UNL DECONVERSION FOR SENTENCE REALIZATION IN TAMIL

D. Josephine Sylvia[1], S. Lakshmana Pandian[2]
[1,2]Dept. of Computer Science and Engineering
Pondicherry Engineering College
Puducherry, India
Josephinesylvia07@gmail.com, lpandian72@pec.edu

Abstract. — Natural language processing (NLP) is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages. Universal Networking Language (UNL) is an intermediate representation for Interlingua based machine translation. The Universal Networking Language (UNL) is an electronic language in the form of a semantic network for computers to express and exchange every kind of information. The UNL approach is a hybrid approach of the rule and knowledge-based approaches to machine translation. The Deconverter is a Language Independent Generator (LIG), which provides a Framework for Syntactic and Morphological generation of Native Language. In this proposed work, Deconvertor can convert UNL expressions into native language (Tamil), using a language specific set of Word Dictionary, Grammatical Rules and Co occurrence Dictionary by ensuring that the meaning of the input sentence is preserved.

Keywords: Natural language processing (NLP), Universal Networking Language (UNL), Morphological generation, Deconvertor

1 Introduction

Natural Language Processing (NLP) is a field of computer science, artificial intelligence and linguistics and 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. Natural Language Processing (NLP) is a sub-field of Artificial Intelligence that is focused on enabling computers to understand and process human languages, to get computers closer to a human-level understanding of language. Natural language processing (NLP) aims to design and build software that will analyze, understand and generate languages that humans use naturally [1]. Machine translation is a very important application in natural language processing.

In order to overcome the language barrier, many attempts have been made in the past. Professional translators have been bridging such a communication gap. The quantity of translation by human however is rather small as compared to the required communication needed for the different languages. The main reason for such a limitation is high cost involved in the translation work. In addition, the number of translators for minor language is rather small[2]. Translation made by human being, thus has its limitation in terms of cost and human resources. Computer translation systems have made significant progress. Some of them are now being incorporated in network browsers.

The two demands for these systems indicate how large the language problem is among Internet users. Computer translation systems are useful under limited conditions. For instance, the user can evaluate and modify a translated document in his own language, but seldom in the other language. However, after translating with a computer, the user has to work to edit the output document. In addition it would require language knowledge to edit the translation of the document in the other language. In sending information throughout the world, the sender normally does not know the language of the recipient. In this case, the sender is bound to use a computer translation system blindly, because the user cannot check whether the translated results are correct or not. This is a serious limitation in current computer translation system.

The Universal Networking Language (UNL) is an electronic language for describing, summarizing, refining, storing information in a machine and natural-language-independent form. UNL, as a language for expressing information and knowledge described in natural languages, has all the components corresponding to that of a natural language [5]. UNL represents sentences in the form of logical expressions, without ambiguity. These expressions are not for humans to read, but for computers. Thus, UNL is an intermediate language to be used through the Internet, which allows communication among people of different languages using their mother tongue.

2. Related Work

Rajeswari Sridhar et al.[1] proposed English to Tamil machine translation system using universal networking language. The system aims at translating a given English sentence to a Tamil sentence, which conveys the meaning of the input, and ensures a grammatically correct sentence as the output. This system was evaluated using bilingual evaluation understudy (BLEU) score. The system is simple and efficient of using UNL for translation and as the result the system is easily scalable. Phan Thi Le Thuyen et al.[2] proposed Automatic translation for Vietnamese based on unl language. A tool is introduced based on UNL application and reused for the process of encoding a Vietnamese sentence into UNL expression and decoding an UNL expression into Vietnamese. Tools used for EnConverter are IAN tool (Interactive Analyzer)and EnCo tool. Tools used for DeConverter are EUGENE tool and DeCo tool. The converter tools from natural language to UNL are effective and the quality is pretty good and acceptable.

Imane Taghablout et al.[3] designed Amazigh verb in the Universal Networking Language. The system focus on presenting inflectional paradigms, and the lexical mapping stage needed for building an "Amazigh dictionary" for the verbal category according to the UNL specifications. It aims at eliminating linguistic barriers and, furthermore, promoting the access to information in the autochthone languages. Aloke Kumar Saha et al.[4] designed Design and Implementation of an Efficient DeConverter for generating Bangla Sentences from UNL Expression. The system focuses on the Linguistic Analysis of Bangla Language for the DeConversion process. A set of DeConversion rules have been burgeoned for converting UNL expression to Bangla. The system is tested with more than 2000 UNL Expressions. The System achieved accuracy as 89%, which can be marked outstanding in this Field of Study.

Biji Nair et al.[5] proposed Language Dependent Featues for UNL-Malayalam Deconversion. The deconverting generator for Malayalam language is done using Universal Networking Language (UNL) for Machine Translation. The deconversion is tested against 100 Malayalam Sentences that has achieved an appreciable F-measure score of 0.978. The system is efficient in generating syntactically unambiguous semantically equivalent target sentence for the UNL source sentences. Ananthi Sheshasaaye et al.[6] tackled The Role of Morphological Analyzer and generator for Tamil language in Machine Translation Systems. Statistical machine translation plays a predominant role in machine translation of larger vocabulary tasks. Achieving this goal is not an easy task especially when it comes to languages like Tamil which are agglutinative in nature. Deep analysis is needed at the word level to confine the correct meaning of the word from its morphemes and categories. The computational implementation of analysing natural language is done by Morphological analyzer. This formed a ground work for better understanding of various approaches that are used to develop morphological analyzer and generator of Tamil languages.

S. Lushanthan et al.[7] proposed Morphological Analyzer and Generator for Tamil Language. The system illustrates how the lexicon and the orthographic rules of the Tamil language have been written as regular expression using finite state operations. This model is built using Xerox toolkit which uses “two level morphology”. The model may produce several results or forms in the look down process for nouns. This model uses a common fully automated transliteration scheme and implemented. Jisha P.Jayan et al.[8] proposed Morphological Analyser and Morphological Generator for Malayalam. Morphological Analyzer is a program for analyzing the morphology of an input word and the analyzer reads the inflected surface form of each word in a text and provides its lexical form while Generation is the inverse process. Both Analysis and Generation make use of lexicon. The suffix stripping method has been used for developing the morphological analyser and for developing the morphological generator the suffix joining method has been used.

