CHAPTER 1

Introduction

1.1 Background and Motivation

As a human computer interface, Natural Languages (NL) are the easiest to use for human beings, and they are the most intuitive interface for conversing with human. The system with such an interface could be a chatting toy or a conversational agent, enabling human users to communicate with a computer, using everyday spoken languages. Such systems need to have a Natural Language Processing (NLP) component to process and extract useful information from input sentences such as topic, sentiment or instructions. The main challenge is that inputs in natural languages do not follow any standard format, therefore, extracting useful information by NLP is still a challenging research topic.

There are a few natural language-based conversation systems that are worth mentioning. Façade is firstly one of the best natural language based game that use NLP for human computer interaction. It attempts to extract interesting information about players who interact by typing text. According to Mateas and Stern, "The Facąde NLP system accepts surface text utterances from the player and decides what reaction(s) the characters should have to the utterance" [13]. Another system is Alice which is one of most popular chat bot and a predominant conversational software. It uses a specific language that is called AIML (Artificial Intelligence Markup Language), which enables people to insert knowledge into Alice in a machine readable format [22].

1.1.1 Façade

Central to the application of interactive conversations in games and artificial intelligence is field of education and entertainment [5]. Therefore, the developers promote many kinds of applications that can interact with users by texting, voicing or touching. The chat bot examines the players' emotion by extracting some sentences that might indicate the players situation. However, far too little attention has been paid to the researchers perform further searches that are significant in information extraction area, because the sentences that have input do not follow any standard format, and also extracting by Natural Language Processing is still challenging in particular for the recognition of the gamers feeling or mood by only textual input [5].

Mateas and Stern [13] developed the Façade game, which is classified as an interactive artificial intelligence system. It has the capability of interacting with users via textual inputs to influence the game direction. Façade is an instantaneous, first-person natural language-based game that uses the Natural Language Processing technique to enable human computer interaction. In effect, the chat bot of Façade extracts some interesting information from the surface text and determine the level of information input by the player [13].

The game operates by putting the player in the role of close associate with the major antagonists, (Grace and Trip), a couple who invite the player to their home for a drink, refreshments and a conversation. However, the pleasant gathering of the couple and the player is disrupted by domestic conversation between the couple the minute the player arrives to the house. Through the use of a language processing software, the game allows the player to type sentences to communicate with the couple, either to support them and get them to come to terms with their current argument, or to drive them apart. The latter scenario will eventually lead to the player being thrown out of the apartment [4].

The main consideration of the game is the capability of the game to extract useful information from textual sources and makes decision based on analytical results of the sentences. Additionally, the sentences are received from surface text that have been applied in the chat bot in order to encourage user to insert details into the machine. For example, "if the player types "Grace isnt telling the truth", the NLP system is responsible for determining that this is a form of criticism, and deciding what reaction Grace and Trip should have to Grace being criticized in the current context" [13].

There is a large volume of published studies describing the role of Information extraction, thus researchers perform further investigations to increase understanding level of textual natural language in the chat bot. This supports the capability of machine to analyse users' inputs at the lexical, syntactic and semantic level. Furthermore, Façade rises its capability of understanding what the player wants and how to behave in such situation by analysing the input with high level of meaningful methodology. [13].

1.1.2 Discourse Acts

The process of information extraction from textual input is based on two primary stages, which are keyword definition and relationship recognition. In the first stage, the system divides the sentence into individual words to facilitate the analytical process. However, the second stage identifies the relationship between desired words and database of discourse acts, this is to grow the understanding level. The issue of the understanding level appears in the difficulty of making a decision of what is the most appropriate relationship that matches words and the discourse acts [10]. Basically, developers of Façade designed a set of discourse acts that represents particular meaning. The aim of their research is to address identify relations between discourse acts to textual inputs [13].

Hereby, we try to compare Façade and Alice in several aspects. Façade has more understanding of what input is about, but Alice applies syntactical structure by detecting certain sentence. For instance, if you say “I”, Alice will reply with “you”. Moreover, Alice uses POS (Part Of Speech) tagging to identify part of speech, then attempt to replace or construct a reply. For example, when player says “my name is Waleed”, Alice will respond by “Hello Waleed”. However, Façade determines and understands objects by the use of discourse acts. Finally, Alice has a knowledge base that contains retrieved knowledge. It is used by Alice to extract useful information, but in some points it tends to be rigid rather than flexible, whereas Façade is more flexible. this is because Façad models a rather configurable small environment, whereas Alice is general purpose.

1.2 Project Aim

The main aim of this project is to apply the Façade methodology to process inputs for conversational agent. In particular, we attempt to extract useful information about a specific fictional character from an inserted sentence. This style leads the system to understand the commands input by a user, to improve its performance and to construct meaningful conversation.

1.3 Methodology

This project is based mainly on applying Façade methodology which has several aspects such as discourse acts, template and rules, on conversational agent. Indeed, this methodology depends on the sort of user's sentence. The inserted sentence from a chatter will be tokenised and then apply Part Of Speech (POS) tagging to facilitate information extraction.

A system is built with two modules in order to achieve the project aim. The first module is to gather knowledge about fictional characters from Wolframe Alpha knowledge engine website. The other program is to extract specific details that a user searches for by seeking the keyword in the plaintext. For example, if a user types "Who is spiderman's wife?", the system will extract the details about spiderman's wife.

This project implements a web interface in a PHP with a MYSQL database. The PHP coding involves POS as a class to tokenise the sentence. The MYSQL is used to insert or select details by sending some queries to a database.

1.4 Dissertation Structure

The following details the structure of the rest of this dissertation.

- Chapter 2: Literature Review:
  The literature review chapter clarifies several aspects that could be applied in the project. The chapter first explains the Façadeś surface text processing is further explained as in how it applies NLP for extracting information from user. Another aspect is how discourse acts play a significant role in information extraction to understand the user request. The next aspect is an information extraction using database queries for respond to the right goal. Information Extraction System based on Inductive Learning and Meta learning is another aspect in this chapter. Lastly, the chapter presents the information retrieval system which is a technique that uses noun phrases to search for certain information on the Internet.

- Chapter 3: Methodology:
  This chapter, first explains the process of building and populating a knowledge base about fictional characters. Then how the knowledge base can be used to answer user queries.

- Chapter 4: Experimental Result and Analytical Evaluation
  This chapter describes the questions classifications that used to exam the two applications. The chapter then presents the results obtained throughout the experiment, followed by analytical evaluation of the results in terms of recognising performance.

- Chapter 5: Conclusion and Future Work:
  The conclusion chapter summarises the significant points of this dissertation. It start by reviewing the objectives of the project, it methods and results. The chapter then concludes with some suggestions to guide future research.

CHAPTER 2

Literature review

The ease of access and the abundance of free electronic text have, become the driving force for the surge of information extraction research. The goal of information extraction is to discover and extract useful information into structured storage from free-formal text in natural language. For example, the topics, sentiment or instructions need to be deciphered for the computer system to make use of them. A natural application area of using natural language as the primary human-computer interface is chatting toys or game conversational agents [17] [21].

This literature review will investigate the detailed technique of surface text processing methodology. In this section, we separate the techniques into Lexical, Syntactical and Semantic three different levels of text processing.

2.1 Lexical Level

The field of information retrieval plays a crucial role in searching certain information through the Internet. Search engines are keyword-based, which requires a query in order to start. Therefore, matching words in the documents with the words in the query is essential of the information retrieval system [19].

The textual input is created with a set of two aspects which are representation of discourse acts and key words. Theses aspects are designed to be significantly related to each other by their meaning. There are two stages to extract information from a textual input. The first stage is called keyword definition. In this stage, each sentence will be separated into individual words which simplify the process of information extraction. On the other hand, the second stage will actually grow the understanding level by recognizing the bonds between the representation of discourse acts and the keywords. Here we have inserted a table 2.1

that would explain how the database of discourse acts and the designed keywords are related [13].

2.2 Syntactical Level

2.2.1 Part Of Speech Tagging

The Part Of Speech is actually an application that has the functionality to classify a parsed sentence. It has been used in web search queries, which are a fundamental aspect that significantly contributes to information retrieval task, in order to increase the accuracy of web-search queries performance. The sentences are usually written in short forms and their quality depends on their grammatical structure. The reasons behind using the application of POS are to discover two main aspects which Cory Barr, Rosie Jones and Moira Regleson [3] are aiming for in the topic of web-search queries. Firstly, POS will be used to determine the tagging performance of the web-search system queries for typical English language. Secondly, it will be taken to develop the search results by investigating the qualification value of these tags. The method, that is been followed by Cory Barr et al in this paper [3], is to set up part of speech tags which makes it appropriate for search query and quantify their results. These results are apparently caused by the current part of speech taggers which are been chosen in this project. There are two part of speech taggers here. One of them is the Brill tagger. This tagger needs to label all tokens with their part of speech tag and need a lexicon. The reasons behind using Brill tagger are firstly, it automatically generates rules that lead for an easier system to be readable by users. Secondly, it is a popular tagger. The other tagger is Stanford part of speech tagger. This tagger has the best accuracy performance in the field. Those two part of speech taggers can explore the type of search queries and part it into grammatical classes based on the noun-phrases queries [3].

In conclusion, the researchers above have made initial investigative experiments in order to test the performance of the POS tagging. Those experiments indicate that part of speech information plays an important role in the outputs of a machine-learned system. This system actually leads to have beneficial proper nouns in queries. In addition to the experiments, they show how accuracy the POS tagging would react in order to select or substitute the right words into the query reformulation. It has been investigated, through practical analysis and experiments, that the relevance web search results can be improved and

Table 2.1: Discourse Acts adapted from [13]

Representation of Discourse Acts	**Keyword**
DAAgree ?char	Agree, okay, pass on, make a deal and cool.
DADisagree ?char	Disagree, unsuitable, no way and unbelievable.
DAPositiveExcl ?char	Yeah, Wow, interesting, amazing, and wonderful.
DANegExcl ?char	Damn, awful, bad and I do not like.
DAHappyExpress ?char	Cheerful, satisfied and thrilled.
DASadExpress ?char	Cheerless, heartbroken and wistful.
DALaughterExpress ?char	Haha, ha ha, lol, loool.
DAAngryExpress ?char	Hate, pisses me off.
DAUnsure ?char	Maybe, unsure, could be, do not know, guess so.
DAThank ?char	Thank, Ta, Thanks.
DAGreet ?char	Hello, Hi, Good morning, Good evening, whats up.
DAAlly ?char	Like you, love you, you are friend, you are mate.
DAOppose ?char	Hate you, kiss of, get out.
DAMisUnderstand ?char	Not understand, confused, do not get it.
DAApologize ?char	Sorry, forgive.
DAPraise ?char	Cute, sweetheart got good idea.
DAIntimate ?char	Talk to me, Whats wrong.
DAGoodbye ?char	Goodbye, Good night, see you, catch you later.
DAContinue ?char	Continue, keep up, press on and stay.
DAExplain ?char	Explain, illustrate, tell

contributed with query reformulation by the part-of-speech tagging. As a result, entity detection and proper-noun detection are recommended for achieving higher improvements. In particular, they are aiming to determine the accuracy of these improvements, which leads to develop the performance of part of speech information [3].

2.2.2 Noun Phrase Chunking

The most important concepts are noun phrases. Therefore, special Natural Language Processing task focuses on Noun Phrase chunking. Natural Language Processing technique is used here by authors [19] to support an amount number of Natural Languages queries and replace them over the used keywords. The reason why this technique is costumed is because it is known as an easy system that uses key phrases in order to deals with noun phrases from a document. These noun phrases contain three main modules which are been used in this research. Those modules are; tokenisation, part of speech tagging and noun phrase identification using Chunking. Lastly, there is an evaluation method that is used to test the quality of the information retrieval technique and its results were positive. The information retrieval is mainly using the Natural Language Processing to represent the right documents that satisfy the users needs. Hence, the Natural Language Processing might be useful to develop the precision of the internet search as well as the NL system which can also be used in society. One of the most remarkable Natural Language Processing modules that would help to develop the accuracy of the web search is the Chunking. The function of this tool is to extract noun phrases first and then it retrieves them back in order to improve the performance more than using the key phrases [19].

2.2.3 Parse Tree Database

Most traditional Information Extraction system takes a processing pipeline, in other words, the text go through tokenisation, sentence splitting, Part Of Speech tagging and noun phrase chunking etc to extract useful information. This pipeline needs to re-run on the text corpus should any new requirement code in.

Tari et al [20] exhibits a demonstration plan, which depicts an innovation model for information extraction [20]. They use parse tree output to store the result of textual processing. Nevertheless, the proposed model for information extraction is meant to replace the traditional extraction systems, which were executed as a pipeline of processing modules. They argue that the traditional approach is time consuming and cumbersome. According to developers [20], the present approaches are rigid and costly in the face of the needed dynamic application. In essence, the study [20] recommends the development of new extraction systems for the current approaches. However, academics observe that this can be expensive since the new developed extraction system requires the whole corpus to be recomputed from scratch. It is worth noting, that not the entire corpus is affected with the new acknowledged entities, since most of these entities overlap both the primary and enhanced recognizers [20].

