UNL based Document Summarization based on Level of Users

Lakshmana Pandian.S
Assistant Professor
Dept of CSE
Pondicherry Engg College

Kalpana. S
PG Scholar
Dept of CSE
Pondicherry Engg College

ABSTRACT

The primary goal is to develop an NLP system to perform automatic document summarization by converting the English sentences into the expressions of an Interlingua called Universal Networking Language (UNL). UNL has been designed at the United Nations University (UNU)/Institute of Advanced Studies (IAS), Tokyo in 1990. UNL represents knowledge in the form of semantic network. The nodes represent concepts and links represent semantic relations between the concepts. Initially, the complex document is represented in UNL form by undergoing enconversion process and the document is deconverted to produce the summarized document for different levels of users thus reducing the complexity of the document and helps in understanding and decision making. The aim of this NLP system is to represent the exact meaning of the document represented in a usual language. Since UNL is a language independent meaning representation language, summarization is carried out by analyzing and filtering out the UNL Expressions. UNL is used for summarization because if the user wants the document in his own native language, he can deconvert the UNL representation and obtain the source document and the summarized document in his own language. UNL supports multilingualism and reusability. The summarized document thus produced could be understood by all level of people without changing the meaning of the original document. This paper focuses on UNL based summarization for tourism domain.

Keywords

Natural Language Processing (NLP), Universal Networking Language (UNL), Enconverion, Deconversion, Parser.

1. INTRODUCTION

Natural Language Processing (NLP) is a field of computer science (CS), artificial intelligence (AI) and linguistics and it is an area of research and application that analyze how computers are used for understanding and manipulating natural language text or speech to achieve the desired tasks. The challenges in NLP system involve natural language understanding i.e. enabling computers to derive meaning from natural language text [1]. Summarization is a very interesting part and the complexity involved in it is to represent the exact meaning representation of the original document and another useful task is that it gives support to many other tasks as well [2]. Text Summarization is a challenging problem and one of the most commonly researched tasks in NLP is Automatic text summarization which produces a readable summary out of a chunk of text. When summarization is done by means of a computer, i.e. automatically, we call this as Automatic Text summarization. The goal of NLP researchers is to create an appropriate tools and techniques to make computer systems understand and manipulate the natural languages by extracting the knowledge and concept. We introduce our NLP system to retain the exact meaning of the document which is represented in Universal Networking Language (UNL) form.

2. LITERATURE SURVEY

There are frameworks like bridging Bangla language sentence to Universal Networking language, bridging UNL deconvertor for Tamil, UNL deconvertor for Malayalam and so on. This is an attempt of using the same UNL (Universal Networking Language) to perform text summarization. Text summarization process is carried out for complex sentences and the summarization is performed for the whole document. Md. Ershadul H. Choudhury, Nawab Yousuf Ali et al. proposed a framework [3] of bridging Bangla language sentence to Universal Networking Language. In this paper, the bangle UNL system has been developed and UNL representation is given for simple sentence and not for a document of complex sentences. T.Dhanabalan, T.V.Geetha proposed *UNL* deconverter *for Tamil* [4] which is an Interlingua approach to machine translation. In this paper, Universal Networking Language (UNL) has been used as the intermediate representation. This paper deals only with the DeConverter part where deconversion of UNL representation to Tamil language is performed. UNL representation is manually taken and the process is carried out. Bevin Sousheel Bhagianath, S.Lakshmana Pandian proposed UNL deconvertor for Malayalam [5] which performs deconversion of UNL expressions into Malayalam sentences. Dipanjan Das Andre F.T. Martins proposed a work on the survey of automatic text summarization. This discusses about various summarization techniques involved and their drawbacks of each summarization process. This paper discusses about many summarization techniques was available in the field of NLP from past years like Naive-Bayes methods, Rich Features and Decision Trees method, Hidden Markov Models, Log- Linear models, Deep Natural Language Analysis Methods. Usually, the flow of information in a given document is not uniform, which means that some parts are more important than others. The major challenge in summarization lies in distinguishing the more informative parts of a document from the less ones. Though there have been instances of research describing the automatic creation of abstracts, most work presented in the literature relies on verbatim extraction of sentences to address the problem of single-document summarization. Thus the text once converted into UNL can be specified in any different language provided with that particular deconvertor system thus supporting multilingualism [8]. The exact meaning is representation by the identifying the entities and relations between the entities for complex sentence is carried out. The meaning representation is directly available for retrieval and indexing mechanisms and tools for automatic summarization and knowledge extraction and it will be converted to a natural language only when communicating with a human user. UNL has the advantage of working instantly, working in any language, improves productivity and it does not miss important facts. So UNL is used for document summarization.UNL greatly reduces the cost of developing knowledge or contents necessary for knowledge processing, by sharing knowledge and contents. Thus we go for UNL (Universal Networking Language) for document summarization.

Information has to be readable and understandable by users without change in the concept of the original document. This work focuses on it [9]. Language is a major barrier in communication. Documents written in one language remain locked up for the people who don't know the language. Out of the eight hundred million population of India, nearly two thousand is currently excluded from participating actively in this so called information age. So a document written in one language can be represented in UNL an intermediate representation, which can be deconverted into any desired language as required by the user. Since it supports multilingualism the information expressed in UNL can be converted into the user's native language with higher quality and fewer mistakes than the computer translation systems (MT). In addition, UNL is free from ambiguities. Its real strength is to represent knowledge and information. Its primary objective is to serve as an infrastructure for handling knowledge in any given language.

