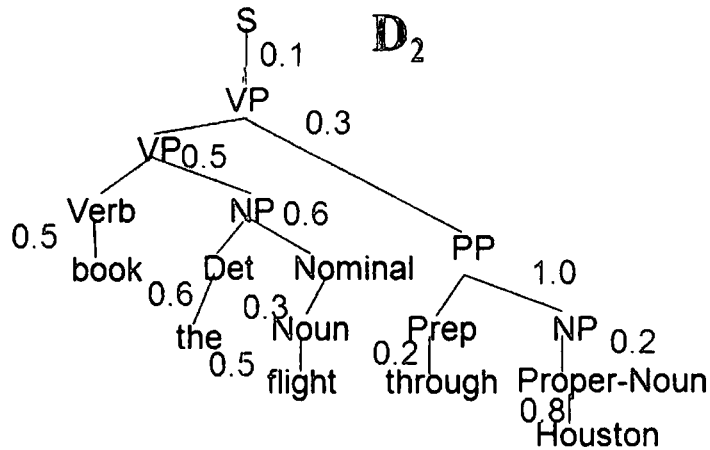


Figure 2.9 - Parse Tree 2 [19]

$$P(D2) = 0.1 * 0.3 * 0.5 * 0.6 * 0.5 * 0.6 * 0.3 * 1.0 * 0.5 * 0.2 * 0.2 * 0.8 = 0.00001296$$

### 2.2.3 Connectionist

This is very much similar to statistical methods. The main difference is this connects statistical learning with various theories.



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

### 2.3 Existing Systems

#### 2.3.1 ELIZA

Eliza first introduced by Professor Joseph Weizenbaum in 1965 as a computerized conversation simulator. It has the facility to plug in different scripts to represent different characters and the most famous character is "Doctor" which simulates a psychotherapist. Eliza cannot understand anything, and it is designed only to search patterns in the entered text and use that to generate responses. Eliza is written in Basic and it runs on 16 Kb of RAM. [10].

Some of the early computer games based on ELIZA are Ecala, Dungeon, Moria, etc.. Programs like Abuse, ZEBEL, Jesus and I Am Buddha are based on ELIZA but use different programming languages.



### 2.3.2 Dr. Sbaitso

This was an AI program for DOS based computers in the early 1990's, created by Creative Labs inc. This program tries to simulate a psychologist. The AI engine is similar to the ELIZA algorithm and can calculate simple mathematics [11].

### 2.3.3 PARRY

PARRY is another early chatter bot which attempted to simulate a paranoid schizophrenic (Mental disordered patient). The code compiles and runs using MLISP language (meta-lisp) on the WAITS operating system running on DEC PDP-10. Some parts of the code are written in assembly code [12].

### 2.3.4 Racter

This is an artificial intelligent computer program, which is used to generate English language prose at random. Written in BASIC on a Z80 with 64k of RAM and run on a CP/M machine and used to compose a book called 'The Policeman's Beard Is Half Constructed'. The macintosh version includes a speech synthesis [13].

### 2.3.5 MegaHAL

MegaHAL is a computer conversation simulator. When the user types a sentence MegaHAL will answer with a sentence. But MegaHAL doesn't understand the conversation or the sentence structure; it generates its answers based on sequential and mathematical relationships.

The procedure MegaHAL uses is, first it takes the user sentence and breaks it into words (series of alphanumeric characters) and non words (series of other characters) to learn new words. Using two 4<sup>th</sup> order Markov model (One to predict which symbol follows the sequence of 4 symbols and the other to predict which symbol precedes the sequence of 4 symbols). It tries to generate the replies based on the key words of the input. Frequently occurring words like 'The', 'And' will be discarded and the remaining words are transformed if necessary e.g. my to your [14].

### 2.3.6 Ultra Hal Assistant

The Ultra Hal Assistant is a chatter bot Loebner prize in 2007. Which is developed by Robert Medeksza of Zabaware. Ultra Hal Assistant has a lot of features which other chatter bot's do not have [15]. A few of them are:

1. It has many animated characters to choose from and it generates sound and the user is able to speak to it. (Windows version).
2. Ultra Hal can remember important details, dial phone numbers, remind important dates.
3. It can find all the windows programs in the start menu and run them.
4. Helps to browse the internet.