The proposed new model of information extraction offers a general-purpose extraction system, which meets varied extraction requirements efficiently [20]. In this regard, investigators distinguished two phases of processing that can protect the corpus from being entirely affected. First is the initial phase, whereby a one-time parse is carried out, to identify candidates for individual entries on the entire corpus, based on the knowledge available. The second phase is the extraction phase, involved circulation of a parse tree database, a storage platform for syntactic parse tree and semantic entity attachment. Consequently, to ease extraction process, a query language named Parse Tree Query Language (PTQL) was designed and implemented, which automatically generates queries for high quality extraction [20].

The paper [20] explained the system architecture of the GenerIE system, in which the initial phase performed for corpus processing, is done by the text processors and stockpiled in the Parse Tree Database (PTDB) [20]. In this part, four modes are generated to enable the user to identify the PTQL to use in the extraction process. Firstly is the text parsing and PTDB, secondly, information extraction by use of PTQL queries, thirdly, pseudo-relevance generation of feedback query and finally, query evaluation and optimization [20]. The approach has proven to be beneficial, because it provides an innovative database central structure for information extraction, in terms of incremental evaluation and database query optimization. [20].

2.3 Semantic Level

Systems used for extracting information perform the analysis for unrestricted text. This is done in order to extract information about pre specified events, entities or relationships. The extracted information is related to a specific domain-ontology [23]. Therefore, there is a large volume of published studies describing the role of ontology in term of information extraction.

Wimalasuriya and Dou [23] describe that ontologies can be used in the process of extracting information. This is because ontologies are precise and have high recall capacity. Their paper focuses on key issues such as:

- Use of multiple ontologies in guiding the information extraction process.
- Challenges involved with multiple ontologies.
- Suitable ontologies and their mappings.
- Solutions to the challenges and experimental results.

Faster creation of content can be achieved through the use of multiple ontologies. The various types of multiple ontologies whose information has been published include; Kylin, C-PANKOW and SOBA [23]. Ontologies should be capable of locating objects from a group of items. Precision and Recall are immensely vital in information extraction. A single ontology in the process of information extraction leads to have less extracting and recall. This problem was solved by following the introduction of multiple ontologies, which provided a myriad of perspectives. The developed multiple ontologies are either involved in; Domains, for example, University domains, or for providing varied perspectives of the same domain [23]. There are two advantages of using multiple ontologies to provide different perspectives in OBIE, which are possible improvement in recall and supporting multiplying perspectives. The extraction based on mappings between concepts of ontologies can be employed in other systems. Multiple ontologies also allows for combination of results that are not necessarily linked [23].

Multiple ontologies that specialize on domains were evaluated using university domains. Based on the training set that is guided by the result and assist to improve ontology. On the other hand, multiple ontologies providing different perspectives were evaluated by the use of the domain of terrorist attacks. The 4th Message Understanding Conference (MUC) corpus was used. Two ontologies obtained from the MUC structure and the Mindswap group of the University of Maryland were also used [23]. Use of multiple ontologies achieved better recall and precision in OBIE. This was verified by the two case studies. A higher figure was obtained when ontologies represented specialized subdomains. This is because a specialized sub domain is made up of two subsets. One subset has terms representing specialty of the domain while the other has text documents related to the domain terms. A lower figure was recorded when different perspectives were provided by the ontologies. Improved precision is as a result of extraction of information by specialized ontologies [23].

The main shortcoming of multiple ontologies is determining the theoretical basis for using the ontologies in extracting information. This can be solved by carrying out intensive research on ontologies and extraction techniques. Finding favourable ontologies and mappings is a problem. Thorough assessment of ontologies to use in the OBIE system should be conducted. Adequate knowledge concerning mappings between the concepts of the selected ontologies should also be acquired. Experimental results from the two case studies were evaluated. Although lower precision was noted in the case of different perspectives, the results proved that multiple ontologies offer better precision and higher recall. The research can be applied in different fields for instance, oil companies. The companies can extract accurate and quality information instead of relying on human interpretation [23].

Wimalasuriya and Dou [24] have provided that, some ontology systems are capable of constructing ontology from their own domain. For instance, an ontology system like OBIE can generate semantic contents automatically. The problem being present in their paper, however, is concerned with ontology technology and its use in information extraction. According to researchers [24] "Multiple ontologies exist for most domains and there is no rule that prevents an OBIE system from using more than one ontology". The paper provides that, the use of multiple ontologies can develop the process of information extraction since multiple ontologies can make more extractions in a single instance. To achieve their objective, researchers have employed two case studies. The challenges experience in this study is concerned with widespread of information extraction. However, this technology is not used widely due to cost expenses and the complex nature of this technology. To address these issues, the authors have suggested the reuse of IE components. Reusing IE components has been tackled in their paper and according to authors it is preferable to other techniques. The authors state that the obtain performance of the study is lower than in comparison with other studies. In addition to that, the paper provides some of the future works required in this study [24].

Macias-Galindo et al [11] defined MKBUILD as "A tool that follows a methodology to create domain-specific ontologies containing related concepts, drawning from existing large-scale resources such as WordNet and Wikipedia/DBPedia" [11]. Consequently, the paper aims at providing a conversational agent from Modular

Knowledge Bases (MKBs) perspective. Nevertheless, the intent of developing MKBUILD is to substitute the hierarchical process of structuring definite domain ontologies that consume both human effort and time. In essence, the paper contours the architecture of the conversation agent by use of an interactive toy. To select a suitable conversational path to pursue, developers applied techniques such as keyword spotting and lightweight semantic parsing. In addition, they pointed that topic module designer generated fragments, whereby the designer ensures the fragments with one node in a topic network. In this regard, each fragment cosiest of a head that provides suitable conditions and body that entail list of the anticipated inputs. According researchers, the Topic Transition Network is utilized to create a consistent dialogue structure, productively used in the selection of the conversational fragments for the next part of conversation [11].

Authers follow strategies to create domain-ontology, then start involve identification of the main concept ontology, finding the higher layer of MKB and expand the domain ontology. The results of the study exhibited that, for domain Internet use of MKBUILD tool was irrelevant due to the available technologies such as televisions and radar [11].

Information extraction (IE) is one of the most common systems that automatically extract structured information in the world. It plays an important role in many of the real world applications in the areas of business intelligence, competitive and military intelligence. XONTO, which is considered to be as a method of the information extraction functions, is a system that has been found from the idea of describing ontology objects and classes which can be dominated by written rules that is called descriptors. These descriptors allow XONTO to accurately determine the requested objects and classes. In other words, this system that is also known as self-describing ontologies uses set of rules to obtain and extract ontology objects and classes which are contained in PDF documents. However, once the system encounter with document in tabular forms, XONTO will qualify its rules and will give an expression to the semantic of the information in order to extract the requested orders [16].

In conclusion, in this task, the XONTO, that extracts information from PDF documents, is used for the ontology based system. It is called self-describing ontology system which enables exploring semantics in (IE) from PDF documents by gathering the power of ontology forms and attribute grammars. This work aims for three major aspects. First of all, It will be used to solve complex issues and analysis them accurately. Secondly, it will gain an extension of the approach to permit exploiting web standard ontology language. Lastly, XONTO will increase its methods and strategies by realising other ontology approaches. [16]

2.4 Rule Based

Muludi et al [14] present a proposal for Information Extraction System based on Inductive Learning and Meta learning, whose performance is exceptionally good. The proposed system seeks to make the search and use of information on the Internet easy. It studies related works and compares them to the Multi Inductive Learning (MIL) system. The state of the art system being compared to MIL is LP2. LP2 is defined by authors [14] as “learning by using symbolic rules for identifying start tag and end tag class of the slot”. The system presented in this proposal seeks to address the problem in the previous existing systems. The problem is that the performance of these systems are not consistent on various information domains. Information extraction can be approached as classifying problems. The text is divided into tokens and classified into related classes [14].

The new idea presented in this paper [14] is the use and introduction of machine learning to improve on information extraction. MIL concept is inspired by the idea of how to use document training to look for best classifier for each slot in a certain domain. MIL is considered better as compared to natural language processing, because of its portability, scalability and adaptability. They managed to show the working of the Multi Inductive Learning and compare it with other systems such as RAPIER, LP2 and SNOW. They showed how the best classifier for each slot is chosen to achieve the best performance in extraction of information from the testing document. The authors used a 10 fold cross validation on training document, they associated each base learner with each slot. They obtained results and analyses them. They concluded that Multi Inductive Learning chooses the best learner. It is the best among other state of the art information systems [14].

The researchers have applied many trainingg sets to allow the generating accurate rules of information extraction. The system designed could not provide consistent results. They decided to use the strengths of different systems mentioned in the document to come up with an algorithm that could overcome the limitation. These included the base learner and performance index that were used to design the algorithm for MIL algorithm. According to authors there is no single classifier that can produce consistent performance across all domains and slots. Multi Inductive Learning has the best performance on average as compared to others. The work is beneficial in that it shows that although there is no single classifier that can produce consistent results across all domains and slots, Multi Inductive Learning is much better than other methods that are in use such as single classifier [14].

2.5 Summary

In brief, we have addressed in details very important aspects of surface text processing methodology, which are significantly related to the process of information extraction. These aspects are divided as following; firstly, the lexical level which leads an application to determine the right meaning of a word. It offers two strategies to extract useful information, information retrieval strategy and discourses acts strategy. The second aspect of this literature is the level of syntactic that contains three different methods such as part of speech tagging, noun phrase chunking and parse tree database. If one of these methods is used, the level of syntactical will be reached eventually. The semantic level is the third aspect. It concentrates on extracting information by the implementation of ontology technique, that helps to understand knowledge as well. Finally, rule based aspects utilised for information extraction through the use of multi inductive learning approach. Also it can be used to recognise and classify problems into different categories.

CHAPTER 3

Methodology

3.1 Overview

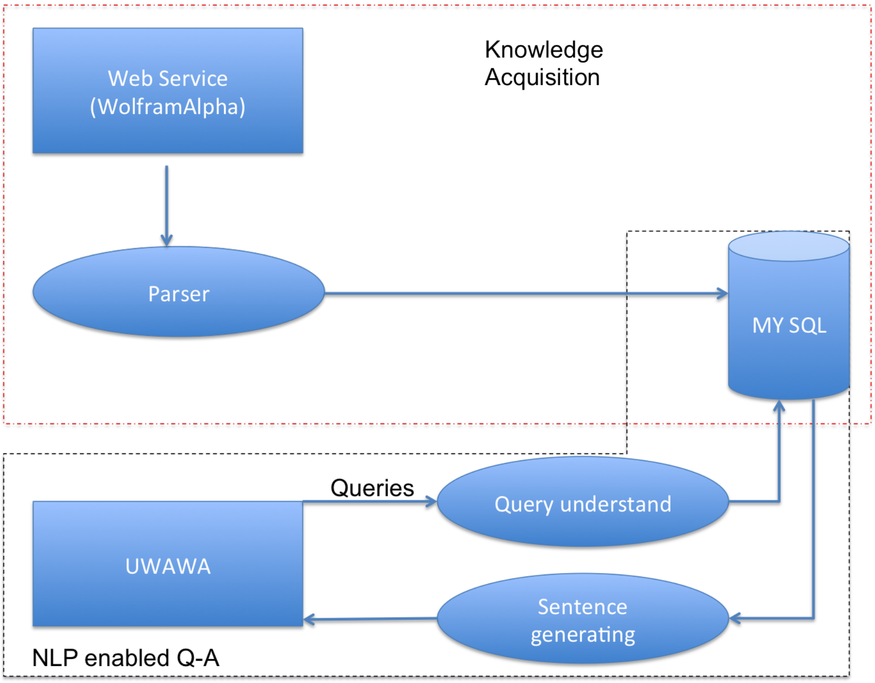


Figure 3.1: Overview of methodology

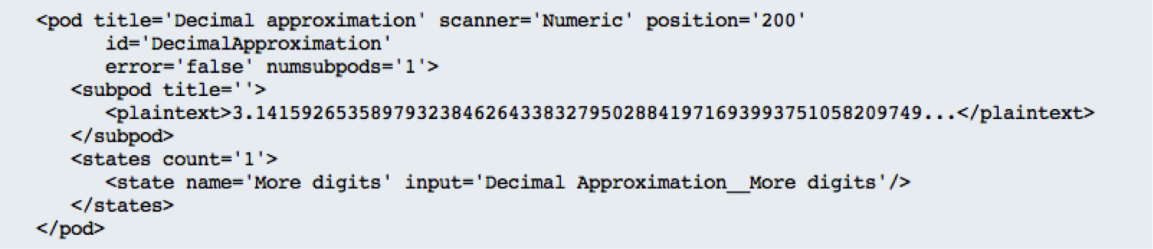
As mentioned earlier, this project aims to extract information from a user input and then applies text processing algorithm in plain text to respond with useful information. We built a system consisting of two modules to achieve our goal. The knowledge acquisition module builds information about fictional characters, while the NLP enables the Q-A module to extract information from the user in two text processing levels (See Figure 3.1).

