Information Retrieval System using UNL for Multilingual Question Answering

Thesis Report

submitted in partial fulfillment of the requirements

for the award of degree of

Master of Engineering

in

Computer Science and Engineering

Submitted By

Kanu Goel
(801432007)

Under the supervision of:

Dr. Parteek Bhatia
Associate Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR UNIVERSITY
PATIALA – 147004

June 2016

Certificate

I hereby certify that the work which is being presented in the thesis entitled, **"Information Retrieval System using UNL for Multilingual Question Answering"**, in partial fulfillment of the requirements for the award of degree of Master of Engineering in **Computer Science and Engineering** submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of **Dr. Parteek Bhatia** and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Kanu Goel
801432007
ME(CSE)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

Dr. Parteek Bhatia
Associate Professor
Computer Science and Engineering Department
Thapar University

Countersigned by

(Dr. Maninder Singh)
Head
Computer Science and Engineering Department
Thapar University
Patiala

(Dr. S. S. Bhatia)
Dean (Academic Affairs)
Thapar University
Patiala

i

Acknowledgement

I would like to express my thanks to my guide Dr. Parteek Bhatia for being an excellent mentor for me during my whole course of thesis.Your encouragement and valuable advice during the entire period has made it possible for me to complete my work.

Sir you gave me complete freedom, right from choosing the topic for thesis until its completion. Timely checks and guidance about how to improve the work done played an integral role. Also his support and patience in times of difficult situations helped me recover when my steps went astray.

Next most importantly all this would not have been possible without the patience and support of my family.Their constant sense of care encouraged me to complete my work.

I would like to acknowledge Dr. Maninder Singh for setting high standards for his students and encouraging them time and again to achieve them as well.

Also I thank Paras Vij for having technical discussions on related topics that helped me improve my knowledge on the same.

Kanu Goel

ii

Abstract

Efficient storage of information in a form which is independent of language forms the basis of enabling information sharing smoothly across various channels. Also the information retrieval systems build so far have the basic tendency to provide the user with the specific information that has been demanded by him. Refined query processing has facilitated easy retrieval across various regions where different languages are spoken using only a single corpus. This can be really helpful for designing a Question Answering system that does not require hard wiring of corpus.

This requirement of having a single corpus for multiple languages is a challenging one and this thesis explains the solution using UNL. UNL is a widely known platform meant to receive,distribute and understand the information on multingual basis. It provides the computer systems with huge amount of knowledge that is accessible as well as understandable. In this thesis a very basic use case that is based on the semantic approach to handle the user based queries has been presented. Instead of returning whole bunch of related information to the user query , the system returns specific answer that gives the most descriptive infromation about the target term. This proposed system has been constructed to cater to the applications of the natural language processing and finally build a model that is language independent.

The thesis describes the process of designing a UNL based QA system which is capable of handling the factoid based user queries such as what, where, what, why and which using a single language independent corpus. This system can service all the language provided their EnConverter and DeConverter are available. After testing the proposed system on the defined corpus it is seen that promising results have been achieved which thereby have proved the utilty of the system.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Introduction

The primary focus for designing any Question Answering(QA) system is to fetch the answers to the queries posted by the users. A QA system is broadly comprising of the user query processing module, refining the processed query according to the corpus and the corpus upon which the whole system is built.

Traditional approaches required replicating the entire corpus in multiple language so as to make the system compatible to different geographies. This introduced a lot of redundancy and was a major limitation. The internet is a biggest example of hyper connectivity, but multilingualism is still something which is a pending issue, and this is more prominent in case of a QA system.

With the thesis a new system is proposed that will reduce the redundancy making the corpus language independent. It shall use language independent UNL to achieve the same. This will consist of a language independent corpus and shall have EnConverters and DeConverters for each language that this system will cater too. This system shall thus be completely language independent.

1.2 Need of multilingual Question Answering System

Question Answering system is a very classic application in the domain of Natural Language Processing. As it seen that it has wide range of practical applications in various fields like health care, education and also personal assistance. So many traditional systems as discussed ahead in this thesis have been built to cater to the needs of the society. Building a question answering system itself involves a large number of steps ranging from identifying the type of questions to generating the answer in the user demanded format. Proper analysing of the questions will enhance the accuracy of the system. Now these systems primarily include the corpus based systems where it becomes essential to have a specified database onto which the queries can be imposed. This serves well for a limited amount of domains where possibly the data can be made available in different languages as per the needs. But if seenfrom a broader prospective this approach fails to serve the greater need where it

is not feasible to have the corpus particular to all the languages. So here comes the need of designinig a multilingual QA system wherein the user will be required to just impose the queries in his choice of language. Futher this queries will be mapped to a corpus that will be language independent. Later the results will be delivered in the desired target language.

1.3 Challenges of the QA systems

There has been a focussed research that has been conducted in the field of existing question answering systems. A significant reflection pertaining to the current scenario of the QA systems has been largely addresed in this thesis work. There are various dimensions attached to the to the Question Answering systems which impose as challenges.First of all the system should conduct a proper analysis to find out the answer type.They may be short or long , or may be narrative.For instance if the user require the justification it would demand a long answer.Next the evaluation of the answer if it is optimum or not , is also very important. The system may give multiple answers but it becomes essential to select the most optimum answer. In some cases the longer answered may be selected while in other cases the shorter ones may be preferred. Third challenge would be to present the answer in an appropriate manner wherein the style of presentation would be decided by the user. The interface should be capable enough to handle user demanded way of depicting the answer.

1.4 Organization of thesis

The solution for performing interlingual question answering is presented in this thesis work.

The next chapter, 2 introduces the concepts of Universal Networking Language (UNL), the approach that is basically used to address to the problem and the used for the process of UNLization and NLization for building up the required system.

In chapter 3, some of the existing prominent QA systems that have been developed so far have been discussed in brief pertaining to literature survey.

Chapter 4 has shown the problem statement .

Chapter 5 presents the basic objective and the methodology that was adopted to cater to the interlingual QA problem.

Chapter 6 explains all the important components of the architecture of the proposed model wherein describing the role of each module briefly. It demonstrates the actual implementation of the system that was developed with the help of an UNL query-answer example that was taken.

Chapter 7 discusses the results and the discussion of the QA system that has been built.

Chapter 2
UNL Framework

Universal Networking Language(UNL) is a medium for representing information in a language independent format. It is recognised as the only living interlingua capable enough for depicting the genral-purpose contents. The following sections will describe the UNL system as a whole explaining each of its modules and components in detail. Launched in 1996 , by the Institute of Advanced Studies (IAS) of United Nations University (UNU) in Tokyo, Japan, UNL is an artificial language which in the form of semantic network has been used for computers to express and exchange information through its interlingual representation.

2.1 UNL System

The UNL system comprises of many components. The language resources are a bundle in any natural language. These are used by the parser while parsing the user inputted language to the desired language. Dependent and Independent are the two classification of the Language resources. Independent resources are independent of the languages and are common to all languages. UNL Ontology (UNLKB) is the common repository for all languages. Language Dependant are different for each languages. They can be assumed to be like the word dictionary. A separate repository is maintained for each language in the Language server. The EnConverter and DeConverter use these modules.

These language resources require to be processed and this is done by the parser. The parser module consists of EnConverter and DeConverter. This parser is in itself maintained using some tools. These form the major heart and soul of a UNL system. The below section shall talk about the various modules that make up a UNL.

2.1.1 Universal words (UWs)

This comprises of the vocabulary of UNL. This is the building blocks of a UNL sentences. This is what gives UNL its language interdependence. They can be considered analogous to words in any normal language, however UW are not confined to a language and they represent more than what ordinary words represent in a natural language. Also a UNL only represent one meaning whereas a natural

language word may represent more than one meaning. UW forms the node of universal expression. They along with attributes and relation represent the words.

A UW of UNL is defined in the format given in (2.1) :

<uw>=<headword>[<constraintlist>] …(2.1)

The headword of a UW is an expression, word or a phase or a sentence in a natural language. The unique meaning headwords themselves become UW. Else to make a more specific UW constraints are attached to the headword. A headword only UW is called a "Basic UW". Now the classification of the universal words is presented in Table 2.1.

Table 2.1: Classification of the Universal words

Type	Concept (in English)	Lexicalization (in English)
Simple UW	above average	big
Compound UW	comparative of above average	bigger
Complex UW	affix a stamp to	stamp
Temporary UW	UNDL Foundation	UNDL Foundation

Universality of UW

UW is expected to be universal and thus they are named so. This however does not mean or guarantee representation of a sort of common lexical denominator to all languages or a semantic primitive. UNL basically assures that is being understood by all. The universality can be assumed in a sense that they are uniform identifiers to the entities defined in the UNL knowledge base, which is expected to map everything that we know about the world, and that is used to assign translatability to any concept.

2.1.2 Universal Relations (URs)

The relationship between Universal words is defined by Universal Relations. Links is the formal word for universal relations. They link two UW together to make an UNL expression. They are like a labeled arc which connects a UW to another UW. UNL consists of 46 relations. Argumentative (agent, object, goal), Circumstantial (purpose, time, place), Logic (conjunction, and disjunction) Relations are the various kinds of

relations. UW and UR together make the semantic network of the UNL. List of universal relations along with their tags is shown in Table 2.2. The edges of the hyper graph or UNL expression are the UR. Taken the case of a sentence ‘ Girl eats pineapple’. Here girl(icl>person)@singular) and pineapple(icl>object) @singular) are the two UWs. Now agt is the relation which connects these two UW, hence it’s a relation.