### 2.3.7 Elbot

Elbot is a chatter bot created by Fred Roberts using artificial solutions technology. It has won some of the artificial intelligence competitions such as Chatterbox Challenge and the Loebner Prize. [17]

## 2.4 Summary

This chapter covers mainly about the various natural language processing techniques and some of the famous chat bots.

The first section discussed about natural language processing techniques such as symbolic approach, where the intelligent behaviour of they system is modelled using symbols; Statistics techniques, which are the most recent trend of natural language processing, usually based on large corpora performing analysis using text characteristics; and Connectionist techniques, which connects statistical learning with various theories.

The second section discussed some of the successful chat bots which currently exist such as Elisa, Parry and MegaHAL with their details.



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

## Chapter 3

### Approach to the Solution

#### 3.1 Introduction

Based on the gathered knowledge the following approach was designed to create the automated chat system. The proposed system has the following three major areas,

- Identifying the question
- Extract the details
- Finding the best prepared answer

Apart from those three major functionalities the system will update the frequencies/probability of questions and its information.

#### 3.2 Design

There are 7 main areas in the Chat bot, such as Identify the question, Extract the details, Database, Prepare the answer, Knowledge capture, Information capture and Knowledge engineer interface. With all these features Chat bot will try to give the best possible answer for customer queries.

##### 3.2.1 Identify the Question.

The first step of the chat bot system is to identify the customer queries. To achieve this chat bot uses a set of sample queries stored in the systems knowledge base and an algorithm is used to determine the most suitable query. Refer Appendix A.

##### 3.2.2 Extract the Details

After identifying the most suitable sample question which matches to the customer query, the chat bot tries to extract useful information from the queries. This is also achieved with the help of the sample questions stored in the knowledge base.



### 3.2.3 Database

The database of the chat bot has 2 parts. The first part is the knowledge base which maintains all useful information to identify the questions, extracted details from them and answer templates. The other section is the company database where the useful stock information and other relevant information are stored.

### 3.2.4 Prepare the Answer

With all the required information available, the chat bot moves to the next step, which is to prepare the answer. Based on the captured knowledge and the company database information it will select the most suitable pre defined answer (also stored in the knowledge base), to build the final answer filling the required information and produce to the customer.

### 3.2.5 Knowledge Capture

While providing answers to customer queries chat bot also has the ability to update its knowledge base by updating the frequencies to increase its accuracy for future queries.

### 3.2.6 Information Capture

This tool can be modified to capture key words in client queries with time and to save in the database and can be used in forecasting purposes and to increase high quality decision making.

### 3.2.7 Knowledge Engineer Interface

This is a separate application which shares the chat bot database and the company database and helps the Knowledge Engineer to maintain the chat bots' knowledge/data.

### 3.3 Scope of the project

- This application is developed only to handle queries related to a Book Shop.
- The system will identify only some of the queries based on the provided knowledge base.
- The system can only alter the probabilities of existing queries and answers, based on the usage of the system.

### 3.4 Limitations of the scope

- This system depends on the domain knowledge, which has to be provided by an external party (Knowledge Engineer) to the system.
- An operator or a Knowledge Engineer's support will be required to feed unhandled information as new knowledge to the system.

### 3.5 Technology Used



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

Following are the summary of technologies which are being used in developing chat bot application.

Operating System	Windows XP
Programming Languages	Java, MySQL
Database server	MySQL
IDE	Eclipse, Net Beans

Table 3.1 - Technology Stack

Windows XP Operating System:

This is the most famous operating system produced by Microsoft. It is common among majority of the computer users and comes with two main versions; home and professional. To design the chat bot the Windows XP Home service pack 3 was used.

### **Java:**

Java, created in 1991 by James Gosling of Sun Microsystems and initially called Oak, before being renamed to Java was originally designed for mobile cell phones. Java comes under the GNU General Public License.

It works as:

- Programming language - It can create any application a conventional language can.
- Development environment - It provides compiler, interpreter, documentation generator, class file package tool, etc.
- Application environment - Runs on any machine where the Java runtime environment is installed.
- Deployment environment - JRE and the environment in the web browser.

It has features such as, JVM (Java Virtual Machine an imaginary machine emulating the software or a real machine), Garbage collection, Code, etc. [23].

