ISSN No. 0976-5697

Volume 6, No. 2, March-April 2015

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

E-Content Based UNL Deconverter Framework for Tamil and Hindi Language

K. Sathiyamurthy
Assistant Professor, CSE
Pondicherry Engineering college
Puducherry, India

Dj. Panimalar
Student, CSE
Pondicherry Engineering college
Puducherry, India

P. Dhivya
Student, CSE
Pondicherry Engineering college
Puducherry, India

Abstract: The enormous growth in E-content information in the web requires the content to be available in the natural language. The demand in converting the E-content from the one natural language to another natural language have been increased recently. E-content information was obtained in the target language from the Interlingua representation such as Universal Networking Language (UNL) rather than from the source language. A framework towards conversion of a UNL expression related E-content into the tamil and hindi language was presented in this paper. This approach uses the dictionary to select target language words for UWs in the UNL expression. The morphological rules were created to modify the headwords owing to the target language. Also the Hidden Markov Model technique has been adopted to define the word order in the generated sentence.

Keywords: E-content, tamil, hindi, deconverter, dictionary, morphological rules, Hidden Markov Model.

I. INTRODUCTION

Education was being improved by various new technologies, particularly by the emerging of computer related information technology. In recent years, the interest in E-learning has been increased because of this lifelong learning context. These systems allow people to ''learning far away'', and they have been frequently designed and used in higher education.

However, the distance learning process still presents some difficulties to be overcome by like providing universal access of the e-learning materials. The complexity faced by the traditional e-learning system is multilinguality, Now-a-days the delivery of technical E-content text material, are available only in the English language predominately. To offer multilinguality in E-content, machine translation (MT) is considered as an important tool. There are many possible approaches to machine translation. One approach is the Interlingua based machine translation systems where the source language is transformed into the intermediate representation.

The interlingua technique such as Universal Networking Language (UNL) has been used in this work to convert the source language content into the intermediate format such as UNL expression. UNL stimulates the sharing of information and knowledge in their native language to support multilingual platform.

The translation system using UNL as intermediate, requires the enconverter to convert the source content into the UNL expression and the deconverter to convert the UNL expression into the target language.

This work is focussed towards the development of deconverter for tamil and hindi language using UNL framework. Deconverter converts the UNL document related to E-content information into tamil and hindi languge. Tamil, hindi language is a relatively free word order language. Tamil being a morphologically rich language requires the syntactic categorization large amount of information. The deconverter perform the translation by considering the attributes and the UNL relation associated with the UNL expression. The proposed system comprises of: (i) UNL-Tamil, UNL-Hindi dictionary (ii) Morphological rules related to the tamil and hindi language for generation of appropriate words (iii) Hidden Markov Model (HMM) technique for sentence ordering. A hidden Markov model (HMM) is a probabilistic model of a multiple sequence alignment of words. A syntactic ordering of the generated target language words is done using the HMM technique. The next section discusses about the related work done on deconversion of a UNL expression into various languages.

II. LITERATURE REVIEW

The system for converting the UNL expression to Tamil language has been described by T.Dhanabalan, T.V.Geetha (2003). In their work the sentence generating from UNL structure is tackled in morphological and syntactical level itself. The syntactic generator has been designed in their work to extract the required syntactic and semantic information in order to build the complete sentence. Syntactic generation in the case of Tamil DeConverter is generating Tamil sentence with the help of binary relation and morphological rules. Also, relation table was used to find out the words or endings for the specified binary relation. This table information plays an important role in both Tamil EnConverter and Tamil DeConverter [1].

A work on development of deconverter for generating Punjabi from Universal Networking Language was attempted by Parteek kumar, Rajendra Kumar sharma (2012). In their work they designed and developed a Punjabi DeConverter with a special focus on syntactic linearization. Syntactic linearization is the process of defining arrangements of words in generated output. They describe phases of the proposed Punjabi DeConverter such as, UNL parser, lexeme selection, morphology generation, function word insertion, and syntactic linearization. The algorithms and pseudocodes for implementation of syntactic linearization of a simple UNL graph, a UNL graph with scope nodes and a node having un-traversed parents or multiple parents in a UNL graph have been discussed in their work [2].

