Language Dependent Features for UNL-Malayalam Deconversion

Biji Nair
College of Engineering, Perumon, Kollam

Rajeev R. R.
Indian Institute of Information Technology and Management-Kerala (IIITM-K)

Elizabeth Sherly
Indian Institute of Information Technology and Management-Kerala (IIITM-K)

ABSTRACT
This paper presents a deconverting generator for Malayalam language using Universal Networking Language (UNL) for Machine Translation. UNL being an Interlingua representation, conveyed as directed hyper graph with relations and attributes of source language sentence. A set of Universal Words are generated from the source language with its semantic representation, are mapped to UNL features. The work involves identifying the dependent features like syntactic, semantic and lexical features of target language. UNL Relations, UNL Attributes and Universal Word (UW), which are the building blocks of UNL are identified and mapped to the dependent features of Malayalam. Lexical mapping of UWs to root words of Malayalam was done through UNL-Malayalam Word Dictionary. The deconversion is tested against 100 Malayalam Sentences that has achieved an appreciable F-measure score of 0.978. .

General Terms
Malayalam Deconversion, Universal Networking Language, Interlingua Machine Translation.

Keywords
Machine Translation, Universal Networking Language, Malayalam, Grammar Codes, POS, Tagset, suffix, inflection, interlingua, F-measure, lexeme, semantic network.

1. INTRODUCTION

1.1 UNL Features [1]
UNL is an electronic language which extracts semantic data from natural languages and expresses it using semantic network made of a set of binary relations. An UNL Document is represented using semantic network composed of (i) UNL Relations (ii) UNL Attributes (iii) Universal Word (UW). UNL Relations are binary relations between two UWs that semantically connect them in the sentence of source language. It establishes the objectivity of the UWs in a sentence. The subjectivity information of these UWs is indicated through UNL attribute.

[S:1]

{org}

I arranged a meeting yesterday.

{/org}

{unl}

agt (arrange(agt>thing,obj>thing):01.@entry .@past, I :02)

obj(arrange(agt>thing,obj>thing):01.@entry.@past, meeting(icl>occassion) :03.@indef)

tim(arrange(agt>thing,obj>thing):01.@entry.@past, yesterday(icl>day) :04)

{/unl}

[/S]

Figure 1: Sample UNL Document

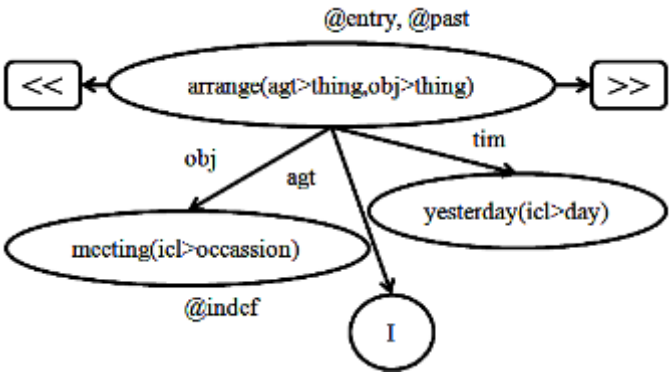

Figure 2: Semantic network for UNL document in Fig:1

2. RELATED WORKS
UNL based projects have been dealt in various Indian languages namely, Hindi, Punjabi, Bangla, Tamil. Parteek Kumar et al [2] published design and development of a Punjabi DeConverter. Both enconversion and deconversion in UNL has been dealt in Punjabi. Analysis and generation of Bangla [3] language for UNL has also been done. Significant amount of work related to UNL is going on in the Centre for Indian Language Technology, IIT Bombay. Their work HinD[4] is a Hindi Deconverter and uses a semantically rich lexicon, a priority-matrix of syntax plan, and elaborate morphology synthesis rules. Deconverter framework for Tamil [5], was implemented by Geetha T.V. et al in Anna University. Malayalam was incorporated by Hameed M. S et al [6] which materialized in Java platform using eclipse IDE.