### **MySQL:**



University of Moratuwa, Sri Lanka.

Electronic Theses & Dissertations

www.lib.mru.ac.lk

My SQL is a popular open source relational database management system (RDMS) which runs as a server providing multi user access. Some of the large scale web products like Flickr, Nokia, YouTube, Wikipedia and Face book use the MySQL database. My SQL is a RDBMS and comes without GUI tools to administrator and manage data, but various third party frontends are available.

### **Eclipse:**

This is a multi language software development environment with an integrated development environment (IDE). Eclipse released under the terms of Eclipse Public Licence is a free and open source software.

### **Net Beans:**

This is an integrated development environment (IDE) and a platform framework for developing language such as Java, JavaScript, PHP, Python, etc. This framework was designed using Java and with the availability of the Java Virtual Machine (JVM) can be used in any operation system.

### 3.6 Software Licensing Issues

Since the whole system is developed using open source tools, spending large sums of cash purchasing license for software development tools, database management systems is not required.

### 3.7 Summary

This chapter covered the design issues, decisions taken in developing the proposed system and the technology proposed to design of the chat bot.



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

## Chapter 4

### Analysis and Design

#### 4.1 Introduction

This chapter discusses the main requirements arranged through a thorough research on natural language processing, and the design of the proposed prototype and its functionality.

#### 4.2 Identified Functional Requirements

- System shall provide an interface to customer to make his/her queries. (Similar to a chat application).
- System shall have facilities to get information from the organization data base.
- System shall have a knowledge base.
- Knowledge engineer shall have facilities to include new information, update knowledge base.
- System shall have the facilities to identify the customer queries.
- System shall have facilities to obtain the operator help if required.

#### 4.3 Identified Non Functional Requirements

- Knowledge engineer should provide correct information.
- Chat bot should have a relatively large knowledge base to achieve higher level of accuracy in question identification.
- An operator's assistance is required to handle queries which chat bot fails to handle.

## 4.4 Proposed System

### 4.4.1 Main areas of the Chatbot

Figure 4.1 illustrates the proposed system design.

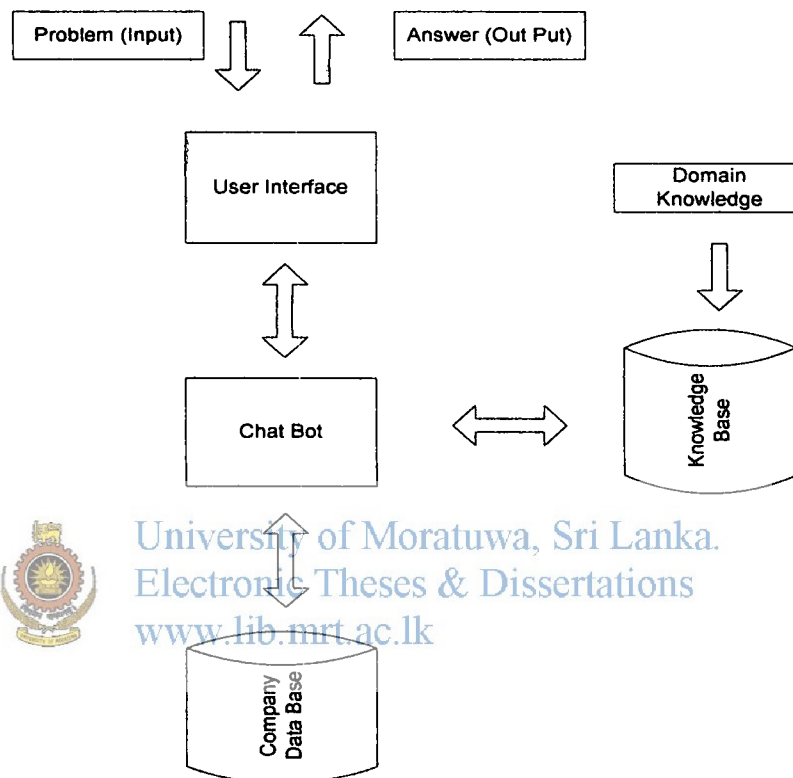


Figure 4.1 - Proposed System

The proposed system has four main areas, User Interface, Chat bot, Knowledge Base and Company Database.