Velliangiri Dhanalakshmi et al.[9] proposed Grammar Teaching Tools for Tamil language. The tools like Parts of speech Tagger, Chunker and Dependency parser for the sentence level analysis and Morphological Analyzer and Generator for the word level analysis were developed using machine learning based technology. The tool is developed for Malayalam, Kannada and Telugu languages. It also improves vocabulary, improve writing and reading skills and also speed up the learning of second language. T.Dhanabalan et al.[10] proposed UNL deconverter for tamil. The DeConverter from UNL to Tamil language is done using Universal Networking Language (UNL) as the intermediate representation. The information needed to generate the Tamil sentence is available at different linguistic levels. UNL greatly reduces the cost of developing knowledge or contents necessary for knowledge processing, by sharing knowledge and content.

3. Universal Networking Language

The Universal Networking Language is a computer language that enables computers to process information and knowledge. It is designed to replicate the functions of natural languages. Using UNL, people can describe all information and knowledge conveyed by natural languages for computers. As a result, computers can intercommunicate through UNL and process information and knowledge using UNL, thus providing people with a Linguistic Infrastructure (LI) in computers and the Internet for distributing, receiving and understanding multi-lingual information. Such multilingual information can be accessed by natural languages through the UNL System [7]. UNL, as a language for expressing information and knowledge described in natural languages, has all the components corresponding to that of a natural language.

UNL project is aimed at elimination of the language barrier. Main approach of this system is to represent information in the form of knowledge, using language independent Interlingua to represent knowledge. With this characteristic in mind, Universal Networking Language (UNL) is developed. UNL intermediates understanding among different natural languages. UNL represent sentences in the form of logical expression without ambiguity. It is an intermediate language, which allows communication among people of different language using their mother tongue. The UNL is a language specification for the exchange of information over the Internet [8]. The motivation behind UNL is to develop an Interlingua representation such that semantically equivalent sentences of all language have the same Interlingua representation. UNL plays the role of an interface between different languages to exchange information. UNL represent each sentence in the given text as set of relation.

4. Proposed Work

The Deconverter is a language independent generator, which provides a framework for syntactic and morphological generation synchronously. It can convert UNL Expressions into a variety of natural languages, by using respective word dictionaries and sets of grammar rules of deconversion of the target languages [10]. A word dictionary contains the information of words that correspond to UWs included in the input of UNL Expressions and grammatical attributes (features) that describe the behaviors of the words. Deconversion rules (grammar rules of deconversion) describe how to construct a sentence using the information from the input of UNL Expressions and defined in a word dictionary. The Deconverter converts UNL Expressions into sentences of a target language following the descriptions of Deconversion rules.

A "Deconverter" which generates natural language from UNL, plays a core role in the UNL system. The UNL Representation is given as input text [11]. Using the UNL Representation, a UNL Graph is generated using UNL Relations such as agt, mod, man etc. The Graph Analyzer analyzes the UNL graph to get tiling of text using Tamil words equivalent to the universal words are collected from the word dictionary. The Text Tiling is to arrange the sequence of terms to realize the sentence. Tamil is a morphologically rich language and hence a large amount of information can be generated in the morphological generation phase with the help of analyzing the UNL words and binary relations. Morphologically formed words with the relation and syntactic rules are used for the sentence formation process. This sentence formation process generates the Tamil language sentence. The Proposed Work Architecture is shown in figure 1.

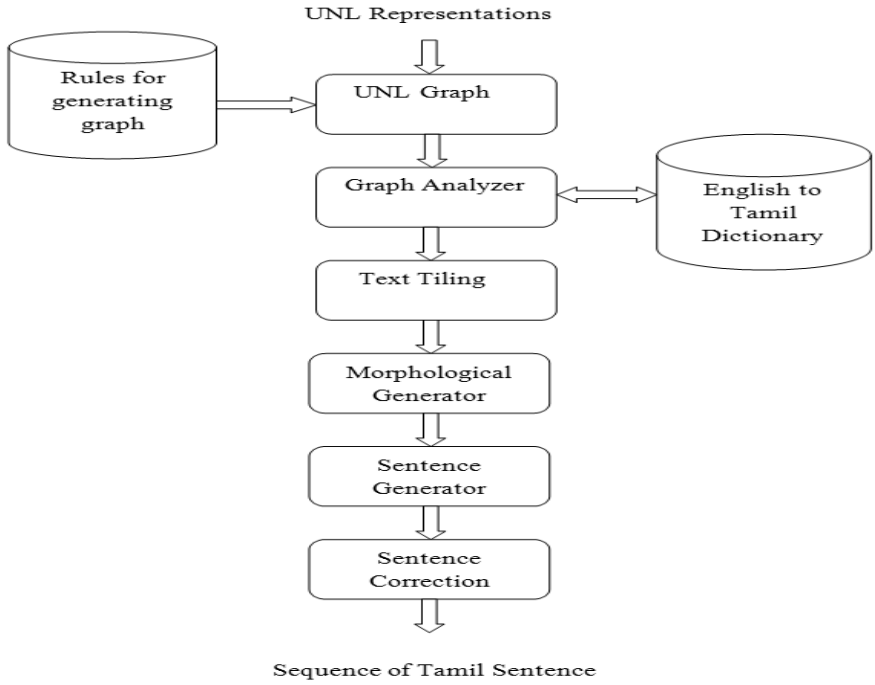

Figure 1. Proposed Work Architecture

4.1 UNL Representation

Universal Networking Language (UNL) is a computer language that enables computers to process information and knowledge across the language barriers. It is composed of UNL Expressions, Universal Words (UWs), Relations, Attribute.

•UNL Expression: UNL expresses information and knowledge in the form of semantic network. The semantic network of the UNL is a directed graph [12]. Such a semantic network of the UNL is called a "UNL Expression" or "UNL Graph". A UNL expression therefore consists of a UW or a set of binary relations. In UNL documents, a UNL expression for a sentence is enclosed by the tags {unl} and {/unl}.

•Universal Words (UWs): UWs are words of the UNL, constitute the UNL vocabulary. A combination of a set of UWs linked with each other through relations and modified by attributes expresses the meaning of a sentence.

•Relations: There are 46 relations in the UNL, such as 'agt', 'gol', 'obj', etc. They are used to connect every two UWs or scopes to construct the semantic networks of UNL Expressions. The relations are edges in the UNL graphs that constitute UNL Expressions.

