UNL Enconversion framework for Machine Translation

Athira.K

Assistant Professor
Department of Information Technology
MES Engineering college
Kuttippuram

Abstract- **This paper presents the building of a language independent Enconversion framework of UNL. Universal Networking Language is a relatively new interlingua of machine translation which is an important application of NLP .The purpose of framework is to localize the software applications and platforms and thus trying to bridge the great digital divide that has created due the linguistic barriers.It is designed to perform the task of converting source languages into UNL format; i.e. UNL expressions. The process of building the framework of Enconverter system includes the task of identifying the category of words in a corpus by doing the morphological analysis. These tagged words are used as input by the UNL Enconverter system to find the semantic relation existing between the words and to generate UNL relations . Enconverter framework maps all the 46 UNL relations and then tested with Dravidian language Malayalam . Later the Enconversion module is altered to include the global language English also.A resultant hyper graph is also generated with concepts as the nodes and arch connecting them showing the relation existing between them.The language corpora once converted to UNL, can be translated to any other language given UNL system built for that language.**

Index Terms- **UNL,Enconverter,hypergraph,morphology, analyser.**

I. INTRODUCTION

Computational linguistics is an interdisciplinary field concerned with the processing of language by computers. Today, the area of natural language processing has emerged as a vibrant field where work is on in developing computational tools in many languages of the world. Universal Networking Language is a relatively new interlingua which was proposed in mid 90s and was undergoing revisions till 2005. The process of converting a source language (natural language) expression into the UNL expression is referred to as Enconversion and of converting UNL expressions into a target language representation is called Deconversion.

The UNL system revolves around a unique artificial language (Universal Networking Language) that pretends to capture the meaning of written documents. This language is based on the representation of concepts and its relations. The main goal of the project is to build a language independent enconversion framework for UNL . The goal of using UNL is to eliminate the massive task of translation between two languages and reduce language to language translation to a one time conversion to UNL. In a normal machine translation system to translate to and from 10 languages ,10 grammers,10 lexicons,90 translation dictionaries,90 translation rules are needed along with the semantic processing needed for each language .For example Malayalam corpora once converted to UNL, can be translated to any other language given UNL system built for that language.

The Enconverter is a language independent parser that provides a framework for morphological, syntactic and semantic analysis synchronously. Malayalam is a morphologically important language in which most of the morphemes co-ordinate with root words in the form of suffixes. Lots of work has been done on POS tagging for English. But there has not been much work done in POS tagging or morphological analysis for Indian Languages like Malayalam. Thus building an efficient analyser plays an important role as it leads to the success of building a proper UNL Enconverter system.

Morphological analyser is built in used in a way to provide a language independent framework which helps other similar languages also to work by altering the language dependent module. Here the grammar rules, lexicon and Unicode conversion belongs to the language dependent module.

UNL is a digital meta language for describing, summarizing, refining, storing and disseminating information in a machine independent and human language neutral form.

The core structure of UNL is based on the following elements:

Universal Words: Nodes that represent word meaning. Example: play(icl>do),woman(icl>person),movie(icl>art).

Attribute Labels: Additional information about the universal words.

Example:@entry,@past,@pl,@interrogative....

ISSN: 2278 – 1323

Relation Labels: Tags that represent the relationship between Universal Words.

Universal words are the character-strings which represent simple or compound concepts. They form the vocabulary of UNL and represent the concepts in a sentence without any ambiguity. Attributes of Universal Words describe the subjectivity of the sentence. They provide information about how a concept is used in a given sentence.

Binary relations of the UNL expressions represent directed binary relations between the concepts of a sentence. There are a total of 46 relation labels defined in the UNL specifications. Thus through the process of Enconversion the source language concepts could be represented in a global UNL format and later can be translated to any other language given UNL system built for that language.

II. RELATED WORKS

Universal Networking Language (UNL) which was started in 1996 as an attempt to cross the language barrier on the web. UNL has been developed and is managed by the Universal Networking Digital Language (UNDL) foundation, an international organization of Institute of Advanced Studies of United Nations University, Tokyo, Japan [12]. 15 language groups from different parts of the world were involved in this endeavour .The idea was to encode (enconversion in UNL) thesentences of a language L1 into the UNL form and then generate (deconversion in the UNL) the sentences of L2 from the UNL form. It should be evident that both the languages must use the same pivot dictionary.

