Multilingual Text Summarization

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering

in

Computer Science and Engineering

Submitted By

Sherry

(801332023)

Under the supervision of:

Dr. Parteek Kumar

Assistant Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA – 147004

July 2015

Certificate

I hereby certify that the work which is being presented in the thesis entitled, "***Multilingual Text Summarization***", 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 Kumar*** 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.

Signature:

(Sherry)

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

(Dr. Parteek Kumar)

Assistant Professor,

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

Countersigned by

(Dr. Deepak Garg)

Head

Computer Science and Engineering Department

Thapar University, Patiala

(Dr. S. S. Bhatia)

Dean (Academic Affairs)

Thapar University

Patiala

ii

Acknowledgement

The task is not successfully terminated without acknowledging the people who have been give their best efforts and precious time to complete the task. There is no volume of words to describe their guidance and support.

First of all I wish to acknowledge the Courteous God for the motivation and inner strength to complete my task and give me courage to overcome from obstacles.

I would like to express my deepest gratitude towards my guide **Dr. Parteek Kumar**, Assistant Professor, Computer Science and Engineering Department, Thapar University for his great support, efforts, positive attitude, keen interest and co-operation. I appreciate his knowledge and skills in my area. He has helped me to explore the topic in organized manner and given me guidance on work towards research oriented manner.

I would like to thank **Dr. Deepak Garg**, Head of Department and **Dr. Ashutosh Mishra**, P.G. Coordinator, Computer Science and Engineering Department, Thapar University for their inspiration and motivation to complete the task.

A special thanks to my parents and friends for showing me always the right path and for their hearted co-operation.

Sherry

801332023

Abstract

The volume of information has been increased almost in all fields like education, medical and science *etc*. This high volume is due to social networking and other activities. The management of data becomes a huge problem. It consumes a lot of time and effort to read the whole information and conclude the results. So the interest in automatic text summarization systems has been increased. It is applied in every field to reduce the information to save time and efforts for analyze the information. Multilingual text summarization is required to provide the summary to user in the desired language. So the all people can take the advantage of information sources which was not possible earlier.

The project work has been carried out in multilingual text summarization. The hybrid approach has been used in the system by extracting the best features of existing algorithms; by providing the new features also. The Universal Networking Language (UNL) has been used in system for the language transformation. The natural language document in English is provided as input. The natural language document is converted into UNL. The system produces the summary of the UNL file. The score computation is a major parameter for the sentence selection. The summary is refined in each step by different UNL heuristics. The language transformations are carried out by UNL tools *i.e.* IAN and EUGENE. At the end summary in Punjabi language is produced.

Table of Contents

List of Figures

List of Tables

Chapter-1
Introduction

Text Summarization represents important text information by leaving the irrelevant one, reduces the details and contents them in compressed way that meets with requirements of user. Text Summarization may define on the basis of three important features which are as follows:

- Text Summaries should be short.
- Text Summaries should conserve the important information.
- Text Summaries may generate from single or multi documents.

In a single document summarization summary is generated from single source document while in multi document summarization more than one source document are provided to generate summary. Although text summaries have traditionally focused on text as input, the input to the automatic summarization can be images, multimedia, video or audio as well as hypertext or online information [1]. Due to the increase in the quantity of data almost in every field, the data management has become one of the most challenging issues. Hence, the interest of researchers in the automatic summaries generation has been increased. Text summarization has become the timely tool for interpreting the important information.

1.1 Challenges of Text Summarization

A big challenge is how to evaluate the summaries. After applying algorithm one must be confident that summaries should be relevant and considering all the important factors of the main document, *i.e.*, if the resultant summary is textual it should preserve the main idea. Irrelevant information should not be the part of document after filtering. The union of conciseness, readability and completeness would always give good summary [2]. It becomes also a challenge that proposed algorithm should produce meaningful and timely summaries.

1.2 Applications of Text Summarization

Text Summarization is used is in medical field, in multimedia news summarization, in producing intelligent reports, in text for hand held devices, in text-to-Speech for blind people, in education and in summarizing meetings [3]. Many other scenarios use text summarization. For example, an information retrieval system uses automatic summarization to produce the list of retrievals. Now a day's summary of the email messages and news articles is sent to mobile devices as Short Message Service (SMS). Search engines also use summary mechanisms. The summary of the web pages is shown on the screen as a result of particular search.

1.3 Approaches of Text Summarization

There are different approaches used for text summarization based upon their method of summary generation. The different approaches of text summarization are shown in Figure 1.1.

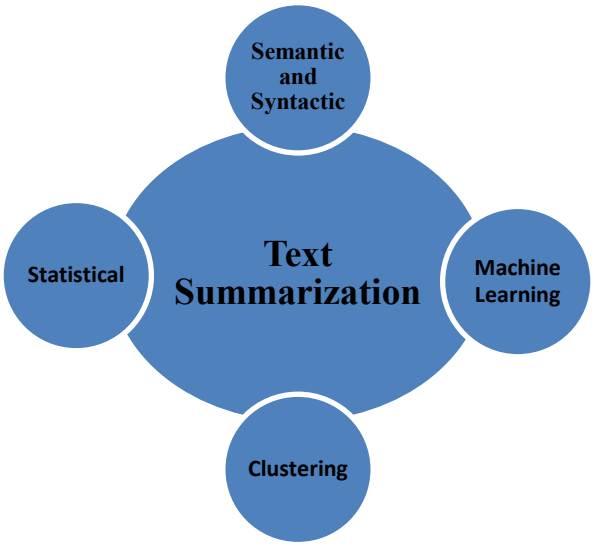


Figure 1.1: Text Summarization approaches

The semantic analysis techniques are applied on text summarization to find the relation between different sentences. The different rules are applied for sentence selection. The Statistical techniques use statistical methods like Binomial distribution to select the sentences for summary. Machine learning approaches are naïve Bayes, log linear model

etc. for summary generation. In clustering method different objects are grouped together based upon the properties *i.e.* similar sentences are part of one cluster. The detail of these approaches will be discussed in chapter 2.

1.4 Role of UNL in Multilingual Processing

Universal Networking language (UNL) is intermediate language used for language translation. Language translation using UNL is more optimal than other techniques. The language translation of UNL is shown in Figure 1.2.

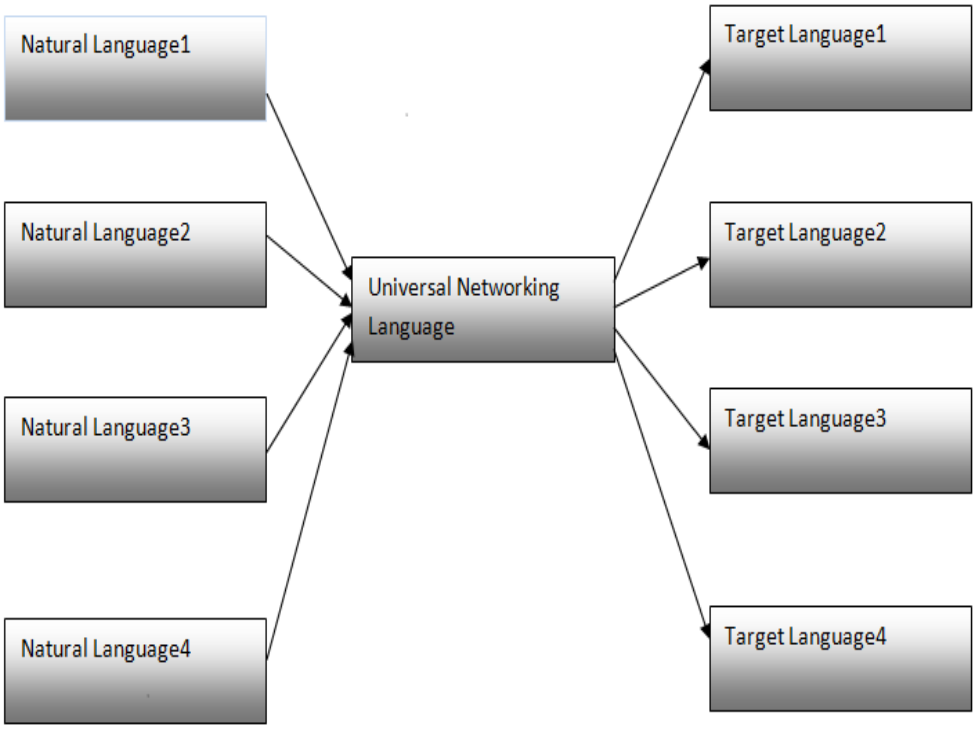


Figure 1.2: Language translation using UNL

For n different languages, using UNL approach conversion of one language to other $(n-1)$ languages can be done in $2 * n$ steps. This is because only two conversions are required to convert a natural language to other natural language, *i.e.,* one for natural language to UNL then UNL to other natural language. But without using UNL approach $n(n-1)$ steps are required for conversion of one language to other $(n-1)$ natural languages.

1.5 UNL as an important tool for Text Summarization

Universal Networking language (UNL) is used in multilingual text summarization. It is intermediate language which is used for language translation, *i.e*., with the help of UNL it is possible to produce a summary of a source natural language document to reduced destination natural language document. First of all source document is converted into UNL document by UNLization process. Summary of this UNL document is produced by using algorithmic approach. Then resultant summary is converted into destination natural language using NLization process. UNL-based summary provides the following important features like unambiguity, multilinguality which are beneficial for multilingual summary point of view.

1.5.1 UNL Representation

The information provided by natural languages is expressed in UNL in the form of semantic network. Semantic network comprises various discrete semantic units like Universal Words (UW), Universal Relations and Universal Attributes. Universal Words form nodes in a graph. Universal Words are the simple English word represents nouns, pronouns, and adjectives *etc*. Universal Attributes are annotations made to nodes in a semantic network. Links in a semantic network is represented by Universal Relations.

1.5.2 Universal Words

Universal Words represents the content conveyed by natural languages in the form of nouns, verbs, adjectives and adverb *etc*. They are called Universal because they comprise the lexicon of the "Universal Language", *i.e*., they convey the ideas that can be expressed in each and every language [4]. Universal Words are expressed by nodes in a semantic network. Universal Words are divided into two categories.

1.5.2.1 Permanent Universal Words

They are included in Universal Dictionary and include the concept that has been already lexicalized in at least one natural language. This means that words are recognize as single lexical item and therefore included in the dictionary of the natural language.

Simple Words

Simple words are represented by isolated nodes in a graph. This refers to a case when Universal word represents concepts that are not compositional. For example, Universal Word "big" is no longer composite.

Compound Words

These words are also represented by isolated nodes but combine with attributes. In compound word concept is fully derived from a combination of a Universal Word and attribute. For example "bigger" can be expressed as a combination of Universal Word "big" and attribute "@more".

Complex Words

These are represented as hyper-node*, i.e.,* a sub graph in a UNL graph. They follow the structure defined for UNL sentences. They are used when the concept is fully derived from the combination of Universal words, Universal attributes and Universal relations.

1.5.2.2 Temporary Universal Words

It includes a entities that are in the processing state of lexicalization ("googlers"), entities that are too specific to be included in UNL Dictionary ("Leon Werth"), entities that are not translatable *etc*.

1.5.3 Universal Attributes

Universal Attributes are annotations made to nodes in a semantic network. They actually denote the circumstances under which these nodes are used like tense, mode, aspect *etc.* [5]. Attributes may convey three different types of information which are as follows:

i. Information about the role of the node
ii. Information conveyed by morphemes and classes such as affixes, determiners, conjunctions, degree adverbs *etc*.
iii. Information on the external context*, i.e.,* non verbal elements of classification. "@past", "@present"," @def", "@indef "are some examples of attributes. Some of the attributes are shown in Table 1.1.

Table 1.1: Some UNL Attributes

Concept	**UNL Attributes**
Time Representation	@past, @present, @future
Reference Representation	@generic, @def, @indef
Quantity Representation	@multal, @extra
Number Representation	@pl
Gender Representation	@male, @female
Logical Representation	@transitive, @symmetric, @identifiable, @disjointed
Reference Representation	@habitual, @perfective and @progressive
Feelings and judgments	@ability, @grant, @wish, @will, @obligation
Attitude	@affirmative, @imperative, @interrogative, @request

1.5.4 Universal Relations

Universal Relations are known as "links", are labeled arcs which are used to connect one node to another in a semantic network. UNL relations are used to represent the semantic cases or roles such as agent, object and instrument *etc.* between Universal Words [6]. Relation is represented by three letter words that specify the kind of semantic relationship between the two Universal Words such as

<name of relation>(<source>;<target>) … (1.1)

Let us consider the example sentence,

Peter is a human being. … (1.2)

UNL of (1.2) is

iof (human being, Peter) …(1.3)

Sentences (1.3) convey the information that "iof" is name of relation. There is "instance of" relationship between Universal Word "Peter" and "human being". Table 1.2 represents some UNL relations and their definition.

Table1.2: Some UNL Relations [7]

Relation	**Meaning**	**Definition**
agt	Agent	A participant in an action that provokes

		change of a state.
and	Conjunction	Used to state conjunction between two entities.
aoj	object of an attribute	Used to express the predicative relation between subject and the predicate.
ben	Beneficiary	A participant who is advantaged/disadvantaged by an event.
cnt	Content	The theme of an entity.
con	Condition	Condition of an event.
dur	Duration	Duration of entity or event.
exp	Experience	A participant in an action who receives a sensory impression.
gol	Final state/place/destination	Final state or place or destination of an entity or event.
Icl	Is a kind of	Used to refer to a subclass of a class.
iof	Is an instance of	Used to refer to an instance or individual element of class.
lpl	Logical place	A non physical place where a event occur
man	Manner	It indicates how a action or process of an event is carried out.
mod	Modifier	A general modification of an entity.
nam	Name	Name of an entity.
obj	Patient	A participant in an action or process.

or	Disjunction	Used to indicate disjunction between two entities.
plc	Place	The location of entity or event.
pos	Possessor	The possessor of a thing.
pur	Purpose	The purpose of an entity.
qua	Quantity	The quantity of an entity.
tim	Time	The temporal placement of an entity.
tmf	Initial time	The initial time of an entity.
tmt	Final time	The final time of an entity.
via	Intermediate state	The intermediate state or place of an entity.

1.5.5 UNL Sentence

UNL sentence is a basic block for expressing the UNL framework. It consists of Universal Words related by Universal Relations and modified by Universal Attributes. When UNL sentence is represented in the form of semantic network consisting of nodes and arcs it is called UNL graph. Consider the English sentence (1.4) for its UNL representation.

Input Sentence: - all my books. … (1.4)

UNL Sentence:

{unl}

pos (**book**:51.@all, 00:02.@1) … (1.5)

{\unl}

In (1.5) sentence “pos” is a possessor relation ,”book” is a Universal word, “@all” is used to represent that books are plural,”@1” represents that “my” is a first person and “00”,“51” represents id of the Universal Words.

1.5.6 UNL Tools

UNL tools are the software programs which are available for transformation, *i.e*., tools which convert the natural language sentence to UNL and then again UNL to target language. The process of conversion of natural language to UNL is called UNLization and for conversion of UNL to target language is called NLization. UNL contains following two tools for UNLization and NLization.