•Attributes: Attributes are mainly for the purpose to describe subjectivity information. It includes time, aspect, emphasis, focus, topic, attitude, feeling and judgment.

4.2 UNL Graph

UNL expresses information and knowledge in the form of semantic network.The semantic network of the UNL is a directed graph.Such a semantic network of the UNL is called a "UNL Expression" or "UNL Graph".

4.3 Graph Analyzer

The Graph Analyzer analyzes the UNL graph to get tiling of text using Tamil words equivalent to the universal words are collected from the word dictionary. The Text Tiling is to arrange the sequence of terms to realize the sentence. Tamil words equivalent for the universal words are collected from the word dictionary [13]. For example, 'அவன்', 'மிக', 'ஓடு', 'வேகம்' are the Tamil word equivalents for the universal words

'he', 'very', 'run', and 'fast' respectively. From the binary relation in UNL, 'அவன்' is agent of the action 'ஓடு' and the verb 'ஓடு' is mentioned as past tense in UNL format. If the agent is third person singular means, the gender marker 'ஆன்' is added to the verb 'ஓடு' and the past tense marker 'n' is also added. 'வேகம்' defines the way to carry out an event 'ஓடு' and it is adverb. So 'ஆக' is added to the adverb 'வேகம்'. The word 'மிக' is intensifier and it modifies the adverb 'வேகம்' and it is simply added before adverb 'வேகம்'. So the generated sentence from the above information is 'அவன் மிக வேகமாக ஓடினான்'.

4.4 English to Tamil Dictionary

The word dictionary consists of about 1 Lakh English words and its equivalent Tamil words. The first step is to retrieve the Tamil word corresponding to every head word (HW) in the UNL [14] . The dictionary is used as a Universal Word dictionary. The query to retrieve the Tamil word matches the Headword. The primary task is to retrieve the relevant dictionary entries from the Tamil language word dictionary corresponding to the words in the word part of the UNL structures. The word entries of each language are stored in the Word Dictionary. Each entry of the Word Dictionary is composed of three kinds of elements: Headword, Universal Word (UW) and Grammatical Attributes. The headword is a notation/surface of word of a natural language. UW expresses the meaning of the word, which is to be used as a trigger or link for obtaining equivalent words or expressions. Grammatical Attributes are the information about the behaviour of the word in a sentence, which is to be used in deconversion rules.

4.5 Morphological Generator

A morphological generator is a program that performs the task of morphological generation. Morphological generation may be considered an opposite task of morphological analysis [15]. A morphological generator needs to be designed to tackle the different syntactic categories such as nouns, verbs, adjectives, adverbs. The general format of the morphological generator is Stem/root + suffixes · Word. The morphological generator generates morphological forms for nouns and verbs when the root word is given. The morphological generation mainly deals with the concatenation of corresponding suffixes with the root word to form a word of specific grammatical category.

The Morphological Generator takes lemma and a Morpho-lexical description as input and gives a word-form as output. The aim in morphological generation is to produce the inflected form of a word according to the features and values in the Feature Structure. It is also necessary to reuse the linguistic resources created for analysis purpose. From practical point of view, morphological generation is the inverse process of analysis, namely the process of converting the internal representation of a word to its surface form [16]. The same rule definitions can be used to generate the desired word form as used for analysis. The only difference will be the direction of execution order of the elements in the rule definition. The morphological generation mainly deals with the concatenation of corresponding suffixes with the root word to form a word of specific grammatical category.

The input of the morphological generator would be the root word which then inflects this word to the morphology of the respective language and gives as the output the target forms of the word [17]. The Morphological structure of Tamil verb is quite complex since it caters to person, gender, and number markings and also combines with auxiliaries that indicate aspect, mood etc. While morphologically generating the verb, the gender, number and person of the subject is necessary in order to select the appropriate suffix catering to the selected tense.

Tamil is a morphologically rich language and hence a large amount of information can be generated in the morphological generation phase with the help of analyzing the UNL words and binary relations. The word level Morphological generator for Tamil generates the derivative word for the root word and the various features conveyed by the suffixes.

4.6 Sentence Generator

Morphologically formed words with the relation and syntactic rules are used for the sentence formation process. This sentence formation process generates the Tamil language sentence. After adding the suffixes to the noun and the verb forms of the root words, the words need to be framed into a sentence. Tamil grammar is used for the creation of Tamil sentences [18]. The verbs in the input sentence are considered as central nodes.

The subject that lies just before the verb is added to the verb. Finally, the sentence is formed just by concatenation, i.e. the subject comes first, the phrase in the reversed order, then the verb, and this might be followed by any number of similar patterns. In the case of compound/ complex sentences, Tamil words corresponding to a conjunction are inserted.

4.7 Sentence Correction

The sequence of generated Tamil sentences is compared with the existing sentences in corpus, if both the suffix get matches then the output of the generated Tamil sentence is correct. Otherwise the more relevant sentences of the corpus information are used to correct the generated sentence.

5. Experimental Result

The UNL represents information sentence by sentence. Each sentence is converted into a directed hyper graph having concepts as nodes and relations as arcs [19]. The knowledge within document is expressed in three dimensions: Word knowledge is expressed by Universal Words (Uws). Concept Knowledge is captured by relating UWs through a set of UNL relations. The UNL Representation is given as input text. Using the UNL Representation, a UNL Graph is generated using UNL Relations such as agt, mod, man etc.

He ran very fast
[W]
He (icl>person) @generic:0
very (icl>concept) @generic:1
run (icl>do) @past .@entry:2
fast (aoj>thing) @generic:3
[/W]
[R]
2 agt 0
3 man 2
1 mod 3
[/R]

Here ‘agt’, ‘man’ and ‘mod’ are the relation labels.‘He(icl>person)’,‘very(icl>concept)’,‘run(icl>do)’ and ‘fast(icl>thing)’ are the Uws. ‘@entry’ is used to indicate entry or main UW of a sentence. ‘@generic’ is used to indicate generic concept. ‘@past’ is used to indicate the time with respect to the speaker.

The figure 2 shows the UNL Graph. UNL expresses information and knowledge in the form of semantic network. The semantic network of the UNL is a directed graph. Such a semantic network of the UNL is called a “UNL Expression” or “UNL Graph”.