Many foreign languages such as Arabic, Russian, English, Chinese, French, Spanish, Italian, Brazilian, and Portuguese have already been included in the UNL platform.

3. PROBLEM STATEMENT
The UNL Machine Translation system provides tools for language independent machine translation of UNL document to target language. The problem concerned is in dealing with the language dependent part of the deconversion process.

The major challenge is to represent the linguistic features of Malayalam language in a form processable by the Deconverter tool of UNL. The representation should be such that it can be mapped to the UNL features for accurate

translation. Another challenging aspect is the inherently morphological rich property of Malayalam language. It also adopts grammatical features of other languages like Sanskrit mainly in Sandhi rules and accepts words of other languages in usage.

The linguistic paradigm of Malayalam has to be deeply investigated and implemented through the features of UNL for Malayalam Sentence Generation. The syntactic categories [7] in Malayalam language are Noun Morphology, Verb Morphology and Modifiers. These categories are associated with specific semantic role in Malayalam sentences.

Noun Morphology gives the general format for noun word formation [8] as given below

W = noun root + derivational suffix + [plural suffix] + [case suffix] +...

Verb Morphology involves inflection of function suffixes for tense, aspect and mood. Modifiers include adjectives, adverbs and postpositions. Modifiers are invariant and do not inflect.

In this paper we formulate the mapping of UNL features to the syntax and semantics of Malayalam language. The mappings are represented as, Suffix Marker-Tagset table, Tagset Corresponding to UNL Concepts table, and UW Word Order-Case suffixes table. The more precise the mapping the more accurate would be the building of deconversion rules.

4. DESIGN AND IMPLEMENTATION

4.1 Stages in Generation Process

There are three main stages in Malayalam sentence generation process using the Deconveter tool DeCo. They are (i) Lexeme Selection (ii) Generation Rules (iii) Post-editing Rules. Figure 3 shows the flow chart containing the main phases in generation process.

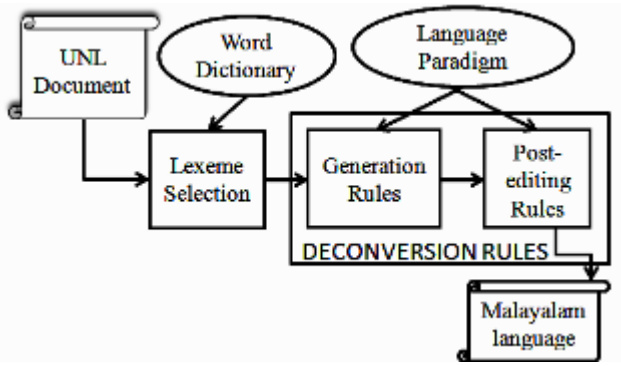

Figure 3: Deconversion Process

The lexeme selection for the words to compose the target Malayalam sentence is done by obtaining the equivalent headword from the UNL-Malayalam Word dictionary for the UW in the UNL document. Universal concepts are represented through UW in UNL expression/document. Every UW is associated to a headword in the UNL-Malayalam Word dictionary. Thus the UW in the UNL document serves as a trigger for surface word selection of the target sentence. The other field in a headword-UW entry in the Word Dictionary is the grammar code. These are attributes that define the subjectivity information of the UW in the sentence and are mainly used in deconversion rules.

Deconversion rules are used for morphological generation of surface words of target sentence from the lexemes selected. These words are then inserted in the right syntactic order to generate target sentence. There are two kinds of deconversion rules, generation rules and post-editing rules. Generation rules is used to formulate how various UNL expressions should be expressed in a native language. Generation rules are sufficient for any natural language generation, while post-editing rules are optional. Post-editing rules include functions for editing the generated results.