An UNL expression for this English sentence is stated in (2.2):

{unl}

agt(eats(icl>event).@entry.@present)

Girl (icl>person).@singular)

obj(eats(icl>event).@entry.@present)

Pineapple ((icl>object).@singular)

{/unl} …(2.2)

Table 2.2: List of Universal relations with their tags[1]

Tag	**Relation**	**Definition**	**Example**
agt	Agent	A process or participant provoking change in state	Rahul killed Jim. agt(kill, Rahul)
ant	Concession or opposition	For indictaing that two entries are not sharing same meaning	Sumit is not Harry. ant(Sumit, Hary)
cnt	Content	It is the object in a experimental state	Sim has two brothers. cnt(has, two brothers)
con	Condition	An event’s ondition	If it rains, crops die. con(crops die, it rains)
dur	Duration	An event’s duration	She sang for two hours. dur(sang, two hours)
equ	Synonym	For indicating that two entries share same meaning	The evening star is the morning star. equ(morning star, evening star)
ben	Beneficiary	Entity advantaged by some event	She gave the book to me. ben(gave, me)
aoj	Object of the attribute.	Subject of the verb.	Book contains pictures. aoj(conatins, book)

Tag	Relation	Definition	Example
and	Conjunction	Used to express conjunction between the entries.	Ram and Sita. and(Ram, Sita)
gol	Final state ,destination	Receipient of the entity	I gave her the book. gol(gave, her)
icl	Hyponomy,a kind of	For referring the subclass of a particular class	Cats are mammals. icl(mammals, cats)
lpl	Logical place	Non-physical place	She is in love. lpl(She, love)
man	Manner	Way of carryong out of the event	She ran very quickly. man(ran, quickly)
iof	Instance of	Refer to an individual or instance of the class	She is a girl. iof(girl, She)
fld	Field	Semantic domain bearing the entity	Meaning(language) fld(meaning, language)
exp	Experiencer	The one reveiving the sensory impression.	I saw her. exp(saw, I)
ins	Instrument	Agent to implement the action.	She cut the cake with knife. ins(cut, knife)
or	Disjunction	Acts as disjunt between the entities.	Book or pen. or(book, pen)
per	Proportion or distribution rate.	For indicating the measure	Thrice a week. per(thrice, week)
rsn	Result	A referent resulting form the entity	The mother cooked rice. rsn(cook, rice)
seq	Consequence	Used to express consequence	We think so I am . seq(We think, I am)
src	Initial origin,place	Initial placee of event	Nima came from Paris. src(came, Paris)
ptn	Partner	A non focussed participant	She lives with Ram. ptn(live, She)
pof	Is part of	Referring to the part of the whole	She is part of the family now. pof(family, She)
tim	Time	Temporal placement of event	She came today. tim(came, today)

Tag	Relation	Definition	Example
tmf	Initial time	Initial time of the event or entity	She danced since decade. tmf(danced, decade)
tmt	Final time	Final time of the event	She cooked untill late. tmt(cooked, late)
via	Intermediatory place	Intermediate state of the entity	She went home via bus. via(went, home)

2.1.3 Universal Attributes

The subjectivity of the sentence is described by the attributes. Subjectivity can be with respect to time(past, present, future), of aspect(begin, continue, complete), of reference(specific, non-specific), focus(emphasis, theme, title), attitude(confirmation, exclamation), feeling etc. They are attached with the UW. In the sentence 'I can eat a pineapple', ability of eating is being referred so one of the attribute is @ability.

2.1.4 Semantic Network

Semantic Network or Universal Expression can be assumed to be a graph. The Universal nodes form the nodes of such a graph. The relation between the UW are the edges of this graph. Figure 2.1 shows the semantic graph for the english sentence, "Tyrioin Lanister killed Tywin Lannister yesterday with an arrow in the house because of Shae" .

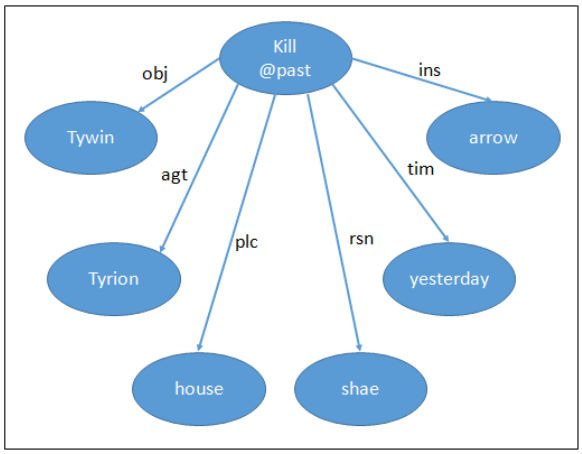

Figure 2.1 Semantic network of English sentence

In the semantic graph shown in Figure 2.1, "Tywin", "kill", "Tyrion", "yesterday", "Arrow", "House" and "Shae" are Universal Words. Next, "agt" (agent), "obj"

(patient), "tim" (time), "ins" (instrument), "plc" (place) and "rsn" (reason) are Universal Relations and "@past", "@def" and "@indef" are Universal Attributes.

2.2 Commitments of the UNL

The UNL programmes main aim is to construct the UNL, an artificial language that can be used to process information across various language. The major commitments of the UNL are stated next.

i. The UNL must represent information

   The UNL is first and foremost a knowledge representation language. The most important corollary of this first commitment is that UNL is not a meta-language as it is not intended to describe or represent natural languages. On the contrary, it is used to represent the information conveyed by natural languages.

ii. The UNL must be self-sufficient

   The UNL must be self-sufficient and must not depend on any implicit knowledge, and should explicitly codify all information. UNLization must be completely seprate from the NLization, and vice-versa.

iii. The UNL must be general-purpose

   UNL must be able to be used for various application and not confined to a particular application.

iv. The UNL must be independent from any particular natural language

   The UNL being the language of United Nations has to assure that it is not confined to any particular language.

2.3 UNLization and NLization

Along with the various applications of UNL like knowledge representation and knowledge management UNLization and NLization are attracting the researchers towards UNL. Many of the current researches are on this module. UNLization means to convert from any native language to UNL. While NLization is to convert a UNL to a native language. Primarily researches have used three approaches to achieve UNLization and NLization**.**

i. Using common EnConvertor and DeConvertor tools provided by the UNL center.

ii. Integration of UNL into pre-existing Machine Translations.

iii. Creating new architecture from scratch.

IAN for UNLization: Interactive Analyser(IAN) a UNLization tool is a Java based web application which takes a natural language as input and delivers the language independent module UNL form. In the system provided here the syntactic processing is done automatically through the grammar and dictionary rules provided for the specific natural language(NL). Figure 2.2 shows the mechanism for the processing of the NL sentence using the IAN module.

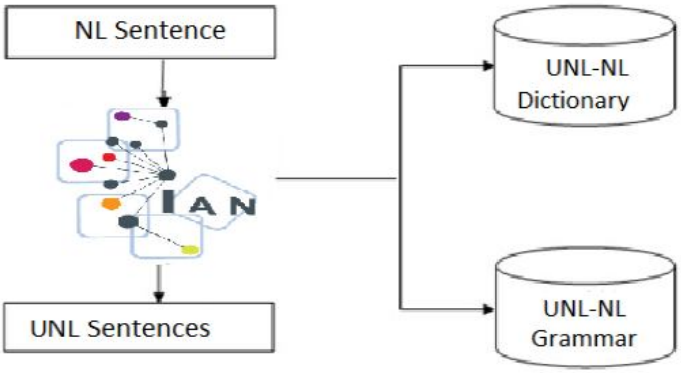

Figure 2.2 Mechanism for the processing of the NL sentence [2]

EUGENE for NLization: An online tool which was developed by the UNDL organisation like IAN and released in 2012 [2], it provides a fully automatic system that takes the UNL as the input and outputs the natural language that is desired without any human intervention. Here, corresponding to the natural language that is desired the dictionary and the grammar rules are supplied in form of separate interpretable files[3]. Figure 2.3 shows the mechanism for the processing of the UNL sentence using the EUGENE module.

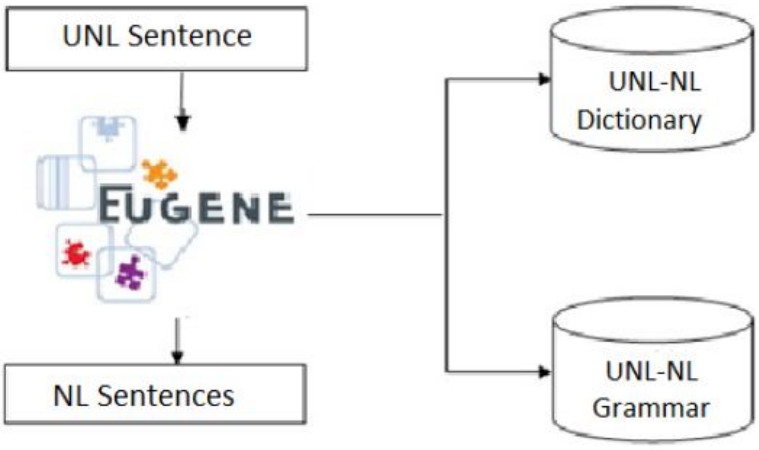

Figure 2.3 Mechanism for the processing of the NL sentence [2]

2.4 Tools, Methodolgies and material for UNL

A lot of manuals, books, papers, slides and examples can be accessed from the official website of UNL: http://www.undl.org. Apart from these UNL Data in chronological order can be obtained in the form of UNL Documents. These constitute various UNL converted from natural language catering to topics like Biographies, Wikipedia Articles, UNESCO Documents etc. These can be used for various research related activities. Next discussed are the tools built for the UNL system.

UNL Editor

It is a UNL authoring tool which provides the the graphic interface to allow the users to manipulate high-level graphs instead of traditional low-level UNL statements. However the language specialists are required to upload the texts which are meant to be analysed, thereby selecting the corresponding UWs (the nodes in the graph); creating semantic relations between nodes; and assigning attributes to nodes.This whole process leads to production of results which help in the decision making process .

UNLweb

It is a web portal that is created and maintained by the UNDL foundation to facilitate the research carried in the field of UNL development and reduce the language barriers that come in the way. Being a property of United Nations it serves as an asset to mankind to bring together all the developers and researchers onto one integrated platform.

It consists of five basic systems that help in this process described as following.

- The **UNL**$^{\textbf{arium,}}$ which provides development environment to create and edit language resources.
- The **UNL**$^{\textbf{dev}}$, which is meant to facilitate the applications and computer programs related to UNL framework.
- The **UNL**$^{\textbf{wiki}}$, providing the necessary documentation and instructions for carrying out UNL related activities.
- The **UNL**$^{\textbf{forum}}$, a social portal to discuss the issues relating to UNL framework and spread the information.

- **VALERIE**, a virtual learning environment for the purpose of teaching and training the UNL developers and researchers[2].

Material for UNL

Material in the form of manuals, books, papers, slides and examples can be accessed from official website of UNL : http://www.undl.org. Apart from these the data can be obtained from in the form of documents in the chronological order of when they were made.It constitutes various converted natural language to UNL form for topics like biographies,Wikipedia articles,UNESCO documents which can be used for various research related activites.

2.5 Applications of UNL

Various researchers have tried their hands on catering different natural language processing problems using the UNL. Table 2.3 states the work of such researchers ranging applications from text translation, text summarization to clustering of data and so on.