1.5.6.1 IAN (Interactive Analyzer)

IAN is a tool used to convert natural language to UNL. It contains the grammar for natural language analysis .Word Sense Disambiguation (WSD) is carried out by language expert, but the system can also have a set of Disambiguation rules (D-rules) [8]. IAN deals with natural language rules (N-rules), transformation rules (T-rules) and disambiguation rules (D-rules). IAN performs the conversion using three steps.

- **Segmentation:** It is a task of dividing the input document into series of different processing units (sentences). IAN processes these units one by one.
- **Tokenization:** It is a process of identification of different lexical items known as tokens from the input document.
- **Transformation:** They are the transformation rules of the grammar which are applied on lexical items or tokenized sentences in order to convert them to UNL graph.

1.5.6.2 EUGENE

It is known as dEep to sUrface natural language GENErator because it converts UNL language to a desired target natural language. It is language independent and it uses dictionary and grammar rules for its conversions [9]. UNL performs the process of NLization in three steps.

- Segmentation is a task of dividing the input UNL document into series of graphs which are then processed one by one.
- Tokenization is a process of finding out the tokens in a graph.
- Transformation is a process of generating transformation rules for the grammar which are required during the UNL to natural language conversion.

1.5.7 UNL Architecture

UNL architecture is shown in Figure 1.3. UNLization and NLization processes are carried out with the help of UNL tools. Natural language sentences are converted by IAN to UNL during UNLization process. For UNLization process Transformation Rules (T-Rules), dictionary entries, Disambiguation Rules (D-Rules) are required. The NLization process converts the UNL to the other natural language. In this process language grammar, dictionary rules, transformation rules and disambiguation rules are used.

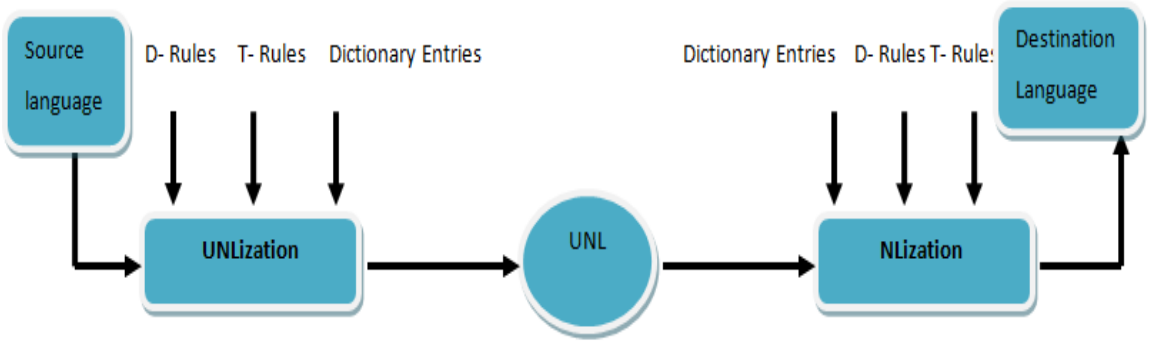


Figure 1.3: UNL Architecture

1.6 Thesis Outline

Thesis has been divided into six chapters. Chapter 1 contains introduction to Text Summarization. It covers the challenges, approaches, applications and role of UNL in Text Summarization. Chapter 2 describes the various approaches of Text summarization like machine learning approach, deep natural language analysis method, Multilingual Text Summarization *etc*. Chapter 3 describes the problem statement, objectives and methodology for the development of the Multilingual Text Summarization system. Chapter 4 includes the implementation of proposed Test Summarization system.

Algorithm and detailed working has been explained in this chapter. Chapter 5 includes results of the proposed system. The conclusion and the future scope has been described in chapter 6.

Chapter Summary

In this chapter text summarization is introduces along with its various applications, need, challenges and approaches like machine learning, deep natural language analysis methods, multi document summarization and Multilingual Text Summarization. The role of the UNL in Multilingual Summarization has been described. The UNL architecture has been described in detail.

Chapter-2
Literature Survey

The research on the automatic text summarization has been started from 1950's.After a gap of decade's progress in the field of language processing was done due to the growing volume of online text. So, the large amount of electronic documents which are available in Internet has stimulated the development of excellent information retrieval systems. For example, Google always shows some part of the text corresponding to the query of the user. The user has to decide whether the document is interested or not only on the basis of extracted text. The user has to browse all the documents until a right document come. The solution is a use of automatic text summarizer which displays only the important and essential information about the document. Hence, it becomes very easy for the user to choose the right document.

The demand of the text summarization has observed in various domains like education, medical, government offices, research, business *etc*. So, the development of automatic summarization systems has importance due to research point of view. There are different approaches used for text summarization on the basis of single document, multi document summarization [10]. The different approaches are shown in the Figure 2.1.

2.1 Single Document Summarization

In single document summarization only one document is provided for summary generation. It is a simple and earliest approach for summarization. Extractive and abstractive both summaries methods can be applied on single document summarization.

2.2 Multi Document Summarization

Multi document summarization is also very important part of summarization. More than one information sources are provided for summary generation. Many web based clustering systems like news were inspired from multi document summary. But task of multi document technique is more difficult and complex than single document techniques. The real aim is not only to remove redundancy and indentify correct text for summary but also to provide novelty and ensuring that final summary should be coherent

and complete in itself. So it was a challenge for them to consider all the documents and relate the summary.

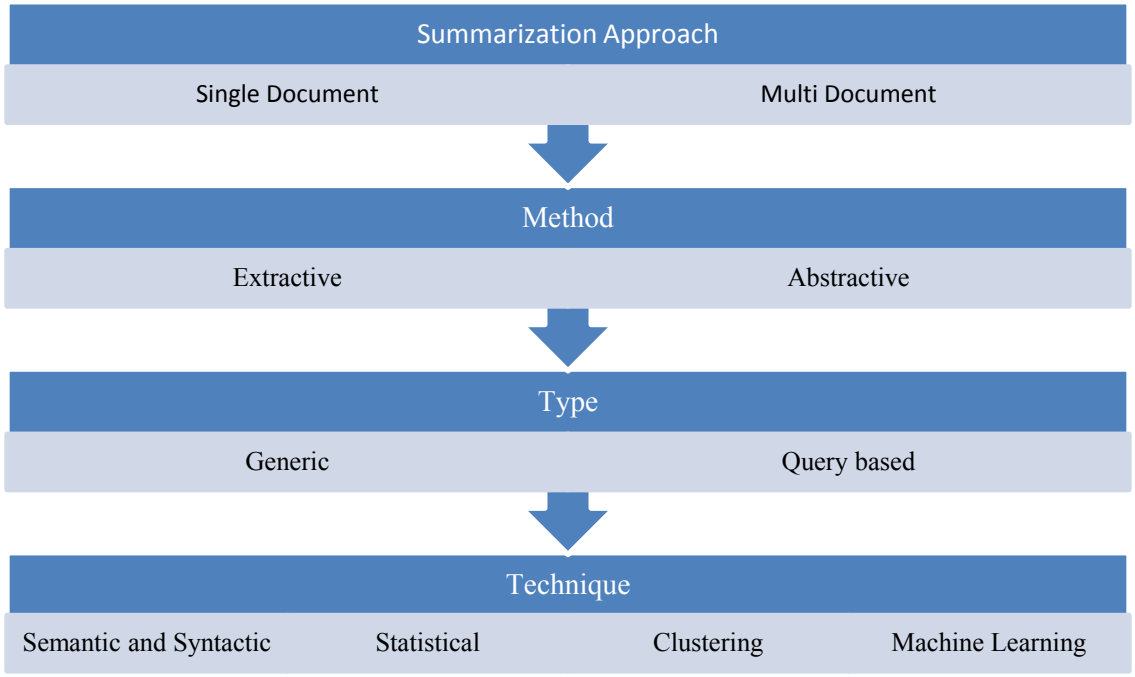


Figure 2.1: Summarization approaches

The work was initiated by the NLP group in the University of Columbia where summary system is called SUMMONS. At start procedures and challenges were different but later on people from different communities added their own perspective to the problem. Some approaches use clustering to indentify common themes and later on each cluster is represented by a sentence others produce more than one sentences from a cluster, while some uses maximal marginal relevance to work dynamically to include a passage if it is novel with the previous passages [11].

2.3 Extractive Summaries

Extractive summaries are simplest form of summaries. These approaches select the important information in the form of sentences, paragraphs *etc.* from the source document; combine them to generate short sentences. The selection of the sentences is made on the basis of different features like linguistic, statistical features *etc*. No synonyms of the words are used in extractive summaries for simplification [12]. Extractive summaries are not suitable for multi documents because sometimes these are biased towards some information sources.

2.4 Abstractive Summaries

Abstractive summaries are complex as compare to extractive summaries. These summaries consider the proper understanding of the source document and redefining it into new simple words [13].Actually it generates internal semantic representation; then use the natural language techniques for the summary creation which is close to human mind summary generation. Abstractive summaries add synonyms for the more simplification of the summaries.

2.5 Query based Summaries

Query based summaries are applicable in case of both single document and multi document summaries. Query based summaries are produced on the basis of information required to user. The main aim is to retrieve sentences which satisfy user query. The score of the sentences are calculated on the basis of frequency of words in a document. A sentence with the query phrase provided a high score than others. The sentences with the high value of the score are extracted for summaries. Partial sentences may also be extracted and further union of them is carried out [14].

2.6 Generic Summaries

Generic summaries are not query based. Query based summaries are biased because they do not provide the overall review of the source document. They deal with the user queries only hence not suitable for content overview. To define the category of the document and to describe the main key points of the document generic summaries are required. A best generic summary considers the main topics of the documents and tries to minimize the redundancy as less as possible [15].

2.7 Summary Techniques

There are four summary techniques which will be described as follows.

- Semantic and Syntactic (Rule-based)
- Statistical Technique
- Clustering Technique
- Machine Learning Technique

2.7.1 Semantic and Syntactic (Rule-based)

There are many semantic analysis techniques which are applied on text summarization to find the relation between different sentences. Following are the three semantic and syntactic summary techniques.

- Graph Representation
- Lexical Chains
- Natural language processing

In the graph representation lexical graphs, Graph matching, weighted graphs and unweighted graphs are used for summarization. In lexical chains word net, co-reference chains and lexical semantics *etc.* are used for summarization. Natural language processing used information extraction, part of speech tagger for the summarization. Summarization techniques under NLP are divided into two categories as follows.

- Plain Text Summarization
- Multilingual Summarization

In Plain text summaries resultant summary is in the same natural language but in multilingual text summarization resultant summary is in different natural language. Initial work in plain text summarization was started in 1950's. Most initial work on the text summarization was targeted on technical documents. Luhn (1958) proposed the first algorithm for text summarization at IBM. The author proposed text summarizer which was based upon frequency of a particular word in a document [16].The main motivation was to short the news information, biographical information. According to the Luhn summary has different categories, some of the summaries are difficult to generate than other. Different categories are Extractive, Abstractive, Indicative, Informative and Critical. Extractive summaries are simplest. These summaries contain the sentences which have already presented in text. Abstractive summaries contain some new text also. Indicative summaries represent the scope of the whole document without including whole content. Informative summaries represent the important factual content of the text document. Critical summaries represent reviews on scientific papers about their work and results. Baxendale (1958) also did work related to extractive summaries at IBM. The

author more focused on the “sentence position”. Approximately 200 documents are analyzed for research [17]. Edmundson (1969) proposed the system for document extraction. The proposed algorithm was the first algorithm for extractive summaries. Two previous features sentence frequency and position of the sentence were used along with the new features like cue words and skeleton. Cue words are words like hardly, significant etc. Skeleton define the heading of the document. After evaluation 44% results matched with the manual results [18].

Multilingual text summarization is come into existence in 2005. This technique is still in early stage but this different framework has many advantages in the newswire field in which information is combined from different foreign news agencies. Evans (2005) described the scenario in which there is always a preferred language in which summary is required, different multiple source documents are in demand and in different languages are available. They preferred English as a source language and documents are from the news articles in English language and Arabic. The logic was to generate the summary of English articles without discarding the details contains in Arabic. IBM’s machine is used to do a transformation of Arabic language to English. The system checks the transformed document in Arabic corresponding to a document of English for each sentence. If match is found then sentence is found relevant for summary. Hence more grammatical summary is found this way, since machine translation is still not perfect of that. To find out the similarities between sentences Simfinder tool was used. This is a clustering based tool based upon similarity over different semantic and lexical features which is using long linear regression model. Universal Networking Language is mostly used in multilingual summarization.

Martins and Rino proposed algorithm for the text summarization using UNL. They presented UNLSumm model to prune the UNL text by means of heuristics that totally focus upon unnecessary binary relations. The system used decoder to produce corresponding summary in Brazilian Portuguese. Their pruning heuristics are based upon the relations of UNL. Although each relation is not candidate for pruning because some relations like “agt” or “obj” convey important information [19]. Only some of the relations are candidates for pruning. According to this algorithm initially there was 84

heuristics were divided into two groups A and B shown by Figure 2.2 and Figure 2.3. Group A considers 39 heuristics. It also called as single pruning and removes the independent binary relations one by one. Group B heuristics are complex than the Group A heuristics. Group B heuristics are called chained pruning, *i.e.,* once the binary relation is excluded the interconnected binary relation is also excluded.

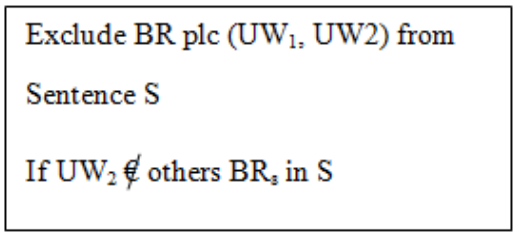


Figure 2.2: Group A heuristics

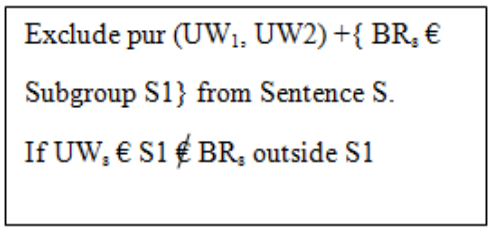


Figure 2.3: Group B heuristics

According to Figure 2.2 Group A heuristic delete the place relation from the UNL document, provided frequency of UW_2 is one means UW_2 should not be a part of any other relation in the same UNL sentence. While applying Group B heuristics frequency of UW_2 should be 2. These heuristics are more complicated because deleting a desired relation containing UW_1, UW_2 leaves blank [ ] in any other relation where UW_2 is placed. Hence to avoid this situation the relation containing [ ] is also removed from the UNL document. For example, if purpose relation as shown in Figure 2.3 is deleted containing UW_2 then any other relation in same UNL sentence containing UW_2 no more will the part of UNL document.

The serious problem regarding these heuristics is to decide the heuristics application order when considering both type of pruning. By default Group A heuristics are always applied and in case of interdependency when dangling of binary relations occurs, Group B are applied. However, Group A and B work on the same binary relations but sometimes after applying Group B heuristic results into more than one dangling relations. Hence, to give a priority to Group A or Group B heuristics is one of the major issues. The precision of the Heuristics is calculated represented by (2.1).

$$Precision(H) = Sat\_Num/Total\_Num \quad \ldots (2.1)$$

$$Sat\_Num = No\ of\ applications\ of H\ leading\ to\ satisfactory\ results \quad \ldots (2.2)$$

$$Total\_Num = Total\ No\ including\ satisfactory\ and\ unsatisfactory\ results \quad \ldots (2.3)$$

There are some limitations of approach which are as follows.