4.1.1 Lexeme Selection

Lexeme Selection phase involves substituting the UW referred in the UNL Document by its equivalent headword from the UNL-Malayalam Word Dictionary. Hence it should contain entries for root words, case suffixes, function words, along with their UW and the grammar attributes. The grammar attributes are listed using grammar codes. The grammar codes elaborate on the behavior of the headword in the target sentence and are used in deconversion rules for syntax planning. Figure 4 shows a sample of the UNL-Malayalam Word Dictionary.

```
[സംഘടിക്]{} "arrange(agt>thing,obj>thing)"
(V,MVERB, BAS, OBJ.DO) <M,1,1>;
[ഞാന്]{}  "I" (1SG, HPRON, NOM,PPRON, SUBJ)
<M,1,1>;
[സ്മോളനം]{} "meeting(icl>occassion)" (BAS, NCOM,
NOU, SG,PAST) <M,1,1>;
[ഒരു]{} "a" (DET, INDEF) <M,1,1>;
[പ്പിച്ചു]{} "do" ( VZ, PAST) <M,1,1>;
[ഇന്നലെ]{} "yesterday(icl>day)" ( BAS, PNOU, TIME)
<M,1,1>;
```

Figure 4: Sample of UNL-Malayalam Word Dictionary

The UNL-Malayalam Word Dictionary formed for an UNL document is processed using the DicBld tool, and creates two files with ".dic" and ".pix" extensions. The file with .dic extension is used by the DeCo tool for deconversion, and selects all possible word entries matching the UWs in the UNL documents. These selected candidate words are enumerated in word list. Figure 5 shows A sample of Word list to be used in deconversion rules.

```
========== WORD LISTS : 04 ==========
01  02  00:<agt
[ഞാന്]{}  "I" (1SG,HPRON,NOM,PPRON,SUBJ)
<M,1,1>;
02  01  00:,@entry ,@past,>agt,@entry,t,>obj,>tim
[സംഘടിക്]{}  "arrange(agt>thing,obj>thing)"
(V,MVERB,BAS,OBJ.DO) <M,1,1>;
03  03  00:,@indef,<obj
[സ്മോളനം]{}  "meeting(icl>occassion)"
(BAS,NCOM,NOU,SG,PAST) <M,1,1>;
04  04  00:<tim
[ഇന്നലെ]{}  "yesterday (icl>day)" (BAS,PNOU,TIME)
<M,1,1>;
=================UNL ==============
```

Figure 5: A sample of lexemes selected in the word list

4.1.2 Generation Rules

Rules need to be written for defining morphological and syntactic structure for generation of Malayalam sentence. Generation Rules are broadly classified into three (i) Morphological Generation (ii) Syntax Planning (iii) Blank Insertion as shown in Figure 6.

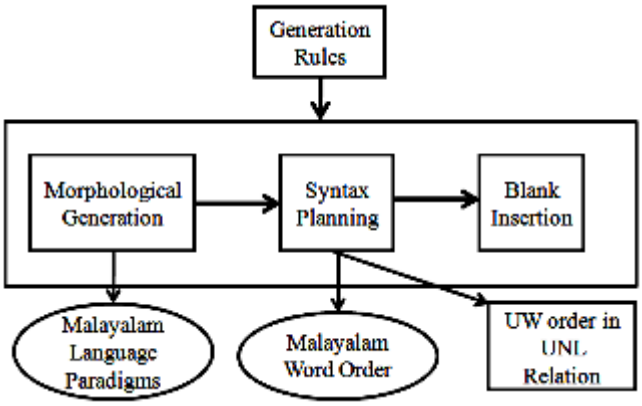

Figure 6: Classification of Generation Rules

4.1.3 Morphological Generation

It is the most challenging part of Malayalam language generation, the main reason being it's inflective nature. The Morphological Generation Rules include case suffix identification and insertion, functional suffix identification and insertion and Postposition selection and insertion.