Generally, text summarization is the process of extracting the important information from the document by leaving out the irrelevant information like "because, but, for, to, has been, had been, as, on, since, is, was et.al" and to reduce the details and collects them in a compressed way. This may not produce the exact meaning of the original document, instead produces the important words alone as the summarized output. But the UNL way of representation of the document gives the expressions, which will collect the exact meaning of the document including the exact meaning of the words like "but, for, since" and also considers the tenses like present, past, future into account for performing summarization. Here, by using the Stanford University Parser and by performing enconversion the complex sentences, lengthy sentences are given the UNL representation. This is an attempt of performing both enconversion and deconversion together and also performing summarization of document for various levels. The knowledge representation graph is generated to represent the knowledge of the whole document.

3. SYSTEM ARCHITECTURE

This NLP system performs Automatic text summarization by converting the complex English sentence into expressions of Interlingua called Universal Networking Language (UNL). UNL is an Interlingua based Machine Translation. The source document is thoroughly analysed and it is represented in an interlingual representation and it undergoes generation process to get the desired output. Fig 1 shows the UNL interlingual pyramid.

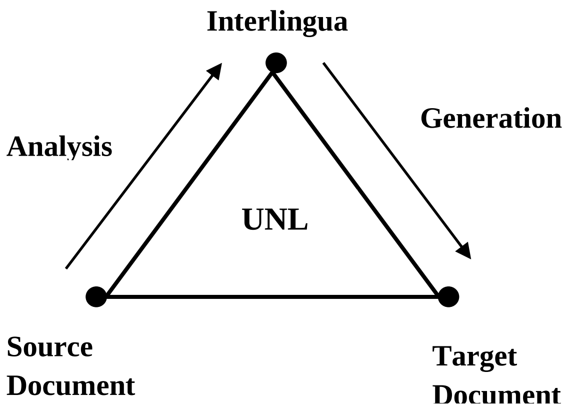

Fig 1: UNL Interlingua pyramid

To perform summarization process, the source document undergoes a process called "enconversion" which involves various steps like Parts-of-Speech Tagging, Parts-of-Speech Parsing, Entity Identification, Relation identification, building Dictionary, Generation rules thus producing UNL expressions as output. So the source document is sent to an enconversion phase where the analysis of the entire source document is carried out and by building the UNL dictionary, the UNL representations are obtained. Then these expressions are passed through the deconvertor and for the summarization process to produce the summarized document as output for three levels (level1, level2, level3) of users. The deconvertor deduces the expressions and produces the summarized document as output. The level of distribution of the summarized document is based on the IQ levels. Fig 2 shows the overall architecture of the system. Fig. 3 shows the architecture of the enconversion process. In the enconversion phase, the parts of Speech tagging and parsing process is carried out by Stanford Parser. The output of the parser is taken as the input so as to find the entities and relations between the entities. Then the rules are generated, knowledge base is constructed to produce the UNL expressions as output.