- Sometimes it covers non-relevant information.
- There is an upper bound to the number of heuristics applied for each entry.
- Application order is relevant and providing satisfactory results or not.

Managaikarasi and Gunasundari (2012) proposed an idea of text Summarization. The most important work they have done is improved methodology, which scans the document and transform into UNL graph. The system introduced UNL as a language for knowledge representation and information representation that can be describe in natural language conversation. They proposed method to find the summary of UNL document. The documents are collected from websites based upon the education domain. These documents contain images and unwanted information also. In the First Step stop words are removed. The sentence splitter is used to split document into sentences. The delimiter used is blank space here. In the next step sentences are again splitted into words. Then Morphological analyzer analyzes these words to find out the root word. These root words provide to UNL dictionary. Tenses and heuristic relations of the root words are indentified. The graph is constructed from given information. During graph construction counter field is also updated. Counter field is provided to find the important concepts, based upon threshold. Highest concepts sentences are finally picked for the resultant summary of the document. The system is tested on education domain document for summary. There was manually tested on the summary with experts. During summary preparation the data is collected from the news service providers. Each document includes the irrelevant information like images, tables *etc*. So, there is a need of creation of ideal summary for evaluation of results. For the ideal results the documents are distributed to three judges and rank is given to the sentences according to their importance in text document. The future work for project is develop a well managed tool for evaluations and updating of UNL dictionaries with the help of root words provided by

Morphological analyzer. It also identifies more and more UNL relations with the help of heuristic rules [20].

Pandian and Kalpana (2013) also proposed text Summarization mechanism using UNL. They focused on the tourism domain document which is UNL based. The Bengali UNL system is developed by them. UNL representation used by the system was for simple sentences not for complicated sentences. The main focus was mostly on DeConversion part which converts the universal networking language to Tamil language. The source document is scanned and it is converted into intermediate language. It further undergoes generation process for final output. For the summary process the source document undergoes a process of EnConversion which includes the steps like parts of speech tag, parts of speech parsing, identification of entity, identification of relation, creation of dictionary and generation rules. Source document is converted into UNL document of UNL expressions. In the first phase parts of speech tagging and parts of speech parsing is carried out with the help of Stanford Parser. The outcome of the parser is used to find the entities and relationship between entities. Further rules are constructed and knowledge base is obtained for the generation of UNL expressions. UNL document containing UNL expressions are passed to the DeConvertor for the generation of the final summary and final output for three levels of users (level1 user, level2 user and level 3 users). The DeConversion module is constructed in such a way that it will perform the function of both summarizer and DeConvertor. To obtained the summary DeConversion module scan the word dictionary and find the relation between the different universal words, attributes of the universal words are collected and relation between universal words are taken. Further the unnecessary information like determiners, prepositions are reduced to obtained the final summary document. Final summary document is produced for the different levels of user's base upon the classification of ages. The distribution level of the summary document based upon the IQ level. DeConversion module produce summary in three steps which are as follows.

- Analysis and preparations of the dictionary information,
- Preparing DeConversion rules and
- DeConversion to produce the output summary document.

The experimental analysis is carried out using NetBeans IDE. The analysis obtained different levels of users based upon their level of IQ to access intelligence. The overall population is considered for experimental analysis. The performance of the overall system is analyzed and considered in the form of Decisiveness for all levels. It is defined by number of words compressed at different levels [21]. It is calculated for all the levels by using the given formula which is shown in 2.4.

$$Decisiveness\ for\ the\ level\ User(DLU) = \left\{100 - \frac{No\ of\ words\ in\ level\ summarized\ document}{Total\ No\ of\ words\ in\ original\ document} * 100\right\} \ldots (2.4)$$

In (2.4) decisiveness is find out and graph is plotted against decisiveness ratio and document. After plotting a graph it is observed that the compression for original document for level 1 user is more than rest level users. Same is true for rest two level users.

Sornlertlamvanich *et al.* proposed an approach for Summarization using Universal Networking Language. While producing summary this approach considers surface and semantic information of the UNL. The multilingualism can also be realized using DeConvertors from the summarized UNL document to the resultant target natural language document under the framework of UNL. Algorithm consists of four steps. In the first step the score of each UNL sentence is calculated. Score of the sentence is calculated by using weight of each universal word. Weight of each universal word is calculated by using the factor of frequency and inverse document frequency. After the score calculation some top most sentences based upon score are chosen for the future summary. By using the semantic information of the UNL the redundant words are removed from the summary in third step. This is mathematically calculated by using contribution function. The values obtained through contribution function are compared with the threshold value 1.5. To make summary more natural and real different sentences are merged based upon the head of the sentence and no of words in the sentence in fourth step. This algorithm is applicable for multiple document summarizations. Their experiment proved that use of

the UNL improves the summary quality as compare to the plain text summarization. The semantic information of the UNL can also be applied to improve the naturalness in sentence level of summary [22].

2.7.2 Statistical Technique

To extract relevant sentences for summary some summarization systems use statistical techniques. Binomial distribution, sentence compression and relevant scores these are statistical method used for summarization. Hidden Markov model also uses this approach.

Conroy and O'Leary (2001) applied hidden Markov model approach for plain text summarization. They used sequential model for the local independence. The system had three features: position of a particular sentence in document, number of different terns in a particular sentence, likeliness of particular sentence terms.

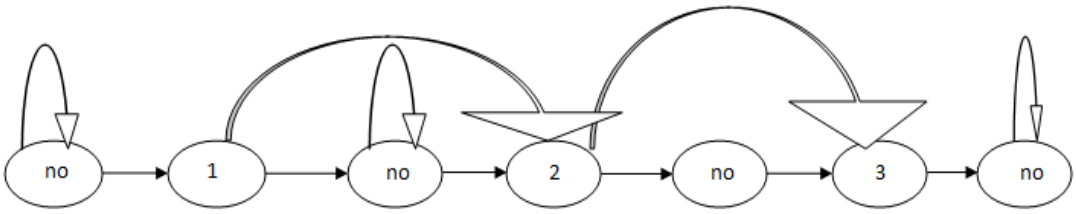


Figure 2.4: Markov model to extract the three summary sentences [23]

In Figure 2.4 Markov model is shown for $2s + 1$ states. Summary states and non summary states are alternated. In this "no" represents the non summary states and numerical numbers represents the summary states. There is a jump to next state in case of summary state. The Figure 2.4 represents the model with 7 nodes corresponds to $s = 3$.

2.7.3 Clustering Technique

Clustering is a process in which different objects are grouped based upon their properties, *i.e.,* objects with the similar properties belongs to same group. In a document different topics are arranged in a particular sequence. In cluster based summaries sentence selection is based upon the cluster Ci. The second factor is location of a particular sentence Li. The factor which increases the score of a particular sentence is its similarity

to already present sentence in a document. The overall score of a sentence depend upon these three factors.

$$S_i = W1 * C_i + W2 * F_i + W3 * L_i \quad \ldots (2.5)$$

In (2.5) Si is score of sentence, C_i is cluster to which sentence belongs, Fi is document to which sentence belongs and L_i is location of a particular sentence.W1, W2 and W3 represents the weights [24].

2.7.4 Machine Learning Technique

Machine learning techniques are very effective for automatic text summarization. The some of the machine learning approaches are discuss as follows.

2.7.4.1 Naive Bayes Approach

Kupiec (1995) described a method for summarization. He described a classification function known as naïve Bayes classifier which is responsible for the each sentence to be a part of summary. If S denotes the total number of sentences and s denotes a particular sentence with a features F1 to Fk [25]. The formula of naïve Bayes is shown in 2.6:

$$P(s \in S|F1, F2, \ldots . Fk) = \frac{\left(\pi_1^k P(F_i s \in S).P(s \in S)\right)}{\pi_1^k P(F_i)} \quad \ldots (2.6)$$

The new features like sentence length and the uppercase words were added. Score is calculated for each sentence and based upon that top most n sentences were chosen. Aone *et al*. (1999) also describe a naïve Bayes classifier with more additional features. He introduced the terms "frequency" and "inverse document frequency" in plain text summaries. The corpus used in the experimental analysis was from newswire. The inverse document frequency was computed from a large corpus of the same area.

2.7.4.2 Rich features and Decision Trees

Lin and Hovy (1997) describe the importance of a feature "sentence position". According to this a weight is provided to sentence based upon its position in the text [26]. This method also called as position method. A newswire corpus was used for experimental

analysis. The authors measured the yield of every sentence position. They ranked the different sentence positions to produce the "Optimal Position Policy (OPP). They performed the two kinds of evaluations. They test on the unseen text. The first evaluation was exactly like the training documents and the second evaluation considered the word overlap for the manual abstracts was measured. Abstract windows and selected sentence windows were compared and precision, recall values were measured.

Lin (1999) broke away the assumption that the features are independent and tried to model the problem using decision trees instead of naïve-Bayes classifier. The system described lot of features in sentence extraction and their effects. The data set used was publicly available texts classified into various topics. The data set is divided into text fragments which are evaluated by human judges. Some important features were query signature (normalized score of the sentences depending on the number of query words), IR Signature (the salient word like the signature word), numerical data, proper name (Boolean value 1 is given to sentence that had a proper name), pronoun or adjective (Boolean value 1 is given if they appeared), weekday or month, Quotation, query and signature. The system experimented with different baselines like positional feature, simple combination of features. When machine extracted and human extracted sentences were matched, the decision tree was clearly the winner [27].

2.7.4.3 Log Linear Models

Osbrone (2002) described the Log Linear model approach for the plain text summarization. This approach is different than the previous approaches which always assumed feature independence. The system showed that this approach is better than naïve Bayes classifier approach [28]. The model can be stated mathematically as follows.

$$P(c|s) = \frac{1}{Z(s)} \exp\left(\Sigma_i \lambda_i F_i(c,s) \right) \quad \ldots(2.7)$$

Where Z(S) is

$$Z(s) = \sum c\,(\sum i\, \lambda_i F_i(c,s)) \qquad \ldots (2.8)$$

In (2.7) and (2.8) c is a label, s is a item to be labeled, f_i is a feature (i-th feature) and λ_i is weight of the feature. There are two possible labels regarding whether the sentence is to be extracted from the document or not. The weights given to sentences are calculated from conjugate gradient descent. The non uniform prior is added to the model by authors. This model rejects too many sentences during processing. The features included by the authors were word paring, length and position of the sentence and discourse features like inside the introduction, part of conclusion.

2.7.4.4 Neural Networks

DUC (2001) applied the neural network technique for plain text summarization. The system produced a summary of single news article in 100 words. However the best systems in evaluations of experiments could not outperform the baseline analyzed by Nenkova in 2005. After 2002 the task for the single document summarization was dropped by DUC. Svore (2007) produced an algorithm based upon neural networks and used the third party features like dataset to resolve the problem of extractive summarization. The data set consists of 1365 documents collected from CNN.com. The datasets consists of human generated stories, articles, title and timestamp *etc*. For the evaluation two metrics were considered. The first one is to combine the system produced three highlights, combine the human generated three highlights and comparison of these two. The second take care about the ordering and the individual level comparison of the sentences [29].

Strove (2007) trained this model on the basis of labels and featured for each sentence that referred the ranking of each sentence in source document. Ranking was provided to the sentences on the basis of RankNet which was a paired based neural network algorithm. ROUGE-1 is used as a training set. The authors concluded that if a sentence contains keywords regarding new search engines and Wikipedia articles then the probability of a sentence in highlight is more.

ROUGH-1, ROUGH-2 was used for the evaluation purpose and statistically improvements were shown over baseline [30].

Chapter Summary

In this chapter various approaches of summarization like single document, multi document, extractive, abstractive, generic and query based *etc*. The various summary techniques like machine learning, semantic and syntactic, clustering and statistical have been discussed. Machine learning includes naïve Bayes, decision trees, and neural networks. Multilingual text summary approaches also have been described with detail procedures.

Chapter-3

Problem Statement

Automatic Text summarization has become an essential part of our daily routine due to the presence of huge volume of information that required to be summarized for humans so that they can go through essential contents in short duration of time. It has been analyzed from long periods of time that more and more data is becoming available and some tools are required to handle it. Almost in every field like education, medical field, multimedia, Text to speech for blind people, scientific field there is a need to study the important information rather than to study the whole provided information. It becomes important to compact the whole information in order to decrease the overall time to review this large information. So, there is a need for Automatic Text Summarizer. Multilingual feature is required because technology is everywhere but still there are millions and billons of people which are away from present technologies. To convey some information to those people one should convey only in their language and it is only possible by using language translations. Hence there is a need of Multilingual Text Summarizer.

3.1 Objectives

The main objective of this research work is to implement a multilingual text summarization system. In order to complete the task, following objectives were framed.

- To study the different existing approaches for plain text and UNL based summaries.
- To study of the UNL framework to develop the multilingual text summarization system.
- To propose and implement the text summarization based upon UNL framework to summarize the input document.
- To test and validate the proposed system.

3.2 Methodology

To attain the objectives as discussed in the previous section following Methodologies has been used.

- Literature survey has been studied regarding the plain text summarization and multilingual text summarization approaches. The various machine learning techniques like neural network, Log linear has been studied. The different text summarization techniques with UNL have been studied in detail.
- Detail study of UNL and its tools like IAN and EUGENE to understand the language transformations. The UNL relations, attributes have been studied in detail.
- Modification in the existing multilingual text summarization algorithm by adding new features and combining the desirable features of the different algorithms.
- To test and validate the system on hare and tortoise corpus provided by UNDL foundation.

Chapter-4
Implementation

4.1 Architecture of the Proposed Multilingual Text Summarizer System

The proposed system is based upon the multilingual text summarization technique of summarization. It produces the summary of the English text document in Punjabi language. The source UNL document is given to system. Proposed algorithm is applied on UNL file. This algorithm generates summary of the UNL file by applying different heuristics and rules. At last UNL summary document is given to EUGENE for the language transformation and resultant summary in different natural language is produced. Figure 4.1 represents the overall architecture of the proposed system.