UNL Enconversion is basically a semantic analysis problem and it varies according to the linguistic features of each natural language. Morphology, syntax, semantics of Enconverter systems depend on the features of the natural language used. English UNL Enconversion is mainly based on parsers and the resulting dependencies are converted to UNL relations and attributes based on the POS tags. Enconversion system proposed by Jain et al [5] uses a standard syntactic parser with limited morphology and semantics.

The French UNL Enconverter generates UNL expressions using an incremental parser which converts the expressions in to a semantic graph using a rule based approach proposed by Gala et al [10].

Igor M et al [8] proposes Enconversion using a multifunctional linguistic processor ETAP-3 converts natural language into ETAP-3 internal representation that is essentially a normalized syntactic structure which is then converted in to required UNL representations. ETAP-3 makes use of syntactic dependency trees for sentence structure representation instead of constituent, or phrase, structure. This approach is centred on the use of syntactic structures to aid the Enconversion process.

Enconversion process for Arabic which is a highly inflected language and of relatively free word nature requires strong interaction between morphological and syntactic processing. Therefore Samesh et al [7] says a rule based approach for Arabic Enconversion cannot use the normal pipeline model where morphological analysis is followed syntactic analysis for resolving ambiguity.

Another approach for English to UNL Enconversion proposed by Mohanty et al[11] uses a two stage process where the conceptual arguments are first identified in the form of semantically relatable sequences (SRS) which are potential candidates for being linked by semantic relations. These are then mapped to form a parsed output and then UNL expressions.

Many Indian languages do not have a rigid fixed word order. Many of them are partially free word order languages and so dealing with these languages need special concern. Richness of morphology determines the freedom allowed by a language. In languages like Hindi and Bangla for which enconversion is done, isolated case markers with local word groupings, shows the case relations existing between various words in the sentence. Since most UNL relations are based on the case roles, the handling of this approach differentiates the approach to enconversion for these languages.

A framework for designing the EnConverter for Punjabi has been discussed by Parteek Kumar et al [13] with special focus on generation of UNL attributes and relations from Punjabi source text. The input is properly chunked to get meaningful parts and then processed to apply the Enconversion rules. They successfully tested the framework system on several Punjabi sentences.

Nawab et al [3] discusses about a dictionary development procedure of Bangla parts of speech for UNL. They proposes rule based approach to develop the morphological analysis of simple and compound Bangla words that can be used to make UNL-Bangla dictionary for converting the natural Bangla sentences to UNL documents and vice versa. They emphasises the importance of rule based analysis to find the proper tags of language corpora and building UNL relations based on it.

T.Dhanapalan et al [2] proposes the development of Tamil DeConverter . In Tamil most information for generating sentence from UNL structure is tackled in morphological and syntactical level. Ambiguity only makes to go for complete semantic processing. Relation table is used to find out the words or endings for the specified binary relation. Rule based approach is used here.

All the Enconverting systems discussed above are based on a rule based approach. A Statistical approach is also used for the parsing stage with associated morphological and syntactic linguistic features has also been attempted by Nguyen et al[6].

All the UNL Enconverters discussed above use a structural syntactic parser for fixed word order languages and dependency parser for partial free word languages.

III. ASPECTS FOR DESIGNING UNL ENCONVERTER FRAMEWORK

The main aim of this project is to provide, a language independent specification for serving as a common

ISSN: 2278 – 1323

medium for documents in different languages. The process basically involves i) building native language to UNL dictionary and ii) deriving language specific syntactic rules called analysis rules for parsing/translating native language corpora to UNL. The meaning of a native language sentence is expressed in UNL system as a hypergraph composed of nodes connected by semantic relations. Nodes or Universal Words (UWs) are words loaned from English and disambiguated by their positioning in a knowledge base (KB) of conceptual hierarchies.

Enconversion part of UNL is made generic by checking all the Enconversion rules in a language, providing the language rules from outside.The dependencies existing between the words could be properly understood from the analyser output of the language and UNL relations were built based on that. Analyser can be defined as a computer program which takes a word as input and produces its grammatical structure as output.

We have to provide the grammatical rules called analysis rules for languages for parsing them. Analysis rules of malayalam can be uploaded to the UNL server so that conversion can be done between Malayalam and any other language. In the same manner English to UNL dictionary along with their analysis rules are provided to the framework for testing the language. Rules helps to decide whether the root words are to be combined into a single headword or a relation is to be set up between them or an UNL attribute is to be generated. Based on this decision a hyper graph is generated with concepts as the nodes and arch connecting them showing the relation existing between them.