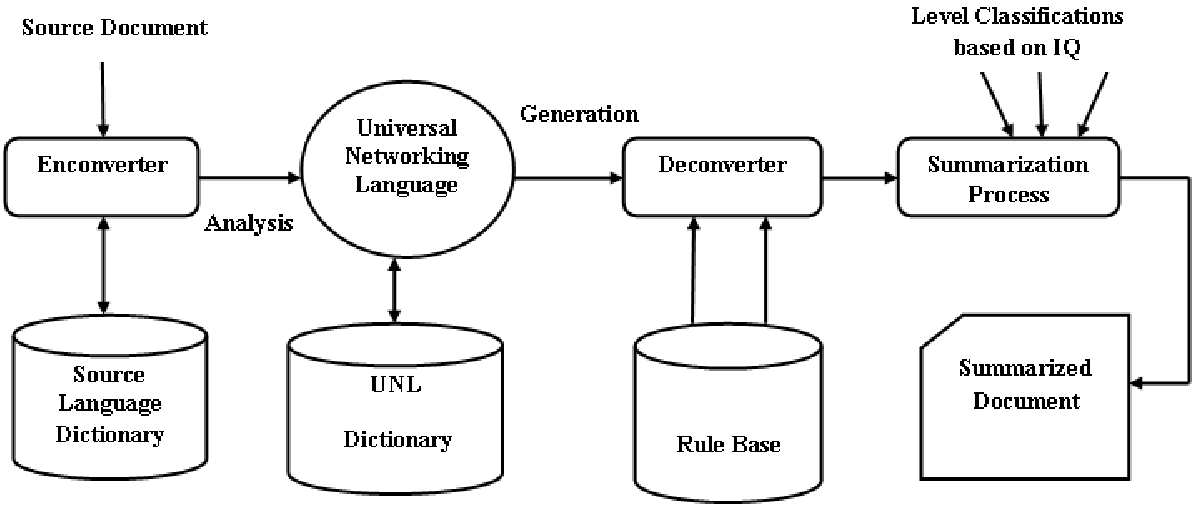

Fig 2: Overall System Architecture

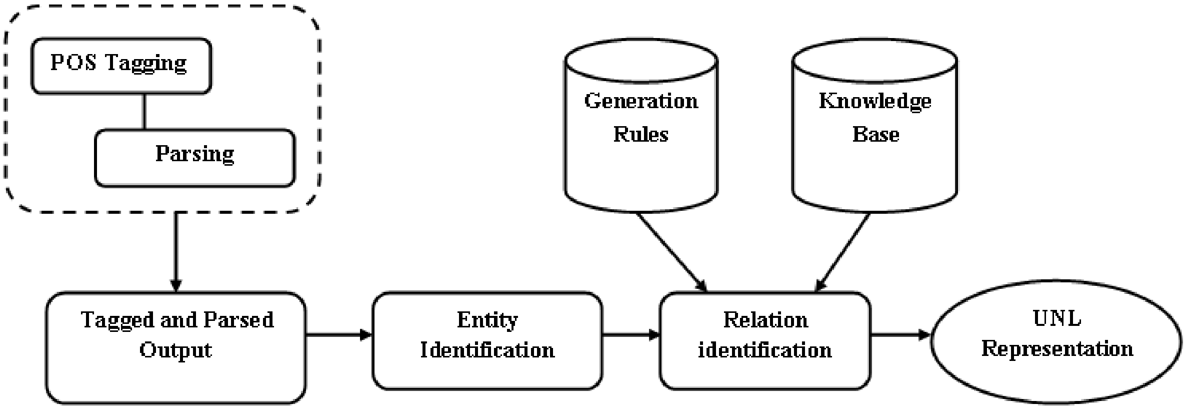

Fig 3: Enconversion process

3.1 Parts Of Speech Tagger and Parser

To perform enconversion process the first step is to do Parts of Speech tagging and parsing. A natural language parser is a program that works out with the grammatical structure of sentences, for instance, which groups of words go together (as "phrases") and which words are the subject or object of a verb. These are the parts of speech tags which are attached to each and every word in the document. For example: for the word “the” the POS tagging will be “(DT The)” where DT is Determiner and for the word “Aurobindo” the POS tagging results in “(NNP Aurobindo)” where NNP is proper noun, singular. Likewise, NNS stands for Noun, plural, CC is Coordinating Conjunction, and the numbers (like age, date, year so on...) will be represented as Cardinal number CD, JJ for adjective and the symbols will be represented as it is. Thus for the entire document the parts of tagging is performed.

The POS tagset for the main phrase includes ADJP for adjective phrase, ADVP for Adverb phrase, NP for Noun phrase, PP for prepositional phrase, S for Simple declarative clause, SBAR for Clause introduced by subordinating conjunction or 0, SBARQ for Direct question introduced by wh-word or wh-phrase, SINV for declarative sentence with subject-aux inversion, SQ sub constituent of SBARQ excluding wh-word or wh-phrase, VP for verb phrase, WHADVP for wh-adverb phrase, WHNP for wh-noun phrase, WHPP for wh-prepositional phrase. These are the possible tags the document can be tagged with. Each and every word in the document is tagged and parsed and the parsed output for the entire document is obtained as the output. In the parser, by loading the file and by selecting the parser the source document is subjected to parsing phase, where tagging and parsing is performed. The output of the document is saved, to perform further actions to produce the UNL representation as output. After obtaining the parsed output, the splitting of each word with its corresponding tag representation is performed, and the tags are associated with the UNL relations and the rules are obtained to find the entities and the relationship between the entities to get the UNL representation. The UNL representation is analyzed thoroughly and deduced to get the summarized document and this implementation is performed inNetBeans IDE 7.1. Fig 4 shows the parsed output of the tourism domain document, obtained after the parsing stage. Then this document undergoes the splitting process, splitting the tags with the corresponding words. The splitted tags with the words are corelated with the UNL relations, rules are formed and the UNL representations are obtained.