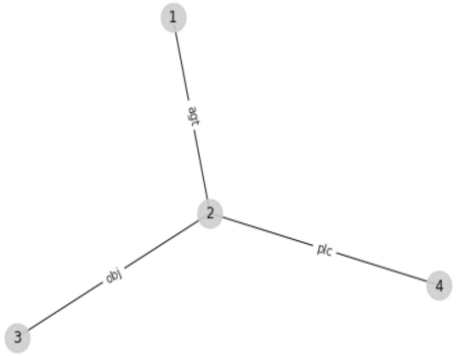

Figure 2. UNL Graph

The Figure 3 shows the Word Dictionary. The Word Dictionary contains English words and its equivalent Tamil words. The dictionary is used as a Universal Word dictionary. The query to retrieve the Tamil word matches the Headword.

Figure 3. Word Dictionary

The Figure 4 shows the Translation of English Words. The word or text is translated from English to Tamil using Word Dictionary.

Figure 4. Translation

The process of converting any word from one language to another without changing its pronunciation and phonetics is known as Transliteration. The Figure 5 shows the Trigram based reward value. Transliteration is the process of transferring a word from the alphabet of one language to another [20]. Transliteration helps people pronounce words and names in foreign languages. In translation transliteration is used for named entities. The Figure 6 shows the Transliteration.

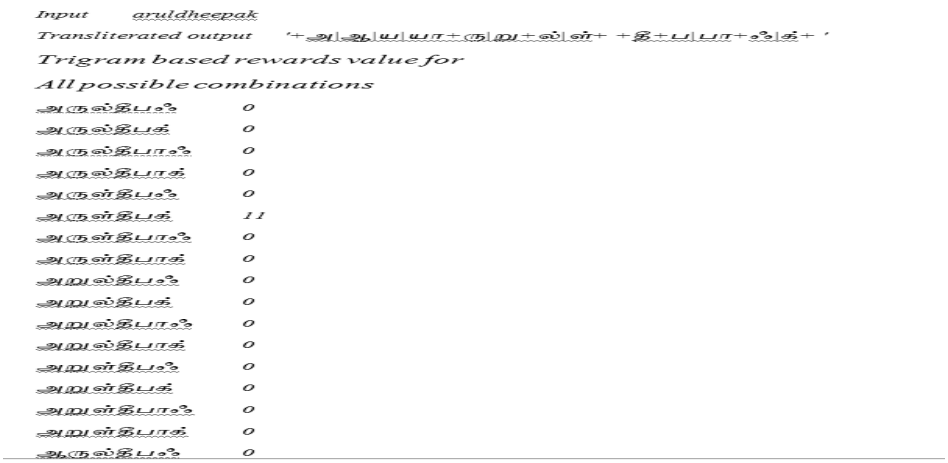

Figure 5. Trigram based reward value

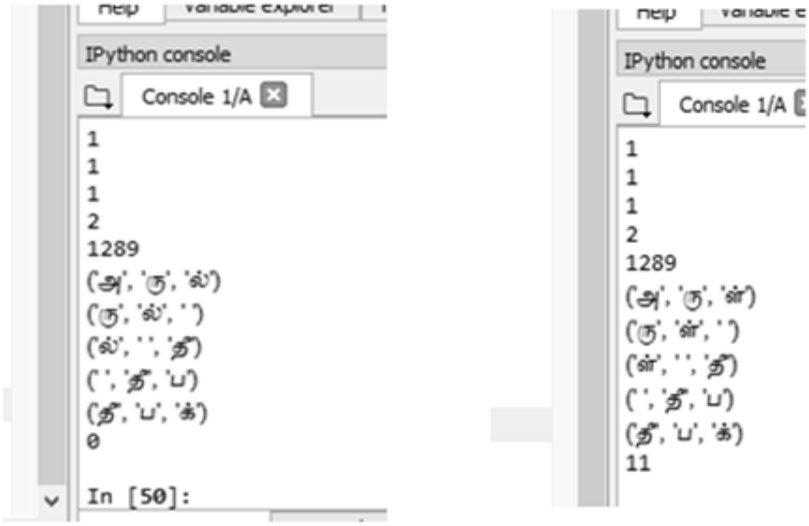

Figure 6. Transliteration

The morphological generator generates morphological forms for nouns and verbs when the root word is given. The input text is given in type of a box. After the necessary Tamil noun rules are applied, the possible combinations of noun words are shown in Figure 7.

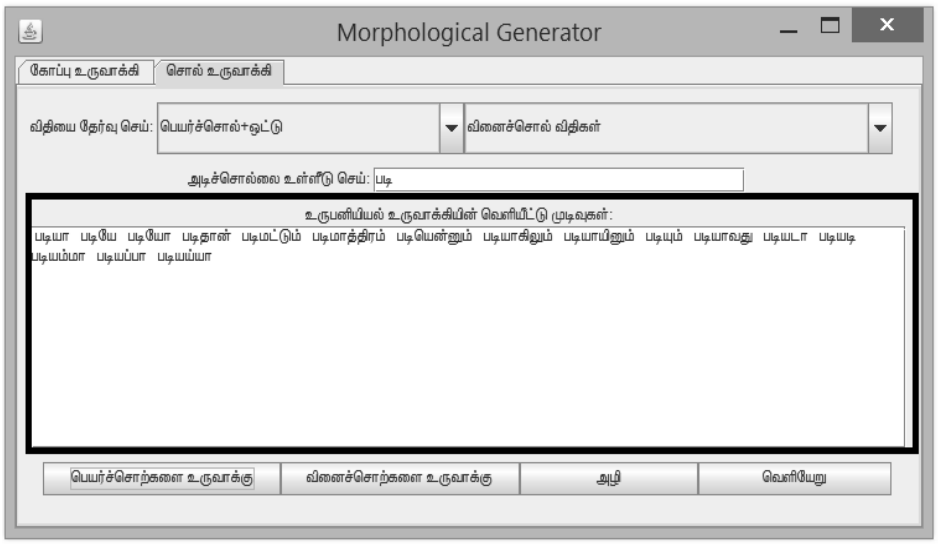

Figure 7. Output for Noun Generator

After the necessary Tamil verb rules is applied, the possible combinations of verb words are shown in figure 8.