Many semantic ambiguities arise in the languages. So special care should be taken for finding the exact relation existing between the concepts. Extra features or attributes should be considered and expressed in the UNL document generated.

IV. LANGUAGE DEPENDENT AND INDEPENDENT MODULES OF UNL ENCONVERTER.

Language independent module helps the framework to act in a language independent manner. Here sentence processing, finding patterns are performing in language independent way. Sentence processing is a language independent module as it involves the pre-processing stages like splitting, segmenting, Root word extraction etc which can be done in a language independent manner. The whole corpus is first split into sentences and then tokenizes providing the language rules from outside.As all of the experiments presented in this work deal with segmented languages,segmenting is carried out based on the delimiter white space which eventually makes this step common. Thus the entire process is enacted regardless of the particular language being used and hence considered to be language independent.

Finding patterns and identifying relations is done in a language independent way after framing the Enconverter rules for the corresponding language. All the morphological or word features are identified from the external input we gave and thus relation vectors are mapped based on that. In the case of Dravidian languages some UNL relations are signalled unambiguously. Morphological features in this language convey the information to build UNL, while in English it also considers the position of sentence constituents. For including English we consider a syntactic parser with limited morphology and semantics.

Each language has its own specific features. Hence demanding a completely language independent scenario is unfeasible. Thus framing the rules based on the categories and language encoding based on Unicode format acts in a language dependent manner. All the UNL enconversion rules will receive corresponding category of language rules from outside. Thus based on the rules we are providing, the enconversion rules are built. A user interface is created for the user to input language features and based on that the enconversion module works.The Fig 1 given below shows the relations that should be given for Malayalam .

Fig. 1.Relation mapping in malayalam

INDEX	REL	MAL
1	and	ഉം
2	aoj	ആയ,ഉള്ള
3	bas	കാൾ
4	ben	കൊണ്ട്
5	cag	ഒപ്പം
6	cnt	എന്നാൽ
7	cob	കൂടെ
8	con	എങ്കിൽ,ആൽ
9	coo	പോലെ
10	dur	ഇടയിൽ,തൊട്ട്
11	fmt	മുതൽ വരെ
12	frm	ഇൽ നിന്ന്
13	icl	എന്നാൽ
14	ins	ആൽ
15	int	പോലെ
16	man	എ
17	mod	ആയ,ഉള്ള
18	nam	പേര്
19	obj	എ
20	or	ഓ,അല്ലെങ്കിൽ
21	pic	ഇൽ
22	plf	ഇൽ നിന്ന്
23	plt	എക്ക്
24	pos	ന്റെ,ഉടെ
25	pur	വേണ്ടി
26	qua	കുറെ
27	seq	പിന്നീട്,മുന്നിൽ
28	src	ഇൽ നിന്ന്
29	tmf	മുതൽ
30	tmt	വരെ
31	to	കുള്ള

ISSN: 2278 – 1323

For morphologically rich languages the UNL relations that obtained through the case suffixes of nouns are “pos”, “ ben” and “obj” .The semantics “mod” and “man ”can be determined from the adjectival suffixes. Certain UNL relations such as “iof” and “nam” connect more than two concepts. Certain case suffixes attached to nouns signal more than one UNL relations. The relations”frm”,”plf,”tmf”,”src” belongs to that category.

The relations of UNL have been used to focus on expressing semantics of a sentence. From a sentence we find out the entry word concept (Wi) and the word (Wi+k) or word (Wi-k) which it is related to.These relations are chosen based on the morphological suffixes, context, connectives,co-occurrence and POS.These relations are chosen based on the morphological suffixes, context, connectives, co-occurrence and POS.

The morphological words that determine the UNL relations for the Dravidian languages follow rule categories as given below. Morphology, connectives, co-occurances and context determines the selection of UNL relations in the case of Dravidian languages. All the rule categories and the corresponding UNL relations it maps are shown in the Fig 2 given below.

Fig. 2.Rule categories

RULE CATEGORY	UNL Relations
Morphology	Obj, ben ,mod, man, pos ,pur.
connectives	and, or, bas, equ, seq, qua,coo,int, cnt,icl,rsn,iof, nam,neg,per,pof,via,dur,met.
Co-occurrence Morphology + POS	cag,cob,ptn
Morphology + Context	frm,fmt,plf,plt,tmf,tmt, src,gol,plc,tim,to,agt,aoj.