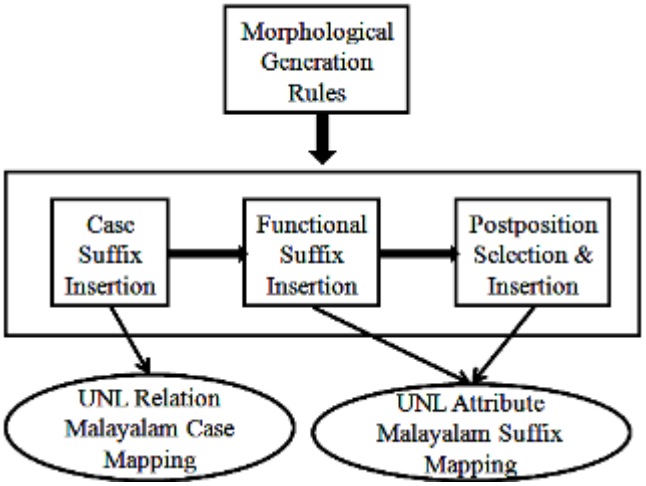

Figure 7: Different types of rules for Morphological Generation

Case, functional suffix identification and postposition selection are language dependent features. After these suffixes are identified, appropriate position for insertion is decided in the syntax planning phase. Our work recognizes more than 460 suffixes including case suffixes, functional suffixes and postpositions, which are tabulated in Suffix Marker-Tagset table. Table 1 shows the screen shot of the tabulation.

Table 1. Suffix Marker –Tagset table

VERB MORPHOLOGY			
Suffix	**Stem / Base**	**Grammar code**	**Attributes**
	aak	V/FINV/CONJV/BE/MVERB	@identity
	aayirunnu	V/FINV/CONJV/BE/MVERB/PAST	@state
	aakunnu	V/FINV/CONJV/BE/MVERB/PRES	@existential
	aayirikkum	V/FINV/CONJV/BE/MVERB/FUT	@attribute, @locative
-atu		V/DO/PAST/PCPL/NOMNL	
	aak	V/FINV/CONJV/BE/CLEF/PRES	

The suffixes are listed along with their grammar codes and attributes. Grammar codes are the tags which define the grammatical relations of the suffix with a root word in a sentence. In-depth analysis of Malayalam language paradigm for Noun, Verb, and Modifier Morphology was done for enumerating these suffixes.

The UNL attributes were mapped to suffixes and postposition, using grammar codes. The UNL attributes play vital role in functional suffix identification and are also the deciding factors for postposition insertion.

All the 46 UNL Relations were used for identifying the case suffixes for noun UWs in each of these relations. We have mapped the UNL features to Malayalam linguistic paradigm which was the driving principle behind the deconversion of UNL sentences to Malayalam.

Table 2. Tagset Corresponding to UNL Concepts

S. No	Concept	POS	Tag
1	Thing	Noun	NOU
2		Concrete	NOU/CONC
3		Abstract	NOU/ASBT
4		Functional	NOU/FUNC
5		Place	NOU/PLC
6		Pronominal	NOU/PRNOMN
7		Time	NOU/TIME
8		Volitional	NOU/VOLI
9	Verbal	Verb	VERB/V
10		be	VERB/BE
11		do	VERB/DO
12		occur	VERB/OCCUR
13	Adjective	Adjective	ADJ
14		Predicative	ADJ/PRED
15		Attribute	ADJ/MOD
16		Attribute	ADJ/QUA
17	Adverb	Adverb	ADV
18		how	ADV/HOW

Using these tables Morphological generation rules can be written for inflections of noun and verb in Malayalam word formations. Figure 8 shows example of Morphological Generation Rule of function suffix insertion for attribute @past.

```
============== APPLIED RULE ==============

:{V,MVERB,BAS,OBJ.DO,@past,^ending:+ending::}"[പ്പി
ച്ചു]",VZ:::"P6;

==========================================

>>>>> lnode

[സംഘടിക്ക്]{}  "arrange(agt>thing,obj>thing)"
(ending,@entry
,@past,>agt,@entry,t,>obj,>tim,V,MVERB,BAS,OBJ.DO)
<M,1,1>;

>>>>> rnode  INSERTED

[പ്പിച്ചു]{}  "do" (VZ,PAST) <M,1,1>;

================ NODE LIST:00 ============
```

Figure 8: Example of Morphological Generation Rule of function suffix insertion

4.1.4 Syntax Planning

Syntax Planning rules are used for arranging morphologically generated surface words in Malayalam to form the target sentence. Malayalam basically follows SOV word order and is relatively free order.