3.2 Phrase Structure Parse

```
(ROOT (NP (NP (NNP Sri) (NNP Aurobindo) (NNP Ashram)) (: :) (S (NP (DT The) (NNP Sri) (NNP Aurobindo) (NNP Ashram))
(VP (VBZ is) (NP (NP (DT a) (ADJP (RB well) (JJ known) (CC and) (JJ wealthy)) (NN ashram)) (PP (IN in) (NP (NNP India)))) (, ,)
(PP (IN with) (NP (NP (NP (NNS devotees)) (PP (IN from) (NP (NNP India)))) (CC and) (NP (NP (DT all)) (PP (IN over) (NP (NP
(DT the) (NN world)) (VP (VBG flocking) (PP (TO to) (NP (PRP it))) (PP (IN for) (NP (JJ spiritual) (NN salvation)))))))))))) (. .)))
(ROOT (S (NP (PRP$ Its) (JJ spiritual) (NNS tenets)) (VP (VBP represent) (NP (NP (DT a) (NN synthesis)) (PP (IN of) (NP (NN
yoga) (CC and) (JJ modern) (NN science))))) (. .)))
(ROOT (S (NP (DT The) (NNP Ashram)) (VP (VBD was) (VP (VBN founded) (PP (IN in) (NP (CD 1926))) (PP (IN by) (NP (NP
(NNP Sri) (NNP Aurobindo) (NNP Ghose)) (, ,) (NP (DT an) (JJ Indian) (NN freedom) (NN fighter)) (, ,) (NP (NN poet)) (, ,) (NP
(NN philosopher)) (, ,) (CC and) (NP (NNS yogi)))))) (. .)))
(ROOT (S (NP (NP (NNP Mirra) (NNP Alfassa)) (PRN (-LRB- -LRB-) (VP (ADVP (RB also)) (VBN known) (PP (IN as) (NP (DT
The) (NNP Mother)))) (-RRB- -RRB-))) (VP (VBD was) (NP (NP (CD one)) (PP (IN of) (NP (NP (NNP Aurobindos) (NNS
followers)) (, ,) (SBAR (WHNP (WP who)) (S (VP (VP (VBD was) (VP (VBN born) (PP (IN in) (NP (NNP Paris))))) (CC and) (VP
(VBD was) (VP (VBN inspired) (PP (IN by) (NP (PRP$ his) (NN philosophy))) (SBAR (IN that) (S (NP (PRP she)) (VP (VBD
stayed) (PRT (RP on)) (PP (IN in) (NP (NNP Pondicherry)))))))))))))))) (. .)))
(ROOT (S (PP (IN After) (NP (NNP November) (CD 24) (, ,) (CD 1926))) (, ,) (SBAR (WHADVP (WRB when)) (S (NP (NNP Sri)
(NNP Aurobindo)) (VP (VBD retired) (PP (IN into) (NP (NN seclusion)))))) (, ,) (NP (PRP she)) (VP (VBD founded) (NP (NP (PRP$
his) (NN ashram)) (PRN (-LRB- -LRB-) (NP (NNP Sri) (NNP Aurobindo) (NNP Ashram)) (-RRB- -RRB-))) (, ,) (PP (IN with) (NP
(NP (DT a) (NN handful)) (PP (IN of) (NP (NP (NNS disciples)) (VP (VBG living) (PP (IN around) (NP (DT the) (NN Master))))))))
(. .)))
(ROOT (S (PP (IN With) (NP (NP (NNP Sri) (NNP Aurobindo) (POS 's)) (JJ full) (NN approval))) (NP (PRP she)) (VP (VBD
became) (NP (NP (DT the) (NN leader)) (PP (IN of) (NP (NP (DT the) (NN community)) (, ,) (NP (NP (DT a) (NN position)) (SBAR
(S (NP (PRP she)) (VP (VBD held) (PP (IN until) (NP (PRP$ her) (NN death))))))))))) (. .)))
(ROOT (S (NP (NP (DT The) (NNP Sri) (NNP Aurobindo) (NNP Ashram) (NNP Trust)) (, ,) (SBAR (WHNP (WDT which)) (S (NP
(PRP she)) (VP (VBD had) (VP (VBN registered) (PP (IN after) (NP (NP (NP (NNP Sri) (NNP Aurobindo) (POS 's)) (NN death)) (PP
(IN in) (NP (CD 1950)))))))))) (VP (VBZ continues) (S (VP (TO to) (VP (VB look) (PP (IN after) (NP (DT the) (NN institution)))))))
(. .)))
(ROOT (S (NP (NP (NP (DT The) (NN idea)) (PP (IN of) (NP (NNP Auroville)))) (CC or) (NP (NP (DT the) (NNP City)) (PP (IN of)
(NP (NNP Dawn))))) (VP (VBD was) (VP (VBN conceived) (PP (IN by) (NP (DT The) (NNP Mother))))) (. .)))
(ROOT (S (S (NP (PRP It)) (VP (VBZ is) (ADJP (JJ open) (PP (TO to) (NP (NP (NP (DT the) (JJ public) (NN daily)) (PP (IN
between) (NP (CD 8) (RB a.m.)))) (CC and) (NP (NP (CD 12) (NN p.m.)) (CC and) (NP (CD 2) (NN p.m.)))))))) (CC and) (S (S (NP
(NP (CD 6) (RB p.m.) (NNP Children)) (PP (IN below) (NP (NP (CD 3) (NNS years)) (PP (IN of) (NP (NN age)))))) (VP (VBP are)
(RB not) (VP (VBN allowed) (PP (IN into) (NP (DT the) (NN ashram)))))) (CC and) (S (NP (NN photography)) (VP (VBZ is) (VP
(VBN allowed) (ADVP (RB only)) (PP (IN with) (NP (NP (NN permission)) (PP (IN of) (NP (DT the) (NN ashram) (NNS
authorities))))))))) (. .)))
(ROOT (S (S (NP (NP (DT Some)) (PP (IN of) (NP (DT the) (NN ashram)))) (VP (VBZ s) (NP (NP (NNS facilities)) (PP (IN like)
(NP (DT the) (NNP Library)))))) (CC and) (S (NP (DT the) (NNP Main) (NNP Building)) (PRN (-LRB- -LRB-) (PP (IN during) (NP
(JJ collective) (NN meditation))) (-RRB- -RRB-)) (VP (MD can) (VP (VB be) (VP (VBN accessed))))) (, ,) (PP (RB only) (IN after)
(S (VP (VBG obtaining) (NP (DT a) (NN gate) (NN pass)) (PP (IN from) (NP (NP (DT the) (NNP Bureau) (NNP Central)) (CC or)
(NP (NP (DT some)) (PP (IN of) (NP (DT the) (NNP Ashram) (NNP Guest) (NNP Houses))))))))) (. .)))
(ROOT (S (NP (PRP It)) (VP (VBZ is) (VP (VBN located) (PP (IN on) (NP (NNP Rue) (NNP De) (NNP La) (NNP Marine))))) (. .)))
(ROOT (S (NP (DT The) (NN ashram)) (VP (VBZ houses) (NP (NP (DT the) (NN samadhi)) (PP (IN of) (NP (NP (NNP Sri) (NNP
Aurobindo)) (CC and) (NP (DT the) (NN mother)))))) (. .)))
(ROOT (S (NP (PRP It)) (VP (VBZ is) (ADJP (JJ open) (PP (TO to) (NP (NP (NN everyone)) (PP (IN for) (NP (VBN fixed) (NNS
hours) (JJ everyday))))))) (. .)))
```