**User Interface** – This is where the users of the Chat bot interact with it. This interface is similar to any chat window. This has 4 main areas as illustrated in Figure 4.2.

- Display area where all the past chats and answers are displayed.
- Chat Area where the user types the queries.
- Exit button to exit from the Chat bot.
- Close and Minimize buttons to close the chat window or to minimize the chat window.
- Submit button to submit the typed queries to the Chat bot.

**Knowledge Engineer Interface** – This is the area where the knowledge engineer manages the knowledge and stock information. The knowledge engineer has the ability to add and edit new knowledge/information to the two databases used by the Chat bot as shown in Figure 4.3.

**Chat Bot** – This is the heart of the chat bot where it identifies the customer queries, and generates answers.

**Knowledge base** – This is where all the supporting knowledge required to chat bot's operations is stored. Chat bot has the ability to update certain data (probabilities) and new knowledge should be given to chat bot through this database.

**Company database** – This is where the chatbot receives the stock and related information to generate answers for customer queries. This same database can be used for knowledge capturing about customer queries against time.

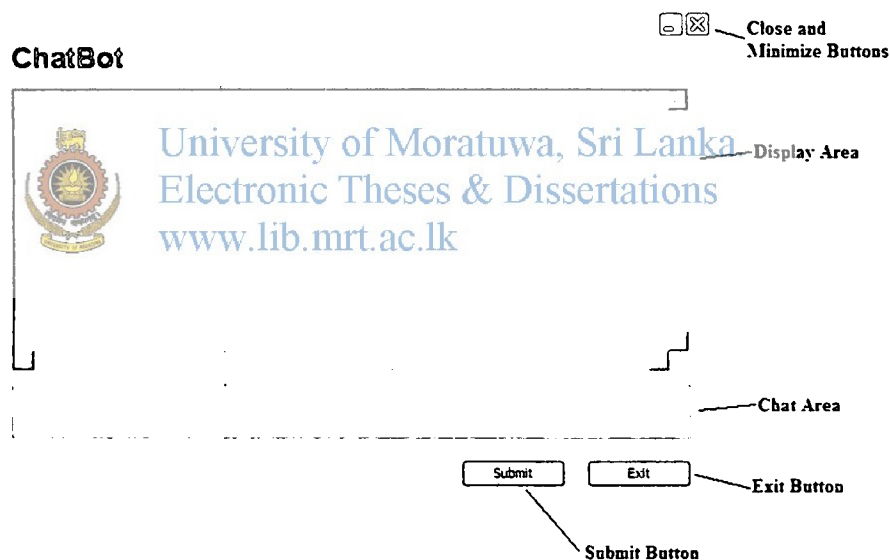


Figure 4.2 - Chat Window

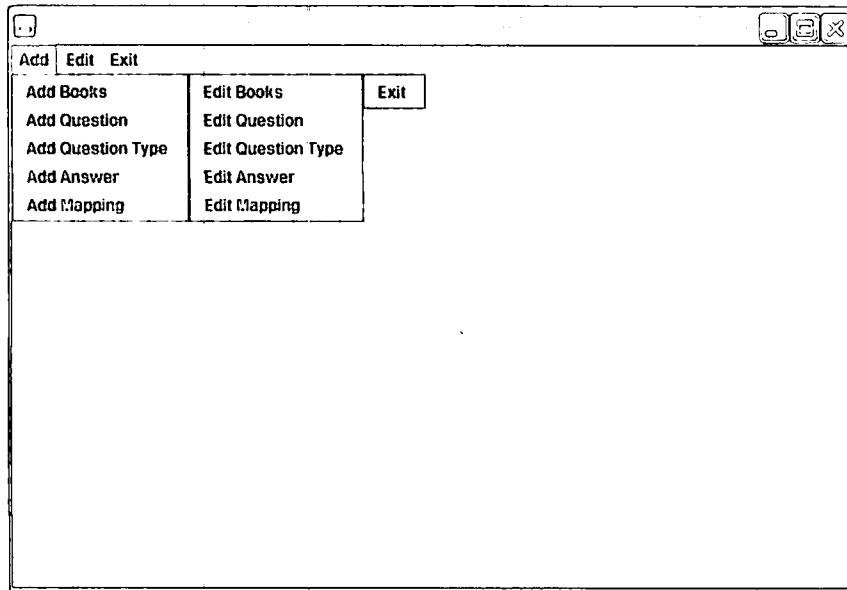


Figure 4.3 - Knowledge Engineer Interface

#### 4.4.2 Architectural Diagram

Figure 4.4 shows the architectural diagram of the system.