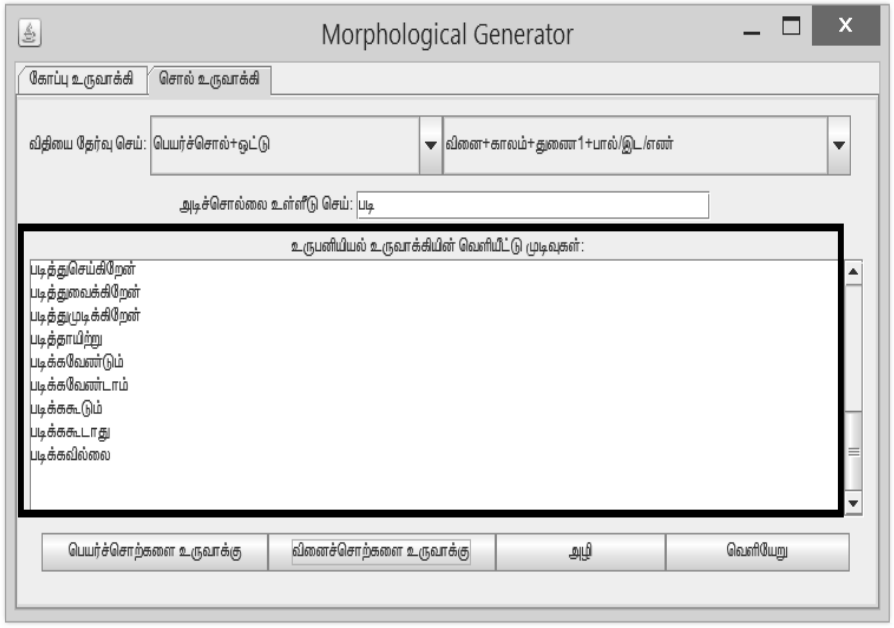

Figure 8. Output for Verb Generator

The figure 9 shows Sentence Generator. Morphologically formed words with the relation and syntactic rules are used for the sentence formation process. After adding the suffixes to the noun and the verb forms of the root words, the words need to be framed into a sentence.

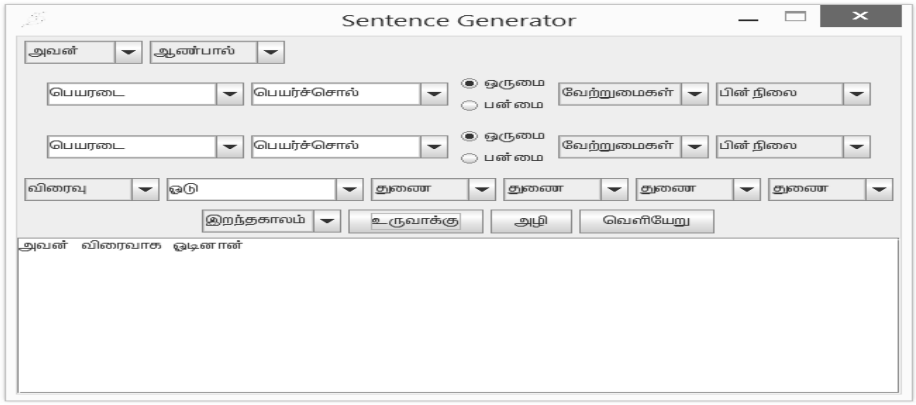

Figure 9. Sentence Generator

The figure 10 shows the Corpus for Sentence Correction. This corpus is compared with the sequence of generated Tamil sentence in the Sentence Generator in Figure 9. If both the suffix gets match then the output of generated Tamil sentence is correct.

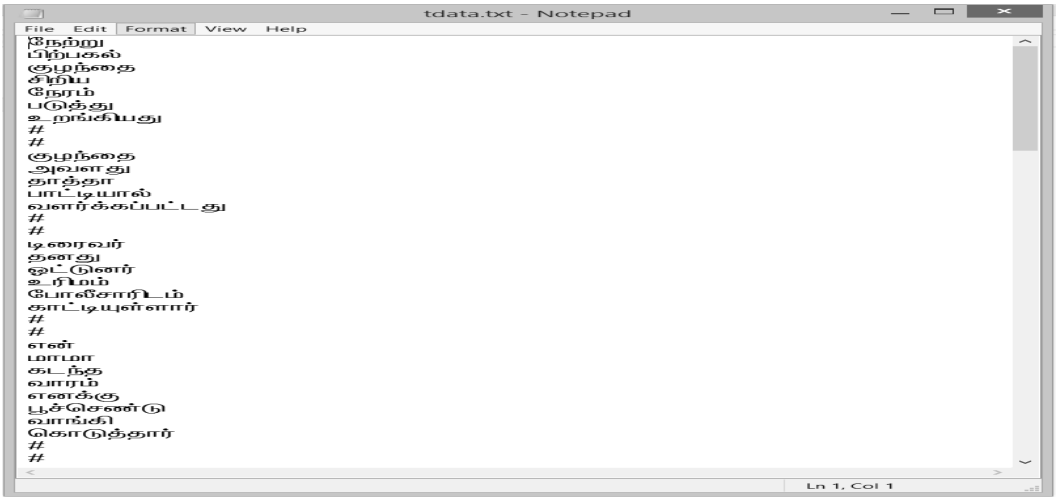

Figure 10. Corpus

6. Performance Evaluation

The performance of the entire system is evaluated using the parameter is described below.

Word error rate (WER): WER is a common metric to evaluate the performance of an MT system. It is determined by calculating the Levenshtein distance between those words in the candidate translation and the reference translation, which have a common prefix of at least three. As discussed already, the Levenshtein distance essentially gives the number of additions, deletions and modifications of words between the candidate and reference translations. WER is calculated for every word in the sentences of the enconversion. The results are then averaged over all the sentences in the corpus.