Fig 4: Phrase structure parser

On considering the complex sentence, The Aurobindo Ashram is a well known and wealthy ashram in India, with devotees from India and all over the world flocking to it for spiritual salvation. The knowledge representation graph is given in figure 5. The nodes represent the entities and the links represents the relationship between the entities. The relationship is found using the UNL relations [11] such as agt (agent), aoj (thing with attribute relation), ben (beneficiary), iof (an instance of), and (conjunction), frm (origin), pur (purpose) and so on. Certains rules are formed and the entities are related with each other using the relation tags. The knowledge representation is implemented using Net Beans IDE 7.1. Table 1 shows the Parts of speech tag set. All the entities in the sentence is being tagged with these specified tags.

Table 1: POS tagset

Tag	Description	Example	Tag	Description	Example
CC	Coordinating conjunction	and, but, or	SYM	Symbol (mathematical or scientific)	**+,%, &**
CD	Cardinal number	one, two, three	TO	to	To
DT	Determiner	a, the	UH	Interjection	ah, oops
EX	Existential there	there	VB	Verb, base form	Eat
FW	Foreign word	mea culpa	VBD	Verb, past tense	Ate
IN	Preposition/subordinating conjunction	of, in, by	VBG	Verb, gerund/present Participle	Eating
JJ	Adjective	yellow	VBN	Verb, past participle	Eaten
JJR	Adjective, comparative	bigger	VBP	Verb, non-3rd ps. sing. present	Eat
JJS	Adjective, superlative	wildest	VBZ	Verb, 3rd ps. sing. present	Eats
LS	List item marker	1,2, One	WDT	Wh-determiner	which, that
MD	Modal	can, should	WP	Wh-pronoun	what, who
NN	Noun, singular or mass	llama	WP$	Possessive wh-pronoun	Whose
NNS	Noun, plural	llamas	WRB	Wh-adverb	how, where
NNP	Proper noun, singular	IBM	**#**	Pound sign	**#**
NNPS	Proper noun, plural	Carolinas	**$**	Dollar sign	**$**
PDT	Predeterminer	all, both	**.**	Sentence-final punctuation	**(. ! ?)**
POS	Possessive ending	‘s	**,**	Comma	**,**
PRP	Personal pronoun	I, you, he	**:**	Colon, semi-colon (Mid-sentence punctuation)	**(: ; ... - _)**
PP$	Possessive pronoun	your, one’s	**(**	Left parenthesis	**([, (, { , <)**
RB	Adverb	quickly, never	**)**	Right parenthesis	**( ] ,) , }, >)**
RBR	Adverb, comparative	faster	**“**	Left quote	**(“ , ‘)**
RBS	Adverb, superlative	fastest	**”**	Right quote	**(” , ’)**
RP	Particle	up, off			

Figure 5 shows the knowledge representation for the first sentence in the document. The sentence is “The Aurobindo Ashram is a well known and wealthy ashram in India, with devotees from India and all over the world flocking to it for spiritual salvation.” Here the word Sri is NNP, Aurobindo is NNP, and Ashram is also NNP. So when two or more NNP comes together, it has to be junked together to produce a single entity. Now Sri Aurobindo Ashram is the name of the place and for summarization purpose it should be considered as three entities, instead it should be chunked and it should be taken as one single entity. For name of the person, it may have first name and last name, so it should not be considered as two words; instead it should be considered as one single entity. Likewise, certain rules are formed and the knowledge representation for the whole document is performed. Entire knowledge of the document is obtained in the knowledge representation where the parser tags and UNL relations are taken into consideration.