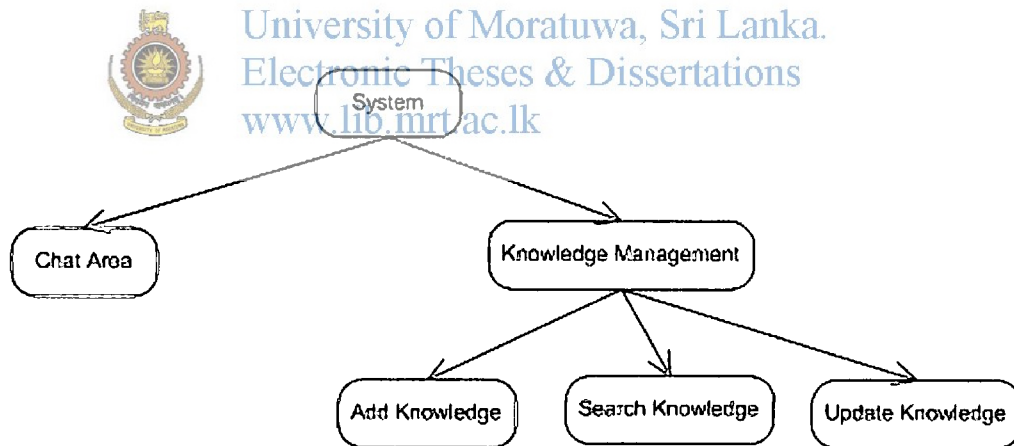


Figure 4.4 - Architectural Diagram

**Chat Area** – This is where the customer interacts with the system. Customer can type queries and will get answers through this window.

**Knowledge Management** – This is where the knowledge engineer maintains the knowledge of the system. He/she can add, search and update knowledge. Access to this section is only for selected persons.



### 4.4.3 Use Case Diagram

Figure 4.4 Figure 4.5 shows the use case diagram of the system.

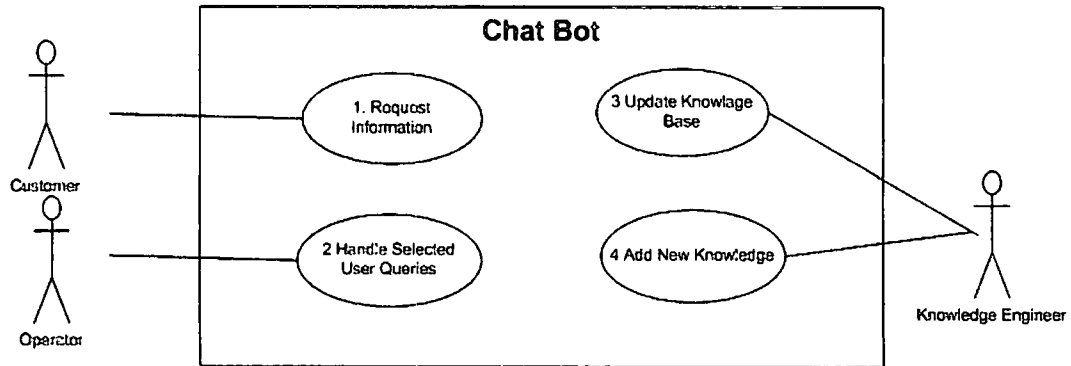


Figure 4.5 - Use Case Diagram of Proposed System

### 4.4.4 Data Base Design

The Following ER diagram (Figure 4.6) shows the RDBMS tables and their relationships.