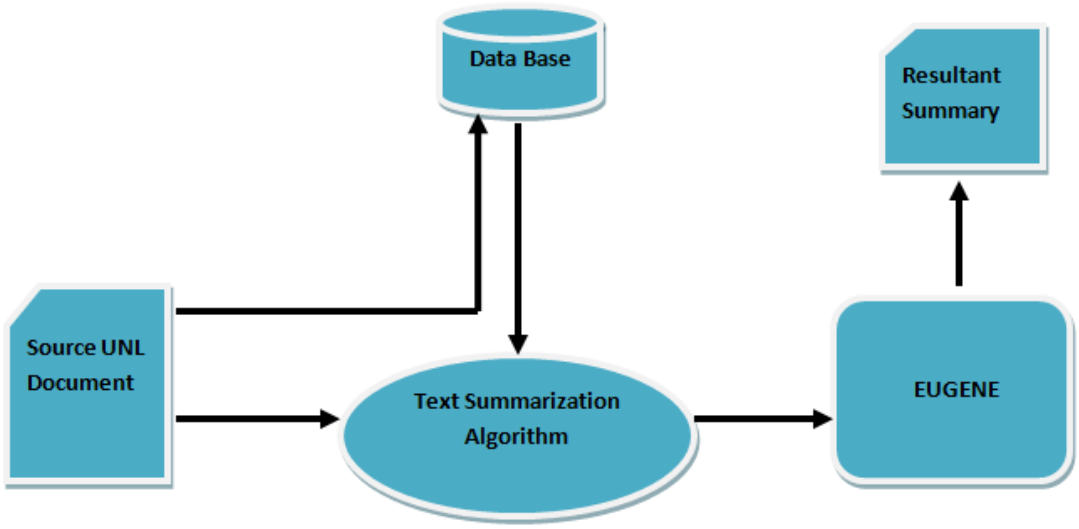


Figure 4.1: Proposed Architecture of Multilingual Summarization System

4.1.1 Text Summarization Algorithm

The proposed algorithm 4.1 is used for the implementation of the system. It consists of six steps for the generation of the summary of UNL document. This algorithm is applied for single document summarization.

Description of algorithm:

Step 1: UNL Input File:

UNL file acts as an input document for text summarization algorithm. The file contains the UNL sentences consists of Universal words, Universal relations and attributes along with all the essential specifications.

Step 2: UNL Heuristics:

UNL Heuristics deals with the relations. UNL heuristics are set of rules applied to UNL document. There are two types of heuristic rule used as heuristic rule1 and heuristic rule2. There are different kinds of relations between two Universal words like "aoj", "rsn", "man", "mod" *etc*. Some relations are very important hence they are always the part of summary. Relations like "agt", "obj", "aoj" are very important relations because they convey about the main processes of the text document.

Algorithm 4.1: Text Summarization Algorithm

Step 1: Input UNL File
 	
Step 2: Apply two UNL Heuristics
• Independent binary relations
If(frequency (UW2)) = =1)
Remove relation
Else
Relation should be a part of summary
 	
• Connected binary relations
If(frequency (UW2)) = =2)
Remove both relation

Else

Relations should be a part of summary

Step 3: Compute the score of the UNL sentences based upon

Weight of UNL sentence = Weight of Universal words * Weight of relation

Weight of Universal word= frequency * inverse document frequenc

Inverse document frequency = log(Total number of sentences / Number of sentences containing UW)

Step 4: Select the n best sentences for summary generation. The value of n depends upon

- The size of the document
- The level of abstraction

Step 5: Compute the value of contribution function for the modifier relations.

If value(Contribution function) > Threshold Value, then:

Sentence is a part of summary

Else:

Sentence is no longer part of summary document.

Step 6: Combine the two UNL sentences based upon

- Same Universal Word as head
- Less no of words in sentences

Step 7: Final processing to remove the unnecessary universal words.

There are some relations also which are not essential to be a part of summary. The heuristic rules are applied because summary is calculated for different domains like education, medical, tourism, science *etc*. In the tourism domain "plc" relation represents a particular place in UNL is very important but in other domains it can be simply ignored.

Similarly in case of critical real time scientific applications the value of "tim" relation (time) is very important but it can be not valuable for other summary domains. So, in the process of summary generation there is a need to remove those relations which are not important at that time. The removal of relations in the first step also decreases the overall processing time of the algorithm.

$$Exclude\ binary\ relation\ rel(UW_1, UW_2) from\ sentence\ S$$

$$if\ UW_2 \nexists\ other\ binary\ relations\ in\ S \quad \ldots (4.1)$$

According to rule described in (4.1), UW_1, UW_2 are Universal words of the binary relation "rel". The binary relation "rel" is removed from the document if Universal Word2 (UW_2) of binary relation does not exist in other sentences, *i.e.,* it is independent from other sentences with a frequency value 1. So the removal of this relation does not affect the efficiency of summary document. For example, a natural language and UNL sentence in given in the following:

The Tortoise had already won the race. … (4.2)

UNL: agt (win.@past.@perfect, tortoise.@def)

cnt (win.@past.@perfect, race.@def)

tim (win.@past.@perfect, already) … (4.3)

UNLization convert the sentence (4.2) to (4.3). In sentence (4.3) "tim" relation is not necessary because "The Tortoise had won the race" is itself a complete sentence. To remove "tim" relation here UW_2 should not be a part of any other relation. Here UW_2 = "already" which is not a part of any other UNL sentence, so according to (4.1) relation can be removed from UNL document.

Second rule is complex as compare to rule described in (4.1). Rule 2 is defined as following.

$$Exclude\ binary\ relation\ rel(UW_1, UW_2)\ and\ rel\ (UW_3, UW_2)\ from\ sentence\ S$$

$$if\ UW_2 \in S, \nexists\ other\ binary\ relations \quad \ldots(4.4)$$

Rule described in (4.4) removes more than one binary relation. This is applied in case of connected binary relations. According to (4.4) rule binary relations can be removed from a UNL sentence if Universal Word of a desired binary relation also appeared in other relation. In this scenario there is a need to remove two connected relations because if one relation is removed then empty space is created in the connected binary relation corresponding to that word. To remove this ambiguity of space that relation should also be removed. For example, consider the natural language and UNL sentence.

John went to university with the collage bus. … (4.5)

UNL: agt (go.@past, John)

plt (go.@past, university)

met (go.@past, bus.@def)

mod (bus.@def, collage) … (4.6)

In (4.6) sentence met is not necessary relation as summary point of view because "John went to university" is itself a complete sentence. Relation "met" contains two universal words "go" and "bus". The frequency of the UW_2 is "2" here. By removing this sentence from the UNL document an empty space contains in next sentence like mod ([ ], collage). To remove ambiguity this connected sentence also has to be removed from UNL document. Hence two binary relations met and mod removed here. These heuristic removes the relations in first step and hence they will not be the part of next upcoming steps. This decreases the overall complexity of the system.

Step 2: Score Computation:

In order to choose important sentences for the summary, score of each UNL sentence is calculated. It is numerical value calculated for every sentence with the help of

mathematical formula. Each UNL sentence consists of universal words and universal relations. Score of the UNL sentence is sum of all the weights of Universal words. The weight of a universal word is calculated by using the frequency, weight of the relations and inverse document frequency as represented by (4.7), (4.8) and (4.9). The term frequency is denoted by Tf and it is mostly used in automatic text summarization. The idea behind is the term with more frequency can better reflect the idea of document than the term with the less frequency value. Hence on the behalf of frequency more weight is provided to universal words which are more frequent. The inverse document frequency is based upon the fact that sometimes the universal words which are more frequent are less useful than those words which are less frequent. It is defined in (4.9). When both Tf and Idf are used together it is called Tf-Idf weighting. This weighting is represented by (4.8). The important thing to be noted is that weight is also provided to universal relations. Some relations are very important and should always be part of summary. To increase the score of those UNL sentences which has important relations more weight is provided to those relations. The mathematical formula is as follows:

$$S(s) = \sum Universal\ Word * Weight\ of\ relation \quad \ldots(4.7)$$

$$W(Universal\ Word_i) = Tf * Idf \quad \ldots(4.8)$$

$$Idf = \log\left(\frac{Total\ number\ of\ sentences\ in\ a\ UNL\ document}{Number\ of\ sentences\ in\ which\ UW\ appears}\right) \quad \ldots(4.9)$$

Where

S=Score computation function

s = UNL sentence whose score is computed

Tf = Frequency of Universal Word in UNL document

Idf = Inverse document frequency

W (Universal Word) = Weight of Universal Word

Weight of rel = Weight of relation in particular sentence

Table 4.1: UNL Corpus

UNL Corpus	
[S]	and(tortoise.@def,hare.@def)
[S]	[/S]
agt(reply.@past,tortoise.@def)	[S]
[/S]	agt(ridicule.@past,hare.@def)
[S]	cnt(ridicule.@past, :01)
and(:01.Null, :02.@although)	mod(:01.Null, tortoise.@def)
agt:01(beat.@future, 00.@1)	and:01(pace.@def, foot.@def.@pl)
obj:01(beat.@future, 00.@2)	mod:01(pace.@def, slow)
lpl:01(beat.@future, race.@indef)	mod:01(foot.@def.@pl, short)
aoj:02(swift, 00.@2)	[/S]
bas:02(swift, wind.@def.@equal)	[S]
[/S]	agt(start.@past, :01)
[S]	man(start.@past, together)
and(:02.Null, :01)	tim(start.@past,:01.Null)
exp:01(believe.@past, hare)	and:01(hare.@def, tortoise.@def)
cnt:01(believe.@past, :03)	obj:01(appoint.@past,day.@def.@topic)
agt:02(assent.@past, 00.@3)	pur:01(appoint.@past,race.@def)
cnt:02(assent.@past, proposal.@def)	[/S]
aoj:03(impossible, assertion)	[S]
man:03(impossible, simply)	agt(stop.@past.@not,tortoise.@def)
mod:03(assertion, tortoise.@def)	[/S]
[/S]	[S]
[S]	agt(go on.@past, tortoise.@def)
exp(agree.@past, 00.@3.@pl)	met(go on.@past, pace.@indef.@with)
obj(agree.@past, :01)	man(go on.@past, straight)
agt:01(:02.@decision.Null, fox.@def)	plt(go on.@past, end.@def)
and:02(choose, fix)	mod(end.@def, course.@def)
cnt:02(choose, course.@def)	mod(pace.@indef.@with, :01)

cnt:02(fix, goal.@def) [/S] [S] agt(win.@past.@perfect, tortoise.@def) cnt(win.@past.@perfect, race.@def) tim(win.@past.@perfect, already) [/S] [S] met(win, progress.@generic) cnt(win, race.@generic) mod(progress.@generic, :01) and:01(slow, steady.@but) [/S]	and:01(slow, steady.@but) [/S] [S] and(:01.Null, :02) agt:01(lay down.@past, hare.@def) plc:01(laydown.@past, wayside.@def.@by) exp:02(take a nap.@past, hare.@def) plc:02(takeanap.@past,tree.@indef.@under) [/S] [S] and(:01.Null, :02) and(:02.Null, :03.@but) exp:01(wake up.@past, hare.@def) man:01(wake up.@past, at last) agt:02(run.@past, 00.@3) man:02(run.@past, fast) bas:02(fast, :04.@equal) aoj:03(late.@extra.@past, 00) agt:04(run.@ability, hare.@def) [/S]

For example, consider a following natural language and UNL sentence represented in (4.10) and (4.11). The example sentence is taken from the UNL corpus 'Hare and Tortoise' provided by UNDL. The corpus is shown in Table 4.1.The overall weight of the sentence is sum of the weight of all sentences. The individual weight of the sentence is calculated by using formula (4.7). But first of all universal word are extracted from the UNL sentences by using the splitting logic.

The Tortoise had already won the race. … (4.10)

UNL: agt (win.@past.@perfect, tortoise.@def)

cnt (win.@past.@perfect, race.@def)

tim (win.@past.@perfect, already) … (4.11)

The weight of the first sentence is = (weight of the universal word "win" + weight of the universal word "tortoise") * weight of the relation "agt". Weights are assigned to different sentences based upon their priority level or importance. In UNL corpus the frequency value of the "win" is '5', "tortoise" is '5', "race" is '3' and 'already' is '1'. The total number of sentences in corpus is 13. The weight of "agt" is '10', "cnt" is '9' and "tim" is '2'.

Score of UNL sentence = (Score of "win" +Score of tortoise) * weight of "agt" +

(Score of "win" + Score of "race") * weight of "cnt"+

(Score of "win" + Score of "already")* weight of "tim"

Score of "win"= $5 * \log_2 (13/5) = 6.89$ … (4.12)

Score of "tortoise"= $5 * \log_2 (13/5) = 6.89$ … (4.13)

Score of "race"= $3 * \log_2 (13/3) = 6.34$ … (4.14)

Score of "already"= $1 * \log_2 (13/1) = 3.7$ … (4.15)

Score of UNL sentence = $(6.89 + 6.34)*10 + (6.89 + 6.34)*9 + (6.89 +3.7)*2$

$= 273.18$ … (4.16)

Hence score of the UNL sentence (4.11) is calculated by using the formulas (4.7), (4.8) and (4.10). At last the calculated score is represented in (4.16).

Step 3: Sentence Selection

After the computation of the score for the UNL sentences, sentences should be selected for the further summary. The sentences are arranged in descending order starting from the sentence with the highest score. Some top best sentences say n is chosen for further processing. The number of sentences that is n depends upon the two factors. First is size of the input document and secondly on the level of abstraction. If the size of the document is very large say 200 then value of n may be 50 but if document size in 20 lines then 7 or 8 sentences can be chosen for summary. It also depends upon the user also. If user wants a totally abstract view of document then value of n should be very less otherwise it is large. So, the value of n varies and it depends upon user.

Step 4: Remove redundant words:

Summary document still contain the redundant words which should be removed for efficient summary generation. Mostly modifier relations are removed. Modifiers can be easily found out by taking UNL semantic into consideration. Relations such as "man" (manner), "mod" (modifier) and "ben" (beneficiary) are modifier relations. In a UNL graph there is a head node which is just like a subject of natural language sentences and there are other nodes auxiliary nodes also. Auxiliary nodes can easily remove if they do not provide support in head clarification. Removal of auxiliary node does not destroy the real meaning of sentence and hence summary become more concise and efficient. The contribution provided by auxiliary node to the head node in UNL graph is determined by value of the contribution function. The formula for the contribution function is shown in (4.17). One threshold value is set. After calculating values of the contribution function for the modifiers values are compared with threshold value. If the auxiliary node has value of contribution function less than threshold value then it will be removed otherwise it will be a part of the summary document.

$$Con\big(rel(UW_1, UW_2)\big) = \frac{Weight(UW_1)}{Weight(UW_2)} \quad \ldots (4.17)$$

Where

Con = Contribution Function

rel = UNL modifier relation

Weight = Weight Computation Function

For example, consider the UNL sentences with value of contribution function. If in a summary document modifier relations are “mod”, “man” and threshold value is 1.25. The value of contribution function is determined by using the (4.17). In UNL corpus provided in Table 4.1, the frequency of ‘pace’ is ‘3’, ‘slow’ is ‘2’, ‘wake up’ is ‘2’ and ‘at last’ is ‘1’. The total number of sentences in corpus is 13.

(['mod' ,'pace', 'slow']) … (4.18)

(['man', 'wake up', ' at last']) … (4.19)

Weight (pace) = 3* $\log_2$ (13/3) = 6.34 … (4.20)

Weight (slow) = 2* $\log_2$ (13/2) = 5.4 … (4.21)

Weight (wake up) = 2* $\log_2$ (13/2) = 5.4 … (4.22)

Weight (at last) = 1* $\log_2$ (13/1) = 3.7 … (4.23)

Con (mod (pace, slow)) = (6.34 / 5.4) = 1.17 … (4.24)

Con (man (wake up, at last)) = (5.4 / 3.7) = 1.45 … (4.25)

In (4.18) and (4.19) the values of the contribution function are 1.17 and 1.45. The threshold of the (4.18) and (4.19) is compared with the threshold value 1.25. Now First value is less than threshold and second values is greater than threshold. Hence (4.18) can be removed from UNL summary (4.19) cannot be removed from UNL document.