First of all,The data used was extracted from WolframAlpha website [1]. We extracted about 110 fictional characters. A MySQL database is used to store structure and organise information collection from WolframAlpha website. Database[1] plays a very important role in computing by enabling easy access and manipulation of amount of information. [15]

3.2 Web service by WolframAplha

3.2.1 Introduction to WolframAlpha web services

WolframAlpha is a computational search engine used for answering a query. It concentrates on knowledge-bases rather than on normal search engine in terms of search processing. In fact, WolframAlpha enables developers to utilise its facilities for different types of development purposes such as information extraction. One of the main outputs of WolframAlpha is a pod, which is an area to express the results. It can be more than one pod in the result that depends on result classification. Furthermore, the pod has many options to gather a result such as plaintext for easy copying. The pod has at least one sub pod for revealing the actual content. The results from WolframAlpha are provided in an XML document. It contains the pod and sup pod, which has the actual content that can be comfortable method for recognizing the answer (See Figure 3.2) [1, 18].



```
<pod title='Decimal approximation' scanner='Numeric' position='200'
      id='DecimalApproximation'
      error='false' numsubpods='1'>
   <subpod title=''>
      <plaintext>3.1415926535897932384626433832795028841971693993751058209749...</plaintext>
   </subpod>
   <states count='1'>
      <state name='More digits' input='Decimal Approximation__More digits'/>
   </states>
</pod>
```

Figure 3.2: Example of XML and pod obtained from WolframAlpha [1]

3.2.2 Build Knowledge from WolframAlpha web services

The code following is applied as a query. This code is run in a sequence of packages from 1 to 10, 11 to 20 until 110. We implemented this method

[1] Wikipedia, Relational database, this is available from: https://en.wikipedia.org/wiki/Relational_database

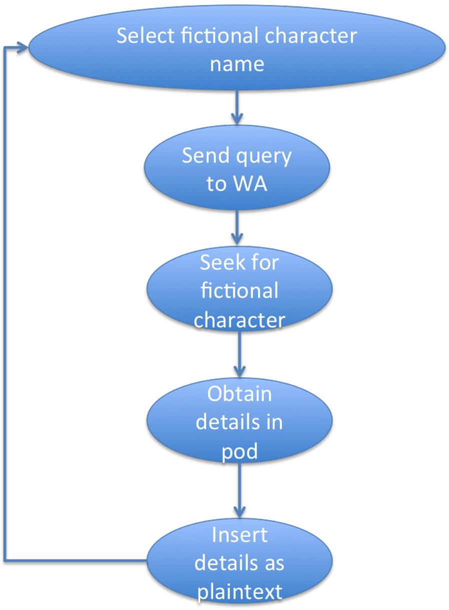

Select fictional character name
Send query to WA
Seek for fictional character
Obtain details in pod
Insert details as plaintext

Figure 3.3: Process to collect knowledge about fictional characters

to avoid execution time error, which may occur when it approaches 110 run iterations. Each query was sent to WolframApha to extract details about a specific character by reading his/her name from a question table. The details about fictional characters are prepared for insertion into the answer table (See Figure 3.3).

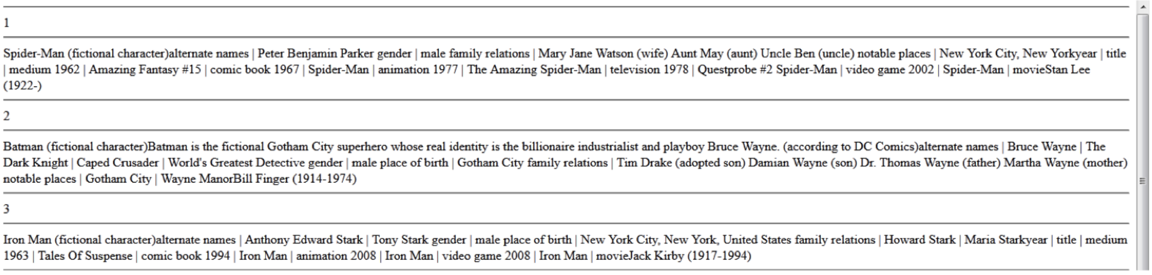
```
$query_question = mysql_query("SELECT * FROM question WHERE id_q
BETWEEN 1 AND 10 ORDER BY id_q");
```

The next step, in building the knowledge about the fictional characters, is to store each answer as a plaintext in the answer_dcpt field by virtue of the coding below.

```
$insert_result = mysql_query("INSERT INTO answer SET
answer_dcpt =’".$description."’, id_q=’".$id_q."’");
```

The description variable is used to store the result extracted from WolframApha. This provides a complete information database used in the project. This information can be accessed, changed and manipulated using the various methods provided by SQL.

The schema of pod for each fictional character's details has several aspects like "alternate names","gender"and "family relation". Those aspects are stored as plaintext in the answer table to prepare them for parsing process (See figure 3.4).



1

Spider-Man (fictional character)alternate names | Peter Benjamin Parker gender | male family relations | Mary Jane Watson (wife) Aunt May (aunt) Uncle Ben (uncle) notable places | New York City, New Yorkyear | title | medium 1962 | Amazing Fantasy #15 | comic book 1967 | Spider-Man | animation 1977 | The Amazing Spider-Man | television 1978 | Questprobe #2 Spider-Man | video game 2002 | Spider-Man | movieStan Lee (1922-)

2

Batman (fictional character)Batman is the fictional Gotham City superhero whose real identity is the billionaire industrialist and playboy Bruce Wayne. (according to DC Comics)alternate names | Bruce Wayne | The Dark Knight | Caped Crusader | World's Greatest Detective gender | male place of birth | Gotham City family relations | Tim Drake (adopted son) Damian Wayne (son) Dr. Thomas Wayne (father) Martha Wayne (mother) notable places | Gotham City | Wayne ManorBill Finger (1914-1974)

3

Iron Man (fictional character)alternate names | Anthony Edward Stark | Tony Stark gender | male place of birth | New York City, New York, United States family relations | Howard Stark | Maria Starkyear | title | medium 1963 | Tales Of Suspense | comic book 1994 | Iron Man | animation 2008 | Iron Man | video game 2008 | Iron Man | movieJack Kirby (1917-1994)

Figure 3.4: Data collection from WolframApha

3.3 Parser

The aim of the parser is to facilitate a keyword matching. The present application has a two level parsing process. The first level is called "Basic Processing", which tokenises a query into a parts of speech tagging to identify noun words in a sentence. It then store each identified noun word into an array for word matching purpose. The next step in Level 1 of parsing process is matching the first noun word with name of a fictional character, which can only be found in the question table. If it is not matched with the content of the table, the application tries to use another noun word in the sentence in the same level 1 of parsing. The application will loop the process until it finds the keyword matching. However, if it cannot match with the question table, the application will print "Sorry! I don't know the Answer". An id_q will be provided when the noun word matches with one of fictional characters names. It is used to determine details of the fictional character in the answer table.

The second level of parsing is called "Attribute Processing", which mines in plaintext what is in the answer table. It has the details of the fictional character extracted from level 1 of parsing. In this level, the application parses plaintext into lines to facilitate the matching process. The other noun word in a sentence will be inserted into this level of parsing, because the application seeks each line for this noun word to find a match for it in the plaintext. Consequently, if the system matches the keyword in plaintext, it will then extract the exact information the user is reaching for. Otherwise, the application prints "Sorry! I don't know the Answer".

3.4 Database design

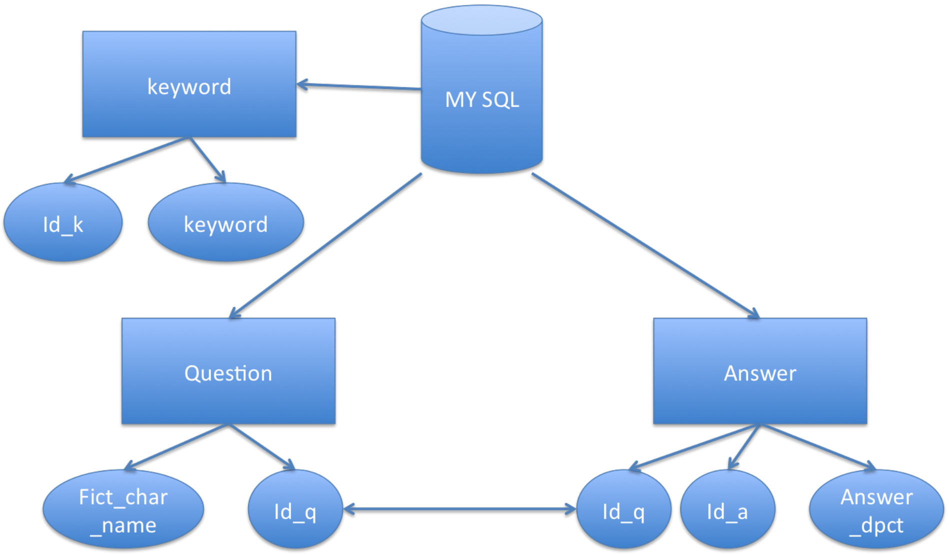


Figure 3.5: Database Design

A database is established for using SQL queries. It consists of three tables in which the data is stored (See figure 3.5). The tables are created with a unique primary key field that uniquely identifies each entry to the database. A question table is created with two fields. The first field deals with an integer number type. This field is named id_q. It is set to generate an auto unique number for each fictional character as the primary key. This field provides a unique identity for each entry that is made in the table. The second field concentrates on data type of tiny texts. It is named fict_char_name which is used to store the name of each fictional character.

On the other hand, the answer table contains three fields. The first field is named answer_dpct with type long text to store data from WolframApha as a plain text for each character. The second field is named id_q with a type integer number. This field stores the value that determines the relationship between the two tables; question and answer tables. The last field is named id_a. It is generated automatically as a unique value that identifies each detail of a character into the answer table. Id_a is a primary key for the answer table.

The information inserted into answer_dpct is obtained from WolframAlpha website [1]. WolframAlpha is a search engine that gives access to the world facts and data[2]. In other words, It is a collection of the world data where users can query for answers [12]. In this paper, 110 fictional characters details were extracted from WolframApha. These details are entered into the chat_bot.

A keywords table is established in chat_bot database for keyword matching with two fields. An id_k is the first field in keyword table, it is an integer number type that is generated automatically. The id_k determines each keyword that stored in the database. The second field is the keyword itself, which stores some specific keyword in order to make it useful for an application. It has different keywords like "alternate names","gender"and "family relation". Those keywords are obtained in the system because the schema for fictional characters details have similar segmentation of them.

3.5 Question Understanding

The idea of extracting information in this project (See Figure 3.6) is to let a user inserts a question about a fictional character into the text area. If the user does not input any question, the system will require to write question. UWAWA is a system that is already connected with chat_bot database. The method of this system is to tokenise a sentence into parts of speech tagging according to Ian Barber algorithm[3]. Consequently, each noun word that caught from the sentence will be kept in an array (see the code below), and then insert it into layered surface text processing approaches, which are the level 1 is called "Basic Processing"and the level 2 is called "Attribute Processing".

```
$question_noun_array[$question_noun_counter] = $t['token'];
                    $question_noun_counter++;
```

3.5.1 Level 1 of text processing

The text processing approach in UWAWA adopts a two level (steps) approach to extract information. In the the first level, following parsing the sentence,

[2] Wikipedia, WolframAlpha, this is available from: http://en.wikipedia.org/wiki/Wolfram_Alpha

[3] Ian Barber, Part Of Speech Tagging, this is available from: http://phpir.com/part-of-speech-tagging/