One of the main advantages of UNL is the Universal Word (UW) lexicon, which enables us to specify word meanings at a deep level and to perform lexical disambiguation in a semantic oriented formalism. Malayalam to English dictionary and Tamil to English dictionary is the source to build Malayalam and Tamil to UNL dictionary correspondingly as universal words are English words mandated by UNL. Such dictionaries also provide all attributes along with meaning of a word. Any entry in the dictionary is put in the following format:
[HW] {ID} “UW”
Here,
HW - Head Word (Natural language word)
ID -Identification of Head Word (omitable)
UW - Universal Word

Some example entries of dictionary for Malayalam are given below:
kerala {} “city(icl>region)”
valuthu {} “huge(icl>big)”

In the example, “icl” in the constraint list enables us to define a sub-concept of a basic UW. “agt” and “obj” are Relation tags, which indicate dependency relations between a head word in linguistic categories and other words, based on a case grammar type specification. “.@entry”, “.@past” and “.@def” are called Attribute tags, which indicate the grammatical conditions of a given utterance.

V .IMPLEMENTATION DETAILS OF ENCONVERTER

The UNL enconversion process makes use of the morphological features conveyed by the word. UNL relations that are simple can be unambiguously determined from the morphological suffixes such as case suffixes, adverbial suffixes, adjectival suffixes or by the presence of certain standalone connective words.

A. Algorithm for Processing Input Sentence

Algorithm for the processing of input sentence involves the following steps:

i)Accept the input sentence.
ii)Process each word with the help of the morphological analyser.
iii)Replace each word in the input sentence with the root and suffix obtained from step ii.
iv)Retrieve UWs corresponding to each analysed word from the source language-UW dictionary.
v)Extract all the concept nodes with the help of step iv.'

B. Algorithm for Relation Resolution

According to the algorithm discussed above, a set of concept nodes are created from the input sentence. Then, Enconverter System invokes the following algorithm for UNL relation resolution and hypergraph construction.

i)Consider each concept node which contain the main concept.
ii)Search the required Enconversion rules based on the Relation mapping table for each language.
iii)Select the Entry node Wi and the associated Wi+k or Wi-k node based on the Enconversion rule categories.
iv)Resolve the UNL relation according to the rule fixed.

Here we are using five categories for classifying the Enconversion rules based on the noun-noun, noun-verb and verb-verb relationships occur in a sentence. Inorder to map all

ISSN: 2278 – 1323

the 46 relations 56 rules are used in the Enconverter framework.When case suffixes are attached to noun certain UNL relations are generated.The noun to which the case suffix is attached is one of the words taking part in the UNL relation. The UNL relations obtained through the case suffixes of nouns are “pos”, “ben”,”obj”.

The UNL expresses information or knowledge in the form of semantic network.UNL semantic network is made up of a set of binary relations where each binary relation is composed of a relation and two UWs that hold the relation. A binary relation of UNL is expressed in the following format:
<relation> (<uw1>, <uw2>).In <relation>, one of the relations defined in the UNL specifications is described. In <uw1> and <uw2> the two UWs that hold the relation given at <relation> are described.

Working of the UNL framework is illustrated below by inputting simple sentences:

Malayalam:

“[illegible] padashalayil poi.

The corresponding UW and head word of each Malayalam word is given below:

Raman{} “raman(icl>person)”
padashala {} “school (icl>institution)”
poi{ } “ go(icl>do)”\

UNL Expression:

agt(Raman(icl>person),go(icl>do))
plt(go(icl>do),school(icl>institution))

English:

“Raman wrote a letter to Seetha”.

The corresponding UW and head word of each English word is given below:

Raman{} “raman(icl>person)”
Seetha{} “seetha(icl>person)”
write{} “house(icl>action)”
letter{} “letter(icl>thing)”

UNL Expression:

agt(Raman(icl>person),write(icl>action))

to(letter(icl>thing),seetha(icl>person))

UNL relations with the noun suffixes are usually with the corresponding mainverb. The verb nominally occurs at the end of the sentence. So if wi+k word is a verb then that word is connected to the case suffixed(dative case) noun wi with the UNLrelation “pos”. An UNL relation with adjectival and adverbial suffixes depends on theconcept to which adverbs and adjectives are connected. The adjective or the adverb wi is associated with the wi+k word which is a noun or verb respectively
.