Step 5: Combine Sentences:

To make summary more real and efficient different sentences of UNL can be combining for the generation of single UNL sentence. Sentences which contain same universal words as a head can easily be merged to reduce the total number of sentences in a

summary document. Small sentences in which numbers of words are very less can easily be merged.

For example, consider the two sentences (4.26), (4.27) which can be easily merged to form a (4.28) sentence.

English is a global language. … (4.26)

The language exists on network. … (4.27)

English is a global language existing on the network. … (4.28)

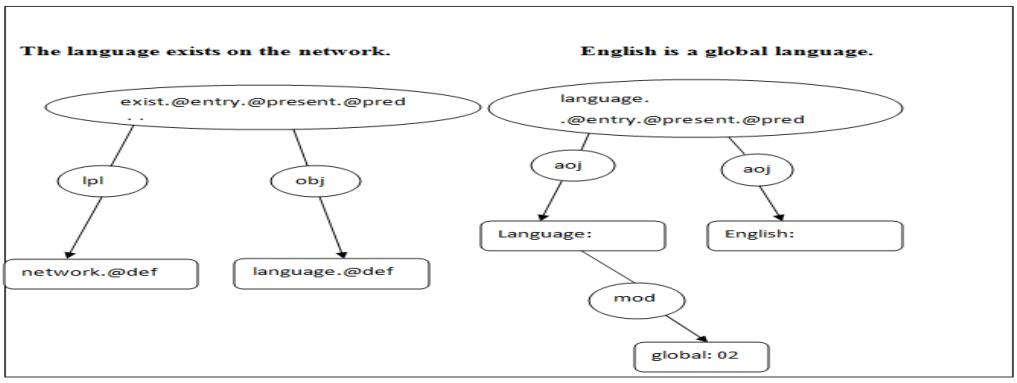


Figure 4.2: UNL graph for (4.26) and (4.27) sentences

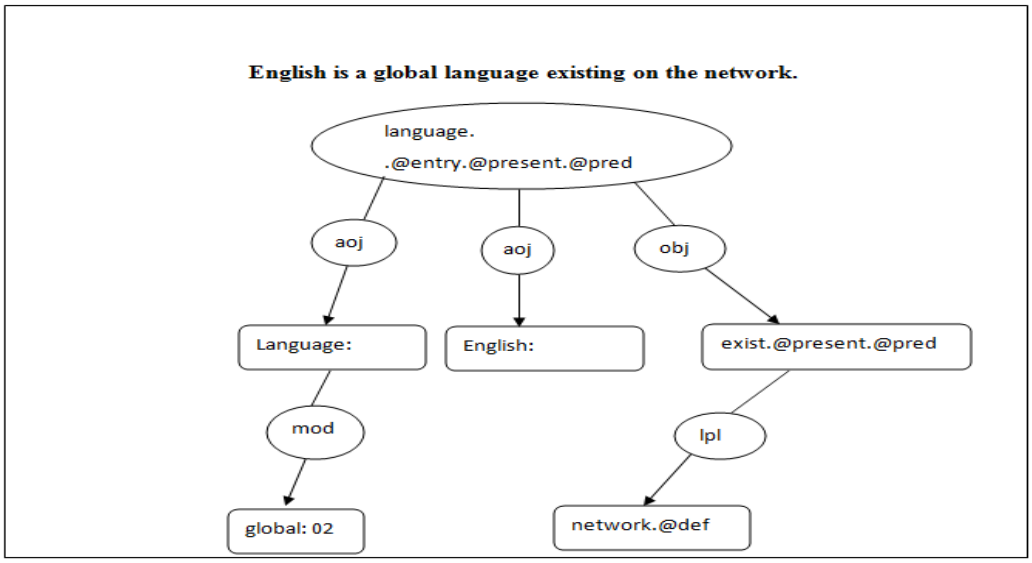


Figure 4.3: UNL graph for (4.28) sentence

Step 6: Further Removal:

In this obtained summary document is again processed and the words like “additionally”, ”because”, ”further”, “already” which are not important as summary point of view are removed from summary document.

Step 7: Output UNL File:

After processing all the steps of the summarization algorithm successfully summary is produced in the form of UNL file. The numbers of relations, universal words are now less as compare to source document. Now this output file is passed to DeConvertor module for language transformation.

4.1.2 EUGENE:

EUGENE is automatic tool for the UNL to natural language generation. It is language independent tool. It is open source NLizer developed by UNDL foundation [31]. EUGENE is actually a web application developed in java. It generates the output with the help of dictionary entries, transformation and disambiguation rules. UNL performs the process of NLization in three steps.

- Segmentation is a task of dividing the input UNL document into series of graphs which are then processed one by one.
- Tokenization is a process of finding out the tokens in a graph.
- Transformation is a process of generating transformation rules for the grammar which are required during the UNL to natural language conversion.

EUGENE architecture is described in Figure 4.6. The UNL input document is provided to EUGENE for the process of NLization. The UNL-NL dictionary is a database provided where the Universal words are mapped with the natural language entries. Entry is done for each Universal word and along with all the features. UNL-NL transformation grammar is a set of transformation rules which are required for the language translations. Disambiguation rules are also required in language transformations for improvement of results.

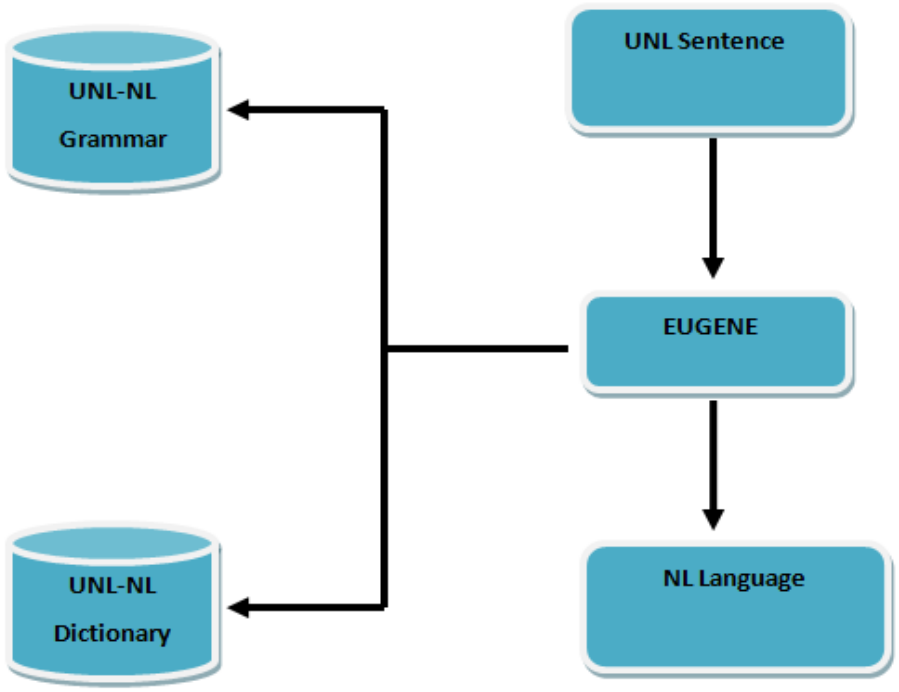


Figure 4.4: EUGENE Architecture

4.1.3 Database:

MySQL database is used in the system for storing the frequency values. The source UNL document is scanned and the frequency of the Universal word based upon the whole document and in the same UNL sentence is calculated. These two different frequency values are stored in a two different tables of SQL. These stored values are retrieved by the text summarization algorithm during the processing of the UNL heuristics.

4.1.4 Resultant Summary:

UNL summary document is provided to EUGENE for language transformations. After the language transformations of the EUGENE resultant summary document is produced which contain the summary in different natural language which is required by the user. This summary document is abstract of the source natural language document and conveys the same information as conveyed by the original document.

4.2 Python as an important tool for Text Summarization

Python is object oriented script language. Python has a very powerful approach to OOPs and high level effective data structures. It has a dynamic typing and refined syntax features which make python ideal language for script. The interpreter of python has extended the many functions and data types which are already implemented in C or C++ [32]. Due to the extension features of python it is suitable for mostly personalized applications.

Python has an important role in automatic text summarization. There are various modules and functions of python which are used to extract data and summary refine process. There are various modules provided by python for graphical user interface (GUI), *e.g*., *Kivy, PyQt, PyGUI, libavg*, *wxPython, JPython, Tkinter and EasyGUI etc*. Multilingual text summarizer uses *Tkinter* for the development of GUI. *Tkinter* provides the efficient and powerful object oriented interface to *TkGUI* toolkit. There are 15 types of control provided by *Tkinter*. A module *askopenfilename* is used to browse the desired files from the computer and *Urllib2* is used to retrieve data from the URLs. Different file functions like *File.open( ), file.seek( ), file.tell( ), file.readline( )* and *file.write( )* are used by the system for open the file, for go to particular location, to find the current location to read and write the data on file. Tkinter provides the *Font* function to change the size, underline style, weight, foreground and background colors of the text. *Tkmessagebox* is used to display the required information to user screen. The Method *re* is used for the regular expression operations. This module provides the regular expression same as Perl language. Other functions like *len( ), range( ), findall( ), split( )* and *strip( )* are also used by the system. The *mysqlconnector* is used by the system to establish the connection with database. The functions *mycusor.fetchall( )* is used to retrieve all the rows and *mycursor.execute( )* is used to run a particular query.

4.3 Working of the Proposed System

Multilingual text summarization system is used to produce the summary of text document. At each step summary is more refine than previous state and final output is produced at the end.

Step 1: Selection of the documents

A text document is selected whose summary is to be produced. The source document is selected with the *askopenfilename* method of python as shown in sentence (4.21) which can be used after including (4.19) and (4.20). The source text document is displayed after clicking the "source document" button. UNL document can be browsed with the help of "UNL document" button".Figure 4.7 and 4.8 shows the English document and UNL text document.

From *Tkinter* import * … (4.29)

From *Tkinter* import *Tk* … (4.30)

Import from *tkFileDialog* import *askopenfilename* … (4.31)

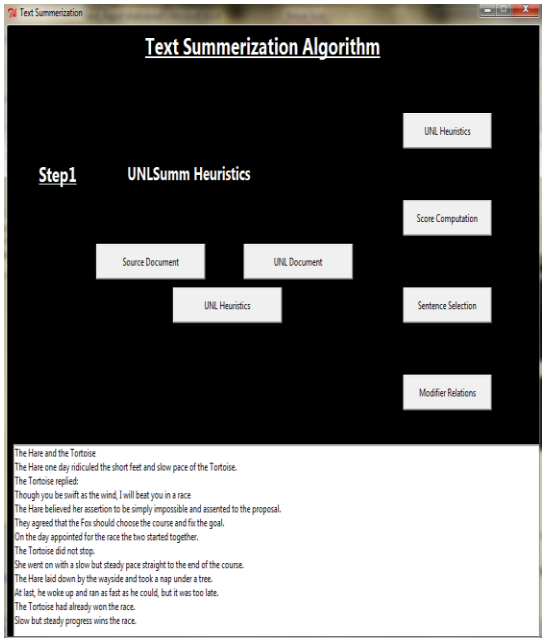


Figure 4.5: English language document

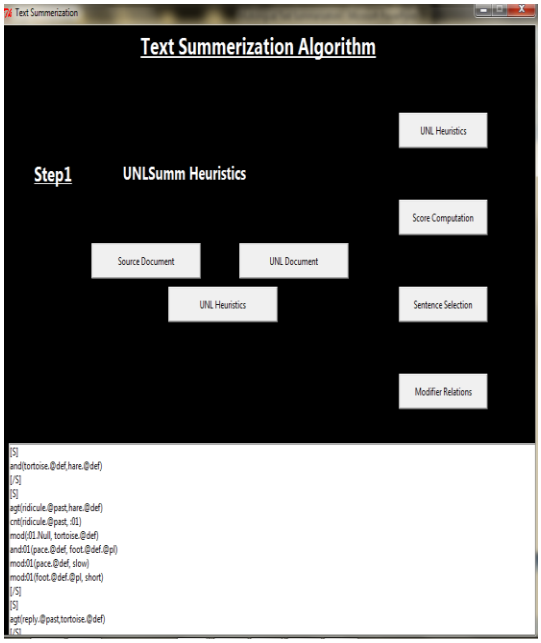


Figure 4.6: UNL document

Step 2: UNL Heuristics:

UNL heuristics are applied on the UNL document. In UNL heuristics data is extracted from the database and according to stored frequency values heuristics are applied. Connection with the database is established by using (4.32).

Import *mysql.connector* … (4.32)

Further the connectivity is done with the desired data base by using (4.33), (4.34) and (4.35). Afterwards the various query statements are executed according to requirement.

conn = mysql.connector.connect( ) … (4.33)

mycursor= conn.cursor( ) … (4.34)

mycursor.execute ( ) … (4.35)

Figure 4.9 and 4.10 frequency values of the universal words based upon the count in whole document and in a particular sentence.