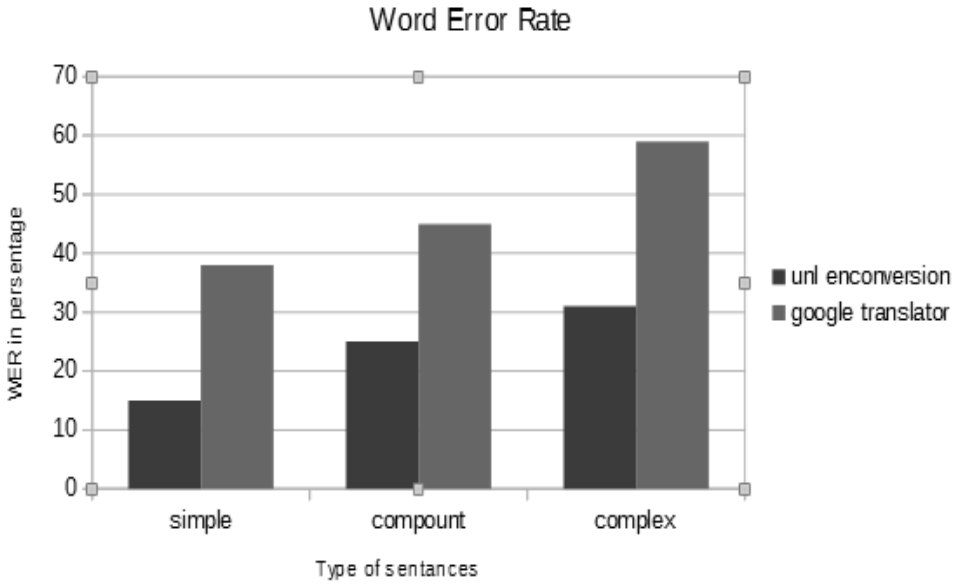

Figure 11. Bar chart showing the percentage of WER for different types of sentences

WER for our system using the UNL enconversion was calculated to be 25%, whereas for the Google translation it was 59%. Figure 5 gives the comparison of WER for simple, compound and complex sentences in the dataset, for our translation system as against the Google translation. The data in figure 11 clearly shows the advantages of incorporating UNL graph and a morphological generator in the system. The Google translation just uses a corpus based approach, and hence, complete sentences are not formed that convey the tense and gender of the subject. These results show that our newly created UNL graph and morphological generator are quite competent, in that it generates words much closer to the ones in the reference translation. However, the morphological generator is prone to a few sandhi errors and also some conjunction errors. Errors in the POS tagger also caused wrong suffixes to be added to the root word. The efficiency of our translation system could be improved to reduce the WER, by further enhancing the rules in the morphological generator.

7. Conclusion and Future Enhancements

In this work, a UNL deconversion for sentence realization in Tamil from UNL representation is developed. It deals with many issues and the required solutions are introduced. The UNL eliminates the need to study both target and source language by the programmer. Once a document is converted into UNL it can be converted into native language using the corresponding deconverter for that language. Thus it eliminates the need for individual machine translation development. In Tamil most information for generating sentence from UNL structure is tackled in morphological and syntactical level. First, the UNL representation is given as input and the UNL Graph is generated and analyzed. The Tamil words equivalent for the universal words are collected from the word dictionary and arranged in a sequence order. The use of UNL as intermediate representation makes translation of Tamil language available worldwide uses a standardized format.

In the future, we can improve the efficiency of the deconverter by introducing the number of heuristics that are currently determining the syntax plan and making them more precise by testing there results on large corpus. To extend the syntax planning, lexical knowledge of the universal words are to be used. In the current implementation, the Syntax Plan is generated purely on the basis of relation labels. But the belief is that in large complicated sentences the lexical information of the UW will have a significant effect on the syntax plan and to make the system more linguistically more strong. This system can be further

improved by removing disamibuigty of words by employing word net and extending the corpus dictionary to include more number of words in it.

REFERENCES

1. Rajeswari Sridhar, Pavithra Sethuraman and Kashyap Krishnakumar, "English to Tamil machine translation system using universal networking language",Vol.No.4, Issue No.6, pp no.607-620, March 2016.
2. Le Thuyen, Phan Thi, and Vo Trung Hung, "Automatic translation for Vietnamese based on UNL language", International Conference on Electronics, Information, and Communications (ICEIC), pp no.1-5, Jan.2016.
3. Baljeet Kaur Dhindsa and Dharam Veer Sharma,"Translation Challenges and Universal Networking Language", International Journal of Computer Applications(IJCA),Vol. No. 133, Issue No.15,pp no. 36-40, January 2016.
4. Imane Taghablout, FadouaAtaa Allah and Mohamed marraki, "Amazigh verb in theUniversal Networking Language", IEEE Conferences of Computer Systems and Applications (AICCSA),Vol. No.4, Issue No.6,pp no. 1-4, November 2015.
5. M. F. Mridha, Aloke Kumar Saha, Md. Akhtaruzzaman Adnan, MollaRashied Hussein and Jugal Krishna Das, "Design and Implementation of an Efficient Enconverter for Bangla Language", ARPN Journal of Engineering and Applied Sciences(ARPN),Vol. No. 10, Issue No. 15, pp no. 6543-6548, August 2015.
6. Biji Nair, Rajeev R and Elizabeth Sherly, "Language Dependent Features for UNL-Malayalam Deconversion", International Journal of Computer Applications (IJCA),Vol. No.100, Issue No.6,pp no.37-41, August 2014.
7. Ananthi Sheshasaayee and Angela Deepa V.R, "The Role of Morphological Analyzer and generator for Tamil language in Machine Translation Systems", International Journal of Computer Science and Engineering (IJCSE), Vol. No.2, Issue No.5, pp no.107-111, 2014.
8. S.Lushanthan, A. R. Weerasinghe and D. L. Herath, "Morphological Analyzer and Generator for Tamil Language", IEEE International Conference on Advances in ICT for Emerging Regions (ICTer), Vol No.7, Issue No.5, pp no.190-196, December 2014.
9. Kumar, Parteek, and Rajendra Kumar Sharma. "Punjabi Deconverter for generating Punjabi from universal networking language", Journal of Zhejiang University Science C , pp no.179-196, March 2013.
10. Hameed, M. S., Subalalitha and C. N., Geetha T. V, "A deconverter framework for Malayalam", In Proceedings of the International Conference on Advances in Computing, Communications and Informatics, pp no. 847-856, August 2012.
11. Bhattacharyya, "Multilingual Information Processing Using Universal Networking Language", IndoUK Workshop on Language Engineering for South Asian Languages (LESAL), April 2001.
12. Velliangiri Dhanalakshmi, M Anand Kumar and R U Rekha, "Grammar Teaching Tools for Tamil language", International Conference on Technology for Education, pp. 85-88 , July 2010.
13. Md. NawabYousuf Ali, Jugal Krishna Das and S. M. Abdullah Al- Mamun, "Specific Features of a Converter of Web Documents from Bengali to Universal Networking Language", IEEE International Conference on Computer and Communication Engineering,Vol No. 8,Issue No. 2, pp no. 726 - 731,2008.
14. Dhanabalan, T., & Geetha, T. V, "UNL deconverter for Tamil", International Conference on the Convergences of Knowledge, Culture, Language and Information Technologies, December 2003.
15. M.N.Y.Ali, A.M.Nurannabi, G. F. Ahmed, J.K.Das, "Conversion of Bangla Sentence for Universal Networking Language", International Conference on Computer and Information Technology (ICCIT), Dhaka, pp.108-113, 2010.
16. Aloke Kumar Saha, M. F. Mridha, Jahir IbnaRafiq and Jugal Krishna Das, "Data Extraction from Natural Language Using Universal Networking Language",IEEE International Conference on Current Trends in Computer, Electrical, Electronics and Communication" (ICCTCEEC) ,Vol No.2, Issue No. 7,pp no.24-29, September 2017.
17. Hiroshi Uchida, Meiying Zhu, "The Universal Networking Language beyond Machine Translation", UNDL Foundation, September 2009.
18. Dey, K., and Bhattacharyya, P, "Universal Networking Language based analysis and generation of Bengali case structure constructs", Universal Network Language: Advances in Theory and Applications, pp no. 215-229, 2005.
19. For Universal Networking Language: Universal Networking Digital Language Foundation. http://www.undl.org/
20. AjiNugraha , SantosaKasmaji and AyuPurwarianti, "Employing Natural Language Processing to Analyse Grammatical Error in a Simple Japanese Sentence", IEEE International Conference on Electrical Engineering and Informatics(ICEEI), Vol No. 3, Issue No. 7, pp no. 82-86, August 2015.