The UNL-Malayalam Deconverter was developed by Biji Nair, Rajeev R. R, Elizabeth Sherly (2014). The work involves identifying the dependent features like syntactic, semantic and lexical features of the target language. UNL Relations, UNL Attributes and Universal Word (UW), 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 three main stages in the Malayalam sentence generation process using the Deconveter tool include (i) Lexeme Selection (ii) Generation Rules (iii) Post-editing Rules.The system is efficient in generating syntactically unambiguous and semantically equivalent target sentence for the UNL source sentences [3].

A work on generation of Hindi sentences from an interlingua representation called Universal Networking Language (UNL) was explained by Smriti Singh. et al. (2007). The generation process consists of three main stages like morphological generation of lexical words, function word insertion, and Syntax Planning. The UNL phenomena have been meticulously handled, a relation by relation, and attribute by attribute. The system has been tested on an agricultural corpora, and the system generated sentences were scored by a team of evaluators. The *BLEU* scores against the reference sentences have been computed [4].

A work on development of Bangala deconverter was attempted by Aloke Kumar Saha et.al (2012). The system takes a set of UNL expression as input and with the help of language independent algorithm and language dependent data generates corresponding Bangla sentence. The process of deconversion involves syntax analysis and morphology phase. The syntax analysis phase is aimed at generation of proper sequence of words for the target sentence[5].

However, there exists no work done in the literature for converting the UNL expression related to the E-content into the tamil and hindi language. The next section describes our proposed architecture of the conversion module (i.e.) translating from the UNL expression into tamil and hindi language.

III. ARCHITECTURE DIAGRAM

The architecture of our proposed system is shown in Figure 1. It makes use of language-dependent components during the generation process. To achieve the mutilinguality in E-content using UNL as an intermediate, needs to have an both enconverter and deconverter. Our work focus on developing the deconverter. This paper explains the conversion from the UNL expression into the tamil and hindi language.

To the deconverter process the UNL expression was given as an input. The UNL expression represents information, sentence-by-sentence. Sentence information is represented using Universal Words and relations. The UNL expression also contains an attribute to describe about words and to specify the correct perspective of the word behaving in a sentence.

For Example consider a sentence “Data structure is a specialized format for organizing and storing data”, its binary relation of UNL expression is expressed in the following way.

agt(format-6,datastructure-2)
cop(format-6,is-3 .@present)
det(format-6, a-4)
aoj(format-6, specialized -5 .@past)
pur(format-6,organizing-8 .@present .@progress)
and (organizing-8 .@present .@progress, storing-10 .@present .@progress)
obj(organizing-8 .@present .@progress, data-11)

The obtained UNL expression was processed further for the deconversion process.

A. Lexeme Selection:

Lexeme selection is the process of selecting target language words for UWs present in the input UNL expression. During lexeme selection, UWs are searched in the dictionary along with constraints specified in the input UNL expression. This phase uses a tamil-UNL dictionary and hindi-UNL dictionary for this task. Dictionary entry has the following format of any native language word [6].

[HW]{ID}“UW”(ATTRIBUTE1,ATTRIBUTE2…)<FLG, FRE, PRI>

Here,
HW - Head Word (Tami and Hindi word)
ID - Identification of Head Word (omitable)
UW-Universal Word (English word)
ATTRIBUTE - Attribute of the HW
FLG - Language Flag (we use T,H for Tamil,Hindi)
FRE- Frequency of Head Word
PRI- Priority of Head Word

Some example entries of dictionary for tamil words related to technical text are given below:

[தரவுகட்டமைப்பு]"datastructure(icl>subject)"(noun, thing,)<T,0,0>;
[ஏற்பாடு]"organizing(icl>arrangement)"(verb,present, order,assemble)<T,0,0>;
[சேமித்து]"storing(icl>method)"(verb,present)<T,0,0>;
[தரவு] "data(icl>information)"(noun)<T,0,0>;
[ஸ்டாக்]"stack(icl>subject)"(verb)<T,0,0>;
[கணினி] "computer(icl>device)"(noun)<T,0,0>;
[நினைவகம்]"memory(icl>information)"(noun)<T,0,0>;