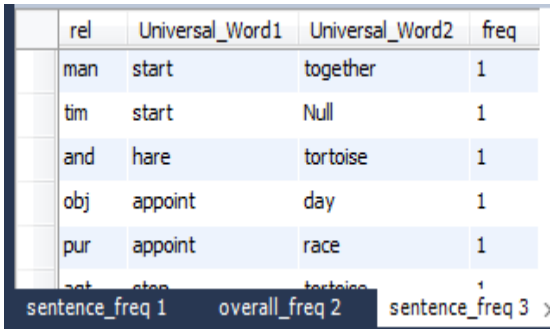


Figure 4.7: Frequencies in a sentence

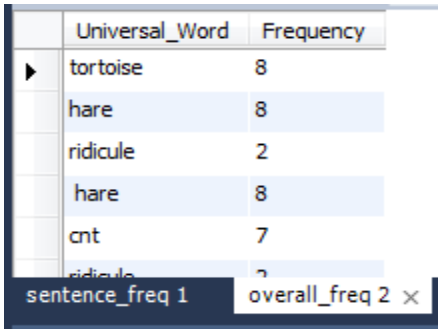


Figure 4.8: Overall Frequencies

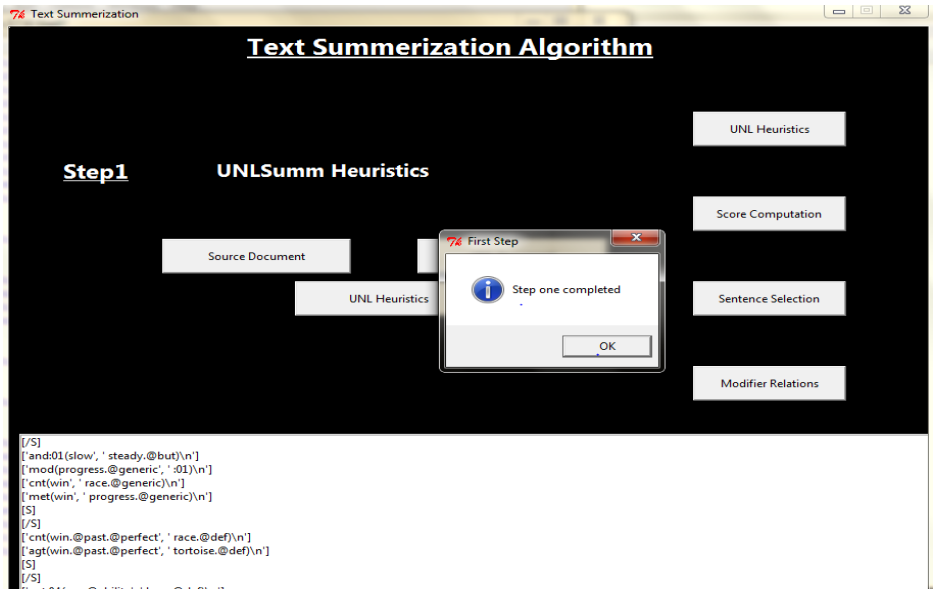


Figure 4.9: UNL heuristics

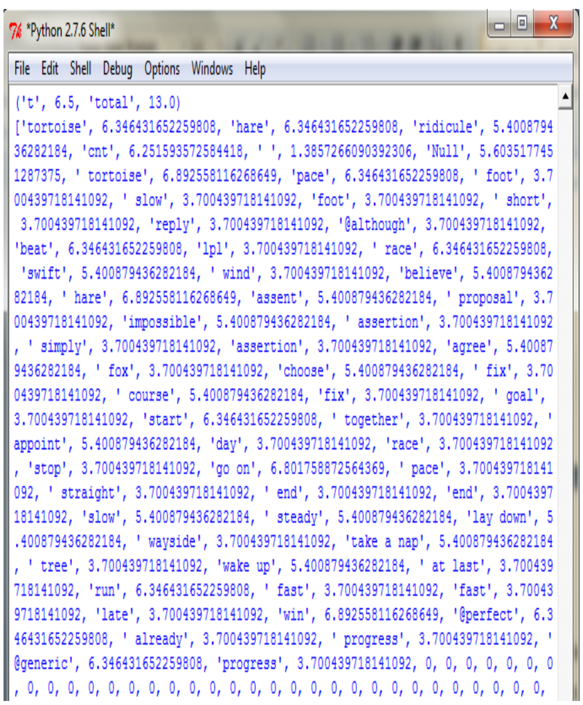
UNL heuristics are applied after fetching the desired values from the tables. Heuristic rules are applied after clicking “UNL Heuristics” button. The UNL sentences after applying UNL heuristics are shown in Figure 4.11.

Step 3: Score Computations of UNL sentences

Score of the individual universal words and UNL sentences is computed by using frequency and inverse document frequency. For the calculation of the score for universal words, UNL sentences are splitted on the basis of delimiters. Score computation function uses the *re.findall* for the regular expression operations. For all the regular expression operations sentence (4.36) is included.

Import *re* … (4.36)

Score is computed by clicking the “Compute Score” button. Figure 4.12 shows the universal words and their score in UNL document. The various UNL sentences and their score are shown in Figure 4.13.



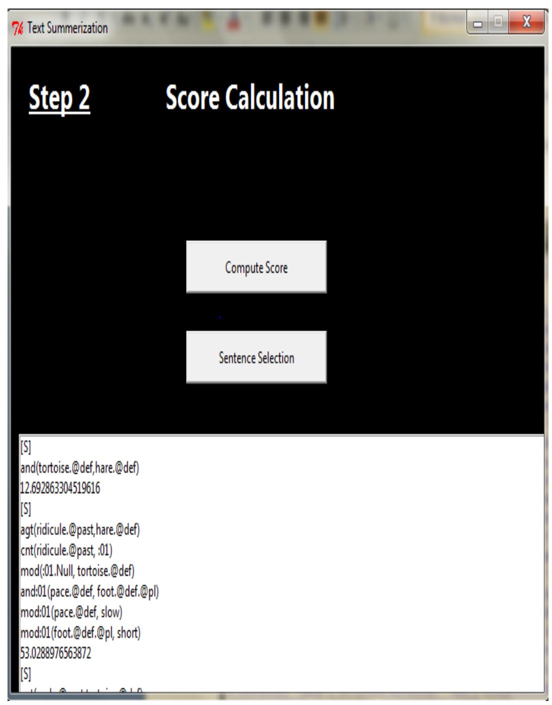


Figure 4.10: Universal Words and their score

Figure 4.11: UNL sentences and their score

Step 4: Selection of sentences for summary

Some best score sentences are selected for summary process. According to the score it is concluded that high score sentences convey important information of the text document. Hence these high score sentences are the part of the further summary process and they can be seen after clicking the "Sentence Selection" button. It returns some high score UNL sentences which are shown in Figure 4.14.

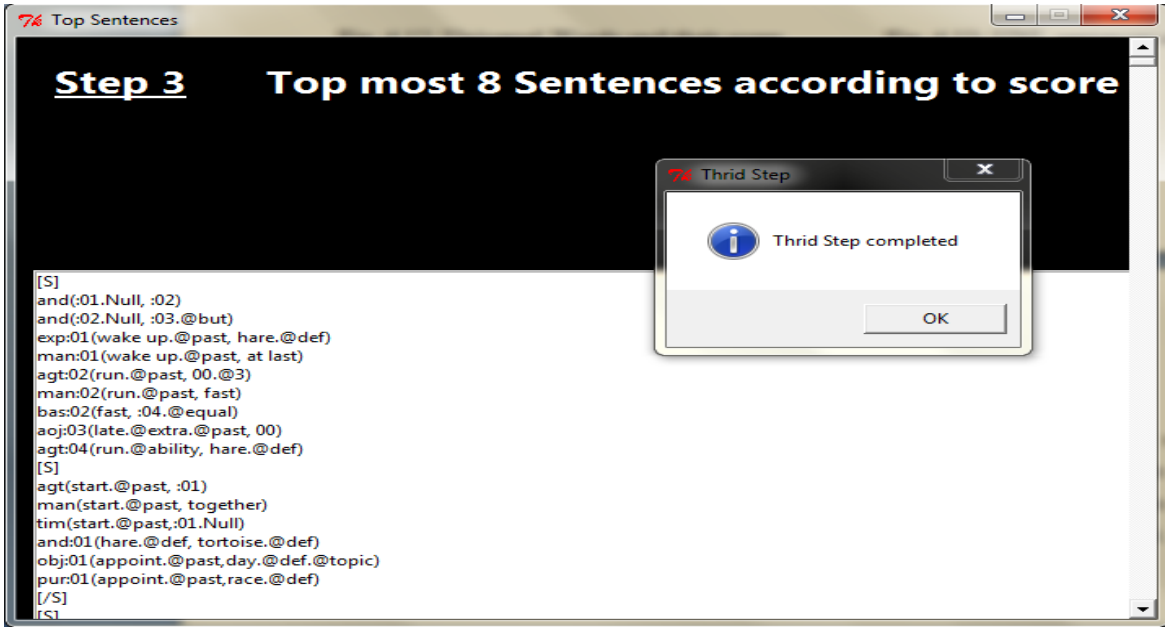


Figure 4.12: Top sentences according to score

Step 5: Role of Contribution function

Contribution function is applied on refined summary document. When the button "modifier relations" is clicked, then it returns the list of modifier relations as shown in Figure 4.16. On the basis of these modifier relations contribution function is called by clicking the "contribution function" button. It returns the value of the contribution function corresponding to modifier relations as shown in the Figure 4.15. The value of the contribution function is compared with the threshold value 1.25. The modifier relations with the contribution function value greater than threshold are the part of the further summary process and rest modifier relations are removed.

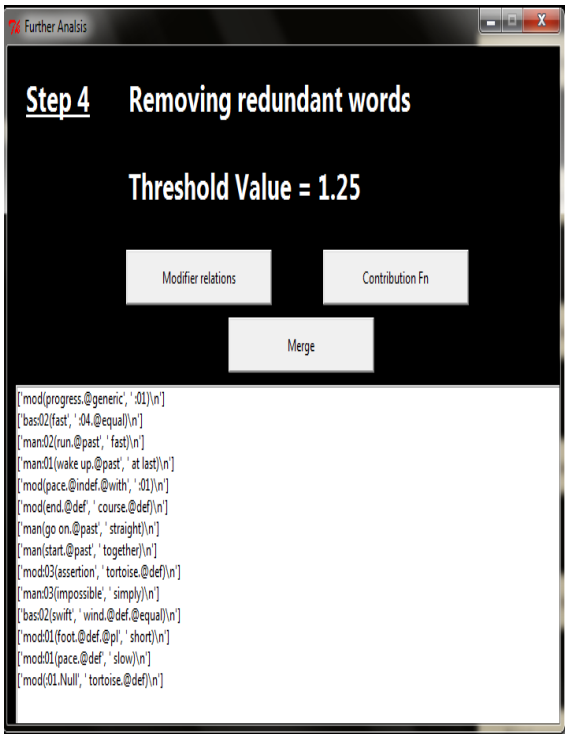


Figure 4.13: Contribution function values

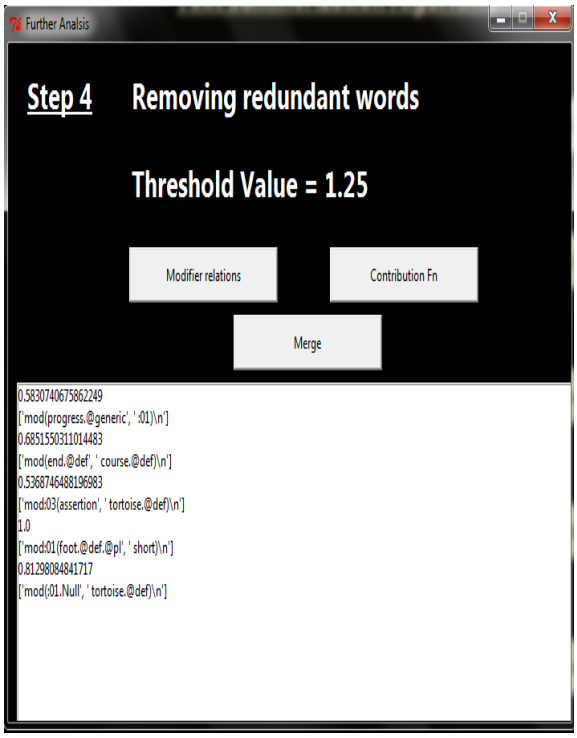


Figure 4.14: Modifier relations

Step 6: Combining UNL sentences

UNL sentences can be combined based upon some rules to form a new UNL sentence. When the recent UNL file is selected and "Merging" button is clicked, then it returns the new reduced UNL sentences which are shown in the Figure 4.17.

Figure 4.15: Merged Sentences

Step 7: Further Removal of Unnecessary words

The Universal words which do not convey important information can be removed in this step. Recent UNL file is selected and "Further Removal" button is clicked. It return the UNL sentences without unnecessary words like "additionaly","basically" *etc* which are shown in Figure 4.18. It is final summary produced by algorithm.

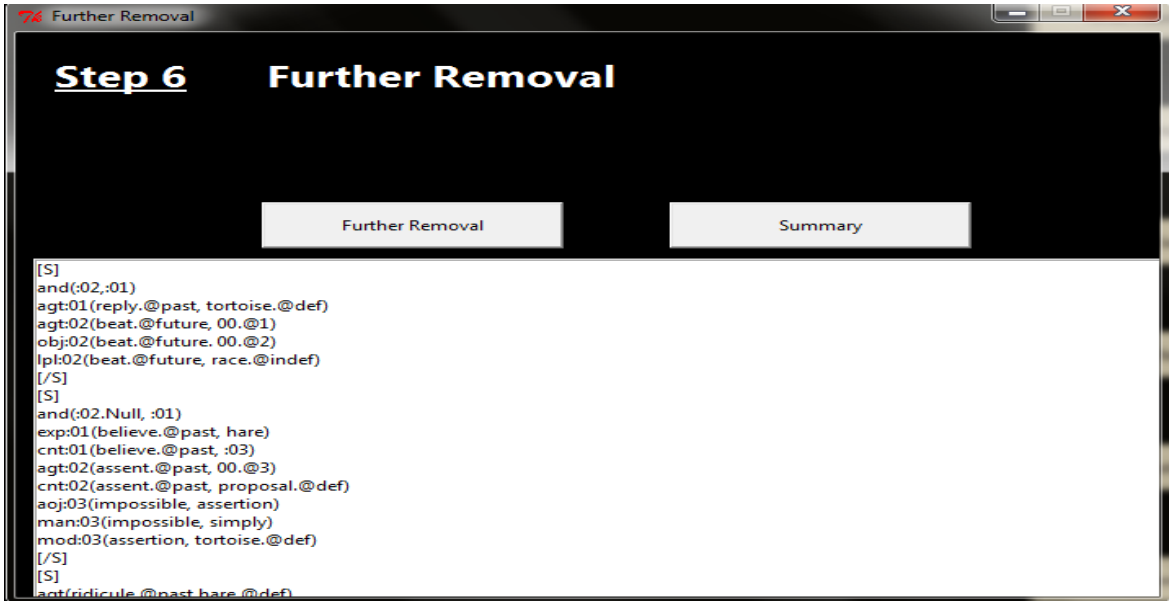


Figure 4.16: Further Removal in UNL summary

Step 8: UNL Summary conversion to natural language

UNL summary document is provided to EUGENE for conversion to desired natural language. In EUGENE dictionary rules and transformation rules are written corresponding to the UNL sentences. Dictionary entries which represent the universal word and its attributes like part of speech, genders *etc* are shown in the Figure 4.19. Transformation rules which are also responsible for language conversion are shown in Figure 4.20. The final outcome in other natural language is shown in the Figure 4.21.