Table 2.3: Applications of UNL

Authors	Approach /Description	Corpus used	Accuracy
Application : Machine Translation			
[6] Etienne Blanc (2005)	Representation of the text meaning as a Dependency Tree, which is further processed into an equivalent UNL graph.	NA	NA
Application : Language Independent question answering system			
Cardeñosa *et al.* (2005)	Cross Lingual information extraction.	UNL coded document base	82.6%
Mukeree *et al.*	Structure matching approach.	English based corpus built from documents obtained from the official websites of EPA and WHO	NA
Application : Text Translation			

Pandian *et al.*	SVM and seed term based Search method.	Biographies of persons from various positions and designations the seed information categories taken as Corpus	Average F-Score of or proposed method is 0.7597
Application : Text Clustering using Semantics			
[7] Choudhary and Bhattacharyya (2002)	Semantic relations capturing and Neural network based technique for clustering	NA	Accuracy of 92.3%
Application : Text Summarization			
[4] Martins *et al.* (2002)	Heuristics based pruning of sentences	ONU and Booklet	Average compression rate of 44.58% and 40.9% respectively for corpus ONU and Booklet
[5] Sherry and Bhatia (2015)	Frequency value based heuristics	UNL corpus “Hare and Tortoise” provided by UNDL	Summary document is 73.38 % of original
[8] Mangairkarasi and Gunasundari (2012)	Heuristic based identification of conceptual relations	NA	97% from human generated summary
[9] Martins and Rino (1997)	Heuristic based sentence pruning	Bureaucratic texts encoded by diverse UNL teams and the THESES Corpus	Achieving compression rates of 44.58% and 40.9% respectively.
Application : Semantic Textual Entailment Recognition			
[10] Pakray *et al.* (2011)	Implementation of matching module on identified UNL relations of text and hypothesis	RTE-3 test annotated set as a development set which includes 800 text-hypothesis pairs	Overall 60.3% precision
Singh et al (2012)	Graph matching strategy	The UNL system available online at: http://www.cfilt.iitb.ac.in /UNL enco	For MSRVid the results obtained were 0.5504

Application : Semantic Textual Similarity			
[11] Nguyen and Ishizuka (2006)	Application of statistical techniques on several lexical and syntactic features.	Dataset supplied by UNDL organization	79% accuracy
[12] Iyer and Bhattacharyya (2005)	Use of wordnet and Semantic information to improve Case retrieval.	NA	NA
Application : Sentiment Analysis of Natural Text			
[13] Rani and Kumar (2014)	Entity identification by UNL UW's and relations. Extraction of sentiment using the attribute information	UC-A1 Corpus given by UNDL foundation	74% accuracy
Application : Multilingual Cross-Domain Client Application Prototype for UNL-ization and NL-ization			
[23] Agarwal and Kumar (2016)	UNL based cross-domain client application	UC-A1 Corpus given by UNDL foundation	100% accuracy, F-Measure of 0.95 on scale(0-1) for UNL-ization and NL-ization modules

Chapter 3
Literature Survey

This chapter gives a brief idea of the Question Answering systems in place currently. This represents the famous existing QA systems related to the thesis. It puts the described multilingual QA model in viewpoint of the field.

3.1 Existing QA systems

The activity in the question answering systems has been due to the combination of the demand of user and the promising results. Many different approaches have been used to construct the query processing modules. The various models discussed here are the corpus based, web based or be it the commercial ventures. The need of the multilingual systems was very much there because none of the systems built so far has catered to the language independent mechanism.

3.2 Corpus Based QA systems

This represents the traditional approach to solve the QA system problems. All of these systems are based on a prepared corpus which is usually in a single language. This has been used quite extensively. The coming section describes the prominent QA systems that are primarily based on the role of the corpus.

3.2.1 Log Answer

Log answer is a open domain QA system. This answers a Natural language question after referring to the knowledge base it has acquired. This project combines NLP, MultiNet knowledge base, based on snapshot of wikipedia and deduction. MultiNet knowledge base consists of tools moslty in German language. It consists of 12 million sentences. This also includes 12,000 background knowledge axioms. The query inputed to the system is first parsed and then converted to the First Order Logic sentences. Then the knowledge base is hit using information retrieval and machine learning techniques[14]. Figure 3.1 shows the system flow of Log Answer system.

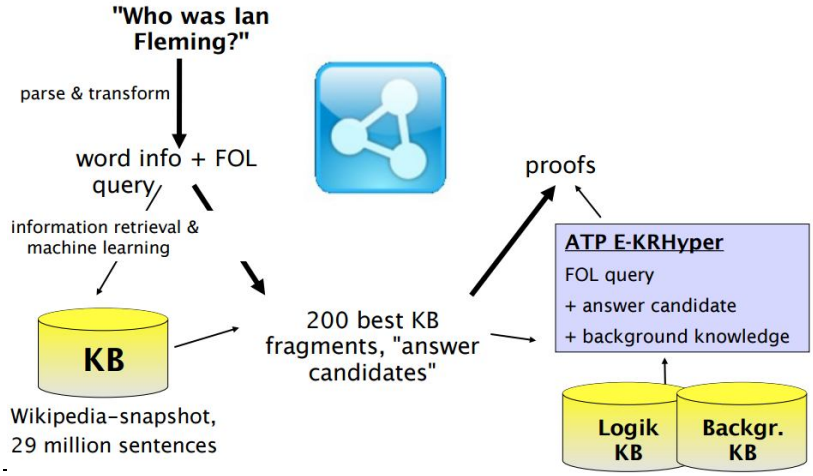

Figure 3.1 System flow of Log Answer system[14]

3.2.2 Text Retrieval Conference (TREC)

TREC is based on the traditional approach of matching the query with a user defined corpus. It provides answer to a Natural Language query based on a document corpus. The first stage involves indexing of document wherein they are cleaned from a markup format and their raw contents is obtained. Then these raw contents is split into paragraph which is then passed to managing gigabytes search engine. The second stage involves question processing. Here pearl modules which uses Extensive Markup Language (XML) comes into picture. They are executed as a large pipeline. Figure3.2 shows a structure of data involved and the architecture of the TREC system. The next module is the sentence splitter and tokenizer. This consists of two modules. The individual sentence are marked up using a set of heuristics by the first module. The next module Link parser is used to annotate the structure of the question. The focus of the question that is whether the question is of who, what or how type is determined first. Then based on this answer type is determined.Then these are followed by process of keyword extraction, paragraph retrieval, candidate answer extraction Answer weighing and Answer scoring[15].

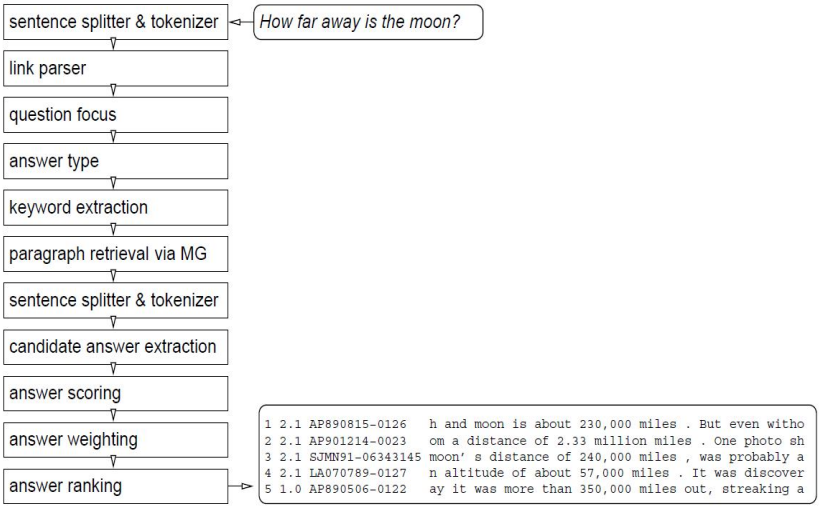

Figure 3.2 TREC system architecture diagram [15]

3.2.3 Pisa Question Answering System(PiQASso)

The Pisa Question Answering System is a semantic filter and modern information retrieval techniques with a combination to select a paragraph containing a justifiable answer. A dependency based parser, a Part of Speech(POS) tagger, a lexical database and a tagger forms the Semantic filtering.

PiQASso is formed of two major components namely a paragraph indexing , retrieval subsystem and a QA system. To process a question first the question is analyzed. This involves parsing the question, the identification of answer type and extraction of relevant words to formulate the paragraph retrieval. Then the query is formulated and paragraph search is done. Then the answer type and relation matching filter are applied. The query built is a target of high precision for retrieval of small number of sentences to be evaluated as candidate answers[16]. PiQASso is formed by combination of several libraries into a single process, thus is a complete vertical system. Figure 3.3. describes the architecture of the PiQASso system.

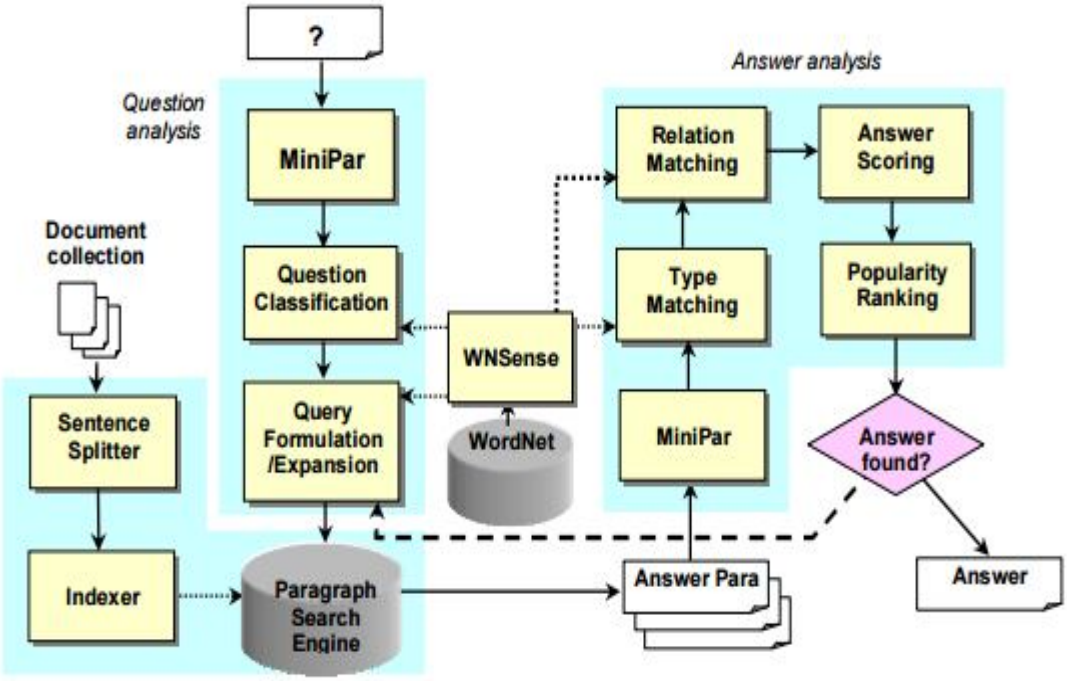

Figure 3.3 Architecture of PIQASSO system [16]

3.3 Web Based

Most of the attempts made at making QA systems are based on a predefined corpus or knowledge base. If a single source of knowledge like wordnet is used, it could act as a force multiplier and improve performance. If now information is clubbed from several sources then this is far better.This conceptual concept representation and relation with domain is known as Ontology.