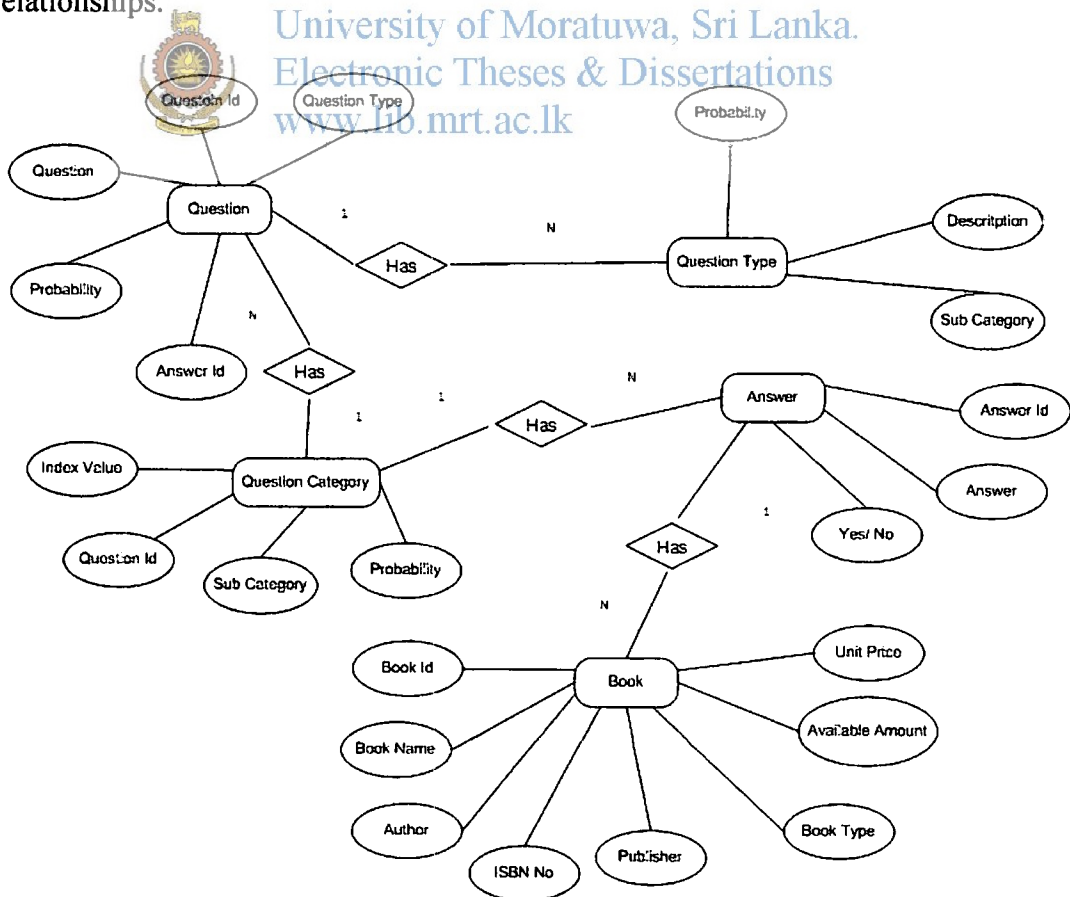


Figure 4.6 - ER Diagram


The system maintains a database to store details about the domain. In-addition to details of the domain, the same database is used to store sample questions, sample answers and other related data required by the application to work smoothly.

The same database keeps domain information (Book shop) separate, which will be used to generate answers to customer queries.

#### 4.5 How Chatbot Works

##### 4.5.1 Identify the Question.

The chat bot will maintain a sample set of questions related to the domain. Each of those questions has its own probability which will be changed/updated on each customer query. The following steps will be followed to identify the customer query:

- 
- University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
www.lib.mrt.ac.lk
- When the Chat bot receive a customer query it will select the sample predefined question with the highest probability and tries to match it with the customer query. If the matching results are lower it will select the next probable sample question and match it.
  - If the Chat bot identifies a successful match its next step is to extract the core details of the customer query. This is achieved by removing the identical words of the sample question and the customer query and taking the remaining word/s.
  - Each sample question has a predefined sub category. With the help of the sub category the Chat bot tries to find a suitable matching with the core details in the company database.

As an example if we take the bookshop domain.

If the selected sample question is “Do you have \_?”, the most suitable sub category will be “Book Name”, “ISBN Number” or “Book Type”. Other categories like “Author”, “Publisher”, etc are not common with the selected sample question.

Each sub category has their unique probabilities (Updated as the sample question probability after each successful matching).

Using the same example if “Book Name” has the highest probability for the selected sample question, the Chat bot will search under the “Book Name” column of the company database to find any matching.

If a successful matching is found by the Chat bot, it will update the probabilities for the selected question by adding 0.01 and subtracting 0.001.