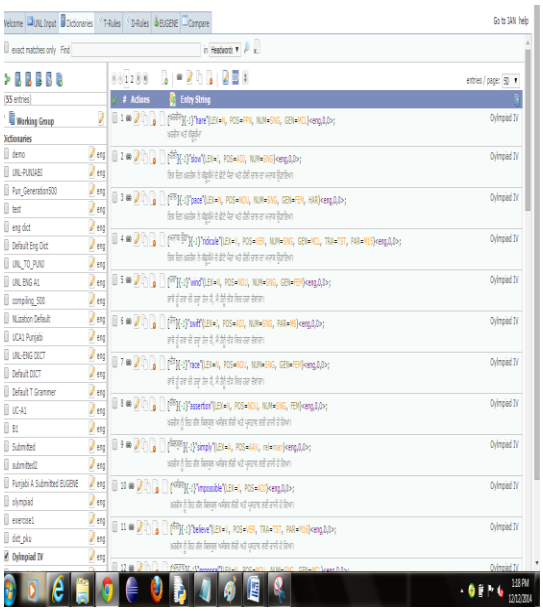


Figure 4.17: Dictionary Rules

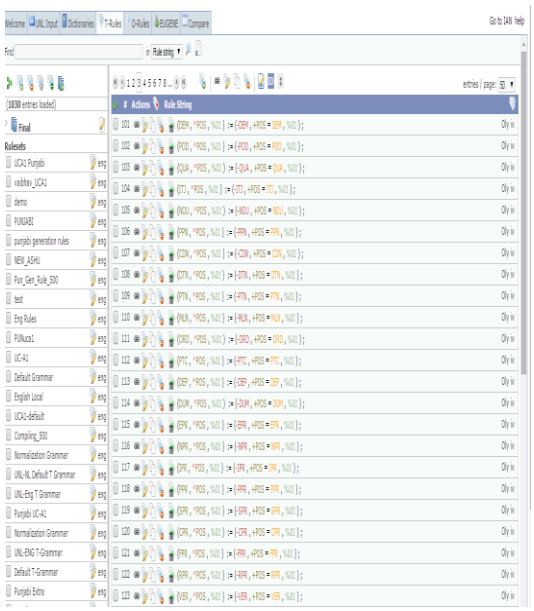


Figure 4.18: Transformation Rules

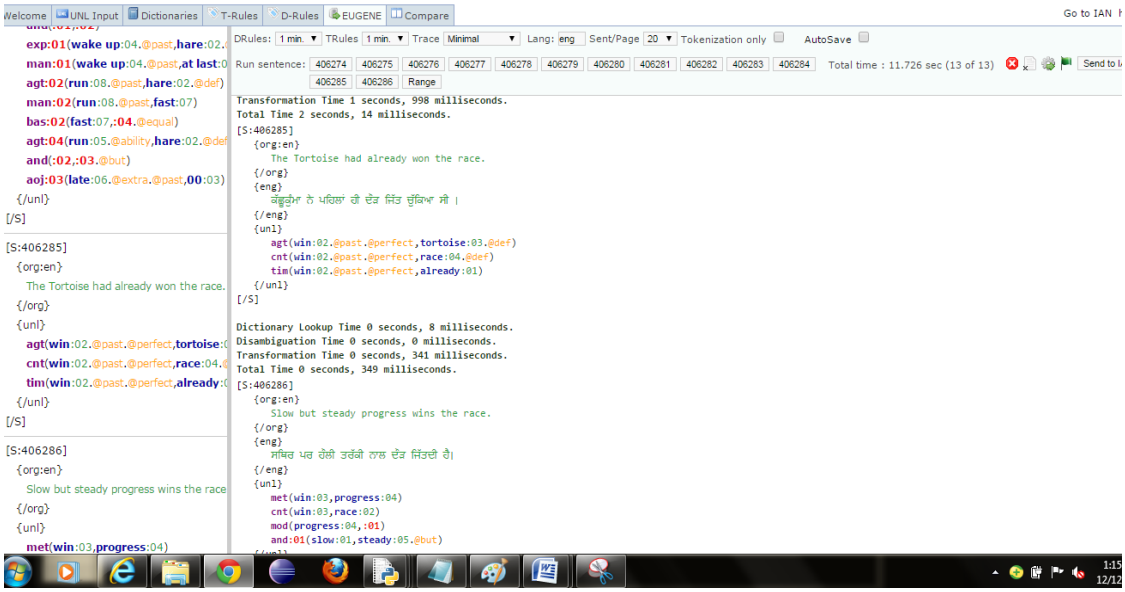


Figure 4.19: Final Outcome

Chapter Summary

The proposed system has been explained in this chapter along with all components. The six step algorithm described in proposed system has been explained. The role of python along with its methods and functions has also described. The proposed system produces the summary of the UNL document by applying the UNL heuristics. In second step score

of UNL sentences in calculated and some best sentences are chosen for further process. Later on Contribution function is applied on the best sentences and merges them according to requirement. Further analysis of reduced document is done for efficient summary. The output UNL summary is provided to EUGENE for conversion into desired language transformation.

Chapter-5

Results and Discussions

5.1 UNL Heuristic Rules:

The proposed system has been tested on hare and tortoise corpus. The corpus is shown in Table 4.1. The results after applying the UNL Heuristics are shown in Table 5.1.

Table 5.1: Results of UNL Heuristics given by proposed system

Sr. No	Before Applying UNL Heuristics	After Applying UNL Heuristics
1.	The Hare one day ridiculed the short feet and slow pace of the Tortoise. agt(ridicule.@past,hare.@def) tim(ridicule.@past,oneday) obj(ridicule.@past,:01) mod(:01,tortoise.@def) and:01(pace,foot.@pl) mod:01(pace,slow) mod:01(foot.@pl,short)	The Hare ridiculed the short feet and slow pace of the Tortoise. agt(ridicule.@past,hare.@def) obj(ridicule.@past,:01) mod(:01,tortoise.@def) and:01(pace,foot.@pl) mod:01(pace,slow) mod:01(foot.@pl,short)
2.	The Tortoise had already won the race. agt(win.@past.@perfect, tortoise.@def) cnt(win.@past.@perfect, race.@def) tim(win.@past.@perfect, already)	The Tortoise had won the race. agt(win.@past.@perfect, tortoise.@def) cnt(win.@past.@perfect, race.@def)

5.2 Score of Universal Words

The score of the different universal words of the document is calculated. The computed score is shown in the Table 5.2.

Table 5.2: Score of the Universal Words given by proposed system

Sr. No.	Universal Word	Score
1.	Tortoise	6.34
2.	Hare	6.34
3.	Ridicule	5.40
4.	Pace	6.34
5.	Foot	3.70
6.	Slow	3.70
7.	Short	3.70
8.	Reply	50.0
9.	Beat	6.34
10.	Race	6.34
11.	Swift	5.40
12.	Wind	3.70
13.	Believe	5.40
14.	Assent	5.40
15.	Proposal	3.70
16.	Impossible	5.40
17.	Assertion	3.70
18.	Simply	3.70
19.	Agree	5.40
20.	Fox	3.70
21.	Choose	5.40
22.	Fix	3.70

23.	Course	5.40
24.	Goal	3.70
25.	Start	6.34
26.	Together	3.70
27.	Appoint	5.40
28.	Day	3.70
29.	Race	3.70
30.	Stop	3.70
31.	Go on	6.80
32.	Straight	3.70
33.	End	3.70
34.	Steady	5.40
35.	lay down	5.40
36.	Wayside	3.70
37.	take a nap	5.40
38.	Tree	3.70
39.	wake up	5.40
40.	at last	3.70
41.	Run	6.34
42.	Fast	3.70
43.	Late	3.70
44.	Win	6.89
45.	Already	3.70
46.	Progress	3.70

5.3 Score of Sentences

The proposed system calculates the score of individual UNL sentences. The English sentences their UNL and score is shown in Table 5.3.

Table 5.3: Score of the UNL sentences given by proposed system

Sr. No.	English Sentence	UNL Sentence	Score
1.	The Hare and the Tortoise.	and(tortoise.@def,hare.@def)	12.69
2.	The Hare one day ridiculed the short feet and slow pace of the Tortoise.	agt(ridicule.@past,hare.@def) cnt(ridicule.@past, :01) mod(:01.Null, tortoise.@def) and:01(pace.@def, foot.@def.@pl) mod:01(pace.@def, slow) mod:01(foot.@def.@pl, short)	53.02
3.	The Tortoise replied:	agt(reply.@past,tortoise.@def)	56.34
4.	Though you be swift as the wind, I will beat you in a race.	and(:01.Null, :02.@although) agt:01(beat.@future, 00.@1) obj:01(beat.@future, 00.@2) lpl:01(beat.@future, race.@indef) aoj:02(swift, 00.@2) bas:02(swift, wind.@def.@equal)	36.18
5.	The Hare believed her assertion to be simply impossible and assented to the proposal.	and(:02.Null, :01) exp:01(believe.@past, hare) cnt:01(believe.@past, :03) agt:02(assent.@past, 00.@3) cnt:02(assent.@past, proposal.@def) aoj:03(impossible, assertion) man:03(impossible, simply) mod:03(assertion, tortoise.@def)	58.03

7.	On the day appointed for the race the two started together.	agt(start.@past, :01) man(start.@past, together) tim(start.@past,:01.Null) and:01(hare.@def, tortoise.@def) obj:01(appoint.@past,day.@def.@topic) pur:01(appoint.@past,race.@def)	43.07
8.	The Tortoise did not stop.	agt(stop.@past.@not,tortoise.@def)	10.04
9.	She went on with a slow but steady pace straight to the end of the course.	agt(go on.@past, tortoise.@def) met(go on.@past, pace.@indef.@with) man(go on.@past, straight) plt(go on.@past, end.@def) mod(end.@def, course.@def) mod(pace.@indef.@with, :01) and:01(slow, steady.@but)	52.43
10.	The Hare lay down by the wayside and took a nap under a tree.	and(:01.Null, :02) agt:01(lay down.@past, hare.@def) plc:01(lay down.@past, wayside.@def.@by) exp:02(take a nap.@past, hare.@def) plc:02(take a nap.@past, tree.@indef.@under)	32.08
11.	At last, he woke up and ran as fast as he could, but it was too late.	and(:01.Null, :02) and(:02.Null, :03.@but) exp:01(wake up.@past, hare.@def) man:01(wake up.@past, at last) agt:02(run.@past, 00.@3) man:02(run.@past, fast) bas:02(fast, :04.@equal) aoj:03(late.@extra.@past, 00)	40.43

		agt:04(run.@ability, hare.@def)	
12.	The Tortoise had won the race.	agt(win.@past.@perfect, tortoise.@def) cnt(win.@past.@perfect, race.@def)	36.43
13.	Slow but steady progress wins the race.	met(win, progress.@generic) cnt(win, race.@generic) mod(progress.@generic, :01) and:01(slow, steady.@but)	45.42

5.4 Best Score Sentence

The proposed system selects the sentences for the on basis of score. The proposed system select top 9 sentences for summary they are shown in Table 5.4.

Table 5.4: Best sentences selected by proposed system

Sr. No	**English Sentence**	**Score**
1.	The Hare believed her assertion to be simply impossible and assented to the proposal.	58.03
2.	The Tortoise replied:	56.34
3.	The Hare one day ridiculed the short feet and slow pace of the Tortoise.	53.02
4.	She went on with a slow but steady pace straight to the end of	52.43

	the course.	
5.	Slow but steady progress wins the race.	45.42
6.	On the day appointed for the race the two started together.	43.07
7.	At last, he woke up and ran as fast as he could, but it was too late.	40.43
8.	The Tortoise had won the race.	36.43
9.	Though you be swift as the wind, I will beat you in a race.	36.18

5.5 Contribution Function Values

The Contribution function is calculated for the modifier relations like man, mod, bas *etc*. The values of the contribution function computed by proposed system are shown in the Table 5.5. The value of the contribution functions are compared with threshold value 1.25. The modifier relations more than threshold value are part of summary document.

Table 5.5: Contribution Function values computed by proposed system

Sr. No	**Modifier Relation**	**Contribution Function values**

1.	mod(:01, tortoise .@def)	0.81
2.	mod: 01(foot. @def.@pl, short)	1.00
3.	mod:03(assertion, tortoise .@def)	0.53
4.	mod(end .@def', course .@def)	0.68
5.	mod(progress .@generic, :01)	0.58
6.	man:03(impossible, simply)	1.45
7.	mod:01(pace .@def', slow)	1.71
8.	man(go on .@past, straight)	1.83
9.	mod(pace .@indef .@with, :01)	4.57
10.	man(start .@past , together)	1.71

11.	man:01(wake up .@past, at last)	1.45
12.	man:02(run .@past, fast)	1.71
13.	bas:02(fast, :04.@equal)	2.67
14.	bas: 02(swift, wind. @def. @equal)	1.45

5.6 Merging Results

The proposed system combines the different sentences to form a new sentence. The merging result is shown in Table 5.6.

Table 5.6: Merged sentences by proposed system

Sr. No	**Sentence1**	**Sentence 2**	**Merged Sentence**
1.	The Tortoise replied: agt (reply. @past, tortoise. @def)	Though you be swift as the wind, I will beat you in a race. and(:01.Null, :02.@although) agt:01(beat. @future, 00.@1) obj:01(beat. @future, 00.@2) lpl:01(beat. @future, race. @	The Tortoise replied that I will beat you in a race. and(:02,:01) agt:01(reply.@past, tortoise. @def)

		indef) aoj:02(swift, 00.@2) bas:02(swift, wind. @def. @equal)	agt:02(beat.@future, 00.@1) obj:02(beat.@future. 00.@2) lpl:02(beat.@future, race. @ indef)

5.7 Summary Results:

The proposed system calculates the output summary in the form of UNL sentences. The summary output in natural language and UNL is shown in Table 5.7

Table 5.7: Summary produced by proposed system

Summary in English Language	The Hare ridiculed the short feet and slow pace of the Tortoise. The Tortoise replied that i will beat you in a race. The Hare believed her assertion to be simply impossible and assented to the proposal. They agreed that the Fox should choose the course and fix the goal. On the day appointed for the race the two started together. She went on with a slow but steady pace straight to the end of the course. He woke up and ran as fast as he could, but it was too late. The Tortoise had won the race. Slow but steady wins the race.	
	[S] agt(ridicule.@past,hare.@def) cnt(ridicule.@past, :01) mod(:01.Null, tortoise.@def) and:01(pace.@def, foot.@def.@pl) mod:01(pace.@def, slow) mod:01(foot.@def.@pl, short)	[S] and(:02.Null, :01) exp:01(believe.@past, hare) cnt:01(believe.@past, :03) agt:02(assent.@past, 00.@3) cnt:02(assent.@past, proposal.@def) aoj:03(impossible, assertion)

Summary in UNL	[/S] tortoise.@def) [S] exp(agree.@past, 00.@3.@pl) obj(agree.@past, :01) agt:01(:02.@decision.Null, fox.@def) and:02(choose, fix) cnt:02(choose, course.@def) cnt:02(fix, goal.@def) pur:01(appoint.@past,race.@def) [/S] [S] agt(go on.@past, tortoise.@def) met(go on.@past, pace.@indef.@with) man(go on.@past, straight) hare.@def) plt(go on.@past, end.@def) mod(end.@def, course.@def) mod(pace.@indef.@with, :01) and:01(slow, steady.@but) [/S] [S] agt(win.@past.@perfect, tortoise.@def) cnt(win.@past.@perfect, race.@def) [/S]	man:03(impossible, simply) mod:03(assertion, [/S] [S] agt(start.@past, :01) man(start.@past, together) and:01(hare.@def, tortoise.@def) obj:01(appoint.@past,day. @def.@topic) [/S] [S] and(:01.Null, :02) and(:02.Null, :03.@but) exp:01(wake up.@past, man:01(wake up.@past, at last) agt:02(run.@past, 00.@3) man:02(run.@past, fast) bas:02(fast, :04.@equal) aoj:03(late.@extra.@past, 00) agt:04(run.@ability, hare.@def) [/S] [S] met(win, progress.@generic) cnt(win, race.@generic) mod(progress.@generic, :01) and:01(slow, steady.@but)

	[/S]
Summary in Punjabi	ਖ਼ਰਗੋਸ਼ ਨੇ ਕੱਛੂਕੰਮੇ ਦੇ ਛੋਟੇ ਪੈਰਾ ਅਤੇ ਹੋਲੀ ਚਾਲ ਦਾ ਮਜਾਕ ਉਡਾਇਆ।ਕੱਛੂਕੰਮੇ ਨੇ ਜਵਾਬ ਦਿਤਾ ਮੈ ਤੈਨੂ ਦੋੜ ਵਿਚ ਹਰਾ ਦਵਾਗਾ।ਖ਼ਰਗੋਸ਼ ਨੂੰ ਇਹ ਗੱਲ ਬਿਲਕੁਲ ਅੰਸਭਵ ਲੱਗੀ ਅਤੇ ਪ੍ਸਤਾਵ ਲਈ ਰਾਜ਼ੀ ਹੋ ਗਿਆ। ਉਹਨਾ ਨੇ ਮੰਨਿਆ ਕਿ ਲੂੰਬੜੀ ਪਥ ਚੁਣੇਗੀ ਅਤੇ ਟੀਚਾ ਮਿਥੱਗੀ। ਦੋੜ ਲਈ ਨਿਰਧਾਰਤ ਦਿਨ ਦੋਵਾ ਨੇ ਇੱਕਠੇ ਸ਼ੁਰੂ ਕੀਤਾ। ਉਹ ਹੋਲੀ ਪਰ ਲਗਾਤਾਰ ਪਥ ਤੇ ਅਖੀਰ ਤਕ ਸਿੱਧਾ ਜਾਂਦਾ ਰਿਹਾ। ਆਖਰ ,ਉਹ ਉੱਠਿਆ ਅਤੇ ਦੋੜਿਆ ਜਿੰਨੀ ਤੇਜ਼ ਉਹ ਦੋੜ ਸਕਦਾ ਸੀ,ਪਰ ਇਹ ਬਹੁਤ ਦੇਰ ਸੀ।ਕੱਛੂਕੰਮਾ ਦੋੜ ਜਿੱਤ ਚੁੱਕਿਆ ਸੀ। ਹੌਲੀ ਪਹ ਲਗਾਤਾਰ ਤਰੱਕੀ ਦੋੜ ਜਿੱਤਦੀ ਹੈ।

5.8 Analysis of Proposed Algorithm

Efficiency of proposed algorithm depends upon the abstraction of summary document, *i.e,* depends upon the number of words left in summary as compare to total number of words in source document. The efficiency of proposed algorithm has been calculated by using the Table 5.8.