3.3.1 ASKMSR

The traditional approach of using as self-made corpus was challenged by the researches at Microsoft and they devised a system which shall use the gigantic information available at web to cater to the question posed at them. They named their system ASKMSR**.** This system uses the vast information available at web instead of preparing a corpus. Thus they have a virtual corpus of billions of pages of electronic text. Their system encourages redundancy and fetch multiple answers of the same question and then determine the best among them. For the sentence 'Nathu Ram Godse altered the history with a bullet. Thus the Mahatma breathed his last' corpus would not be able to answer the question who killed Gandhi accuretly. However if net

is searched then system might get a closer answer. The Figure 3.4 represents the system architecture of the ASKMSR system[17].

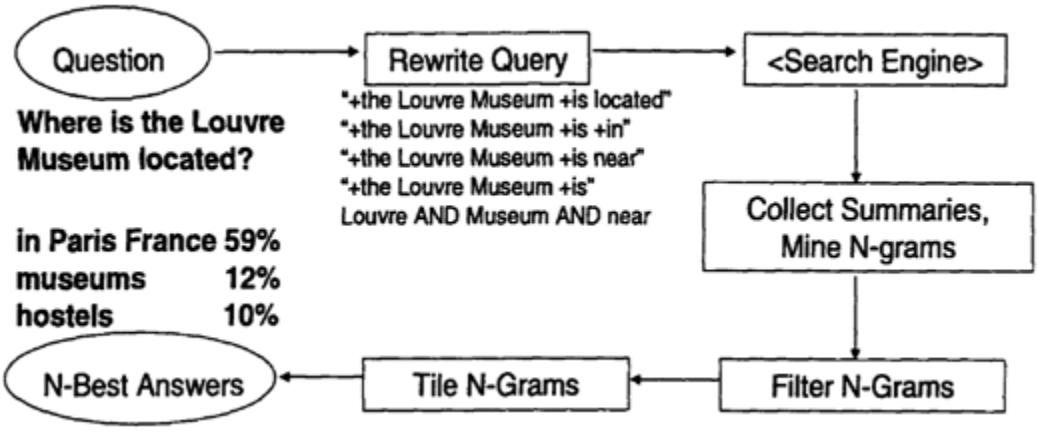

Figure 3.4 ASKMS system architecture diagram [17]

3.3.2 QASYO

QASYO also known as Question Answering System for Yet Another Great Ontology(YAGO) is another attempt using web based technologies. This system uses YAGO Ontology. Yago Ontology is an integration of conceptual hierarchies of WordNet dictionaries and Wikipedia. It achieves an accuracy of 95% after uses both. Currently there are 1.7 million entities and 15 million facts.The input to QASYO consists of natural language processed queries, which it handles using powerful techniques like WordNet by mapping to a semantic markup. WordNet is used to make sense of NL queries with respect to target knowledge base[18]. Figure 3.5 shows the QASYO system architecture.

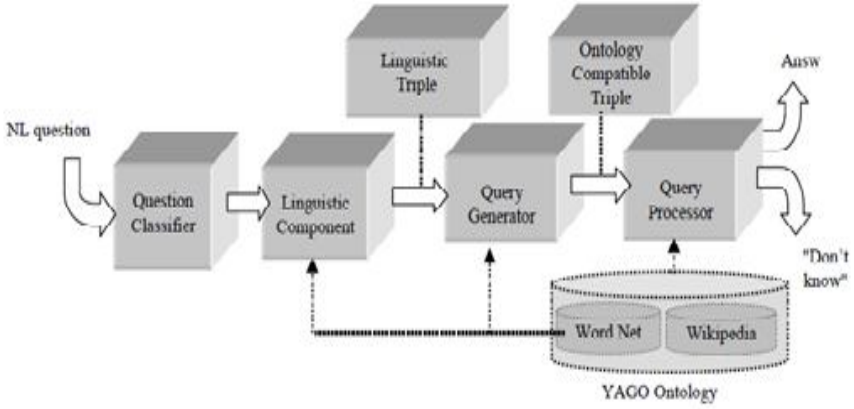

Figure 3.5 QASYO system architecture diagram [18]

QASYO has four phases namely question classifier, linguistic component, query generator and query processor. The QA process is divided into two phases namely question analysis and answer retrieval. This is basically converting a natural language query to a YAGO compatible form. Classification and parsing the question are two of the most phases in this section. In answer retrieval we actually get the answer.The natural language query is mapped to subject, relation and object also known tighter as the linguistic triplet using linguistic component. Figure 3.6 represents the QASYO parse tree.

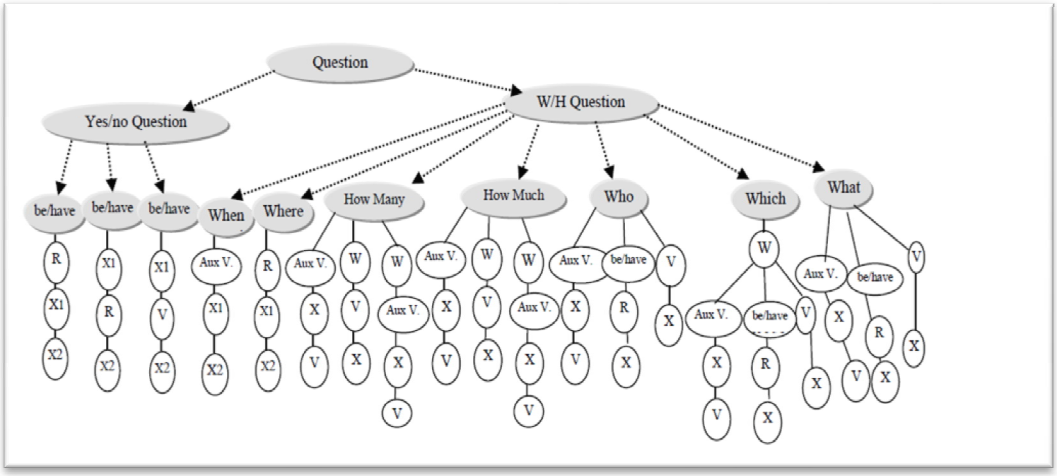

Figure 3.6 QASYO Parse Tree [18]

3.4 Commercial Ventures

There are been many commercial ventures into the question answering system domain and all major players are competing to get the lions share. According to reports in the era where the smart phones are like virtual assistants, shall determine the success of the operating systems. So while Apple's champion into the doomain of question answering systems is SIRI, google flagship is Google Now and Microsoft is represented by its banner man Watson.

3.4.1 Apple Siri

An abbreviation for Speech Interpretation and Recognition Interface is the apple attempts to win the battle of operating systems. SIRI works as a virtual digital assistant. SIRI is a voice based system and its unique selling property is providing acting as a actual assistant by speaking real time [19].

3.4.2 Google Now

Google Now is an virtual assistant developed by Google. Like Siri Google Now uses a NLP to do various tasks for user. Information proactively predicted by Google Now is also delivered to users. Google now however lacks the voice.

3.4.3 IBM Watson

IBM Watson is the flag bearer of Microsoft in the virtual assistant arena. Like Siri and google Now Watson uses natural language processing to handle its day to day activities.

3.5 UNL Based system

A new domain area catering to the existing question answering systems would be the one build using the UNL as the intermediatory language. Here briefy described are some of information retrieval systems that have been built using the UNL as the platform. This can greatly help in constructing the UNL base question answering systems as well.

3.5.1 Agro-Explorer

Agro-Explorer uses a UNL based corpus as opposed to a normal text based corpus. The usage of UNL makes this system language independent.

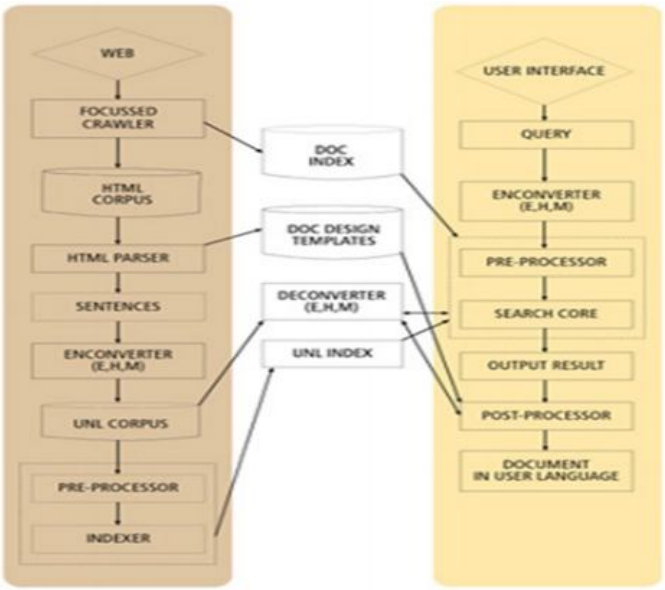

Figure 3.7 AGRO-Explorer system architecture diagram [20]

Agro-Explorer before performing the search first extracts its meaning and then used this information to optimize the search. Agro–Explorer has a dedicated focused crawler which scans the web and collects all pages related to agriculture and creates a HTML corpus.The EnConverter process the raw text extracted from this HTML and converts it to UNL. The indexer module receives this processed UNL and creates an inverted index out of it. The query entered by the user is first converted to UNL form by the encoder. The inverted index created earlier is used in the search module to which the expression is passed next. The DeConverter module then acts upon the UNL output of search module and converts it to a form that user has selected [20]. Figure 3.7 depicts a system architecture diagram of an Agro-Explorer.

Chapter 4
Problem Statement

The aim of the thesis is to design a QA system which is capable of handling multiple languages using a single corpus. Traditional approaches to design QA system have a major limitation as they are confined to catering the user queries only in a particular language. This greatly limits the usage of a QA system and also the reach of the product.