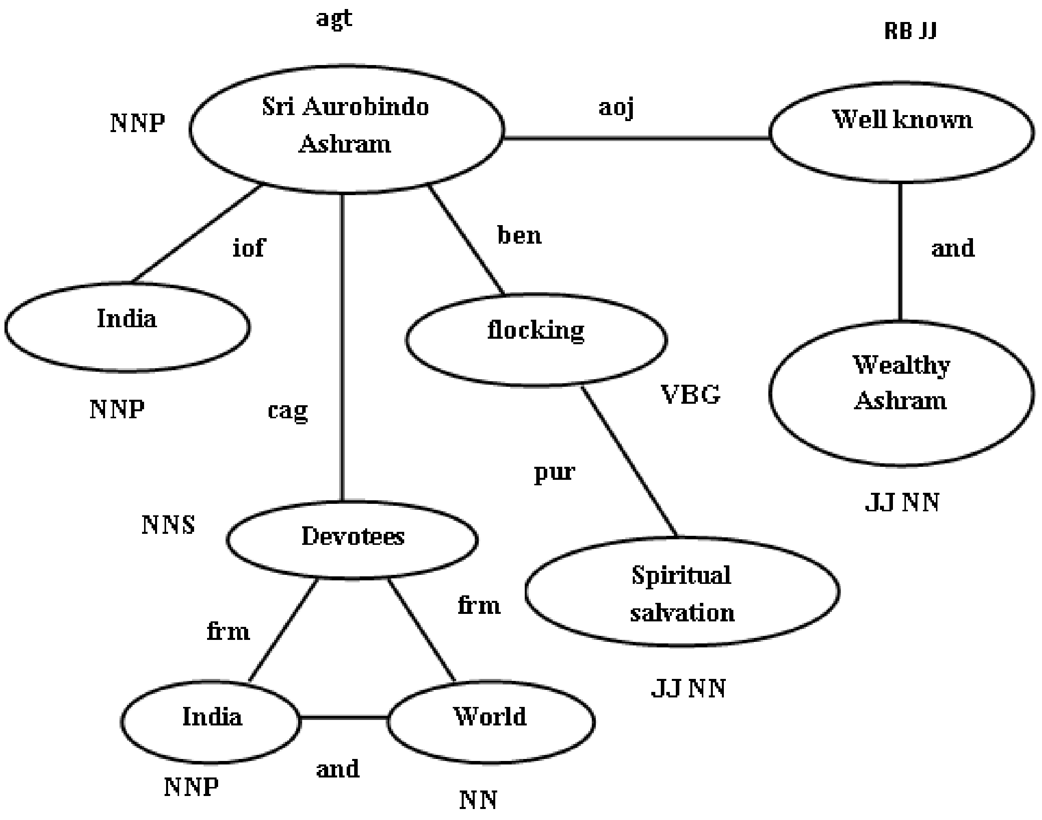

Fig 5: Knowledge Representation

3.3 Entity Identification and Relation Identification

UNL vocabulary consists of Universal Words, Relations and Attributes. Labels that represents word meaning (word knowledge) is universal words (UW's).Tags that represent the relationship between Universal Words (concept knowledge) are UNL relations [10]. Further definition or additional information (Speakers view aspect, time of event etc.) which appears in the sentence are considered as UNL attributes. After identifying the entities and relationship between the entities the UNL representation is obtained. A UNL expression for a sentence is enclosed by the tags {unl} and {/unl}. Thus, the UNL expression of a sentence will be in the format of opening and close tag as shown in figure 6.

```
{unl}
<Binary Relation>
...
{/unl}
```

Fig 6 UNL representation tags

3.3.1 Universal Words

Universal Words form the vocabulary of UNL. A UW is a character string (an English language word) followed by a list of constraints. The syntax used for describing a UW is:

<UW>::= <headword> [<constraint list>]

For example: aurobindo(icl>name), deepavail(icl>festival)', 'bhangra(icl>dance)', 'saag(icl>food)'.

3.3.2 UNL Knowledge Base

UNL Knowledge Base (UNLKB) defines every possible relation between concepts. UNLKB not only provides linguistic knowledge in the form that computer can understand but also provides the semantic background of UNL expressions, that is the UNLKB ensures the meanings of UNL expression

3.3.3 UNL Relations

Relations are basic building blocks of UNL sentences. UNL expressions are formed using binary relations and a binary relation includes a UNL relation and two UWs [11]. A UNL binary relation is of the form **rel (arg1, arg2)** where arg1 is UW1 and arg2 is UW2. There are 46 relations that are defined f for UNL. They are agt, and, aoj, bas, ben, cag, coo, pur, fmt, gol, icl, ins, man, met, plc, plf and so on. Some of the relation tags are shown in the table 3.

Table 2: UNL Relations

UNL relation	Description	Constituent elements	Examples
icl	Defines an upper concept or a more general concept	Included/a kind of	A bird is a (kind of) animal. icl(bird(icl>animal), animal(icl>living thing))

gol	defines a final state of object or a thing finally associated with the object of an event	Final state of verbs of change like ‘give’, ‘send’, etc.	… gave … to Mary.gol(gave, Mary) … sent … to Mary. gol(sent, Mary)

agt	Defines a thing that initiates an action	Agent, Unergative verbs (intransitive verb semantically have an agent subject) like ‘sleep’, ‘snore’, ‘cough’, ‘run’, etc.	John slept …
			agt(slept, John)
			John killed Mary.
			agt(killed, John)
			… arrival of John …
			agt(arrival, John)
			…play by Shakespeare
			agt(play, Shakespeare)

3.3.4 UNL Attributes

Table 3: UNL Attributes

Concept	**Attributed as**
Time with respect to the speaker	@past, @present, @future
Speaker’s view on aspects of event	@begin, @complete, @continue, @custom, @end, @experience, @progress, @repeat, @state @just, @soon, @yet
Speaker’s view of reference to concepts	@generic, @def, @indef, @not, @ordinal

Attributes are used to describe subjective information in a sentence. There are 87 attributes (which can be augmented with the user defined ones) to express the semantic content of a sentence. This is some of the UNL attributes. Some of the UNL attributes are shown in table 3.

3.4 UNL Representation

For example, on considering the tourism domain document, a sentence “Sri Aurobindo Ashram is a well known and wealthy ashram in India with devotees from India and all over the world flocking towards the spiritual salvation.” has the UNL representation as shown in figure 8. The list form of UNL representation is performed. In the list format of UNL sentence representation the UWs are sequenced with UW-ID from ‘01’ to ‘08’ it is listed under the [R] and [/R] tags. Likewise all the sentences in the document are deconverted to produce the UNL representation as the output. The UNL representation is fed as the input for the deconvertor.

3.5 Deconversion

A deconvertor system is developed which takes the UNL representation as input and for the deconvertor system to perform the summarization process.