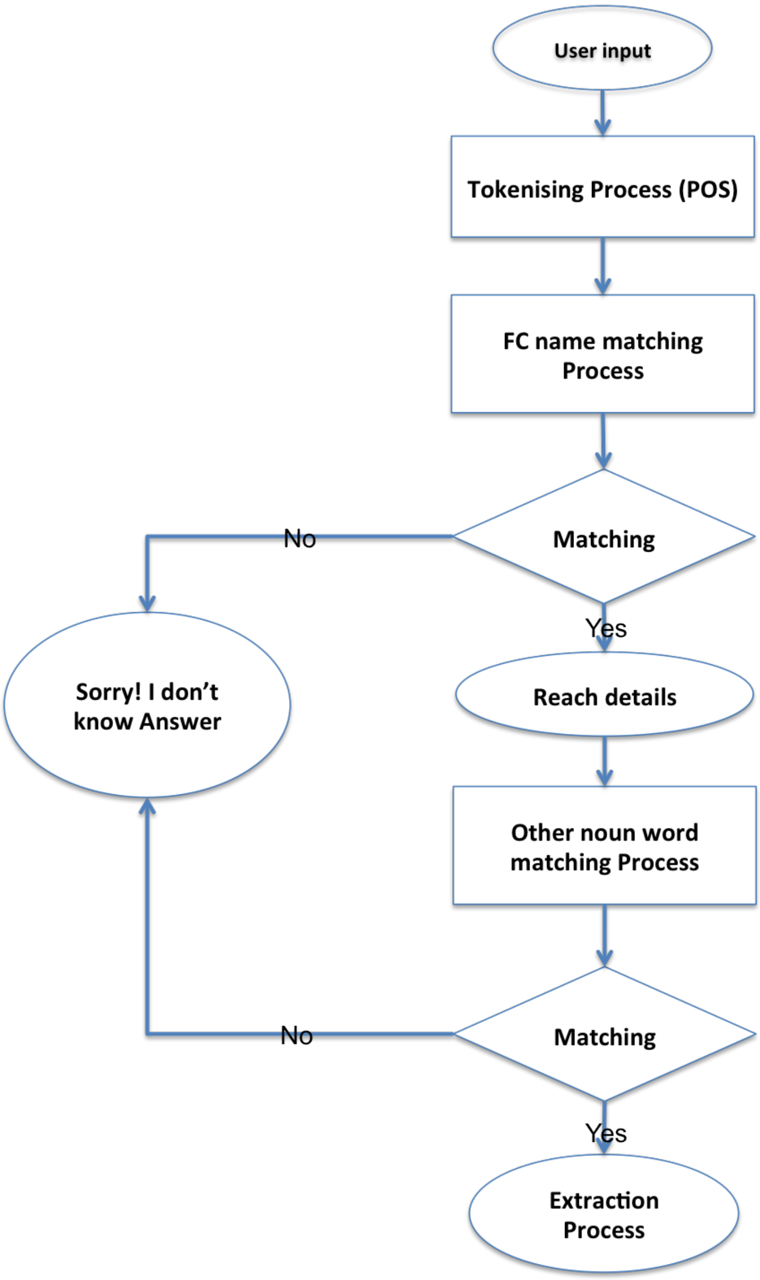


Figure 3.6: Process to extract information about fictional characters

the system will process the exising noun words in that sentence separately. It will then attempt to match each noun word (but not all at once) within the question table. When the first noun word is matched with one of the fictional character's names, it will return to an id_q fictional character in the question table. Hence, this id_q will be inserted into the answer query in order to determine the requested fictional characters details. This step assists with the preparation of the information extraction for the next processing level.

On the hand, if the initial chosen noun word is not matched with any fictional character, the system will repeat the process using another word from the users sentence. For example the user may write “who is wife of Spiderman?”, according to Part Of Speech tagging (POS) the wife and Spiderman are nouns, so system puts the word “wife”into first level of text processing. When system could not match “wife”with question table, because the table does not include “wife”, the system will ignore it and attempt to put Spiderman in same process. If the system could not match all of them, UWAWA system will print “Sorry! I don’t know the Answer”(see code below).

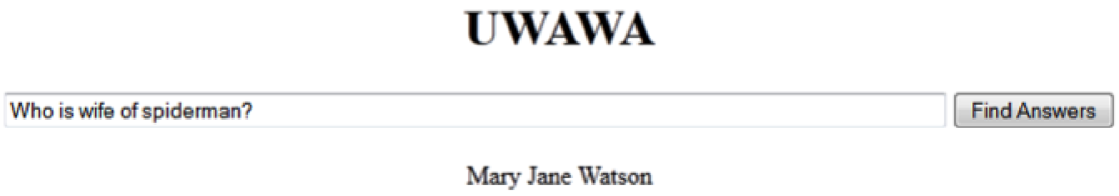
```
if (sizeof($question_noun_array) > 0) {
$query  = setupQuery($question_noun_array);
$result = mysql_query($query);

if (mysql_num_rows($result) == 0) {
echo "Sorry! I don’t know the Answer";
}
else {
$qdata = mysql_fetch_array($result);
$plainText = $qdata[’answer_dcpt’];
```

3.5.2 Level 2 of text processing

The second level of the text processing part is called “Attribute Processing”which aims at mining the plaintext to extract whatever requested by the user. In this case, the keyword table is established in the chat_bot database for keyword matching. It has different keywords like “alternate names”,“gender”and “family relation”. Those keywords are obtained in the system because the schema for fictional characters details have similar segmentation of them. In this level, UWAWA will parse the plaintext into lines in order to facilitate the keyword matching. Subsequently, The UWAWA searches the requested keyword on each single line to extract useful answer for chatter. In the previous example, when the system matched spiderman within database and arrived at the details of “Spiderman”, the word “wife”will be extracted from plaintext using the method mentioned before (See Figure 3.7). Otherwise, if UWAWA cannot match the keyword in plaintext, it will display “Sorry! I don’t know the Answer”. For this algorithm the system is coded as following:

```
$AnswerArray = explode("\n", $qdata['answer_dcpt']);
foreach ($QuestionArray as &$word) {
$kwQuery=CheckKeyword($word);
$kwResult = mysql_query($kwQuery);
if (mysql_num_rows($kwResult) != 0) {
$kwData = mysql_fetch_array($kwResult);
```



UWAWA

Who is wife of spiderman?

Mary Jane Watson

Figure 3.7: Attribute Processing

3.5.3 Spelling Mistake Correction

In addition to Q-A capability of the UWAWA scheme, it provides an intelligent suggestion method for keyword matching. Accordingly, UWAWA recommends a word that is close to that intended by the user, but was spelt incorrectly. The recommendation word is the result of a Soundex function. The Soundex function offers opportunity to select suitable word from the question table in the first level of the text processing or the answer table in the second level of text processing. Obviously, the recommended word matches the noun words in the original sentence with words in chat bot database according to the word pronunciation[4]. For this stage, UWAWA implements the code below for keyword matching with fictional characters name and keyword in plaintext.

```
//Fictional character matching
$query .= " OR SOUNDEX(‘fict_char_name‘) = SOUNDEX(’$array[$i]’)";
```

```
//Keyword matching in plaintext
$kwQuery  ="SELECT  ‘Keyword‘ FROM ‘keywords‘ WHERE SOUNDEX(‘Keyword‘)
like  CONCAT(’%’,SOUNDEX(’$word’),’%’);";
```

[4] PHP website, PHP manual, this is available from: http://php.net/manual/en/function.soundex.php

CHAPTER 4

Experimental Results and Analytical Evaluation

4.1 Overview

In this chapter we provide results of our experiments, which illustrate the effectiveness of our system in terms of information extraction at two levels “Basic Processing”and “Attribute Processing”. These experiments will compare the accuracy of our system UWAWA against WolframAlpah. We conducted more than 60 questions about the fictional characters and their attributes.

4.2 Test Question Preparation

We established approximately 60 questions about fictional characters. The queries are divided into three parts for each level of text processing. The classifications of questions are standard questions, grammatical mistake questions and questions with typos.

- Standard Questions.

  A set of questions that are in standard format, which are prepared with no grammatical mistakes and attempt in asking questions about attributes (See Tables 4.1, 4.2).

- Grammatical Mistake Questions.

  A set of questions with grammatical mistakes are constructed to test the performance of the two applications. The words in sentence are reordered mistakenly for evaluating the understanding of the applications for each level of text processing (See Table 4.3, 4.4).

Standard questions for level 1	UWAWA		WolframAlpha	
Question	YES	NO	YES	NO
Who is spiderman?	1	0	1	0
Who is Batman?	1	0	1	0
Who is Ironman?	1	0	1	0
Who is Hulk?	1	0	1	0
Who is Superman?	1	0	1	0
Who is Faust?	1	0	1	0
Who is James Bond?	0	1	1	0
Who is Sherlock?	0	1	1	0
Who is Snoopy?	1	0	1	0
Who is Uncle Sam?	1	0	1	0

Table 4.1: Standard questions for level 1

Standard questions for level 2	UWAWA		WolframAlpha	
Question	YES	NO	YES	NO
Who is wife of Spiderman?	1	0	0	1
What is gender of Batman?	1	0	1	0
What are alternate names of Ironman?	1	0	1	0
Who is father of Hulk?	1	0	0	1
where is mother of Superman?	1	0	0	1
who is heir of Faust?	1	0	0	1
Who is James Bond's father?	0	1	0	1
Who is brother of Sherlock?	0	1	1	0
Which is gender of Snoopy?	1	0	1	0
What is alternate names of Uncle Sam?	1	0	1	0

Table 4.2: Standard questions for level 2

Grammatical Mistake Questions for level 1	UWAWA		WolframAlpha	
Question	YES	NO	YES	NO
Who Spiderman?	1	0	1	0
Batman Who is?	1	0	1	0
What is Ironman?	0	1	1	0
which is Hulk?	0	1	1	0
Who are Superman?	0	1	1	0
Faust Who is?	1	0	0	1
who is James Bond?	0	1	1	0
Sherlock who?	0	1	0	1
Who are Snoopy?	0	1	1	0
Who Uncle Sam?	0	1	1	0

Table 4.3: Grammatical Mistake Questions for level 1

Grammatical Mistake Questions for level 2	UWAWA		WolframAlpha	
Question	YES	NO	YES	NO
Who is wife Spiderman?	1	0	0	1
What Batman gender?	1	0	1	0
who is alternate names of Ironman?	1	0	1	0
father of Hulk?	1	0	0	1
mother of Superman?	1	0	0	1
How is heir of Faust?	1	0	0	1
James Bond's father?	0	1	0	1
what is brother of Sherlock?	0	1	1	0
who is Snoopy's gender?	0	1	1	0
Alternate names of Uncle Sam?	0	1	1	0

Table 4.4: Grammatical Mistake Questions for level 2

Questions with Typos for level 1	UWAWA		WolframAlpha	
Question	YES	NO	YES	NO
Who is Spudernan?	1	0	0	1
Who is Batnan?	1	0	1	0
Who is Eronnan?	0	1	0	1
Who is Halck?	1	0	0	1
Who is Superrnan?	1	0	0	1
Who is Foost?	1	0	0	1
Who is Jimes Bwnd?	0	1	0	1
Who is sharluck?	0	1	0	1
Who is Snwopy?	1	0	1	0
Who is ancle Saam?	0	1	0	1

Table 4.5: Questions with Typos for level 1

- Questions with Typos. Wrongly spelt words are quite often present in natural-language driven systems, so we built test sentencing with typos as well (See Tables 4.5, 4.6).

4.3 Experimental Results and Analytical Evaluation

4.3.1 Level 1 of text processing

In this section, the data analysis shows that two experiments are used to examine the two applications.

Figure 4.1 illustrates how our application (UWAWA) is performing, comparing to the application of WolframAlpha in terms of the text processing in level one. It is clearly that from the 10 standard questions, UWAWA has successfully completed 80% of them, whereas WolframAlpha reached 100%. As an initial prototype, this is actually a positive results indicating that UWAWA is doing reasonably well, but with room for improvement. In fact, as will be shown in the following experiments, when introducing level two text processing, the performance is improved. In regard to evaluate the UWAWA, it reaches 20% of misrecognizing the standard queries, because it encountered some issues to match noun phrase with database. However, the difference between UWAWA and WolframAlpha in term of recognise these questions is less than we expected initially.

Questions with Typos for level 2	UWAWA		WolframAlpha	
Question	YES	NO	YES	NO
Who is wifee of spiderman?	1	0	0	1
What is gindar of Batman?	1	0	0	1
What are alternat names of Ironman?	1	0	1	0
Who is futhar of Hulk?	1	0	0	1
where is mather of supernan?	1	0	0	1
who is heer of Faust?	1	0	0	1
Who is Jimes Bond's futhar?	0	1	0	1
Who is bruthar of sherlock?	0	1	0	1
Which is Snoopy's gundar?	0	1	0	1
What is altrnat nomes of Uncle Sam?	0	1	1	0

Table 4.6: Questions with Typos for level 2

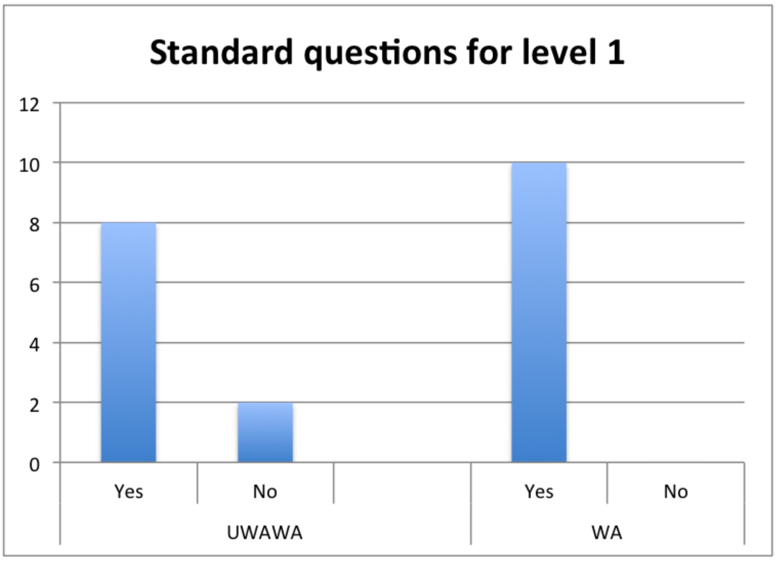


Figure 4.1: Standard questions for level 1

As Figure 4.2 presents, there is a major difference between the two applications. These differences are based on the type of the introduced questions. The questions are written with some grammatical mistakes. The experiment shows that the WolframAlpha application misunderstood 20% of the questions. In

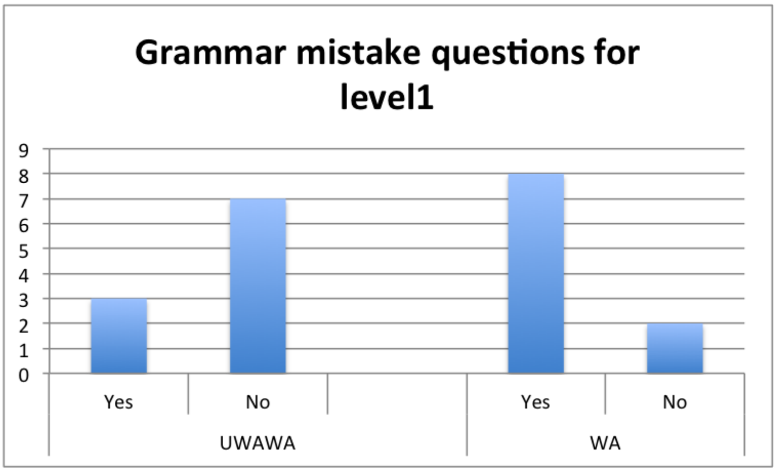


Figure 4.2: Grammatical mistake questions for level1

comparison to that, UWAWA has achieved only 30% because the questions were difficult to be recognized due to their grammartical mistakes. It is clear that UWAWA recognizes 20% of queries, which were set up in grammatical mistake. The algorithm is used at this stage needs to be improved for recognizing purpose. However, WolframAlpha is faster recognizing in this level, due to the application of "anti-phrasing"process as stated by Andersen [2].