In the cases where such a system is to be made useful there is a need to make a replica of the huge corpus in that particular language. This shall be a resource wasting mechanism and also greatly increase the effort of overall phenomenon of building a question answering system.

The simple solution requires making the corpus language independent. The thesis uses UNL to achieve this language independence platform.

The thesis proposes a system which has a UNL based corpus. The system has an UNLizer and a NLizer for each language that it has to process. Thus the proposed system is completely language independent. The proposed system will enable the user to query a database that will be in intermediatory format rather than a specific language and after processing it the final answer will be generated in the desired language as demanded by the user.

Chapter 5
Objective and Methodology

There exists the concern of using multiple languages for the pupose of searching. This facilitates the user to search in any of the already known natural language where the result may complicate thus, pertaining into the higher complex issues. It has been discussed already that using the universal networking language (UNL) has reduced the complexity in the multiple languages that are involved. So far the optimum way out for performing the tasks related to interpretation of the results have been discussed. For including the context of cross-lingual methodolgy, UNL has been found as an optimum option over the others. The approach where the user is allowed to give the queries in any of the possible natural languages, there cross-lingual systems can play a crucial role in a sense that they are very different from multi-lingual systems. After this the system will be allowed to print results in a specific language that has been selected by the user instead of having the output in just one particular language.

5.1 Objectives

To achive the goal of constructing the language independent question answering system following objectives were proposed to be carried out.

i) To perform a systematic review on the existing question answering systems and analyse their techniques.

ii) To build a farmework for the UNL based multilingual question answering system.

iii) To develop the search module to identify the type of questions that are to be handled by the proposed system.

iv) To develop a subgraph matching module which matches the pattern of the query subgraph with the graphs of the UNL coded sentences.

v) To generate the UNL of the one word and the composite answers for the Eugene module for the target language.

5.2 Methodology

For achieving the objectives discussed in section 5.1, next discussed is the methodology that was used to implement the proposed system.

i. To conduct the detailed survey of the existing qusetion answering systems, various QA systems were studied that included the traditional approaches lie corpus based and web based. Also mentioned are the latest commercial techniques and the cross lingual systems.

ii. To build the framework for the UNL based QA system all of the components have been constrcuted namely the search engine, pattern matcher, associated question identifier and finally the output generator.

iii. The search module has been designed specifically to locate the answer to the relevant query addressing the type of questions that were identified for the system on the EOLSS corpus[21].

iv. The subgraph matching module has been developed which uses the concept of generating the triplets and also the dictionary based format for even storing the same content.This content was mainly meant to depict the UNL based information related to the universal words and the universal relations. The constructed algorithm works on these intermediate format and generates the output for the same.

v. The NLizer model has been proposed to convert the system generated UNL output answers to the desired natural language. Here the requirements for completion of this phase would be the grammar rules and the dictionary supported knowledge pertaining to that specific language. Here for the simplicity of the system we have given the final answers in the English language.

Chapter 6
Architecture and working of proposed model

The basic modules of the proposed system are described in this chapter of system architecture. This chapter discusses the working and role of the components in solving the multilingual Question answering problem. Figure 6.1 illustrates the system architecture. With the aim of implementing the model it is assumed that the manually imposed questions have their answers in the provided UNL coded document. The proposed model is implemented using the Python framework.

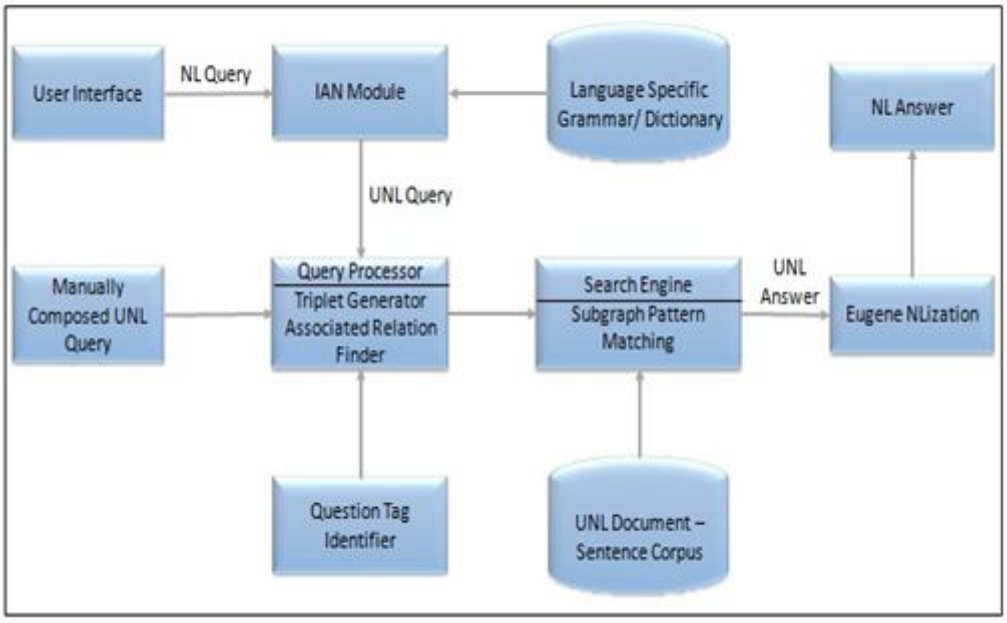

Figure 6.1 System Architecture of proposed model

6.1 UNL-Document Sentence Corpus

The information that has been considered worthy of performing question answering is taken in the form of UNL document. For the purpose of this problem various articles present at the Encyclopedia of Life Support Systems (EOLSS) [21] are taken for query processing. Each UNL sentence is presented in the format given by (6.1). This is basically the collection of the UNL coded sentences that were used as a repository to match the possible answers to the imposed query by the user.

-The formal sentence which is originally present is enclosed in the {org} and the {/org} tags.

-The converted UNL form of the original text is enclosed in the {unl} and {/unl} tags in the page. …(6.1)

6.2 User Interface

Here the file based module is used to enter the basic input to our system that is the query in the natural language.The query as well as the input sentences are fed into the system in th UNL coded format.

6.3 IAN Module

This component basically converts the natural language query into the UNL format query which will be further processed by the system components.

6.4 Query Processor

The user query that is being entered by the user in a specific natural language is fed to the EnConvertor, the IAN module and converted to the desired respective UNL format. In the process of conversion there are multiple pre-processing steps that are undertaken namely removal of unwanted space and delimiters, attributes and replacing the question tag with the possible UNL relations. Figure 6.2 gives the description of the internal query processor that works internally.

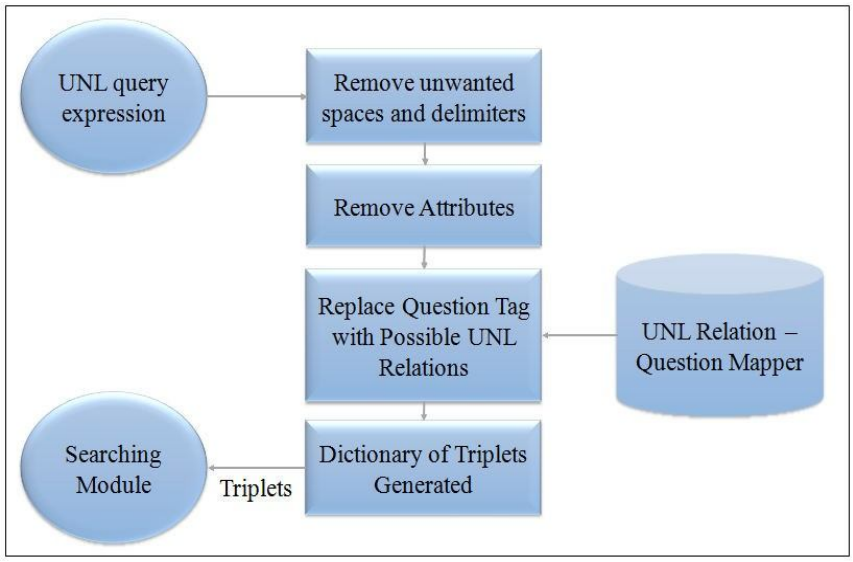

Figure 6.2. Working of Query Processor

A sample English sentence given by (6.1) from the UNL corpus and its corresponding UNL format is shown in Table 6.1.

Natural Language(English) sentence : In 2000 she was elected to the Indian judiciary

however she resigned in 2005 to become president of provincial state. ...(6.1)

Table 6.1**:** UNL of document sentence(6.1)

and (resign(agt>thing):1D.@entry.@past,elect(agt>thing,ob j>thing):0H.@past) agt(resign(agt>thing):1D.@entry.@past,she:0A.@topic) man(resign(agt>thing):1D.@entry.@past,however(icl>in spite of this):10) pur (resign(agt>thing):1D.@entry.@past,become(icl>start to be(gol>thing,obj>thing)):1X) tim (resign(agt>thing):1D.@entry.@past,2005:1P) gol(become(icl>start to be(gol>thing,obj>thing)):1X,president(icl>person):24) tim(elect(agt>thing,obj>thing):0H.@past,2000:05) mod(president(icl>person):24,state(icl>region):2M) mod (state(icl>region):2M, provincial(iof>state):2D) obj (elect(agt>thing,obj>thing):0H.@past,she:0A.@topic) mod(judiciary(icl>council):0Z.@def,Indian(equ>India):0W) gol(elect(agt>thing,obj>thing):0H.@past,judiciary(icl>counci l):0Z.@def)

Now consider a query stated in (6.2) that is imposed on this corpus. The UNL corresponding to the query is represented in Table 6.2.

Natural Language Query: When was she elected to the Indian judiciary? ...(6.2)

Table 6.2: UNL of query sentence(6.2)

obj (elect(agt>thing,obj>thing):0H.@past,she:0A.@topic) *tim*|*tmf*|*tmt*(elect(agt>thing,obj>thing):0H.@past, X) gol(elect(agt>thing,obj>thing):0H.@past,judiciary(icl>council):0Z.@def) mod(judiciary(icl>council):0Z.@def,Indian(equ>Indian):0W)

Now first the identified query word is mapped to the possible universal relations using the information in Table 6.3. In this example “When” is mapped to three relations tim, tmf, tmt as shown here.