VI. RESULTS AND DISCUSSION

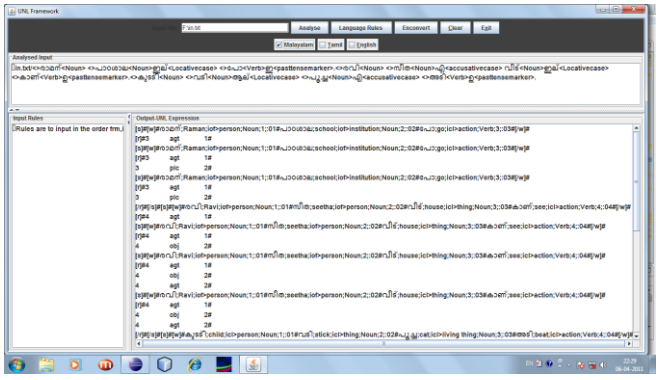
UNL framework is tested for Malayalam and English languages. It has been seen that the system successfully handles the resolution of UNL relations for the two languages.The size of the UNL Dictionary used for the system now contains only limited words.UNL relations are unambiguously generated for simple sentences.Table 1 analyse on the number of rules and relations used.

Table 1. Statistical Analysis

Languages	No: of rules	No:of relations	No:of sentence categories
Malayalam	56	43	5
English	50	46	5

We are using 56 rules to map 43 UNL relations for Malayalam, while for English all the 46 relations were mapped with the help of 50 rules.Sentence structure is classified to five categories for the languages.Fig 3 indicate a sample test output obtained.

Fig. 3 Output of Enconverter Framework



The three relations ”opl”, ”scn” and “coo” are found difficulty to map for Malayalam.These relations are chosen based on the morphological suffixes, context, connectives, co-occurrence and POS. We tested it manually on 10,000 sentences . Enconverter produce UNL expressions as output with a correctness of 90%.

ISSN: 2278 – 1323

VII. CONCLUSION

A language independent framework for UNL has been discussed, developed, implemented and tested. Languages used for testing are Malayalam and English. Emphasis on semantics of natural languages is the most important attribute of UNL. In these languages the dependencies that generated between the concepts are suitable for unambiguously determining the UNL relations occurring between them. Here we discuss about EnConversion analysis rules for the EnConverter and indicates its usage in the generation of UNL expressions. The output of testing the framework using these languages are encouraging as it gives similar UNL expressions for the same input given in two languages.

Presently, the EnConverter handles only simple sentences. Future work involves extending the scope of EnConverter to include clausal, interrogative and compound sentences. Work should be done on implementing the word sense disambiguation module in the proposed EnConverter. The framework can also be made more generic by incorporating more languages to it.

REFERENCES

[1] Ershadul H. Choudhury, Nawab Yousuf Ali,Mohammad Zakir Hussain Sarkar, Md.Ahsan. "Bridging Bangla to Universal Networking Language Human Language Neutral MetaLanguage" .ICCIT2009 .

[2] Gala N. "Using an increment Robust parser to automatically generate Semantic graph." 2004 Proceedings of the third workshop on Robust Methods of Analysis of Natural Language Data.

[3] Hiroshi Uchida & Meiying Zhu, "The Universal Networking Language beyond machine translation," Proc. Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Spain, 2002.

[4] Igor Boguslavsky, Nadezhda, ,eonid I.N, Leonid, hina sagalova, Victor Sizov, "Creating a Universal Networking Language Module within an Advanced NLP System", 2007 Proceedings of International conference on the convergence of knowledge, culture, language and Information Techonologies.

[5] Manoj Jain and Damani O. P. "English to UNL(Interlingua) Enconversion." Indian Institute of Techonology, Bombay.

[6] Nguyen, D.P.T. Ishizuka, M.A 2006,"A Statistcal approach for UNL-based relation extrach Research,Innovation and vision for the future",2006 IEEE International conference.

[7] Pushpak Bhattacharyya .Multilingual. "Information Processing Through Universal Networking Language". 2008 International conference on Advances in Recent Technologies.

[8] Parteek Kumar, R. K. Sharma. "Generation of UNL Attributes and resolving Relations for Punjabi Enconverter". *Malaysian Journal of Computer Science*, Vol. 24(1), 2011, pp 34-46.

[9] Rajatkumar Mohanty ,Anupama Dutta. "Semantically relatable sets: Building blocks for representing semantics". P. B . 2005, Machine Translation Summit.

[10] Samesh Alansary, Magdy Nagi and Noha Adly,"A Library Information System based on UNL knowledge infrastructure".2008 seventh International conference on Computer Science and Information Techonologies.

[11] T.Dhanabalan, T.V.Geetha. "UNL Deconverter for tamil". International Conference on the Convergence of Knowledge, Culture, Language and Information Technologies. December 2003.

ISSN: 2278 – 1323