4.3.2 Level 2 of text processing

In this section, comparisons between the two applications were made for extracting attribute information. Therefore, the figures in this section provide the experimental results on the applications performance. In addition, it will represent the accuracy of the applications for extracting the attribute details.

The results obtained from the primary analysis of UWAWA and WolframAlpha are shown in Figure 4.3. UWAWA has increased to 80% of understanding the introduced standard questions, which are regarding the attributes of fictional characters. But WolframApha responds to only half of the queries. Furthermore, according to the Figure, UWAWA effectively recognizes the queries, which leads to an increase in the accuracy level of extracting the particular information that user seeks for. The variance between UWAWA and WolframAlpha for recognizing standard queries in this level is 30%. It illustrates that the UWAWA applies better algorithm for mining in plaintext rather than WolframAlpha [9].

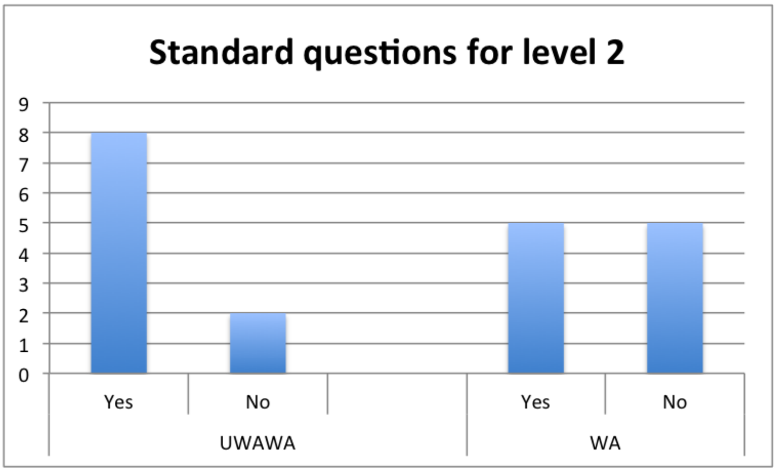


Figure 4.3: Standard questions for level 2

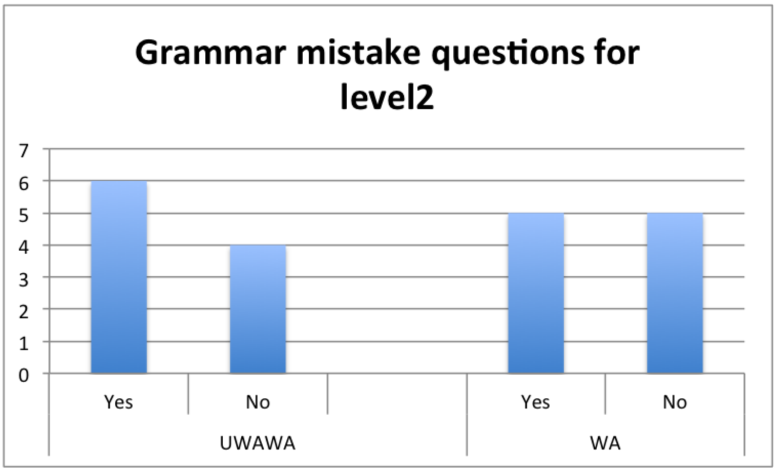


Figure 4.4: Grammatical mistake questions for level2

As shown in Figure 4.4, UWAWA application has interestingly more concentration on the importance of recognising the questions in this level of text processing than WolframAlpha application. The questions are formatted with grammatical error to indicate the recognizing level for each application. UWAWA increased to 60% of accuracy as compared to WolframAlpha 50% of understand the queries. UWAWA has the highest percentage by 60% for recognizing fictional characters attributes, which indicates that it has better performance than WolframAlpha with 10% difference of comprehending. This because of our system implements the keyword matching technique in this level to mine the desired details.

4.3.3 Spelling Mistake Correction

There are variations of questions which are chosen for these experiments in order to differentiate the recognition abilities between UWAWA and WolframApha. In particular, the types of questions which are specialised for this section are the spelling mistake queries. Both applications are presented with more describing details in the following graphs.

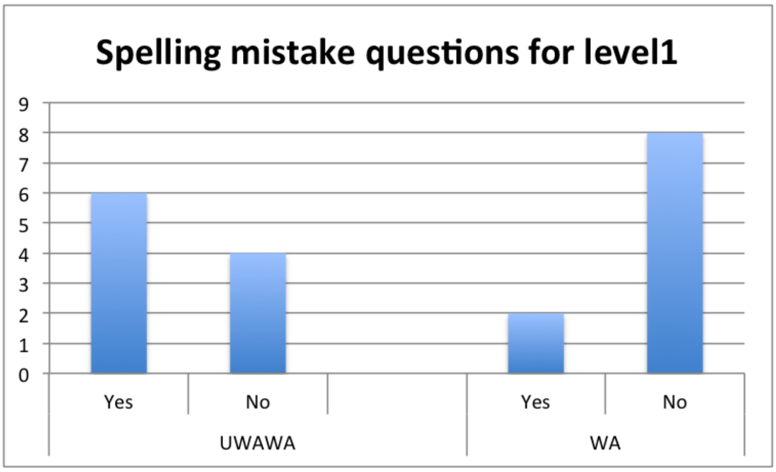


Figure 4.5: Spelling mistake questions for level1

From Figure 4.5, it is apparent that UWAWA managed to deal with questions that have been written with spelling mistake. It comprehended the given questions and responded to them with a percentage of 60%. On the other hand, WolframApha application has failed in this experiment because it was not easy for it to manipulate the questions in order to give a correct answer. As a result, the test shows that WolframApha only recognised 20% and could not read 80% of the questions. These comparisons between the two applications were regarding the test processing in level one of the system.

It can be seen from the data in the Figure 4.6 that the performance of Wolfram Alpha is unexpectedly less than UWAWA performance. Even though WolframApha was acting very well with the type of standard questions and grammar mistake questions, it struggled to deal with the type of spelling mistake questions. Therefore, it could not identify 80% of these questions. However, only 40%, which is half of the failure percentage of WolframApha, of the questions were not understood by UWAWA.

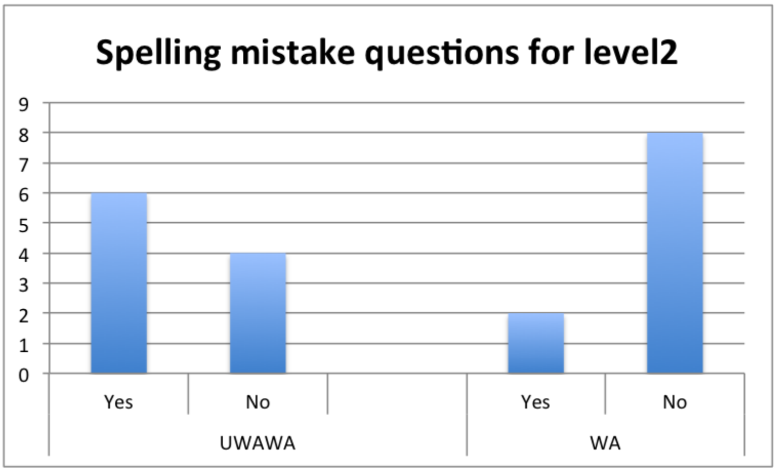


Figure 4.6: Spelling mistake questions for level 2

In terms of evaluating these levels, the Soundex function is applied in the UWAWA to check misspelling. As a result, UWAWA achieves 60% of the entire queries that have misspelling issues. On the other hands; WolframAlpha is failed in recognizing 80% of the same candidate queries [8]. In general, it can clearly be noticed that the UWAWA has privilege results in answering some kinds of questions, whereas WolframApha is not able to answer them.

CHAPTER 5

Conclusion and Future work

This chapter summaries the main points of the research in terms of the initial objectives, method of the study, and the significant findings. In addition, a set of discussion for future work is provided to further develop the proposed system.

5.1 Main contribution

Our project aims to extract the exact information that user requests for. Therefore, our work implemented Façade's surface text processing methodology in term of Natural Language Processing (NLP). As a proof of concept, we focused on a knowledge base about fictional characters. A query from the user is tokenised into several fragments, then Part Of Speech (POS) tagging is applied to facilitate the information extraction process. This process leads the application to interpret commands from a user, and construct a meaningful answer. We built an application consisting two sub-modules in an attempted to reach the project's objective. The first module is to gather knowledge about fictional characters from Wolfram Alpha (WA) knowledge engine web services. The other module is to extract information from the user through two text processing levels. The first level, after the sentence was parsed, the system will use the primary noun phrase in that sentence. The second level of process is to go through the rest of the input to further extract the information, which the user is seeking for. Moreover, the system implements a suggestion method for refine queries. This suggestion method can recommend a closer word when the user has a spelling mistake.

In comparison with WolframAlpha (WA) web services, we found that the proposed system is able to extract more targeted answers for user queries. In fact the system obtains reasonable answers with standard questions, but with grammatical mistake questions in this level, it returns with some errors. However, in level two of text processing with standard queries, the application can extract the desired information about fictional characters attributes, out performed WolframAlpha. WolframAlpha answered only half of the grammatical mistake questions that were introduced. Despite, our application has interestingly more concentration on the importance of recognising the queries in this level. For measuring the recognition abilities between our application and WolframApha, we formatted queries with some misspelling words. UWAWA application deals with these questions in the level one of the text processing to extract meaningful information. It comprehends the queries and responds to them with more willingly than WolframApha. Correspondingly, WolframAlpha did not perform well in the attribute processing coupled with. It struggled to deal with this type of issue for information extraction.

5.2 Future Work

Ontologies are currently considered to be recent development sub-field systems of Information Extraction (IE). Moreover, they play a vital role in generating clear and official conceptions which contribute to increase the performance of the IE. The reasons behind their capacity of improving IE are the speed of reacting and the specific outcomes they can provide. Furthermore, the functionality of ontologies is to search for articles from an existing objects. In other words, ontologies are like a bridge that supports the processing of the information extraction to determine the exact results. However, a single ontology is usually used individually in information extraction, although using multiple ontologies, which as stated by Wimalasuriya et. al. [23] has a significant effect on the accuracy of results. In addition to the multiple of ontologies, they have the ability to combine results from different sources. In future, we will introduce ontologies to our system in order to improve it for better accuracy [23].

More algorithms for correcting spelling mistakes such as edit distance can be used to improve the understanding of the question as stated by Han et. al. [7]. Another approach, noun phrase chunking, for identifying key concepts from the question can also be used for better input processing in UWAWA as stated by Echizen and Araki [6].

Abstract

In the age of technology, many applications, such as chatting toy or conversational agents, contains a natural language based user interface that leads to facilitate the interaction between human and computer. The Natural Language Processing (NLP) component is required in such applications in order to process and extract desired information from input sentences. In this project, we built a question answering system that is named UWAWA, concentrating on processing user input and extracting useful information from a dynamically built database. To achieve this, we introduced a tokenised method that intends to parse sentences. Another component, which is used in our application, is the Part Of Speech (POS) tagging that facilitates the UWAWA information extraction. These methods are implemented as two sub-modules, namely, knowledge acquisition and NLP enabled Q-A. We evaluated UWAWA against the well-known application of WolframAlpha, with a set of systematically constructed test sentences. The result shows UWAWA performs well in the case of attribute processing and questions with misspelling words.