- At this stage the Chat bot has the following details. Sample answer for the identified question, details about the core information, and the core information available in the company database.
- The sample questions used in the chat bot are obtained through a small survey done among a selected set of people. Where the selected people are asked to provide various types of questions they are going to ask in the selected domain (Book shop).

#### 4.5.2 Prepare the Answer

With the identification of the customer queries and the required data the Chatbot has to have the ability to generate successful answers to complete the dialog, and uses the following method to do so.

- At the sample question design stage the knowledge engineer has to identify the possible positive and negative answers related to each sample question. The Chatbot using the assigned answers to the already identified question determines the correct positive or negative answer and displays it to the customer.
- The positive or negative identification are made based on the search results of the company database. If it is having matches for the core information searched in the database will result a positive answer, vice versa will result in a negative answer.

#### 4.5.3 Update Probability Based on Frequency

The Chat bot requires some sort of mechanism to identify the most likely sample question first. It is highly unlikely to obtain a large set of training data, to set the initial probabilities of each sample question which is used to identify the customer queries. Hence, the method used to set initial probabilities of the sample questions is as follows.

- Give same probability “1” or a pre decided value for all sample questions. And provide the Chat bot with a mechanism to adjust each questions probability based on each customer queries. If the Chat bot successfully matches a customer query to a sample question it will automatically increase the probability of that sample question by 0.01, making that question’s chances of getting selected next time higher. At the same time the Chatbot will decrease the probabilities of all other sample questions by 0.001 to reduce their chances of selecting next time. With this method we expect to get more and more stable probabilities for sample questions based on customer queries.
- Sample survey results (frequencies of each question type) not used for setting up the initial probabilities, due to small number of people used in the survey.

The above mentioned methods are used to update the sub category probabilities. This way the Chat bot tries to learn from the experience as discussed under learning networks in Chapter 2.

#### 4.6 Main Issues Faced

- The coding language used is Java and only partially familiar with it. Eclipse and Net Beans are IDEs which require putting an effort to learn them.
- Had to refer larger number of documents/research papers to get a good understanding about natural language processing and its techniques.

- Amzi prolog and SWI prolog languages were learnt as AI languages, but due to the new areas, complexities, and new way of coding used in those languages (logic programming) and from time limitations, this was not used for chat bot design and development.
- Statistical techniques were not used for planning and development since they require larger corpuses and training data, which are harder to obtain freely.
- Most of the existing systems were designed for various competitions and therefore very little technical and design information were available.
- Certain high quality research papers were not freely available which resulted in some limitations of the research of existing technologies.

#### 4.7 Summary

Under the design phase we discussed about the design of the proposed system through activity, use case and ER diagrams. Furthermore, we explained how the Chat bot works to achieve its target and how it learns (update its knowledge base) and the problems faced in developing the system.



# Conclusion and Future work

### 5.1 Conclusion

The main aim of the chat bot is to provide a basic idea of a technique which can be used to improve small scale businesses. With the implementation of the chat bot businesses can increase the customer satisfaction by providing information fast and accurately than the existing e-mailing options. This system is also used to provide 24 hour customer supports in a much cheaper way, which is a major requirement for global companies. With some more improvements we can make this chat bot successful as a commercial product.

### 5.2 Future Work

Although the system has the ability to handle some of the customer queries, there are many areas which can be further improved.



University of Moratuwa, Sri Lanka  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

Currently, the system is using simple matching techniques and that can be improved to use much more complex methods and algorithms, which will provided a much more accurate understanding ability to the system.

Instead of using a predefined common probability, if we can achieve a much more suitable probability for each sample question through a survey, we can get much more accurate results from the beginning.

The initial knowledge base is relatively small and contains only the knowledge acquired by the knowledge engineer and its ability to improve the efficiency of the chat bot is limited. Adding more training data to the knowledge base of the chat bot will lead to more accurate question identifications.

Adding a knowledge capture mechanism of customer queries can add additional value to the chat bot. That is, with the ability to identify the customer inquires about certain goods can be combined with the time period (time and date) and stored in the company database. After a large amount of such data is collected, it can be used to forecasting customer trends/patterns. This can help the vendors in various ways, will save the time and money spent on various marked surveys and sharing those information with vendors can earn extra profits, and allows to be prepared well for future sales.