Depending on each of the UNL Relations, lexemes corresponding to each of the UWs in it are inserted into the Node-list. Initially the only node in the Node-list is the UW which has the @entry attribute associated to it. The insertion rule begins with the relation which contains the UW with @entry attribute. The other UW of the relation (binary) is inserted either to the right or left of the UW already on the Node-list. Position of insertion is found using, Word order and suffixes based on Relations table. Table 3 shows the word order mapping in UNL Relations. The table maps positions of the UWs with respect to each other in each of the 46 UNL Relations. Similarly for each relation, insertion rules are written for one of the UW in the relation that is not present in the Node-list.

Along with the insertion rules corresponding to each of the relations in the UNL document, insertion rules need to be written for case suffix for UW representing noun. The case suffix for UW in a relation is mapped in the Word order-suffixes based on Relations table.

Appropriate Postpositions can also be used instead of case endings in Malayalam. Postpositions are placed after the nominal. They are usually used in relations between two noun UW or between a noun and a verb UW.

Suffix corresponding to each of the UNL attribute associated to a UW is retrieved from the Suffix Marker-Tagset table and inserted to the node list relative to its UW position.

Table 3. Word order- suffixes based on Relations

Rel	Noun	Infl	Verb	Infln	Adj	Infl	Adv	Infl
agt	UW2 (L)	NOM	UW1 (R)	DO				
and		-um		-um		-um		-um
aoj	UW2 (L)	NOM	UW1 (R)	BE				
	UW2 (R)				UW1 (L)			
bas	UW2 (L)	ACC	UW1 (R)	BE				
	UW2 (L)	ACC	UW1 (R)	DO				
	UW2 (L)	ACC					UW1 (R)	HOW
ben	UW2 (L)	DAT	UW1 (R)	BE				
	UW2 (L)	DAT	UW1 (R)	DO				
	UW2 (L)	DAT	UW1 (R)	OCC UR				

4.1.5 Blank Insertion

Blank space need to be inserted between individual words. Inflections are achieved by having no space inserted between the root and the suffix. Therefore rules were written for blank space insertion for each word separation in the target sentence.

4.1.6 Post-editing Rules

Post-editing rules were used to modify the generated results. We have used post-editing rules for applying sandhi while word formation after suffix inflection if required.

4.2 Algorithm for Malayalam Generation

Following are the steps for writing the generation rules for generating Malayalam sentences for an input UNL document.

1. Take the relation with UW having @entry attribute.

2. Insert the other UW, in the selected relation, into the Node-list according to the UNL-Relation UW Word Order table

3. Consider next relation with one of the UWs already in the Node-list and insert the other into the Node-list according to the UNL-Relation UW Word Order and case suffixes table

4. Repeat Step 3 until all relations have exhausted.

5. Insert case marking suffix or postposition for each of the relations to the appropriate UW according to the Word order-suffixes based on Relations table.

6. For each attribute of an UW insert corresponding function suffix or postposition into the Node-list for morphological generation of surface words according to Suffix markers-Tagset table.

7. Repeat step 5 and 6 until all attributes of every UW have exhausted.

8. Insert blank spaces for word separation.

9. Post-editing rules can be written for further refinement.

4.3 Sample Output

The output obtained from the DeCo tool after executing the generation rules for the UNL document in Fig 1.

```
==============NODE LIST:00 ==============

/[ഇന്നലെ|yesterday(icl>day)]/ /ഞാന്|I/ /ഒരു/
/സമ്മേളനം|meeting(icl>occassion)/
/|arrange(agt>thing,obj>thing)/സംഘടി/പ്പിച്ചു/>>/

========================================

lnode >>>>>

rnode >>>>>

[ഇന്നലെ]{}  "yesterday(icl>day)"
(blk,<tim,BAS,PNOU,TIME) <M,1,1>;

ഇന്നലെ ഞാന് ഒരു സമ്മേളനം സംഘടിപ്പിച്ചു

;;Time  0.0            Sec

;;Done!
```