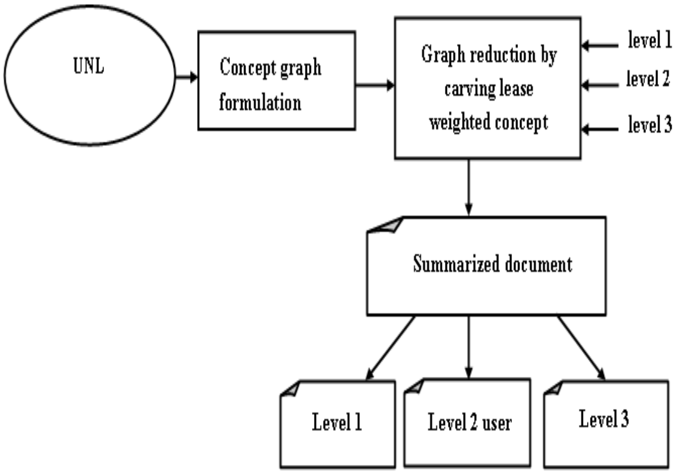
UNL
Concept graph formulation
Graph reduction by carving lease weighted concept
level 1
level 2
level 3
Summarized document
Level 1
Level 2 user
Level 3

Fig 7: Deconversion architecture

To perform deconversion process the word dictionary and deconversion rules has to be created. To perform summarization process the word dictionary and the relation between the UW’s, attribute features of the words are collected and the relation between words are taken and the irrelevant information like prepositional phase, determiners and so on, are reduced to obtain the summarized document. Word builder is also used to represent the exact meaning of the source document.
Step1: Preparations of the dictionary data.
Step 2: Writing deconversion rules.
Step3: Deconvert to produce summarized document.
Thus a deconvertor module will be developed in such a way that will perform a summarization process. The summarized document will be produced for three levels of users according to age classification. The deconvertor system is shown in figure 7.

```
{unl}
[w]
Sri Aurobindo Ashram(icl>place)@entry:01
well known(icl>state)@entry@present:02
wealthy(icl>state)@entry:03
India(icl>location)@generic:04
devotees(icl>people)@entry:05
world(icl>unique)@generic:06
flocking(icl>do)@entry:07
spiritual salvation(iof>thing)@conclusion:08
[\w]
[R]
01aoj02
01aoj03
01plf04
01mod05
05plf06
05pur08
[\R]
{\unl}
```

Fig 8: UNL Representation

4. EXPERIMENTAL ANALYSIS

The experiment is carried out using Net Beans IDE 7.1.1 and the experiment analysis is shown below. The process of summarization is carried out for three levels of users. Level 1, level2, level 3 users based on their IQ level. IQ is used to assess intelligence and on considering the overall population the IQ score lies between 70 and 130. So, when the IQ is greater than 130 they are considered as level 1 users, when the IQ level is between 70 to 130 they are considered as level 2 users and when the IQ level is bellow 70 they are considered as level 3 users. Thus the summarized document is produced based on the levels. Table 4 shows the word count in each document and the word count for different levels of users.

Table 4: Word count in summarization

DNO	Word count-original document	Word count in Summarized Document		
		Level 1 users	Level 2 users	Level 3 Users
1	452	252	310	346
2	460	235	314	348
3	477	270	282	321
4	413	252	295	341
5	493	248	312	344
6	414	252	292	342
7	431	244	296	324
8	447	233	288	342
9	419	244	283	336
10	433	237	287	321
11	438	235	313	322
12	473	231	314	323
13	468	253	301	339
14	465	267	319	320
15	468	252	303	322
16	464	258	310	345
17	428	256	282	330
18	433	259	297	348
19	438	277	305	338
20	434	234	302	345

This figure shows the word count after the summary is performed. Here a sample of 5 documents are taken and the word count of the original document, level 1 document, level2 document, level 3 document is shown.

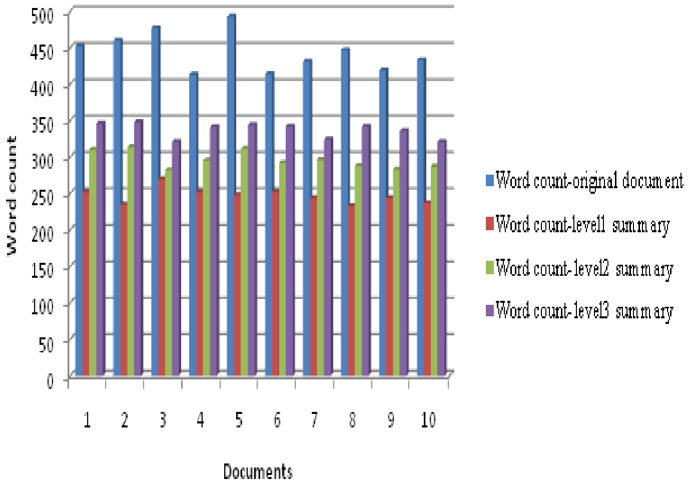

Fig9: Graph showing the word count of the original document with the summarized document for various levels.