Further, including a mechanism to direct customer queries which fails to be identified by the system to an operator can be an added advantage.

### 5.3 Summary

In this chapter we have discussed about the limitations of the proposed Chat bot and the areas of improvement. With the introduction of these proposed new features we can expect to have a much more efficient and user-friendly Chat bot which can handle more complex dialogs.

## References

- [1] <http://seekingalpha.com/article/192498-forecast-online-retail-sales-will-grow-to-250-billion-by-2014>
- [2] [worldaccordingtocarp.wordpress.com/2010/](http://worldaccordingtocarp.wordpress.com/2010/)
- [3] <http://ezinearticles.com/?7-Reasons-Online-Shopping-Malls-Are-So-Popular&id=2377512>
- [4] [http://www.offt.gov.uk/shared\\_offt/reports/consumer\\_protection/oft921.pdf](http://www.offt.gov.uk/shared_offt/reports/consumer_protection/oft921.pdf)
- [5] <http://www.squidoo.com/7dollars-shorturl>
- [6] <http://www.jfsowa.com/pubs/semnet.htm>
- [7] <http://www.cs.cofc.edu/~manaris/publications/advances-in-computers-vol-47.pdf>
- [8] <http://www.webopedia.com/TERM/N/NLP.html>
- [9] <http://www.cnlp.org/publications/03NLP.LIS.Encyclopedia.pdf>
- [10] <http://www.filfre.net/2011/06/eliza-part-1/>
- [11] [http://download.cnet.com/Dr-Sbaitso/3000-2121\\_4-10656151.html](http://download.cnet.com/Dr-Sbaitso/3000-2121_4-10656151.html)
- [12] <http://lazytoad.com/lti/pub/aaai94.html>
- [13] <http://www.robotwisdom.com/ai/racterfaq.html>
- [14] <http://megahal.alioth.debian.org/>
- [15] <http://www.zabaware.com/fb.html>
- [16] <http://www.elbot.com/artificial-solutions>
- [17] <http://www.elbot.com/chatterbot-elbot/>
- [18] <http://itcboisestate.files.wordpress.com/2008/02/fig1cmapaboutcmaps-large.png>
- [19] [userweb.cs.utexas.edu/~mooney/cs388/slides/stats-parsing.ppt](http://userweb.cs.utexas.edu/~mooney/cs388/slides/stats-parsing.ppt)
- [20] <http://www.elbot.com/artificial-intelligence-faq/>
- [21] <http://www.microsoft.com/enterprisesearch/en/us/search-glossary.aspx>
- [22] <http://www.philhist.uni-augsburg.de/lehrstuehle/anglistik/sprachwissenschaft/mitarbeiter/stoll/term/>



[23] <http://www.javapassion.com/javase/javaintro.pdf>



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

## Appendix A

### Algorithms Used in the Research

#### Levenshtein distance

- 1 : initialize integer matrix d
- 2 : initialize integer n, m, i, j
- 3 : initialize character s\_i
- 4 : initialize character t\_j
- 5 : initialize string temp
- 6 : initialize integer cost
- 7 : n equals length s
- 8 : m equals length t
- 9 : if n equals 0
- 10 : return m
- 11 : if m equals 0
- 12 : return n
- 13 : if n greater than m
- 14 : temp equals s
- 15 : s equals t
- 16 : t equals temp
- 17 : n equals m
- 18 : m equals length of t
- 19 : d equals new integer n+1, m+1
- 20 : for i equals 0 to i less than or equals n increment i
- 21 : d[ i],[0] equals i
- 22 : for j equals 0 to j less than or equals m increment j
- 23 : d[0][j] equals j
- 24 : for i equals 1 to i less than or equals n increment i
- 25 : s\_i equals character at i-1 of s
- 26 : for j equals 1 to j less than or equals m increment j
- 27 : t\_j equals character at j-1 of t
- 28 : if s\_i equals t\_j

29 : cost equals zero

30 : else

31 : cost equals 1

32 :  $d[i][j]$  equals minimum of  $d[i-1][j]+1$  or  $d[i][j-1]+1$  or  $d[i-1][j-1]+$  cost

33 : return  $d[n][m]$

Source : <http://www.merriampark.com/ld.htm>



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