Table 5.8: Algorithm analysis

Parameters	**Source Document**	**Summary Document**
Number of Sentences	13	8
Number of Words	139	102

Percentage of abstraction = (102/139) * 100

= 73.38 %

= 100 – 73.38

= 26.619%

The proposed algorithm when applies to UNL corpus provided in Table 4.1 produce a summary document which is 73.38 % of the original document. Hence, the proposed algorithm does 26.619% abstraction.

Chapter-6

Conclusion and Future Scope

6.1 Conclusion

Multilingual Text Summarization is a process of reducing the content of text document with the help of computer program and providing the summary in other natural language. The summary document contains the important information of original document. The volume of data has been increased from decade. The social media and other information sources have the huge contribution to increase the volume of data. The management of huge volume data becomes important task. So, a Multilingual Text Summarization system has been proposed. The proposed system provides the summary in different natural language than source. The main aim of the proposed system is to provide the reduced important information to the people which are still away from technologies and understand their language only. The proposed system used intermediate language UNL for transformation of source language to destination language. The system helps in computing the summary of English document in Punjabi language.

To produce the summary of the English document/ Punjabi document, the UNL file is provided to proposed system. The system computes the score of the UNL sentences. On the basis of score system considers the important sentences for summary. At last, final summary is produced in UNL using text summarization algorithm. The final transformation of UNL summary to desired natural language is done by EUGENE.

The proposed system has been tested on hare and tortoise corpus. The system selects the 9 sentences for summary according to the scores. It is observed that system has correctly identified important sentences for summary generation.

6.2 Limitations and Future Scope

Some limitations of the proposed system and future scope to resolve limitations are as follows.

- The proposed system produced extractive summaries only. It can be extended to produce abstractive summaries also.
- Merging of simple sentences has been done on rule bases. The complex sentences can also be merged by adding new rules.
- The proposed system is not web based application. So, it can be make web based in future.
- The system worked for simple UNL sentences. In future it can be extended for complex UNL based documents.
- Usually long sentences are high score sentences. Some efforts are done to include the short sentences. In future more rules can be added to include the short sentences which are very important.
- The system can be extended by adding IAN module in future.

References

[1] Lloret, Elena, and Manuel Palomar. "Resúmenes de textos: nuevos retos en la Web 2.0." *Subjetividad y procesos cognitivos* 14.2 (2010): 113-126. Mangairkarasi, S., and S. Gunasundari. "Semantic based text summarization using universal networking language." *Int. J. Appl. Inf. Syst* 3.8 (2012): 18-23.

[2] Das, Dipanjan, and André FT Martins. "A survey on automatic text summarization." *Literature Survey for the Language and Statistics II course at CMU* 4 (2007): 192-195.

[3] Lal, Partha, and Stefan Ruger. "Extract-based summarization with simplification." *Proceedings of the ACL*. 2002

[4] "Universal Words," [Online]. Available" http://www.unlweb.ney/wiki/Universal_Words.

[5] "Universal Attributes," [Online]. Available" http://www.unlweb.ney/wiki/Universal_Attributes.

[6] "Universal Relations," [Online]. Available" http://www.unlweb.ney/wiki/Universal_Relations.

[7] "Universal Relations," [Online]. Available" http://www.unlweb.ney/wiki/Universal_Relations.

[8] Dev.undlfoundation.org,. 'UNL Platform - Analysis'. N.p., 2015. Web. 21 May 2015.

[9] Dev.undlfoundation.org,. 'UNL Platform - Analysis'. N.p., 2015. Web. 21 May 2015.

[10] Das, Dipanjan, and André FT Martins. "A survey on automatic text summarization." *Literature Survey for the Language and Statistics II course at CMU* 4 (2007): 192-195

[11] Das, Dipanjan, and André FT Martins. "A survey on automatic text summarization." *Literature Survey for the Language and Statistics II course at CMU* 4 (2007): 192-195

[12] Cheung, Jackie CK. *Comparing Abstractive and Extractive Summarization of Evaluative Text: Controversiality and Content Selection*. Diss. UNIVERSITY OF BRITISH COLUMBIA, 2008.

[13] Gupta, Vishal, and Gurpreet Singh Lehal. "A survey of text summarization extractive techniques." *Journal of Emerging Technologies in Web Intelligence*2.3 (2010): 258-268.

[14] Karmakar, Lad, and Chothani Hiten. "A Review Paper on Extractive Techniques of Text Summarization." (2015).

[15] Gong, Yihong, and Xin Liu. "Generic text summarization using relevance measure and latent semantic analysis." *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2001.

[16] Lal, Partha. "Text Summarization." (2002).

[17] Baxendale, Phyllis B. "Machine-made index for technical literature: an experiment." *IBM Journal of Research and Development* 2.4 (1958): 354-361.

[18] Edmundson, Harold P. "New methods in automatic extracting." *Journal of the ACM (JACM)* 16.2 (1969): 264-285.

[19] Martins, Camilla Brandel, and Lucia Helena Machado Rino. "Revisiting UNLSumm: Improvement through a case study." *the Proceedings of the Workshop on Multilingual Information Access and Natural Language Processing*. Vol. 1. 2002.

[20] Mangairkarasi, S., and S. Gunasundari. "Semantic based text summarization using universal networking language." *Int. J. Appl. Inf. Syst* 3.8 (2012): 18-23.

[21] Kalpana, S. "UNL based Document Summarization based on Level of Users."*International Journal of Computer Applications* 66.24 (2013).

[22] Sornlertlamvanich, Virach, Tanapong Potipiti, and Thatsanee Charoenporn. "UNL Document Summarization." *Proceedings of the First International Workshop on Multimedia Annotation*. 2001.

[23] Conroy, John M., and Dianne P. O'leary. "Text summarization via hidden markov models." *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2001.

[24] Karmakar, Lad, and Chothani Hiten. "A Review Paper on Extractive Techniques of Text Summarization." (2015).

[25] Conroy, John M., and Dianne P. O'leary. "Text summarization via hidden markov models." *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2001.

[26] Das, Dipanjan, and André FT Martins. "A survey on automatic text summarization." *Literature Survey for the Language and Statistics II course at CMU* 4 (2007): 192-195.

[27] Das, Dipanjan, and André FT Martins. "A survey on automatic text summarization." *Literature Survey for the Language and Statistics II course at CMU* 4 (2007): 192-195.

[28] Osborne, Miles. "Using maximum entropy for sentence extraction."*Proceedings of the ACL-02 Workshop on Automatic Summarization-Volume 4*. Association for Computational Linguistics, 2002.

[29] Kaikhah, Khosrow. "Automatic text summarization with neural networks." (2004).

[30] Svore, Krysta Marie, Lucy Vanderwende, and Christopher JC Burges. "Enhancing Single-Document Summarization by Combining RankNet and Third-Party Sources." *EMNLP-CoNLL*. 2007.

[31] Dev.undlfoundation.org,. 'UNL Platform - Analysis'. N.p., 2015. Web. 23 May 2015.

[32] Python.org, 'Welcome To Python.Org'. N.p., 2015. Web. 23 May 2015.

Publications

Paper Published

Sherry, Parteek Kumar," Multilingual Text Summarizer" in **"***International Conference on Advances in Computer Engineering and Applications (ICACEA) IMS Engineering College, Ghaziabad, India 2015.*

Paper Communicated

Sherry, Parteek Kumar, "Multilingual Text Summarization approach for Punjabi language" in "*ICON-2015: 12th International conference on Natural Language Processing*", IIITM-K, Trivandrum, India.

You Tube Video Link

A video describing the “Multilingual Text Summarization System” has been uploaded at www. youtube.com. The URL of the same is as follows:

https://www.youtube.com/watch?v=n6nW8khdR5I&feature=youtu.be

Reflective Diary

The research work of this semester is an extension of major project. My guide Dr. Parteek Bhatia suggested me to study regarding 'Automatic Text Summarization'. The Project was based upon the *Multilingual Text Summarization* which take Universal Networking Language (UNL) document as input and Punjabi language as output. Due to the huge volume of information the interest in the automatic text summary systems has been increased. Te proposed algorithm was based upon the efficient features of the existing algorithms and some new features also; which decrease the complexity and increase the efficiency of algorithm. It was a challenge to implement an algorithm which produces summaries in efficient way and in timely manner.

In third semester detail study of the EUGENE was carried out to understand the concept of NLization in UNL. The dictionary rules, transformation rules were understand and then some rules are written for simple fifty sentences. With the help of EUGENE UNL document was converted into Punjabi language.

January

In starting of the 4th semester there is detailed study on the plain text summarization techniques and multilingual text summarization techniques. The process of automatic summarization was started in 1958. The various systems were based upon plain text summarization. The different approaches like machine learning, clustering, semantic and syntactic are used to produce summary of text document.

Because the multilingual systems were based upon the UNL hence there is a need to understand the UNL in detail. The detail study of the UNL system had carried out with the help of "UNL Web". The website provided the whole information about the intermediate language. UNL sentences were based upon the universal words, relations and attributes. The Universal words represent the noun, pronoun of the natural language

sentences. The Universal relations specify the linking between the universal words. The Universal attributes represent the emotions of the sentences. The various UNL relations, UNL attributes were studied to understand the concepts.CUP-250 certification of UNL had done in this month. EUGENE system again studied in detail to learn the language transformations.

Based upon the previous research work of the 3rd semester and initial work of the 4th semester I wrote a research paper at the end of January. The paper named "Multilingual Text Summarization" had written and sends to international IEEE conference for reviews.

February

Multilingual text summarization with UNL techniques has been studied in detailed with their procedures. Multilingual summarization concept was started in 2005. After studying the detailed procedures of the summaries with UNL techniques it was concluded that previous existing algorithms calculate score on the basis of frequency, inverse document frequency *etc*. Usually the algorithms were based upon on or two language heuristics rules.

I came up with the idea that relation of Universal words in a UNL document was also very important. It should also be a part of score computation. So, the important relations always be the part of the summary document and unnecessary relations can be neglected easily by providing low score value. Each relation was studied again to understand its importance in UNL sentence. Based upon their importance in the sentence and in corpus score was provided to each relation. It was analyzed that same relations had different importance in different corpus.

The all existing algorithms were again analyzed and important properties noted down. A new algorithm was proposed based upon the existing important properties and with new features. The new added feature was score based upon relations. Other heuristic rules were also considered.

My guide helped me a lot during proposing a new algorithm. He gave me new ideas to add new features to algorithm.

At the end of month result paper had accepted in the international IEEE conference, Gaziabad.

March

In this month proposed system is manually tested on 'Hare and Tortoise' corpus. All the score computations and contribution functions were applied manually. On manual basis summary was produced and it was analyzed that system work well by adding new features. Then prerequisite of the proposed system had decided.

Python language was chosen for implementation due to its advance features and modules. Python GUI made with tkinter module. Various inbulid file functions provided by python were used to read, write, open and access files. Other functions like *len( ), range( ), findall( ), split( )* and *strip( )* are also used by the system.

My guide suggested me to add database to store the frequency values of Universal words. MySQL database was chosen. All the frequency values based upon document and sentences were stored here. The *mysqlconnector* is used by the system to establish the connection with database.

In this month I attended the international IEEE conference, at Gaziabad to present my research paper. The implementation of the system started in this month.

April

The implementation of the system continued in this month. The system was implemented step by step. The output after each step had shown in GUI; so that user can analyze the clear process of summary. The first implemented step was regarding the heuristic rules which remove the unnecessary relations from the UNL document.

The second step I implemented was regarding the score computation. In this step UNL sentences was splitted on the basis of delimiters. Then score of the individual Universal Words was computed using various factors like frequency, inverse document frequency, weight of relation *etc*. Finally score of the UNL sentences had computed.

The third implemented step was just to select the top best sentences for the summary. I select the top nine sentences for the corpus for summary.

Fourth step was regarding the contribution function values. The modifier relations were considered for that. I studied the modifier relations and their significance. The contribution function value is compared with threshold value.

To continue with fifth step i studied UNL sentences again in detail. Fifth was regarding the merging of sentences. Then fifth step was successfully implemented.

May

The last step of the algorithm was implemented. The UNL summary of a document was produced. The UNL summary document was provided to EUGENE for conversion to Punjabi language. Various dictionary entries corresponding to Universal words had written. Transformation rules had written which were responsible for conversion of UNL to Punjabi language.

The system was tested on the hare and tortoise corpus. The whole corpus contains 13 sentences. Top score 9 sentences were chosen for summary. At each step we got more and more refined summary.

At the end of month thesis writing had started. In Thesis introduction the need, challenges, application and approaches of proposed system documented. The role and importance of UNL also documented.

June

Thesis writing continued in this month. The detailed literature study about the existing plain text summarization and multilingual summarization techniques had documented. The problem statement with objectives and methodology had documented.

The proposed system architecture had explained in detail along with all prerequisites. The step by step implementation of algorithm had documented with examples. The screen shots of the implementation had shown in report.

Various results were discussed in thesis report. The conclusion was derived and suggestions for future extension documented.

A research paper on “Multilingual Text Summarization system for Punjabi” had written. The paper is still under “Communicated status”. It is communicated in the "ICON-2015: 12th International conference on Natural Language Processing”, IIITM-K, Trivandrum, India”.

Last but not least I would thank to my guide Dr. Parteek Bhatia for his support, positive attitude and for showing me always the right path. The whole research work is a result of his guidance and support.

Plagiarism Report