Table 6.3:Mapping of Universal relations with query words

Question Type Keyword	**Possible main relation Tags**	**Relation**	**Definition**
What	aoj	object of the attribute	subject of the verb
	cnt	content of attribute	object of the verb

	obj	patient	participant who went through change of location/state
Where	src	source,initial state	initial state of entity /event
	plc	place	location of event
Who	agt	agent	participant in action
	aoj	object of the attribute	subject of the verb
	nam	name	name of entity identified
Why	pur	purpose	purpose of event
	rsn	reason	reason of event
When	tmt	final time	final time of event/entity
	tmf	initial time	initial time of event/entity
	tim	time	temporal nature of event
How	man	manner	how action took place

a) UNL Triplet Generator

Here the input UNL form of the query and the UNL document are taken and triplet form of the fed UNL is generated. The expresion (6.2) gives the general format being used to represent the triplet .Table 6.4 gives the UNL sentence triplets for sentence (6.1) and Table 6.5 gives the UNL query triplets for query (6.2).

Triplet Format: <Relation,Universalword1,Universalword2> ...(6.2)

Table 6.4 UNL sentence triplets for sentence (6.1)

UNL sentence	**UNL sentence triplets**
and (resign@entry.@past, elect.@past)	<and, resign, elect>
man(resign.@entry.@past, however)	<man, resign, however>
agt(resign.@entry.@past, she@topic)	<agt, resign, she>
pur (resign@entry.@past, become)	<pur, resign, become>
tim (resign.@entry.@past, 2005)	<tim, resign, 2005>
gol(become, president)	<gol, become, president>

mod(president, state)	<mod, president, city>
mod (state, provincial)	<mod, state, provincial>
tim(elect.@past, 2000)	<tim, elect, 2000>
obj (elect.@past, she.@topic)	<obj, elect, she>
gol(elect.@past, judiciary.@def)	<gol, elect, judiciary>
mod(judiciary.@def, Indian)	<mod, judiciary, Indian>

Table 6.5 UNL query triplets for query (6.2)

UNL query	**UNL query triplets**
obj (elect.@past, she.@topic)	<obj, elect, she>
tmf (elect.@past, X)	<tmf, elect, X>
tim (elect.@past, X)	<tim, elect, X>
tmt(elect.@past, X)	<tmt, elect, X>
gol(elect.@past, judiciary.@def)	<gol, elect, judiciary>
mod(judiciary.@def, Indian)	<mod, judiciary, Indian>

b) Associated Relation Finder

Specific question tag is identified and the possible associated main UNL relations are mapped onto query word. The mapping is done with help of Question tag identifier which contains a predefined set of relations and query words as shown in Table 6.3. For instance the query word ‘where’ would be mapped to two universal relations called source or the initial state (src) and place (plc) in the context.

6.5 Search Engine

Generated query triplet and associated universal relation will be be further processed to reach to the final answer with the pattern matching approach .The triplet query will be searched against the UNL coded corpus to locate the best possible match.

a) Subgraph Pattern Matching Module

According to the system the answer to the query can only be found if the present UNL query is the subgraph of any UNL sentence present in the document. So ,if the match is found then the answer exists. In case the query subgraph matches with more than one graphs in the document then all the answers are fetched as per relevance in the statement.

The dictionary data structure is used to represent the UNL sentence given in (6.1) and query given in (6.2). Here, the keys in the dictionary are used to depict all the possible universal words in the UNL sentence. In the sentence given by (6.1) the universal words identified are 'resign', 'she', 'judiciary', 'however', '2005', 'president', 'become', 'Provincial', 'state', 'elect', '2000',' Indian'. These all have been made the keys and their values will be the subdictionaries which constitute all the relations attached to them as keys and the corresponding universal words as their values. Taking the case of universal word 'resign', the value subdictionary corresponding to it has 'agt' as the key and 'she' as its value. Similarly it has four more such key-value pairs like 'and': 'elect', 'tim': '2005', 'man': 'however' and 'pur': 'become' This format is further depicted in Table 6.6.

Key : 'resign(Universal Word 1) ,

Value: SubDictionary1

Table 6.6 SubDictionary1 mappping for the universal word 'resign'

Key	agt	and	tim	pur	man
Value	she	elect	2005	become	however

Sentence Graph Representation for (6.1) in the Key:Value representation is :

{

'she' :{},

'judiciary' :{'mod': 'Indian'},

'however' : {},

'2005' : {},

'president' :{'mod': 'state'},

'become' : {'gol': 'president'},

'Provincial ': {},

'state' :{'mod': provincial},

'resign' : {'agt': 'she', 'and': 'elect', 'tim': '2005', 'man': 'however', 'pur':

'become'},

'elect' : {'obj': 'she', 'tim':'2000', 'gol': 'judiciary'},

'2000' : {},

'Indian' : {}

}

Query Graph Representation for (6.2):

{

'she' : {},

'judiciary' : {'mod': 'Indian'},

'elect' : {'obj': 'she', 'tim|tmt|tmf': ?,'gol': 'judiciary'},

'Indian' : {}

}

Now here the query subgraph as shown in Figure.6.4 is matched against the UNL sentence graphs in the document. The graph which gives the most appropriate match will finally result into giving the exact answer to the query. Figure 6.3. shows the UNL document graph against which the query will be matched.and Figure 6.4 shows a sample query subgraph. The dotted portion in the Figure 6.3 shows the matched subgraph inside the main graph.

6.6 Answer Generation

Finally the answer is searched upon in the subgraph that has been the exact match as in the former case answer would be: '2000'. This is done by looking for the possible identified relations as tmt, tmf, tim in the given example .

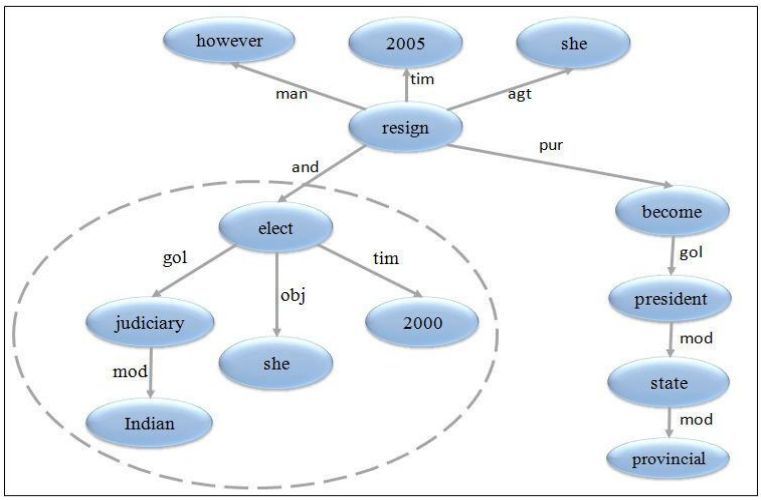

Figure 6.3 Graph representation for a UNL sentence given in (6.2)

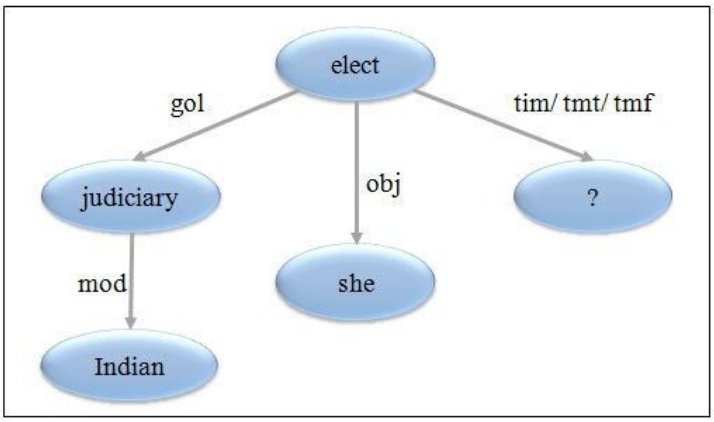

Figure 6.4 Subgraph representation for query given in (6.2).

The correct match is found by detecting the missing node corresponding to the tim relation and thus mapping to the final answer given by the value in subdictionary of the universal word '*elect*' .Now the sub- expression forming answer can be thus given in the target language using the EUGENE or using the ones already been developed for many languages. The algorithm for answer generation is given in Table 6.7.

Table 6.7: Algorithm for answer generation.

Algorithm 6.1

```
dictionariesList<-List of UNL sentence dictionary objects: Dictionary Objects Format:
 UW1:['relation1':'UW2']
 UW2:['relation2':'UW3']
. …..
UWn:['relation.n':'UWn']
otherKeys=queryTriplet.keys()-{'relation','?'}
 subgraph matching :
For each dictionary in dictionariesList: keys=dictionary.keys();
if otherKeys has keys:
        mainDictionary=dictionary findSubGraph(relation)
        break
findSubGraph(subDictionaryRelation):
if mainDictionary[subDictionaryRelation] is empty
        return
else
        value=mainDictionary[subDictionaryRelation]
        universalWords=value.split(":") answerString=
        answerString+subDictionaryRelation(universalWords[0],universalWords[1])
        findSubGraph(universalWords[0])
```

The algorithm for answer generation explains that process of extracting the UW1,UW2 and the relation. It then depicts the process by which the subgraph matching is performed given the sub dictionary for the relations. Finally the answer string is generated using this procedure.

Chapter 7

Results and Discussion

The proposed system is implemented and tested on the datasets “Water supply for agriculture” and “Composition and structure of the atmosphere” belonging to the EOLSS repository [21].The former dataset consisted of total of 495 UNL based sentences and the latter had 603 UNL based sentences. A total of 65 manually constructed questions were imposed on these datasets articles. The test set was composed of factoid (What, When, Where, Why etc.) type of identified questions out of which majority returned precise answers. Some of these are given by (7.1).The output UNL was compared against the actual answers .On testing the system on such questions an accuracy of nearly 82% was achieved as 53 questions were correctly answered out of 65. The rest of the 12 questions that were incorrectly answered were either the ones out of scope like calculation based or the ones to which answers did not exist in the document scope. Table 7.1. gives the detailed observations and findings. For now UNL of English sentences has been tested but later the system can be implemented on other languages as well once their UNL database is available.

The corpus, The Encyclopaedia of Life Support Systems (EOLSS), having 12,916 sentences is divided among 13 documents. Each document contains one or more than one chapter. The sentence is relevant to the query if the query search graph of UNL of query matches to the UNL of the sentence in that document.

Sample Questions:

-What does natural water resources represent?

-How much water consumption on Earth during twentieth century increased to?

-What is the greatest world water consumer?

-Where are complex multifactor water-nutrition systems created?

-When did studies of the atmosphere originated? …(7.1)

Table 7.1: Observations and results of Implementation

Total number of UNL sentences taken for testing	72
Total number of questions framed	65
Questions with one word answers	50
Questions with composite answers	15
Number of answers matched with manually constructed output	53
Number of answers mismatched with actual output	12
Accuracy achieved	82%

7.1 Type 1: One word answer type questions

The system is designed to handle the two type of answer based questions. One of them being the one word answer type questions. The next section describes a case to handle the same.