Keywords: Information Extraction, Web Services, artificial intelligence, Text Mining.
CR Categories: A.2, I.7.2

العنوان:	Building a Question-Answer System from WolframAlpha
المؤلف الرئيسي:	Waleed, Saeed
مؤلفين آخرين:	Liu, Wei(Super.)
التاريخ الميلادي:	2013
موقع:	سيدني
الصفحات:	1 - 42
رقم MD:	615578
نوع المحتوى:	رسائل جامعية
اللغة:	English
الدرجة العلمية:	رسالة ماجستير
الجامعة:	Western Australia University
الكلية:	School of Computer Science and Software Engineering
الدولة:	أستراليا
قواعد المعلومات:	Dissertations
مواضيع:	هندسة البرمجيات، أسئلة الاختبارات
رابط:	https://search.mandumah.com/Record/615578

Contents

List of Tables

List of Figures

العنوان:	Building a Question-Answer System from WolframAlpha
المؤلف الرئيسي:	Waleed, Saeed
مؤلفين آخرين:	Liu, Wei(Super.)
التاريخ الميلادي:	2013
موقع:	سيدني
الصفحات:	1 - 42
رقم MD:	615578
نوع المحتوى:	رسائل جامعية
اللغة:	English
الدرجة العلمية:	رسالة ماجستير
الجامعة:	Western Australia University
الكلية:	School of Computer Science and Software Engineering
الدولة:	أستراليا
قواعد المعلومات:	Dissertations
مواضيع:	هندسة البرمجيات، أسئلة الاختبارات
رابط:	https://search.mandumah.com/Record/615578

Building a
Question-Answer System
from WolframAlpha

Waleed SAEED
STD: 20543932
The University of Western Australia,
School of Computer Science and Software Engineering
Supervisor: Wei Liu

This report is submitted as partial fulfilment
of the requirements for a Masters Degree in the
School of Computer Science and Software Engineering,
The University of Western Australia,
2013

Abstract

In the age of technology, many applications, such as chatting toy or conversational agents, contains a natural language based user interface that leads to facilitate the interaction between human and computer. The Natural Language Processing (NLP) component is required in such applications in order to process and extract desired information from input sentences. In this project, we built a question answering system that is named UWAWA, concentrating on processing user input and extracting useful information from a dynamically built database. To achieve this, we introduced a tokenised method that intends to parse sentences. Another component, which is used in our application, is the Part Of Speech (POS) tagging that facilitates the UWAWA information extraction. These methods are implemented as two sub-modules, namely, knowledge acquisition and NLP enabled Q-A. We evaluated UWAWA against the well-known application of WolframAlpha, with a set of systematically constructed test sentences. The result shows UWAWA performs well in the case of attribute processing and questions with misspelling words.

Keywords: Information Extraction, Web Services, artificial intelligence, Text Mining.
CR Categories: A.2, I.7.2

Acknowledgements

I would like to give special thanks to my special supervisor Prof. Wei Lui who taught me a great deal about this master thesis. Without her golden guidance, invaluable dedication and patient, this project would not have been possible for me to achieve.

Also I wish to express my great thanks to all members of the School of Computer Science and Software Engineering for their advice, encouragements, supports and collaboration.

My deep thanks and gratitude go to my parents, my wife Ahlam ALABBAS and my brother Wael Saeed for their endless love and inspiration through the duration of this dissertation.

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

1.1 Background and Motivation

As a human computer interface, Natural Languages (NL) are the easiest to use for human beings, and they are the most intuitive interface for conversing with human. The system with such an interface could be a chatting toy or a conversational agent, enabling human users to communicate with a computer, using everyday spoken languages. Such systems need to have a Natural Language Processing (NLP) component to process and extract useful information from input sentences such as topic, sentiment or instructions. The main challenge is that inputs in natural languages do not follow any standard format, therefore, extracting useful information by NLP is still a challenging research topic.

There are a few natural language-based conversation systems that are worth mentioning. Façade is firstly one of the best natural language based game that use NLP for human computer interaction. It attempts to extract interesting information about players who interact by typing text. According to Mateas and Stern, “The Facąde NLP system accepts surface text utterances from the player and decides what reaction(s) the characters should have to the utterance” [13]. Another system is Alice which is one of most popular chat bot and a predominant conversational software. It uses a specific language that is called AIML (Artificial Intelligence Markup Language), which enables people to insert knowledge into Alice in a machine readable format [22].

1.1.1 Façade

Central to the application of interactive conversations in games and artificial intelligence is field of education and entertainment [5]. Therefore, the developers promote many kinds of applications that can interact with users by texting, voicing or touching. The chat bot examines the players’ emotion by extracting some sentences that might indicate the players situation. However, far too little attention has been paid to the researchers perform further searches that are significant in information extraction area, because the sentences that have input do not follow any standard format, and also extracting by Natural Language Processing is still challenging in particular for the recognition of the gamers feeling or mood by only textual input [5].

Mateas and Stern [13] developed the Façade game, which is classified as an interactive artificial intelligence system. It has the capability of interacting with users via textual inputs to influence the game direction. Façade is an instantaneous, first-person natural language-based game that uses the Natural Language Processing technique to enable human computer interaction. In effect, the chat bot of Façade extracts some interesting information from the surface text and determine the level of information input by the player [13].

The game operates by putting the player in the role of close associate with the major antagonists, (Grace and Trip), a couple who invite the player to their home for a drink, refreshments and a conversation. However, the pleasant gathering of the couple and the player is disrupted by domestic conversation between the couple the minute the player arrives to the house. Through the use of a language processing software, the game allows the player to type sentences to communicate with the couple, either to support them and get them to come to terms with their current argument, or to drive them apart. The latter scenario will eventually lead to the player being thrown out of the apartment [4].

The main consideration of the game is the capability of the game to extract useful information from textual sources and makes decision based on analytical results of the sentences. Additionally, the sentences are received from surface text that have been applied in the chat bot in order to encourage user to insert details into the machine. For example, “if the player types “Grace isnt telling the truth”, the NLP system is responsible for determining that this is a form of criticism, and deciding what reaction Grace and Trip should have to Grace being criticized in the current context” [13].

There is a large volume of published studies describing the role of Information extraction, thus researchers perform further investigations to increase understanding level of textual natural language in the chat bot. This supports the capability of machine to analyse users’ inputs at the lexical, syntactic and semantic level. Furthermore, Façade rises its capability of understanding what the player wants and how to behave in such situation by analysing the input with high level of meaningful methodology. [13].

1.1.2 Discourse Acts

The process of information extraction from textual input is based on two primary stages, which are keyword definition and relationship recognition. In the first stage, the system divides the sentence into individual words to facilitate the analytical process. However, the second stage identifies the relationship between desired words and database of discourse acts, this is to grow the understanding level. The issue of the understanding level appears in the difficulty of making a decision of what is the most appropriate relationship that matches words and the discourse acts [10]. Basically, developers of Façade designed a set of discourse acts that represents particular meaning. The aim of their research is to address identify relations between discourse acts to textual inputs [13].

Hereby, we try to compare Façade and Alice in several aspects. Façade has more understanding of what input is about, but Alice applies syntactical structure by detecting certain sentence. For instance, if you say “I”, Alice will reply with “you”. Moreover, Alice uses POS (Part Of Speech) tagging to identify part of speech, then attempt to replace or construct a reply. For example, when player says “my name is Waleed”, Alice will respond by “Hello Waleed”. However, Façade determines and understands objects by the use of discourse acts. Finally, Alice has a knowledge base that contains retrieved knowledge. It is used by Alice to extract useful information, but in some points it tends to be rigid rather than flexible, whereas Façade is more flexible. this is because Façad models a rather configurable small environment, whereas Alice is general purpose.

1.2 Project Aim

The main aim of this project is to apply the Façade methodology to process inputs for conversational agent. In particular, we attempt to extract useful information about a specific fictional character from an inserted sentence. This style leads the system to understand the commands input by a user, to improve its performance and to construct meaningful conversation.

1.3 Methodology

This project is based mainly on applying Façade methodology which has several aspects such as discourse acts, template and rules, on conversational agent. Indeed, this methodology depends on the sort of user's sentence. The inserted sentence from a chatter will be tokenised and then apply Part Of Speech (POS) tagging to facilitate information extraction.

A system is built with two modules in order to achieve the project aim. The first module is to gather knowledge about fictional characters from Wolframe Alpha knowledge engine website. The other program is to extract specific details that a user searches for by seeking the keyword in the plaintext. For example, if a user types "Who is spiderman's wife?", the system will extract the details about spiderman's wife.

This project implements a web interface in a PHP with a MYSQL database. The PHP coding involves POS as a class to tokenise the sentence. The MYSQL is used to insert or select details by sending some queries to a database.

1.4 Dissertation Structure

The following details the structure of the rest of this dissertation.

- Chapter 2: Literature Review:
  The literature review chapter clarifies several aspects that could be applied in the project. The chapter first explains the Façadeś surface text processing is further explained as in how it applies NLP for extracting information from user. Another aspect is how discourse acts play a significant role in information extraction to understand the user request. The next aspect is an information extraction using database queries for respond to the right goal. Information Extraction System based on Inductive Learning and Meta learning is another aspect in this chapter. Lastly, the chapter presents the information retrieval system which is a technique that uses noun phrases to search for certain information on the Internet.

- Chapter 3: Methodology:
  This chapter, first explains the process of building and populating a knowledge base about fictional characters. Then how the knowledge base can be used to answer user queries.

- Chapter 4: Experimental Result and Analytical Evaluation
  This chapter describes the questions classifications that used to exam the two applications. The chapter then presents the results obtained throughout the experiment, followed by analytical evaluation of the results in terms of recognising performance.

- Chapter 5: Conclusion and Future Work:
  The conclusion chapter summarises the significant points of this dissertation. It start by reviewing the objectives of the project, it methods and results. The chapter then concludes with some suggestions to guide future research.

CHAPTER 2

Literature review

The ease of access and the abundance of free electronic text have, become the driving force for the surge of information extraction research. The goal of information extraction is to discover and extract useful information into structured storage from free-formal text in natural language. For example, the topics, sentiment or instructions need to be deciphered for the computer system to make use of them. A natural application area of using natural language as the primary human-computer interface is chatting toys or game conversational agents [17] [21].

This literature review will investigate the detailed technique of surface text processing methodology. In this section, we separate the techniques into Lexical, Syntactical and Semantic three different levels of text processing.

2.1 Lexical Level

The field of information retrieval plays a crucial role in searching certain information through the Internet. Search engines are keyword-based, which requires a query in order to start. Therefore, matching words in the documents with the words in the query is essential of the information retrieval system [19].

The textual input is created with a set of two aspects which are representation of discourse acts and key words. Theses aspects are designed to be significantly related to each other by their meaning. There are two stages to extract information from a textual input. The first stage is called keyword definition. In this stage, each sentence will be separated into individual words which simplify the process of information extraction. On the other hand, the second stage will actually grow the understanding level by recognizing the bonds between the representation of discourse acts and the keywords. Here we have inserted a table 2.1

that would explain how the database of discourse acts and the designed keywords are related [13].

2.2 Syntactical Level

2.2.1 Part Of Speech Tagging

The Part Of Speech is actually an application that has the functionality to classify a parsed sentence. It has been used in web search queries, which are a fundamental aspect that significantly contributes to information retrieval task, in order to increase the accuracy of web-search queries performance. The sentences are usually written in short forms and their quality depends on their grammatical structure. The reasons behind using the application of POS are to discover two main aspects which Cory Barr, Rosie Jones and Moira Regleson [3] are aiming for in the topic of web-search queries. Firstly, POS will be used to determine the tagging performance of the web-search system queries for typical English language. Secondly, it will be taken to develop the search results by investigating the qualification value of these tags. The method, that is been followed by Cory Barr et al in this paper [3], is to set up part of speech tags which makes it appropriate for search query and quantify their results. These results are apparently caused by the current part of speech taggers which are been chosen in this project. There are two part of speech taggers here. One of them is the Brill tagger. This tagger needs to label all tokens with their part of speech tag and need a lexicon. The reasons behind using Brill tagger are firstly, it automatically generates rules that lead for an easier system to be readable by users. Secondly, it is a popular tagger. The other tagger is Stanford part of speech tagger. This tagger has the best accuracy performance in the field. Those two part of speech taggers can explore the type of search queries and part it into grammatical classes based on the noun-phrases queries [3].