One-ANAPHORA RESOLUTION IN TAMIL

Vijay Sundar Ram
AU-KBC Research Centre, MIT Campus of Anna University, Chennai
sundar@au-kbc.org

Abstract. Natural language is cohesive. The cohesiveness is brought by various phenomena. Reference is one of the phenomena which brings cohesiveness. Reference includes different anaphoric expressions and *one*-anaphora is one among them. It is less attempted in Indian languages and there is no published work in Tamil. We have described *one*-anaphora in Tamil discourse and presented a rule-based algorithm to identify and resolve the *one*-anaphors. We have evaluated the engine with a set of web News dailies. The results are encouraging.

Keywords: *One*-anaphora, Tamil, Reference, *One*-anaphora resolution

1 Introduction

The cohesiveness in natural language brings elegancy to the text. Reference is one of the major phenomenon which brings cohesiveness between intra and inter sententially. The reference markers are pronominal, reflexives, reciprocal, distributive, *one*-anaphor and noun-noun reference. Resolution of these reference markers plays a vital role in building semantic intensive NLP systems. In the present work, we describe one of the less attempted reference markers, *one*-anaphora in Tamil discourse. We present a rule-based algorithm to identify and resolve*one*-anaphors.Cardinals occur as anaphoric expressions in certain instances, these are defined as *One*-Anaphors.

We have dealt *One*-anaphora resolution in Tamil. Tamil is one of the south Dravidian language. It is morphologically rich and highly agglutinated language. It is a nominative-accusative language and clauses are introduced by non-finite verbs. Though Tamil is a relatively free word-order language, noun phrases and clauses have rigid structures. In Tamil, genitive drop, accusative drop, copula drop, and pro drop are allowed. In the following section, we will explain *one*-anaphora in Tamil.

Consider the following discourse (Ex.1).

maraththil ainthu puRaakkaL uLLana. (1.a)

Tree(N)+Loc five pigeons be(V)+past

(There are five pigeons in the tree.)

iraNtu venniram maRRum muunRu saampalniram. (1.b)

Two white_colour and three grey_colour

(Two are white and three are grey.)

In Ex.1.b, there are two cardinals ‘iraNtu’ (two) and ‘muunRu’ (three). These two cardinals have occurred as anaphoric expressions. Both the cardinals ‘iraNtu’ and ‘muunRu’ refers to ‘ainthu puRaakkaL’ (five pigeons) in Ex.1.a.

The cardinals also occur with person, number and gender as ‘oruvan’ (one person 3 singular masculine), ‘oruththi’ (one person 3 singular feminine), ‘oruvar’ (one person 3singular honorific), ‘iruvar’ (two people plural honorific). Consider the following example.

viruthukkAka muuvar thernthuthetukkappattanar. (2.a)

Award(N)+adv three_people select(V)+past+3ph

(Three people were selected for the award.)

athil oruvar maruththuvar. null (2.b)

In_that one_person doctor (N) (copula)

(In that one is a doctor.)

Here in Ex.2.b, ‘oruvar’ (one person) in the second sentence is anaphoric and it refers to ‘muuvar’ (three people) in the first sentence.

Cardinals such as ‘oru’ (one), ‘iru’ (two) do not occur as anaphoric expressions, where as ‘onru’ (one), ‘iraNtu’ (two) etc can occur as anaphoric. These cardinals such as ‘onru’, ‘iraNtu’ etc also occur in listing of points. Consider the following example.

avan ***oru*** *nalla paiyan.* (3)

He(PN) one good boy.

(He is a good boy.)

Here in Ex.3, the cardinal ‘oru’ occurs as a quantifier preceding the head noun ‘paiyan’ (boy.) And it is not anaphoric.

avarukku ***onru*** *mattum pidiththathu.* (4)

He(PN)+dat one only like(V)+past+3s

(He liked one only.)

In Ex.4., cardinal ‘onru’ occurs as head noun. And it requires a referent.

There are many theoretical studies on one-anaphora such as Halliday & Hasan [1]; Webber [4]. There are very few attempts in computational system development. One of the earliest attempts in Indian languages is Vasisthpublished by Sobha & Patnaik [3]. The authors have classified the *One*-anaphora in Malayalam and Hindi on the basis of countability [+/-C]. The pronoun with [+C] can have features having [+count, +/-animate]. Consider the example pronouns such as ‘one’ is [+C], while ‘little’ is [-C]. The authors have presented a rule to resolve these anaphoric expressions which states that the antecedent NP is the non-subject NP in the immediate clause.