Question: What is the greatest world water consumer?

Identified question type: What

Possible associated universal relations mapped: aoj, cnt ,obj

UNL representation of the query:

aoj(great, consumer)
aoj|cnt(consumer,?)
aoj(world, consumer)
mod(consumer, water)
man(great, most)

Now after the corpus of the UNL coded sentences is searched for to match the relevant answer the following sentence is found to be the best possible match.

UNL of possible answer sentence matched:

{unl}
aoj(consumer.@entry.@def, agriculture.@topic)
aoj(great, consumer@entry.@def)
aoj(world, consumer.@entry.@def)
mod(consumer.@entry.@def, water)

man(great, most)

{/unl}

Triplet generated for the sentence:

<aoj, consumer, agriculture>

<aoj, great, consumer>

<aoj, world, consumer>

<mod, consumer, water>

<man, great, most>

After navigating through the UNL of the desired answer sentence the answer is located and found to be 'agriculture'. Table 7.2 shows some of the sample UNL queries that result into one-word type answers along with the UNL representation of the UNL coded sentences that are matched for locating the answers.

Answer Found: agriculture

Table 7.2: one-word type Question-answering observations

Question	**UNL query**	**UNL coded sentence of document matched**	**Answer**
Question Type:What Possible relations mapped: (aoj,cnt,obj)			
What is the greatest world water consumer ?	aoj(great, consumer) **aoj|cnt|obj** (consumer,?) aoj(world, consumer) mod(consumer, water) man(great, most)	{unl} aoj(consumer.@entry.@def, agriculture.@topic) aoj(great, consumer@entry.@def) aoj(world, consumer.@entry.@def) mod(consumer.@entry.@def, water) man(great, most) {/unl	agriculture
Question Type:How Possible relations mapped: (man)			
How did the water consumption on Earth during the twentieth century increased?	dur(increase.@entry.@past, while) obj(increase.@entry.@past, consumption.@topic) **man**(increase.@entry.@past, ?) plc(consumption.@topic, earth) obj(consumption.@topic, water) dur(earth,century) mod(century, 20)	{unl} dur(increase@entry.@past, while) obj(increase@entry.@past, consumption@topic) man(increase@entry.@past, sevenfold) plc(consumption.@topic, earth) obj(consumption.@topic, water) dur(earth, century) mod(century, 20) obj(while, increase@past) obj(increase.@past, population@topic.@def) man(increase.@past, time@pl) man(increase.@past, only) qua(time.@pl, 3) {/unl}	sevenfold
How important factor is the quality of water used in	aoj(factor@entry.@indef, quality@topic.@def) aoj(important, factor@entry.@indef) **man**(important, ?) scn(use, irrigation)	{unl} Aoj(factor@entry.@indef,quality@topic.@def) mod(factor@entry.@indef,safety) aoj(important@entry.@indef) man(important, very) man(safety, ecologically) obj(safety, system) mod(system,irrigation) mod(quality@topic.@def,water)	very

irrigation?		obj(use,water) scn(use,irrigation) {/unl}	
Question Type:Why Possible relations mapped: (rsn,pur)			
Why was temperature of water used?	mod(temperature.@topic, water) obj(use, water) pur|rsn(use, ?)	{unl} aoj(importance.@entry, temperature.@topic) aoj(as, importance.@entry) aoj(great, importance.@entry) obj(as, factor.@indef) mod(factor.@indef, physiological) mod(temperature.@topic, water) obj(use, water) pur(use, animal.@pl) {/unl}	animals
Question Type:Where Possible relations mapped: (src,plc)			
Where does iron bacteria grow?	obj(growth@topic, iron@pl) **plc|src**(iron @pl,)	{unl} aoj(result in.@entry, growth.@topic) obj(result in.@entry, formation) obj(formation, deposit.@pl) aoj(cause.@may, deposit.@pl) aoj(slimy, deposit.@pl) obj(cause.@may, system.@pl) pur(system.@pl, clog) mod(system.@pl, pipe) man(clog, as well as) obj(as well as, odor.@pl) aoj(unpleasant, odor.@pl) obj(growth.@topic,.@pl) plc(iron bacterium.@pl, pipe) {/unl}	pipe
Where are the standards for pollutant substances for watering of animals?	pur(standard@entry.@pl, substance) **plc|src**(substance.@pl, ?) aoj(pollutant,1 substance@pl) pur(water, watering) obj(watering, animal.@pl)	{unl} pur(standard@entry.@pl, substance@pl) plc(substance@pl, water) aoj(pollutant, substance@pl) pur(water, watering) obj(watering, animal.@pl) {/unl}	Water

7.2 Type 2 :Answer as a phrase

The next type of qusetions to which our system caters are the ones which have their answer as composite ie. The answer is not just one word but a collection of the words. Here in this case the answer returned is the UNL of the phrase located. The following instance gives the use case of this sceanrio.

Question : What does irrigated agriculture cause?

Identified question type: What

Possible associated universal relations mapped: aoj,cnt ,obj

UNL representation of the query:

obj(irrigate, agriculture)
aoj|cnt|obj(cause,?)
aoj(cause, agriculture)

Now after the corpus of the UNL coded sentences is searched for to match the relevant answer the following sentence is found to be the best possible match. Here the query is found out to be the subgraph of the graph generated by this sentence.

UNL of possible answer sentence matched:

aoj(cause@entry, agriculture@topic)
obj(cause@entry, formation)
obj(formation, system@indef)
aoj(complex, system@indef)
mod(system@indef, nutrition)
obj(nutrition, water)
obj(irrigate@past, agriculture@topic)

Triplet generated for the sentence:

<aoj, cause, agriculture>
<obj, cause, formation>
<obj, formation, system>
<aoj, complex, system>
<mod, system, nutrition>
<obj, nutrition, water>
<obj, irrigate, agriculture>

After navigating through the UNL of the desired answer sentence the answer is located and found to be composite answer which constitutes the five UNL phrases. This combined together formulates the final answer in the UNL format. Table 7.3 shows some of the sample UNL queries that result into composite type answers along with the UNL representation of the UNL coded sentences that are matched for locating the answers.

Answer Found:

UNL form:

obj(cause@entry, formation(icl>process):0V)
obj(formation, system@indef)
aoj(complex, system@indef)
mod(system@indef, nutrition)

obj(nutrition, water)

Natural Language Answer:

causes formation of a complex water-nutrition system

Table 7.3: Composite type Question-answering observations

Question	**UNL query**	**UNL coded sentence of document matched**	**Answer**
Question Type:What Possible relations mapped: (aoj,cnt,obj)			
What does Natural water resources represent?	agt(represent@entry, resource@topic.@pl) **aoj|cnt|obj(**represent@ entry, ?) mod(resource@topic.@pl, water) mod(water, natural)	{unl} agt(represent@entry, resource@topic.@pl) obj(represent@entry, component.@indef) cnt(component@indef, resource@indef) mod(component@indef, biosphere) mod(biosphere, vital) mod(resource@indef, economic) aoj(possess, resource@indef) aoj(industrial, resource@indef) obj(possess, property@pl) aoj(indispensable, property@pl) mod(property@pl, consumer) mod(resource@pl, water) mod(water, natural) {/unl}	obj(represent@entry, component.@indef) cnt(component@indef, resource@indef) mod(component@indef, biosphere) mod(biosphere, vital)
What does irrigated agriculture cause?	obj(irrigate, agriculture) **aoj|cnt|obj** (cause,?) aoj(cause, agriculture)	{unl} aoj(cause.@entry, agriculture.@topic) obj(cause.@entry, formation) obj(formation, system.@indef) aoj(complex, system.@indef) mod(system.@indef, nutrition) obj(nutrition, water) obj(irrigate.@past, agriculture.@topic) {/unl}	obj(cause.@entry, formation) obj(formation, system@indef) aoj(complex, system@indef) mod(system.@indef, nutrition) obj(nutrition, water)
Question Type:How Possible relations mapped: (man)			
How do the following principles form the basics of water quality standardization?	mod(standardization, quality) **man**(form@entry, ?) aoj(form@entry, principle@topic.@def.@pl) obj(form@entry, basics@def) mod(basics@def, standardization)	{unl} man(form@entry, in accord with) aoj(form.@entry, principle@topic.@def.@pl) obj(form@entry, basics@def) mod(basics@def, standardization) pur(standardization, use) mod(standardization, water) mod(standardization, quality) scn(use, irrigation) mod(principle@topic.@def.@pl, following {/unl}	man(form@entry, in accord with) obj(in accord with, this)
Question Type:Why Possible relations mapped: (rsn,pur)			
Why it is more difficult to provide delivery	aoj(difficult.@entry, provide) man(difficult.@entry, more) rsn|pur(provide,?) pur(provide, agriculture) obj(provide, delivery) obj(delivery, water) mod(water, quality)	{unl} aoj(difficult.@entry, provide) man(difficult.@entry, more) rsn(provide, because of) pur(provide, agriculture) obj(provide, delivery) obj(delivery, water)	rsn(provide, because of) obj(because of, condition.@def.@pl) aoj(specific, condition.@def.@pl)

of water of required quality for agricultur e than for urban municipal supplies?	aoj(required, quality) bas(agriculture, supplies.@pl) aoj(urban, supplies.@pl) aoj(municipal, supplies.@pl) obj(condition.@def.@pl, supply) aoj(agricultural, supply) obj(supply, water)	mod(water, quality) aoj(required, quality) bas(agriculture, supplies.@pl) aoj(urban, supplies.@pl) aoj(municipal, supplies.@pl) obj(because of, condition.@def.@pl) obj(condition.@def.@pl, supply) aoj(specific, condition.@def.@pl) aoj(agricultural, supply) obj(supply, water)	
Why did he resign?	agt(resign(agt>thing):1D.@entr y.@past, he:1A.@topic) **pur**|**rsn**(resign(agt>thing):1D. @entry.@past, ?)	{unl} and(resign@entry.@past,elect@past) man(resign(.@entry.@past,but) agt(resign@entry.@past,@topic) pur(resign@entry.@past,become) tim(resign@entry.@past,) gol(become,mayor) mod(mayor,city) mod(city,New York) tim(elect@past) obj(elec@past,he@ topic) gol(elect@past,senate@def) mod(senate@def,US) {/unl}	pur(resign@entry.@pa st,become) gol(become,mayor) mod(mayo r, city) mod(city,New York)

7.3 Error Analysis

As stated by Table 7.2 and Table 7.3 it is seen see that the questions that were entitled to the type of (What,where,why,how,when) were answered successfully ,provided there answers actually existed in the database of the UNL corpus.