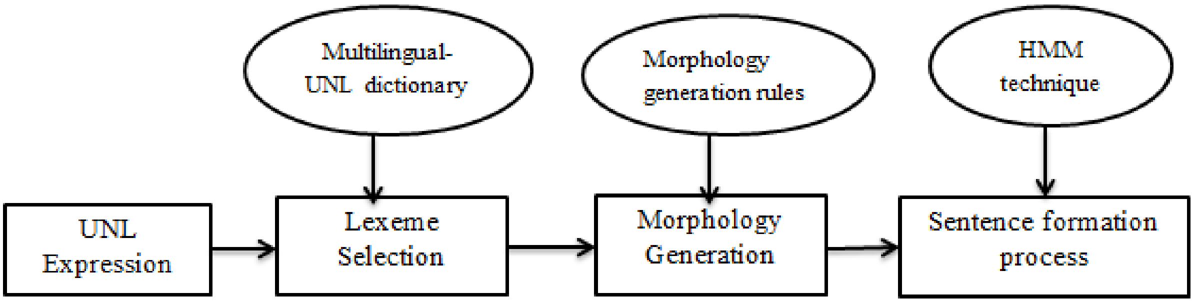
Multilingual-UNL dictionary
Morphology generation rules
HMM technique
UNL Expression
Lexeme Selection
Morphology Generation
Sentence formation process

Figure 1. Archicteture diagram of tamil, hindi deconverter system

Some example entries of dictionary for hindi words related to technical text are given below:

[डेटा संरचना]"datastructure(icl>subject)"(noun, thing,)<T,0,0>;

[का आयोजन]"organizing(icl>arrangement)"(verb,present, order,assemble)<T,0,0>;

[डाटा] "data(icl>information)"(noun)<T,0,0>;

[अनेकता]"stack(icl>subject)"(verb)<T,0,0>;

[याद]"memory(icl>information)"(noun)<T,0,0>;

B. *Morphology Generation:*

The headwords are modified according to the morphology of the target language. The system makes use of morphology generation rules during this process. These generation rules are designed on the basis of analysis of Tamil and Hindi morphology carried out for this purpose. The morphology rule deals with creation of Tamil and hindi words on the basis of UNL attributes attached to a node.

The root words retrieved from Multilingual-UW dictionary is changed in this phase depending on their attributes such as number, tense, aspect.

Table I. Morphology Rules for the Tamil Words Related to E-content

ROOT WORD	*ATTRIBUTES*	*SUFFIX*
ஏற்பாடு(organize)	@present, @progress	ஏற்பாடுக்கும்
கணக்கீடு(calculation)	@plural	கணக்கீடுகள்
சிறப்பு(special)	@past	சிறந்த
பில்லியன்(billion)	@plural	பில்லியன்கள்
சேமிப்பு(storing)	@present, @progress	சேமிப்பதற்கும்

Table 2. Morphology Rules for the Hindi Words Related to E-content

ROOT WORD	*ATTRIBUTES*	*SUFFIX*
ஏற்பாடு(organize)	@present, @progress	आयोजन
கணக்கீடு(calculation)	@plural	गणना
சிறப்பு(special)	@past	विशेष
பில்லியன்(billion)	@plural	अरबों
சேமிப்பு(storing)	@present, @progress	भंडारण के

C. *Sentence Formation Process:*

Sentence formation is a progression of linearizing the words presented in the UNL expression. It is a process to define the word order in the generated sentence. It deals with the arrangements of words in generated output so that output matches with the natural language sentence. Tamil and hindi languae basically follows SOV word order and is relatively free order. The each UWs word in the UNL expression was independent from each other. To formulate the sentence correctly the machine learning technique was included in the work.

A supervised machine learning technique such as hidden Markov model (HMM) can be used to solve the problem of word ordering in sentence formation process [7]. Hidden Markov models are probabilistic models that can assign likelihoods to all possible combinations of matches, and mismatches to determine the most likely sequence alignment or set of possible sequence alignment of words.