HweeTou et al. [2] has classified use of ‘one’ in English into six classes namely Numeric, Partitive, generic, Anaphoric, Idiomatic and Unclassified. They have a presented a computational system using C4.5 Decision tree, a machine learning technique to classify the ‘one’ and to resolve the anaphoric ‘one’.

2 Our Approach

We attempt to resolve *one*-anaphors with a rule-based approach. Here we follow Sobha & Patnaik [3] approach of classifying the *one*-anaphor. We try to resolve the +C *one*-anaphors, whose antecedents will also have +C characteristics. Consider the following examples Ex.5.

raamanitam ***pala cattaikaL*** *uLLana.* (5.a)

Raman(N)+Loc many shirts be(V)+past

(Raman has many shirts.)

athil ***onru*** *civappu niram.* (5.b)

In_that one red colour

(One is red in colour)

In Ex.5, the second sentence (Ex.5.b) has *one*-anaphor ‘onru’. It is has +C characteristics. It refers to ‘pala cattaikaL’ (many shirts) in the first sentence, which also have +C characteristics.

We have performed it in two steps.

1, Identification of *One*-Anaphors

2, Resolution of *One*-Anaphors with a rule-base approach.

2.1 Identification of *one*-anaphors

We try to identify the Cardinals that have occurred as noun phrase. Cardinals occurring as quantifiers in the noun phrase and Cardinals in the listing are not considers as these cardinals are not anaphoric. The algorithm is given below.

1) If Cardinal such as ‘onru’, ‘iraNtu’, etc occurs in a sentence, then step 2.
2) Check if the consecutive sentences have cardinals such as ‘onru’, ‘iraNtu’ in the starting of the sentences. If consecutive sentences do not have cardinals in the beginning of the sentences, then the cardinal is classified as anaphoric.

2.2 Resolution of *One*-Anaphor

After identifying the anaphoric cardinals, we proceed to resolve it using rule-based approach. We look for non-nominative noun phrases with [+C] characteristics in the immediately preceding clause or sentence. Step-wise process is described below.

1) If a sentence with *one*-anaphor with [+C] occurs, then look for non-nomimative NPs in the preceding immediate clause or sentence.
2) Check if the NP has [+C] characteristics, then it is chosen as antecedent NP.

2.2.1 Identification of NP with [+C] characteristics

Characteristic of NP with [+C] is determined based on the following two conditions.

1) If the NP has a quantifier then YES.
2) If the NP is in plural form then YES.

3 Corpus Description

We have collected 400 News articles from Tamil News dailies online versions, containing cardinals. We first scraped the text from the web pages and processed it with a sentence splitter and tokeniser. The sentence splitted and tokenised text is pre-processed with syntactic processing tools namely morphanalyser, POS tagger, chunker, pruner clause boundary identifier. The text enriched with shallow parsed information is fed to Named entity recogniser and thenamed entities are identified. The News articles are from Sports, Disaster and General News. The distribution of the cardinals is given in table 2.

Table 2: Distribution of Cardinals in the Corpus

S.No	Type	Number of Occurrence
1	Quantifier	186
2	Used in Listing	93
3	Anaphoric NPs	42

4. Experiments and Result

The text enriched with shallow parsing and Named Entity information is fed to rule based engine to identify the anaphoric cardinals and then processed with the *one*-anaphora resolution system. The performance measures namely precision, recall and f-measure are presented in the table 3.

Table 3: Performance Measures

S.No	**Precision (%)**	**Recall (%)**	**F-Measure (%)**
1	77.23	64.70	70.31

The output of the rule-based engine is analysed. The observations are as follows.

The accusative drop in Tamil introduces error in identifying the antecedents, as we look for non-nominative NP in the immediately preceding clause or sentence. Consider the following example.

raamu coomuvukku muunRu cattaikaL kotuththaan. (6.a)

Ramu(N) Soomu(N)+DAT three shirt(N)+pl give(V)+past+3sm

(Ramu gave three shirts to Soomu.)

athil onRu mangcaL iraNtu patcai niramaakum. (6.b)

In_that one yellow two green(N) colour be(V)

(In that one is yellow and two are green.)

In Ex.6.b has two cardinals ‘onRu’ (one) and ‘iraNtu’ (two). These *One*-anaphors refer to ‘muunRu cattaikaL’ (three shirts) in Ex.6.a. The rule to resolve *one*-anaphor looks for a non-nominative noun phrase with (+count) in the immediate preceding clause or sentence. But here in the immediately preceding sentence Ex.6.a, as there is an accusative drop, the noun phrase ‘muunRu cattaikaLai’ has occurred as ‘muunRu cattaikaL’ with the accusative marker ‘ai’ being dropped. As the referent NP occurs as nominative NP, the rule fails to identify the correct antecedent. The errors introduced by preprocessing modules namely POS tagger and chunker leads to wrong identification of anaphoric cardinals.

5. Conclusion

We have presented a description on *One*-anaphora in Tamil discourse, one of the reference marker, which brings cohesiveness to the discourse.We have presented a rule-based methodology to identify and resolve *One*-anaphora. We have tested the methodology with a set of web Tamil News dailies and obtained F-measure of 70.3%. The accusative drop in Tamil discourse poses a challenge in identifying the correct antecedents.

Acknowledgement

Authors thank *IMPRINT India initiative* for the support given to carry out this research.

Reference

1. Halliday, M & Hasan, R 1976, Cohesion in English, Longman, London.
2. HweeTou, N, Zhou, Yu, Dale, Robert, Gardiner& Mary 2005, ‘A Machine Learning Approach to Identification and Resolution of One-Anaphora’, Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI 2005), Edinburgh, Scotland, UK, pp. 1105-1110).
3. Sobha, L & Patnaik, BN 2000, ‘Vasisth: An Anaphora Resolution System for Indian Languages’, In Proceedings of International Conference on Artificial and Computational Intelligence for Decision, Control and Automation in Engineering and Industrial Applications, Monastir, Tunisia
4. Webber, B 1979, A Formal Approach to Discourse Anaphora, Garland Publishing Inc., New York &London.