In conclusion, the researchers above have made initial investigative experiments in order to test the performance of the POS tagging. Those experiments indicate that part of speech information plays an important role in the outputs of a machine-learned system. This system actually leads to have beneficial proper nouns in queries. In addition to the experiments, they show how accuracy the POS tagging would react in order to select or substitute the right words into the query reformulation. It has been investigated, through practical analysis and experiments, that the relevance web search results can be improved and

Table 2.1: Discourse Acts adapted from [13]

Representation of Discourse Acts	**Keyword**
DAAgree ?char	Agree, okay, pass on, make a deal and cool.
DADisagree ?char	Disagree, unsuitable, no way and unbelievable.
DAPositiveExcl ?char	Yeah, Wow, interesting, amazing, and wonderful.
DANegExcl ?char	Damn, awful, bad and I do not like.
DAHappyExpress ?char	Cheerful, satisfied and thrilled.
DASadExpress ?char	Cheerless, heartbroken and wistful.
DALaughterExpress ?char	Haha, ha ha, lol, loool.
DAAngryExpress ?char	Hate, pisses me off.
DAUnsure ?char	Maybe, unsure, could be, do not know, guess so.
DAThank ?char	Thank, Ta, Thanks.
DAGreet ?char	Hello, Hi, Good morning, Good evening, whats up.
DAAlly ?char	Like you, love you, you are friend, you are mate.
DAOppose ?char	Hate you, kiss of, get out.
DAMisUnderstand ?char	Not understand, confused, do not get it.
DAApologize ?char	Sorry, forgive.
DAPraise ?char	Cute, sweetheart got good idea.
DAIntimate ?char	Talk to me, Whats wrong.
DAGoodbye ?char	Goodbye, Good night, see you, catch you later.
DAContinue ?char	Continue, keep up, press on and stay.
DAExplain ?char	Explain, illustrate, tell

contributed with query reformulation by the part-of-speech tagging. As a result, entity detection and proper-noun detection are recommended for achieving higher improvements. In particular, they are aiming to determine the accuracy of these improvements, which leads to develop the performance of part of speech information [3].

2.2.2 Noun Phrase Chunking

The most important concepts are noun phrases. Therefore, special Natural Language Processing task focuses on Noun Phrase chunking. Natural Language Processing technique is used here by authors [19] to support an amount number of Natural Languages queries and replace them over the used keywords. The reason why this technique is costumed is because it is known as an easy system that uses key phrases in order to deals with noun phrases from a document. These noun phrases contain three main modules which are been used in this research. Those modules are; tokenisation, part of speech tagging and noun phrase identification using Chunking. Lastly, there is an evaluation method that is used to test the quality of the information retrieval technique and its results were positive. The information retrieval is mainly using the Natural Language Processing to represent the right documents that satisfy the users needs. Hence, the Natural Language Processing might be useful to develop the precision of the internet search as well as the NL system which can also be used in society. One of the most remarkable Natural Language Processing modules that would help to develop the accuracy of the web search is the Chunking. The function of this tool is to extract noun phrases first and then it retrieves them back in order to improve the performance more than using the key phrases [19].

2.2.3 Parse Tree Database

Most traditional Information Extraction system takes a processing pipeline, in other words, the text go through tokenisation, sentence splitting, Part Of Speech tagging and noun phrase chunking etc to extract useful information. This pipeline needs to re-run on the text corpus should any new requirement code in.

Tari et al [20] exhibits a demonstration plan, which depicts an innovation model for information extraction [20]. They use parse tree output to store the result of textual processing. Nevertheless, the proposed model for information extraction is meant to replace the traditional extraction systems, which were executed as a pipeline of processing modules. They argue that the traditional approach is time consuming and cumbersome. According to developers [20], the present approaches are rigid and costly in the face of the needed dynamic application. In essence, the study [20] recommends the development of new extraction systems for the current approaches. However, academics observe that this can be expensive since the new developed extraction system requires the whole corpus to be recomputed from scratch. It is worth noting, that not the entire corpus is affected with the new acknowledged entities, since most of these entities overlap both the primary and enhanced recognizers [20].

The proposed new model of information extraction offers a general-purpose extraction system, which meets varied extraction requirements efficiently [20]. In this regard, investigators distinguished two phases of processing that can protect the corpus from being entirely affected. First is the initial phase, whereby a one-time parse is carried out, to identify candidates for individual entries on the entire corpus, based on the knowledge available. The second phase is the extraction phase, involved circulation of a parse tree database, a storage platform for syntactic parse tree and semantic entity attachment. Consequently, to ease extraction process, a query language named Parse Tree Query Language (PTQL) was designed and implemented, which automatically generates queries for high quality extraction [20].

The paper [20] explained the system architecture of the GenerIE system, in which the initial phase performed for corpus processing, is done by the text processors and stockpiled in the Parse Tree Database (PTDB) [20]. In this part, four modes are generated to enable the user to identify the PTQL to use in the extraction process. Firstly is the text parsing and PTDB, secondly, information extraction by use of PTQL queries, thirdly, pseudo-relevance generation of feedback query and finally, query evaluation and optimization [20]. The approach has proven to be beneficial, because it provides an innovative database central structure for information extraction, in terms of incremental evaluation and database query optimization. [20].

2.3 Semantic Level

Systems used for extracting information perform the analysis for unrestricted text. This is done in order to extract information about pre specified events, entities or relationships. The extracted information is related to a specific domain-ontology [23]. Therefore, there is a large volume of published studies describing the role of ontology in term of information extraction.

Wimalasuriya and Dou [23] describe that ontologies can be used in the process of extracting information. This is because ontologies are precise and have high recall capacity. Their paper focuses on key issues such as:

- Use of multiple ontologies in guiding the information extraction process.
- Challenges involved with multiple ontologies.
- Suitable ontologies and their mappings.
- Solutions to the challenges and experimental results.

Faster creation of content can be achieved through the use of multiple ontologies. The various types of multiple ontologies whose information has been published include; Kylin, C-PANKOW and SOBA [23]. Ontologies should be capable of locating objects from a group of items. Precision and Recall are immensely vital in information extraction. A single ontology in the process of information extraction leads to have less extracting and recall. This problem was solved by following the introduction of multiple ontologies, which provided a myriad of perspectives. The developed multiple ontologies are either involved in; Domains, for example, University domains, or for providing varied perspectives of the same domain [23]. There are two advantages of using multiple ontologies to provide different perspectives in OBIE, which are possible improvement in recall and supporting multiplying perspectives. The extraction based on mappings between concepts of ontologies can be employed in other systems. Multiple ontologies also allows for combination of results that are not necessarily linked [23].

Multiple ontologies that specialize on domains were evaluated using university domains. Based on the training set that is guided by the result and assist to improve ontology. On the other hand, multiple ontologies providing different perspectives were evaluated by the use of the domain of terrorist attacks. The 4th Message Understanding Conference (MUC) corpus was used. Two ontologies obtained from the MUC structure and the Mindswap group of the University of Maryland were also used [23]. Use of multiple ontologies achieved better recall and precision in OBIE. This was verified by the two case studies. A higher figure was obtained when ontologies represented specialized subdomains. This is because a specialized sub domain is made up of two subsets. One subset has terms representing specialty of the domain while the other has text documents related to the domain terms. A lower figure was recorded when different perspectives were provided by the ontologies. Improved precision is as a result of extraction of information by specialized ontologies [23].

The main shortcoming of multiple ontologies is determining the theoretical basis for using the ontologies in extracting information. This can be solved by carrying out intensive research on ontologies and extraction techniques. Finding favourable ontologies and mappings is a problem. Thorough assessment of ontologies to use in the OBIE system should be conducted. Adequate knowledge concerning mappings between the concepts of the selected ontologies should also be acquired. Experimental results from the two case studies were evaluated. Although lower precision was noted in the case of different perspectives, the results proved that multiple ontologies offer better precision and higher recall. The research can be applied in different fields for instance, oil companies. The companies can extract accurate and quality information instead of relying on human interpretation [23].

Wimalasuriya and Dou [24] have provided that, some ontology systems are capable of constructing ontology from their own domain. For instance, an ontology system like OBIE can generate semantic contents automatically. The problem being present in their paper, however, is concerned with ontology technology and its use in information extraction. According to researchers [24] "Multiple ontologies exist for most domains and there is no rule that prevents an OBIE system from using more than one ontology". The paper provides that, the use of multiple ontologies can develop the process of information extraction since multiple ontologies can make more extractions in a single instance. To achieve their objective, researchers have employed two case studies. The challenges experience in this study is concerned with widespread of information extraction. However, this technology is not used widely due to cost expenses and the complex nature of this technology. To address these issues, the authors have suggested the reuse of IE components. Reusing IE components has been tackled in their paper and according to authors it is preferable to other techniques. The authors state that the obtain performance of the study is lower than in comparison with other studies. In addition to that, the paper provides some of the future works required in this study [24].

Macias-Galindo et al [11] defined MKBUILD as "A tool that follows a methodology to create domain-specific ontologies containing related concepts, drawning from existing large-scale resources such as WordNet and Wikipedia/DBPedia" [11]. Consequently, the paper aims at providing a conversational agent from Modular

Knowledge Bases (MKBs) perspective. Nevertheless, the intent of developing MKBUILD is to substitute the hierarchical process of structuring definite domain ontologies that consume both human effort and time. In essence, the paper contours the architecture of the conversation agent by use of an interactive toy. To select a suitable conversational path to pursue, developers applied techniques such as keyword spotting and lightweight semantic parsing. In addition, they pointed that topic module designer generated fragments, whereby the designer ensures the fragments with one node in a topic network. In this regard, each fragment cosiest of a head that provides suitable conditions and body that entail list of the anticipated inputs. According researchers, the Topic Transition Network is utilized to create a consistent dialogue structure, productively used in the selection of the conversational fragments for the next part of conversation [11].

Authers follow strategies to create domain-ontology, then start involve identification of the main concept ontology, finding the higher layer of MKB and expand the domain ontology. The results of the study exhibited that, for domain Internet use of MKBUILD tool was irrelevant due to the available technologies such as televisions and radar [11].

Information extraction (IE) is one of the most common systems that automatically extract structured information in the world. It plays an important role in many of the real world applications in the areas of business intelligence, competitive and military intelligence. XONTO, which is considered to be as a method of the information extraction functions, is a system that has been found from the idea of describing ontology objects and classes which can be dominated by written rules that is called descriptors. These descriptors allow XONTO to accurately determine the requested objects and classes. In other words, this system that is also known as self-describing ontologies uses set of rules to obtain and extract ontology objects and classes which are contained in PDF documents. However, once the system encounter with document in tabular forms, XONTO will qualify its rules and will give an expression to the semantic of the information in order to extract the requested orders [16].

In conclusion, in this task, the XONTO, that extracts information from PDF documents, is used for the ontology based system. It is called self-describing ontology system which enables exploring semantics in (IE) from PDF documents by gathering the power of ontology forms and attribute grammars. This work aims for three major aspects. First of all, It will be used to solve complex issues and analysis them accurately. Secondly, it will gain an extension of the approach to permit exploiting web standard ontology language. Lastly, XONTO will increase its methods and strategies by realising other ontology approaches. [16]

2.4 Rule Based

Muludi et al [14] present a proposal for Information Extraction System based on Inductive Learning and Meta learning, whose performance is exceptionally good. The proposed system seeks to make the search and use of information on the Internet easy. It studies related works and compares them to the Multi Inductive Learning (MIL) system. The state of the art system being compared to MIL is LP2. LP2 is defined by authors [14] as "learning by using symbolic rules for identifying start tag and end tag class of the slot". The system presented in this proposal seeks to address the problem in the previous existing systems. The problem is that the performance of these systems are not consistent on various information domains. Information extraction can be approached as classifying problems. The text is divided into tokens and classified into related classes [14].

The new idea presented in this paper [14] is the use and introduction of machine learning to improve on information extraction. MIL concept is inspired by the idea of how to use document training to look for best classifier for each slot in a certain domain. MIL is considered better as compared to natural language processing, because of its portability, scalability and adaptability. They managed to show the working of the Multi Inductive Learning and compare it with other systems such as RAPIER, LP2 and SNOW. They showed how the best classifier for each slot is chosen to achieve the best performance in extraction of information from the testing document. The authors used a 10 fold cross validation on training document, they associated each base learner with each slot. They obtained results and analyses them. They concluded that Multi Inductive Learning chooses the best learner. It is the best among other state of the art information systems [14].

The researchers have applied many trainingg sets to allow the generating accurate rules of information extraction. The system designed could not provide consistent results. They decided to use the strengths of different systems mentioned in the document to come up with an algorithm that could overcome the limitation. These included the base learner and performance index that were used to design the algorithm for MIL algorithm. According to authors there is no single classifier that can produce consistent performance across all domains and slots. Multi Inductive Learning has the best performance on average as compared to others. The work is beneficial in that it shows that although there is no single classifier that can produce consistent results across all domains and slots, Multi Inductive Learning is much better than other methods that are in use such as single classifier [14].

2.5 Summary

In brief, we have addressed in details very important aspects of surface text processing methodology, which are significantly related to the process of information extraction. These aspects are divided as following; firstly, the lexical level which leads an application to determine the right meaning of a word. It offers two strategies to extract useful information, information retrieval strategy and discourses acts strategy. The second aspect of this literature is the level of syntactic that contains three different methods such as part of speech tagging, noun phrase chunking and parse tree database. If one of these methods is used, the level of syntactical will be reached eventually. The semantic level is the third aspect. It concentrates on extracting information by the implementation of ontology technique, that helps to understand knowledge as well. Finally, rule based aspects utilised for information extraction through the use of multi inductive learning approach. Also it can be used to recognise and classify problems into different categories.