Figure 9: Sample output for UNL Document in Fig:1

5. EVALUATION

The problem being tackled in this work is the generation of Malayalam sentences from UNL document using the dictionary builder and deconverter tool provided by UNL. This work achieves it by mapping UNL features to language dependent linguistic features of Malayalam. The Deconverter tool works for single sentence conversion at a time. It yields accurate translation of single sentence UNL document into Malayalam sentence. When a corpus of 133 sentences was used the rules worked appropriately the same way as it was for these 133 individual sentences per document. But in cases where multiple candidate entries were selected for a UW in the word list from the UNL-Malayalam Word Dictionary in Lexeme selection phase the result was not impressive. The first entry among the candidate entries in the word list was selected each time. The suggested solution for this problem is the usage of co-occurrence dictionary.

6. RESULT AND DISCUSSION

Deconversion rules written according to the mapping of UNL features to linguistic features of Malayalam were accurate according to the evaluation results, provided there is unique word list entries for each UW. Lexeme Selection yielded an F-measure score of .978 when an UNL corpus containing 133 sentences was executed for translation using the Deconverter tool. The single sentence document of these 133 sentences was accurate and specific. By using co-occurrence dictionary further selection from multiple entries can be achieved. We observe that if the mapping of the UNL features to Malayalam language features are done more precisely UNL machine translation can be a break through. Translation of greater number of UNL Corpus would further strengthen our analysis.

7. CONCLUSION AND FUTURE WORK

The aim of this paper was to report our work on mapping language dependent features of Malayalam to UNL features to facilitate generation of Malayalam sentences using the Deconversion tool provided by UNL. The proposed system is efficient in generating syntactically unambiguous and semantically equivalent target sentence for the UNL source sentences. Further refinement of the mapping can be done for automating the production of generation rules rather that manual rule generation. Another suggested improvement is to include co-occurrence dictionary for selection among multiple candidate entries for an UW in the word list.

8. ACKNOWLEDGMENTS

We acknowledge Meiying Zhu, Director UNDL for allowing access to UDS (UNL Development Set), and also for the support and guidance throughout the practical implementation of the generation process using the tools of UNL for the deconversion process.

9. REFERENCES

[1] Uchida H., Zhu M., The Universal Networking Language (UNL) specifications version 3.0, 1998. Technical Report United Nations University, Tokyo, 1998.

[2] Kumar P., Sharma R. K., “Punjabi DeConverter for generating Punjabi from Universal Networking Language”, Journal of Zhejiang University-SCIENCE C (Computers & Electronics), ISSN 1869-1951 (Print); ISSN 1869-196X (Online), www.zju.edu.cn/jzus; www.springerlink.com

[3] Ali M. N. Y., Sarker M. Z. H., Das J. K.,"Analysis and Generation of Bengali Case Structure Constructs for Universal Networking Language", IJCA International Journal of Computer Applications, March Edition 2011, Volume 17, No. 2, pp. 34-41, 2011

[4] Singh S., Dalal M., Vachhani V., Bhattacharyya P., Damani O. P., “Hindi Generation from Interlingua (UNL) “, Indian Institute of Technology, Bombay (India)

[5] Dhanabalan T., Geetha T.V., “UNL Deconverter for Tamil “International Conference on the Convergence of Knowledge, Culture, Language and Information Technologies, December 2 - 6, 2003, Alexandria, EGYPT

[6] Hameed M. S., Subalalitha C. N., Geetha T.V., Parthasarathi R., “A Deconverter Framework for Malayalam” ICACCI '12 Proceedings of the International Conference on Advances in Computing, Communications and Informatics , 2012.

[7] Nair R. S., Language in India, www.languageindia.com, ISSN 1930-2940

[8] Asher R.E, Kumari T.C, Malayalam, Psychology Press 1997-Foreign Language Study