The questions to which the answers were not mapped were of the types for whom the answers didn't exist in the database ,the ones that involved the calculations and also that do not confirm to the factoid based questions. These are described in detail in coming sections.

7.3.1 Answer not existed in the database

Here the sub graph generated for the query did not match with any of the graphs of the UNL coded sentences belonging to the corpus. Table 7.4 shows some of the examples for questions to which the answer didn't exist in the corpus.

Table 7.4 :Questions to which answer doesn't exist in corpus

Question	**Question Type matched (Posiible Relations mapped)**	**UNL query**	**Output**
Where are the pollutants accumulated ?	Where (src,plc)	obj(accumulate.@entry, pollutant@topic.@pl) **src**|**plc**(accumulate.@entry,?)	Answer not found
When did the rural regions need water?	When (tmt,tmf,tim)	aoj(need.@entry, region@topic.@pl) obj(need.@entry, water) **tim**|**tmt**|**tmf**(need.@entry,?) aoj(rural, region.@topic.@pl)	Answer not found
When were complex multifactor water-nutrition systems created ?	When (tmt,tmf,tim)	obj(create.@entry, system.@topic.@pl) **tim**|**tmt**|**tmf**(create.@entry, ?) aoj(complex, system.@topic.@pl) mod(system.@topic.@pl, nutrition) obj(nutrition, water) mod(water, factor) mod(factor, multi)	Answer not found

7.3.2 Question not conforming to the adhered Wh-type

The questions which were not of the specified type of factoid based did not get their actual response. These were the questions that were of the type If,Yes/No type. Following stated are some examples for the same.

- Do natural water resources represent a vital biosphere?
- Is agriculture is the greatest world water consumer ?

7.3.3 Calculation based questions

This is another type of questions which are not addressed by the application yet. Here the questions to which the answer determination involved calculation based operations were addressed. Since the algorithm doesn't involve any mathematical calculation provision yet so they could not be addressed.

- For how long were the pollutants accumulated?
- How much organic substances were found in catchments ?

Chapter 8
Conclusion and Future scope

8.1 Conclusion

Most of the current traditional question answering systems do not provide support for multiple language and greatly increase redundancy by duplicating the corpus if at all they are able to. The thesis gives a very compelling use case where a very complex problem is solved using UNL. Here various existing systems have been discussed thereby explaining the need for the multilingual QA system. The UNL framework that is proposed has effectively handled the sceanrio of the one word answers and the composite type answers. The graph based pattern matching approach discussed have given a vivid description about the working of the system. It is seen UNL greatly reduces data redundancy and promotes cross-lingual communication. This bridges the language barrier. The system has given very promising results. Currently the UNL corpus creation process is not very refined.It can be refined for running advanced web crawlers so as to fetch maximum information and increasing the scope of the system. Thus it is wise to conclude that the system has potential to build a multilingual platform for the question answering system.

8.2 Future Scope

In the future following areas will be looked upon to enhance the capabilities of the system.

i. The proposed system is limited to factoid based questions. The system can in future be expanded to non-factoid based questions also. Also the system can be made compatible to answer compound type questions.

ii. The size of the corpus currently is quite humble as the system was tested on the EOLSS[21] corpus. With the course of time with the availablity of the UNL coded sentence corpus it can be applied on multiple dataset.

iii. With the use of universal word attribute in the document base accuracy can be improved. This is one area where the depth of answer analysis will be greatly enahanced as the attributes more deeply define the universal words taken into account.For example one universal word considered in present tense in one

sentence may be mapped differently to a separate sentence when taken in past tense. So it is seen here the '@time' attribute would play a crucial role.

iv. The system can be integrated with the Eugene framework as a whole to ensure that the system finally gives the output in the natural language rather than in the UNL format.

Certifications

In the process of thesis work following certifications were completed [22].These work majorly required to have a basic level of understanding of the concepts of UNL and its fundamentals.

1) CUP 250 : Certificate of Proficiency in UNL is a certificate issued by the UNDL Foundation.

2) CLEA250: Certificate of Language Engineering Aptitude in UNL is a certificate issued by the UNDL Foundation.

UNDL FOUNDATION

Universal Networking Language

CUP 250

Certificate of Proficiency in UNL

The UNDL Foundation certifies that

KANU GOEL

has successfully completed the exercises leading to the Certificate of Proficiency in UNL (CUP250), offered by the Virtual Learning Environment for UNL (VALERIE).

Geneva, 27 July 2015

Tarcisio G. Della Senta, PhD
President
UNDL Foundation

Ronaldo T. Martins, PhD
Language Resources Manager
UNDL Foundation

UNDL FOUNDATION

Universal Networking Language

CLEA^{250}

Certificate of Language Engineering Aptitude

The UNDL Foundation certifies that

KANU GOEL

has successfully completed the exercises leading to the Certificate of Language Engineering Aptitude (CLEA250), offered by the Virtual Learning Environment for UNL (VALERIE).

Geneva, 23 July 2015

Tarcisio G. Della Senta, PhD
President
UNDL Foundation

Ronaldo T. Martins, PhD
Language Resources Manager
UNDL Foundation

References

[1]. "Universal Networking Language(UNL):Relations" [Online] Available : http://www.unlweb.net/wiki/index.php?title=UniversalRelations[Accessed 10 June 2016]

[2]."Universal Networking Language: UNL development" [Online] Available : http://www.undl.org/ [Accessed 11 June 2016]

[3]. "Universal Networking Language: UNL development index" [Online] Available : http://dev.undlfoundation.org/index.jsp[Accessed 11 June 2016]

[4].R.T. Martins, L.H.M. Rino, M.G.V. Nunes and O.N. Oliveira, "The UNL distinctive features: evidences through a NL-UNL encoding tas," Proc. 1st Int. Workshop on UNL, other Interlinguas and their Applications, Las Palmas, Spain, pp. 8-13,2002.

[5].Sherry and P. Bhatia, "Multilingual text summarization with UNL," In Computer Engineering and Applications (ICACEA), International Conference on Advances IEEE, pp. 740-745, 2015

[6].E. Blanc, "About and around the French Enconverter and the French Deconverter," Universal Network Language: Advances in Theory and Applications, Ed(s) Cardeñosa J, Gelbukh A, Tovar E, México, Research on Computing Science: pp. 157-166, 2005.

[7].B. Choudhary and P. Bhattacharyya, "Text clustering using Universal Networking Language representation," Proc. 11th Int. Conf. on World Wide Web, Hawaii,USA,pp. 1-7, 2002

[8].S. Mangairkarasi and S. Gunasundari, " Semantic based text summarization using universal networking language," Int. J. Appl. Inf. Syst, 3(8),pp. 18-23, 2012.

[9].R.T. Martins, L.H.M. Rino, M.G.V. Nunes and O.N. Oliveira, "The UNL distinctive features: evidences through a NL-UNL encoding tas," Proc. 1st Int. Workshop on UNL, other Interlinguas and their Applications, Las Palmas, Spain, pp. 8-13, 2002.

[10].P. Pakray, S. Poria, S. Bandyopadhyay and A. Gelbukh, "Semantic textual entailment recognition using UNL,"Polibits, (43), pp. 23-27, 2011.

[11].P.T. Nguyen and M. Ishizuka, “A statistical approach for Universal Networking Language-based relation extraction,” Proc. Int. Conf. on Research,Innovation and Vision for the Future, Ho Chi Minh City, Vietnam, pp. 153-160, 2006.

[12].J.A. Iyer and P. Bhattacharyya, “Using semantic information to improve case retrieval in case-based reasoning systems,” Universal Network Language: Advances in Theory and Applications, Ed(s) Cardeñosa J, Gelbukh A, Tovar E, México, Research on Computing Science: pp. 347-358, 2005

[13].S. Rani and P. Kumar, “Rule Based Sentiment Analysis System ,” Second Elsevier Int.Conf. on Emerging Research in Computing Information ,Communication and Applications, NMIT, Bangalore, India, 2014

[14].Dong, Tiansi, U. Furbach, I. Glöckner, and B. Pelzer. "A natural language question answering system as a participant in human Q&A portals." In Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three, pp. 2430-2435. AAAI Press, 2011.

[15].Cooper, J. Richard, and S. M. Rüger. "A Simple Question Answering System" In Text Retrieval Conference **(**TREC). 2000.

[16].Attardi, Giuseppe, A. Cisternino, F. Formica, M. Simi, A. Tommasi, and C. Zavattari. "PiQASso: Pisa Question Answering System." In Text Retrieval Conference **(**TREC). 2001

[17].M. Banko, E. Brill, S. Dumais,& J. Lin, “AskMSR: Question answering using the worldwide Web. In Proceedings of AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases pp. 7-9, 2002.

[18]. M.M. Abdullah and Rehab F. Abdel-Kader “QASYO: A Question Answering System for YAGO Ontology”,International Journal of Database Theory and Application 4(2), June, 2011

[19]. “Apple ios :SIRI” [Online] Available : http://www.apple.com/in/ios/siri [Accessed on 14 June2016]

[20].M. Surve, S. Singh, S. Kagathara, S. Dubey, G. Rane, J. Saraswati et al., “Agro-explorer: A meaning based multilingual search engine,” in International Conference on Digital Libraries (ICDL). Citeseer, 2004.

[21].“Universal Networking Language: UNL development EOLSS” [Online] Available : http://www.undl.org/unl-eolss/unldoc.html [Accessed on 15 June 2016]

[22].“Universal Networking Language: certificates” [Online] Available: http://www.unlweb.net/user/index.php?unlweb=certificates

[23].V. Agarwal and P. Kumar, "A Multilingual Cross-Domain Client Application Prototype for UNL-ization and NL-ization for NLP Applications" Digital Scholarship in the Humanities, 2016

Research Publications

1) K. Goel and P. Bhatia, " Information retrieval system using UNL for multilingual question answering" In "International Conference on Recent Trends in Electronics, Information Tehnology (RTEICT) Bangalore,IEEE", 2016

[Accepted]

2) K. Goel and P. Bhatia, "Universal Networking Language: A framework for emerging NLP applications" In "International Conference on Inventive Computation Technologies (ICICT), IEEE", 2016

[Accepted]

Video Presentation

Video for the thesis presentation can be seen at link:

https://www.youtube.com/channel/UCaHzNex6WfZh-_c2AfZS5zg