The performance of the system is measured using Decisiveness for various levels of user. Decisiveness is the amount of words that has been compressed at various levels. Decisiveness is calculated for level1, level2 and for level3 users using the formula given below. Since the IQ score is greater for level1 users, the decisiveness will be more compared to other levels, and the decisiveness will be greater for level2 than compared to level3. Thus decisiveness will be level1>level2>level3.

$$\text{Decisiveness for level1 user } (DL_1U) = 100 - \left\{\left(\frac{\text{No of words in level 1 summarized document}}{\text{Total No. of words in the original document}}\right) * 100\right\}$$

$$\text{Decisiveness for level2 user } (DL_2U) = 100 - \left\{\left(\frac{\text{No of words in level 2 summarized document}}{\text{Total No. of words in the original document}}\right) * 100\right\}$$

$$\text{Decisiveness for level3 user } (DL_3U) = 100 - \left\{\left(\frac{\text{No of words in level 3 summarized document}}{\text{Total No. of words in the original document}}\right) * 100\right\}$$

The decisiveness is found out and the graph is plotted against the documents and the decisiveness ratio. When the graph is plotted, it is clear that the compression of the original document for the level 1 user is more than compared to others, since their IQ level is more. Similarly, for level2 the decisiveness will be more compared to level 3 users.

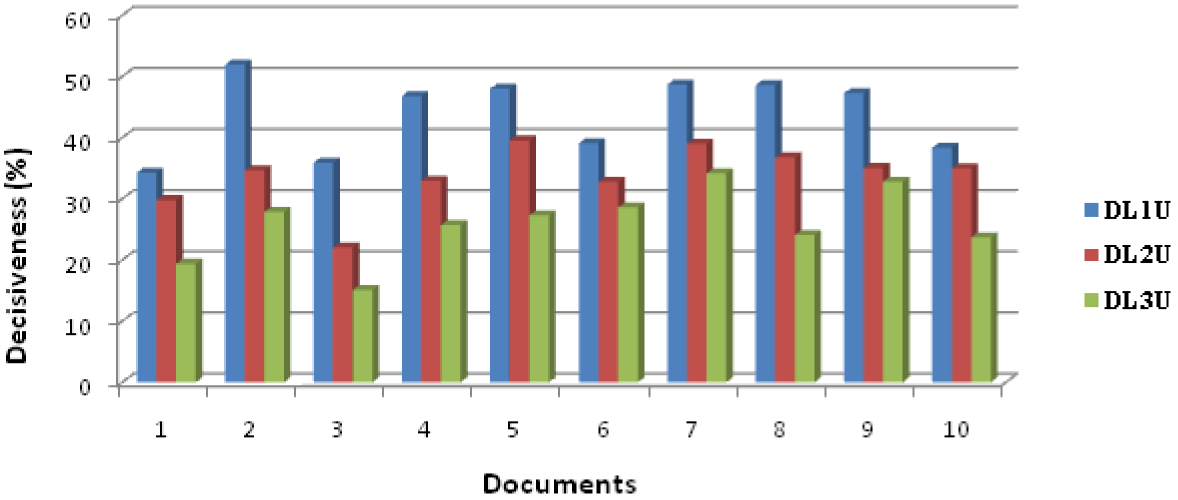

Fig 10: Decisiveness for 3 levels of summarized documents

Thus the summarization is carried out for different levels of users.

5. CONCLUSION

Thus the document is enconverted by adopting various steps like POS tagging and parsing, entity identification, relation identification, building dictionary and so on to give the UNL representation which gives the exact meaning of the original document as output. This UNL representation works in any language, meaning if we want to convert the source document into any other languages we can take this UNL representation and develop a deconvertor system to deconvert into any specific language. For this purpose, we adopt UNL way of representing the source language text. Datasets are collected for the tourism domain and when a document is selected it is subjected to all the process of enconversion and the UNL representation of the document is obtained. Later on, after the development of an enconvertor, a deconvertor system is developed in such a way that performs summarization process which produces the summarized document for three levels of users based on their IQ levels. The Decisiveness is also calculated and the graph is drawn. In future, according to the users, the original document can be given in any desired language by using the deconvertor system. That is the reason behind this NLP system for introducing UNL for the purpose of summarization.

6. REFERENCES

[1] Hiroshi Uchida, Meiying Zhu", Sep. 2009. The Universal Networking Language beyond Machine Translation" UNDL Foundation, Tokyo.

[2] Daniel Jurafsky, James H. Martin, Speech and Language Processing "An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition", September, 2007.

[3] Md. Ershadul H. Choudhury, Nawab Yousuf Ali, Mohammad Zakir Hussain Sarkar, Md. Ahsan Razib, "Bridging Bangla to Universal Networking Language- A Human Language Neutral Meta-Language", December 2004.

[4] T.Dhanabalan, T.V.Geetha, Dec, 2003. "UNL Deconverter for Tamil", International Conference on the Convergence of Knowledge, Culture, Language and information Technologies.

[5] Bevin Sousheel Bhagianath, S.Lakshmana Pandian, "UNL Deconverter for Malayalam", June. 2010.

[6] For Stanford Parser: http://nlp.stanford.edu/software/lex-parser.shtml

[7] For Universal Networking Language: Universal Networking Digital Language Foundation. http://www.undl.org/

[8] Bhattacharyya, "Multilingual Information Processing Using Universal Networking Language", IndoUK Workshop on Language Engineering for South Asian Languages (LESAL), Mumbai, India, April, 2001.

[9] Advances in Theory and Applications", Volume 12, Centre for Computing Research of IPN, ISSN: 1665-9899, ISBN: 970-36-0226-6February, 2005

[10] Uchida H., Zhu M. "The Universal Networking Language (UNL) Specification", Version 3, Edition 3, UNL Centre, UNDL Foundation, Tokyo, December, 2004.

[11] For UNL relations: http://www.undl.org/unlsys/unl/unl2005/relation.htm, UNDL Organization, Tokyo.