CHAPTER 3

Methodology

3.1 Overview

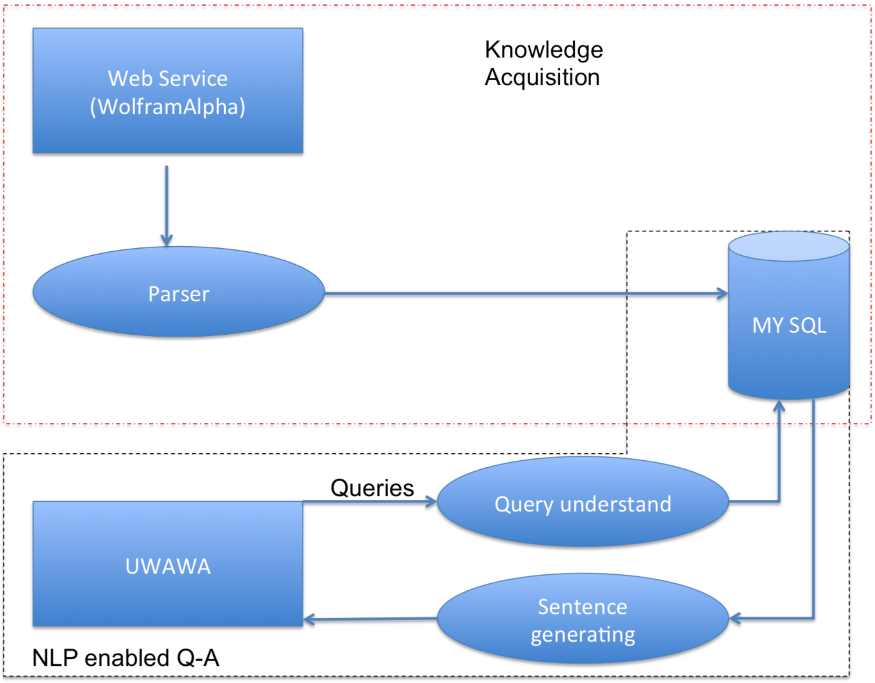


Figure 3.1: Overview of methodology

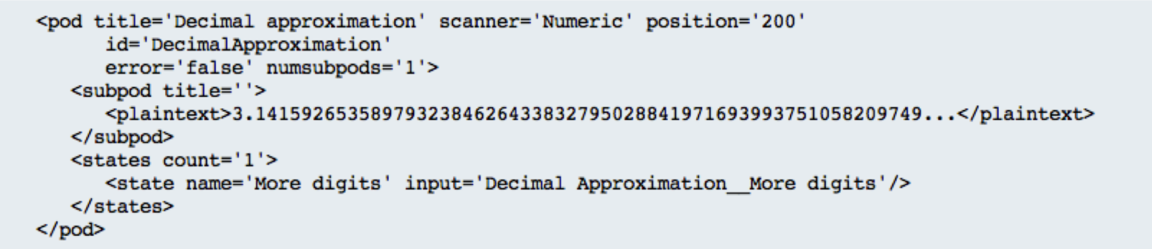
As mentioned earlier, this project aims to extract information from a user input and then applies text processing algorithm in plain text to respond with useful information. We built a system consisting of two modules to achieve our goal. The knowledge acquisition module builds information about fictional characters, while the NLP enables the Q-A module to extract information from the user in two text processing levels (See Figure 3.1).

First of all,The data used was extracted from WolframAlpha website [1]. We extracted about 110 fictional characters. A MySQL database is used to store structure and organise information collection from WolframAlpha website. Database[1] plays a very important role in computing by enabling easy access and manipulation of amount of information. [15]

3.2 Web service by WolframAplha

3.2.1 Introduction to WolframAlpha web services

WolframAlpha is a computational search engine used for answering a query. It concentrates on knowledge-bases rather than on normal search engine in terms of search processing. In fact, WolframAlpha enables developers to utilise its facilities for different types of development purposes such as information extraction. One of the main outputs of WolframAlpha is a pod, which is an area to express the results. It can be more than one pod in the result that depends on result classification. Furthermore, the pod has many options to gather a result such as plaintext for easy copying. The pod has at least one sub pod for revealing the actual content. The results from WolframAlpha are provided in an XML document. It contains the pod and sup pod, which has the actual content that can be comfortable method for recognizing the answer (See Figure 3.2) [1, 18].



```
<pod title='Decimal approximation' scanner='Numeric' position='200'
      id='DecimalApproximation'
      error='false' numsubpods='1'>
   <subpod title=''>
      <plaintext>3.1415926535897932384626433832795028841971693993751058209749...</plaintext>
   </subpod>
   <states count='1'>
      <state name='More digits' input='Decimal Approximation__More digits'/>
   </states>
</pod>
```

Figure 3.2: Example of XML and pod obtained from WolframAlpha [1]

3.2.2 Build Knowledge from WolframAlpha web services

The code following is applied as a query. This code is run in a sequence of packages from 1 to 10, 11 to 20 until 110. We implemented this method

[1] Wikipedia, Relational database, this is available from: https://en.wikipedia.org/wiki/Relational_database

Bibliography

[1] Alpha, W. This is available from http://www.wolframalpha.com, 2013.

[2] Andersen, E. Edging toward the semantic web: Protocols, curation, and seeds. *Ubiquity 2010*, November (Nov. 2010).

[3] Barr, C., Jones, R., and Regelson, M. The linguistic structure of english web-search queries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Stroudsburg, PA, USA, 2008), EMNLP '08, Association for Computational Linguistics, pp. 1021–1030.

[4] Dow, S., Mehta, M., Lausier, A., MacIntyre, B., and Mateas, M. Initial lessons from ar facade, an interactive augmented reality drama. In *Proceedings of the 2006 ACM SIGCHI international conference on Advances in computer entertainment technology* (New York, NY, USA, 2006), ACE '06, ACM.

[5] Dow, S. P., Mehta, M., MacIntyre, B., and Mateas, M. Eliza meets the wizard-of-oz: blending machine and human control of embodied characters. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2010), CHI '10, ACM, pp. 547–556.

[6] Echizen-ya, H., and Araki, K. Automatic evaluation method for machine translation using noun-phrase chunking. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (Stroudsburg, PA, USA, 2010), ACL '10, Association for Computational Linguistics, pp. 108–117.

[7] Han, Y.-S., Ko, S.-K., and Salomaa, K. Computing the edit-distance between a regular language and a context-free language. In *Proceedings of the 16th international conference on Developments in Language Theory* (Berlin, Heidelberg, 2012), DLT'12, Springer-Verlag, pp. 85–96.

[8] Hands, A. Wolfram—alpha http://www.wolframalpha.com. *Technical Services Quarterly 29*, 2 (2012), 171–172.

[9] Hearst, M. A. 'natural' search user interfaces. *Commun. ACM 54*, 11 (Nov. 2011), 60–67.

[10] Kitani, T. Merging information by discourse processing for information extraction. In *Artificial Intelligence for Applications, 1994., Proceedings of the Tenth Conference on* (mar 1994), pp. 412 –418.

[11] Macias-Galindo, D., Cavedon, L., and Thangarajah, J. Building modular knowledge bases for conversational agents. In *Proceedings of the 7th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems* (2011).

[12] Marcus P. Zillman, M.S., A. An internet miniguide annotated link compilation. In *Academic and Scholar Search Engines and Sources* (2012).

[13] Mateas, M., and Stern, A. Natural language understanding in faade: Surface text processing. In *Proceedings of the Conference on Technologies for Interactive Digital Storytelling and Entertainment* (2004).

[14] Muludi, K., Widyantoro, D., Kuspriyanto, K., and Santoso, O. Multi-inductive learning approach for information extraction. In *Electrical Engineering and Informatics (ICEEI), 2011 International Conference on* (july 2011), pp. 1 –6.

[15] MySQL. Mysql reference manual, this is available from http://dev.mysql.com/doc/refman/5.5/en/what-is-mysql.html, 2013.

[16] Oro, E., and Ruffolo, M. Xonto: An ontology-based system for semantic information extraction from pdf documents. In *Tools with Artificial Intelligence, 2008. ICTAI '08. 20th IEEE International Conference on* (2008), vol. 1, pp. 118–125.

[17] Phyu, A., and Thein, N. Domain adaptive information extraction using link grammar and wordnet. In *Creating, Connecting and Collaborating through Computing, 2007. C5 '07. The Fifth International Conference on* (jan. 2007), pp. 47 –53.

[18] Professor Rhoda Joseph, P. Wolframalpha, a new kind of science by bruce walters, 2011.

[19] Subhashini, R., and Kumar, V. Shallow nlp techniques for noun phrase extraction. In *Trendz in Information Sciences Computing (TISC), 2010* (2010), pp. 73–77.

[20] Tari, L., Tu, P. H., Hakenberg, J., Chen, Y., Son, T., Gonzalez, G., and Baral, C. Generie: Information extraction using database queries. In *Data Engineering (ICDE), 2010 IEEE 26th International Conference on* (march 2010), pp. 1121 –1124.

[21] Turmo, J., Ageno, A., and Català, N. Adaptive information extraction. *ACM Comput. Surv. 38*, 2 (July 2006).

[22] Wallace, R. The elements of AIML style. ALICE AI Foundation., 2004.

[23] Wimalasuriya, D. C., and Dou, D. Using multiple ontologies in information extraction. In *Proceedings of the 18th ACM conference on Information and knowledge management* (New York, NY, USA, 2009), CIKM '09, ACM, pp. 235–244.

[24] Wimalasuriya, D. C., and Dou, D. Components for information extraction: ontology-based information extractors and generic platforms. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (New York, NY, USA, 2010), CIKM '10, ACM, pp. 9–18.

APPENDIX A

Original Master Dissertation Proposal

A.1 Background and Motivation

As a human computer interface, Natural Languages (NL) are the easiest to use for human beings, and they are the most intuitive interface for conversing with human. The system with such as interface could be a chatting toy or a conversational agent, enabling human users to communicate with a computer, using everyday spoken languages. Such systems need to have a Natural Language Processing (NLP) component to process and extract useful information from input sentences such as topic, sentiment or instructions. The main challenge is that inputs in natural languages do not follow any standard format, therefore, extracting useful information by NLP is still a challenging research topic.

There are a few natural language-based conversation systems that are worth mentioning. Facąde is firstly one of the best natural language based game that use NLP for human computer interaction. It attempts to extract interesting information about players who interact by typing text. According to Mateas and Stern, "The Facąde NLP system accepts surface text utterances from the player and decides what reaction(s) the characters should have to the utterance" [13]. Another system is Alice which is one of most popular chat bot and a predominant conversational software. It uses a specific language that is called AIML (Artificial Intelligence Markup Language), which enables people to insert knowledge into Alice in a machine readable format [22].

Hereby, we try to compare Façade and Alice in several aspects. Façade has more understanding of what input is about, but Alice applies syntactical structure by detecting certain sentence. For instance, if you say "I", Alice will reply with "you". Moreover, Alice uses POS (Part Of Speech) tagging to identify part of speech, then attempt to replace or construct a reply. For example, when player says "my name is Waleed", Alice will respond by "Hello Waleed". However, Façade determines and understands objects by the use of discourse acts. Finally, Alice has a knowledge base that contains retrieved knowledge. It is used by Alice to extract useful information, but in some points it tends to be rigid rather than flexible, whereas Façade is more flexible. this is because Façad models a rather configurable small environment, whereas Alice is general purpose.

A.2 Project Aim

The main aim of this project is to apply Façade methodology to process inputs for conversational agent. In particular, we attempt to extract useful information about a specific fictional character from an inserted sentence. This style leads the system to understand the commands from a user, to perform well and to construct meaningful conversation.

A.3 Methodology

This project is based mainly on applying Façade methodology which has several aspects such as discourse acts, template and rules, on conversational agent. Indeed, this methodology depends on the sort of user's sentence. The inserted sentence from a chatter will be tokenised and then apply Part Of Speech (POS) tagging to facilitate information extraction.

A system is built with two modules in order to reach the project aim. The first module is to gather knowledge about fictional characters from Wolframe Alpha knowledge engine website. The other program is to extract specific details that a user searches for by searching the keyword in the plaintext. For example, if a user types "Who is spiderman's wife?", the system will extract the details about spiderman's wife.

This project implements a web interface in PHP with a MYSQL database. The PHP coding involves POS as a class to tokenise the sentence. The MYSQL is used to insert or select details by sending some queries to a database.

A.4 Timeline:

Stage 1	Project proposal due to Coordinator.
	Learn PHP and MYSQL.
	Project proposal talk presented to research group.
	Literature Review and Revised project proposal due to Coordinator.
Stage 2	Write code and perform test.
	Draft dissertation due to project supervisor(s) .
	Final dissertation due to Coordinator.
	Seminar presented to seminar marking panel.
	Design Poster .
	Corrected dissertation due to Coordinator

Building a Question-Answer System from WolframAlpha

Waleed SAEED
STD: 20543932
The University of Western Australia,
School of Computer Science and Software Engineering
Supervisor: Wei Liu

*This report is submitted as partial fulfilment
of the requirements for a Masters Degree in the
School of Computer Science and Software Engineering,
The University of Western Australia,
2013*