HMM is composed of a number of interconnected states, each of which emits an output symbol as the UNL expression. A sequence is generated by starting at an initial state and moving from state to state until a terminal state is reached.

IV. EVALUATION METRICS

The UNL expressions are given as input to the DeConverter for their DeConversion to Tamil and hindi language. The output of the DeConverter is compared with the corresponding manually translated sentences from English given in the gold standards. The bilingual evaluation understudy (BLEU) score has also been calculated to evaluate the quality of machine translated output.

BLEU is used to measure the translation closeness between a candidate translation and a set of reference translations with a numerical metric. BLEU's output value ranges between 0 and 1.

The value indicates the resemblance between the machine translated text and the human translated text. The value 1 indicates the more similarity between the translations. The candidate texts must be identical to a reference translation [8].

The BLUE formula is written as follows:

$$BLEU = BP \bullet \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \qquad (1)$$

The weighting factor Wn, is set at 1/*N*

The formula to calculate the brevity penalty is,

$$BP = \begin{cases} 1, & if|c| > |r| \\ e^{(1-\frac{|r|}{|c|})}, & if|c| \leq |r| \end{cases} \qquad (2)$$

Where |*c*| denotes the length of the candidate translation and |*r*| denotes the length of the reference translation.

N-gram precision in BLEU is computed as follows:

$$P_n = \frac{\sum_{C\epsilon(candidates)n-gram\,\epsilon\,C}\sum Count\ clip\ (n-gram)}{\sum_{C\,\in(candidates)n-gram\,\in C}\sum Count(n-gram)} \qquad (3)$$

Where Countclip (n-gram) is the maximum number of n-grams co-occurring in a candidate translation and a reference translation, and Count (n-gram) is the number of n-grams in the candidate translation.

V. CONCLUSION

In this work a deconverter was developed to convert the UNL expression related to E-content into the tamil and hindi sentences. It mainly comprises of the natural language dictionary to choose the appropriate natural language word for the equivalent universal words. The words are selected based on the attributes associated with the universal word. The morphology rules are created to change the natural language verbs. A Hidden Markov Model was implemented in this work to improve the efficiency of word ordering in sentence sequence process. The proposed system is efficient in generating an equivalent target sentence for the UNL source sentences. To evaluate the quality of translated output BLUE evaluation technique was used in this work.

VI. REFERENCES

[1]. V. Geetha and T. Dhanabalan, "UNL deconverter for tamil", Proceedings of 03 International Conference on the Convergence of Knowledge, Culture, Language and Information Technologies, December 2 - 6, 2003.

[2]. Parteek kumar, Rajendra kumar sharma, "Punjabi DeConverter for generating Punjabi from Universal Networking Language", Journal of Zhejiang University-SCIENCE C (Computers & Electronics), Zhejiang University and Springer-Verlag Berlin Heidelberg 2013.

[3]. Biji Nair, Rajeev R. R, Elizabeth Sherly, "Language Dependent Features for UNL-Malayalam Deconversion", International Journal of Computer Applications (0975 – 8887) Volume 100– No.6, August 2014.

[4]. Smriti Singh, Mrugank Dalal, Vishal Vachhani, Pushpak Bhattacharyya, Om P. Damani, "Hindi Generation from Interlingua (UNL)", Indian Institute of Technology, Bombay (India).

[5]. Aloke Kumar Saha , Muhammad F. Mridha Jugal Krishna Das, "Semantic Analysis of Bangla Language for Developing A UNL Deconverter", International Journal of Advanced Research in Computer Science and Software Engineering 2 (12), December - 2012, pp. 273-278.

[6]. Md. Sadequr Rahman, Muhammad Firoz Mridha, Mohammad Nurul Huda, and Chowdhury Mofizur Rahman, "Structure of Dictionary Entries of Bangla Morphemes for Morphological Rule Generation for Universal Networking Language", IEEE, Dhaka, Bangladesh, 2010.

[7]. Stephan vagel, Hermann Ney, Christoph Tillmann, " HMM-Based word alignment in statistical translation", Germany.

[8]. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.