Named Entity Recognition and Transliteration for Telugu Language

Kommaluri VIJAYANAND and R. P. Seenivasan
Department of Computer Science
School of Engineering and Technology
Pondicherry University
Puducherry – 605 014, India.
Email: kvixs@yahoo.co.in, rpsv@yahoo.com

1. Introduction

The concept of transliteration is a wonderful art in Machine Translation. The translation of named entities is said to be transliteration. Transliteration should not be confused with translation, which involves a change in language while preserving meaning. Transliteration performs a mapping from one alphabet into another. In a broader sense, the word transliteration is used to include both transliteration at the micro level and transcription.

Transliteration is a process in which words in one alphabet are represented in another alphabet. There are a number of rules which govern transliteration between different alphabets, designed to ensure that it is uniform, allowing readers to clearly understand transliterations. Transliteration is not quite the same thing as transcription, although the two are very similar; in transcription, people represent sounds with letters from another alphabet, while in transliteration, people attempt to map letters over each other, sometimes with accent marks or other clues to suggest particular sounds.

As we say technically the transliteration is the process of transforming the text in one writing system (Source language) to another writing system (Target Language) without changing its pronunciation. Transliteration is a very good asset for machine translation. Machine translation cannot translate some of the text. Because, there could not be correspond translation word in the bilingual dictionary. Those words are called out of vocabulary words (OOV). To overcome this OOV problem transliteration came into being. The transliteration involves the process of converting the character sequence in the source language to target language on the basis of how the characters are pronounced in source language.

Transliteration needs knowledge of characters in source and target language. Since the pronunciation is the aim goal of transliteration it is difficult to give exact transliteration. Because, the pronunciation of single character of the source language can have multiple character in the target language as the transliteration is done by character wise. In transliteration so far we can give possible transliterations and yet it is the great challenge to the researchers to give exact transliteration in target language.

People try to use standardized trends when they transliterate so that transliterations are uniform, but this does not always happen. Muhammad, for example, is spelled in a variety of ways, including Mohammad and Mahomet. This can be confusing, as changes in transliteration change the way that a word sounds when it is spoken out loud. A good transliteration also employs accent markings to guide people, and it may have unusual combinations of letters in an attempt to convey unique sounds. Transliteration is not the same thing as translation, a process in which

words are made meaningful to speakers of other languages. Translation requires knowledge of languages, where transliteration is more about alphabets.

2. The Origin of the System

Advances information technology leads to the discovery of transliteration. Today transliteration plays a major role in all aspects of the society. There are a number of reasons to use transliteration, but most of them involve conveying information across cultures. Transliteration is needed in our day – to- day life. Even translation cannot be fulfilled without this translation. The translation of named entities cannot be possible in machine translation. In every writings named entities play a major role. So without named entities a text cannot be fulfilled. The named entities can be transliterated and cannot be translated. So the translation system also needs transliteration.

We can explain the use of transliteration using an example. For example when a Telugu man who don't know to read English going to restaurant, if he see menu card which is in English he can't order anything because of his lack of English reading knowledge. Suppose the menu card consists of Telugu transliteration of those menus he can order the food items without knowing what it is.

In literature also transliteration plays a role. When the translator translates the novels or stories they need transliteration in case names of persons and places. Transliteration is also used in language education, so that people can understand how words are pronounced without needing to learn the alphabet as well. Academic papers may also use transliteration to discuss words in various languages without forcing their readers to learn an assortment of alphabets.

In the Internet also the transliteration is applied. Usually the web news is all in English. When we need it in any other language the websites has the facility to display it in that particular language. In that translated web page out of vocabulary words are transliterated.

In the natural language processing applications such as machine translation, cross language information retrieval, question answering system etc., the transliteration is used.

Initially there is a technical motivation of building intelligent computer system such as Machine Translation (MT) systems, natural language (NL) interfaces to database, man-machine interfaces to computers in general, speech understanding system, text analysis and understanding systems, computer aided instruction systems, system that read and understand printed or hand written text. Second, there is a cognitive and linguistic motivation to gain a better insight into how humans communicate using natural language.
For development of any natural language processing system, there are several sources of knowledge that are used in decoding the information from an input. These can be classified as follows:-

- Language knowledge
    (a) Grammar
    (b) Lexicon
    (c) Pragmatic and discourse. Etc.
- Background Knowledge

(a) General world knowledge (including common sense knowledge)
(b) Domain specific knowledge (includes specialized knowledge of the area about which communication is taking place)
(c) Context (Verbal or non-verbal situation in which communication is to take place)
(d) Cultural knowledge

From the various sources of knowledge mentioned above, a hearer (or a reader) can extract information conveyed form a given source (a speaker or writer).

3. The Methodology

In Grapheme-Based method, source words are transcribed to the target words based on grapheme units directly without making use of their phonetic representations. The grapheme based method is called direct method. The grapheme based technique is direct orthographical mapping from source graphemes to target graphemes.

The methods based on the source-channel model deal with English-Telugu transliteration. They use a chunk of graphemes that can correspond to a source phoneme. First, English words are segmented into a chunk of English graphemes. Next, all possible chunks of Telugu graphemes corresponding to the chunk of English graphemes are produced. Finally, the most relevant sequence of Telugu graphemes is identified by using the source-channel model. The advantage of this approach is that it considers a chunk of graphemes representing a phonetic property of the source language word. However, errors in the first step (segmenting the English words) propagate to the subsequent steps, making it difficult to produce correct transliterations in those steps. Moreover, there is high time complexity because all possible chunks of graphemes are generated in both languages. In the method based on a decision tree, decision trees that transform each source grapheme into target graphemes are learned and then directly applied to machine transliteration. The advantage of this approach is that it considers a wide range of contextual information, say, the left three and right three contexts.

Furthermore, they segment a chunk of graphemes and identify the most relevant sequence of target graphemes in one step. This means that errors are not propagated from one step to the next, as in the methods based on the source-channel model. The method based on the joint source-channel model simultaneously considers the source language and target language contexts (bigram and trigram) for machine transliteration. Its main advantage is the use of bilingual contexts.

3.1. The Algorithm

The present transliteration system is implemented using the algorithm narrated step wise as follows:

1. The input for this system is an xml file.
2. This xml file consists of only names in source language.
3. The xml file is read and the source names are extracted and stored in the array list.
4. Source names are retrieved from the array list one by one for the further process.
5. Then the source name is rewritten using rewriting Techniques.

6. The next step is segmentation
7. After segmentation the chunks retrieved from the array list where they are stored one by one for target grapheme retrieval
8. In the target grapheme collection process the source grapheme is compared with the database and all the relevant graphemes are collected and stored it in the array list
9. The target graphemes of first grapheme is stored in one array list and that target graphemes of other source graphemes are stored in one array list.
10. After generation of target names for the source names and it is stored in the xml file.

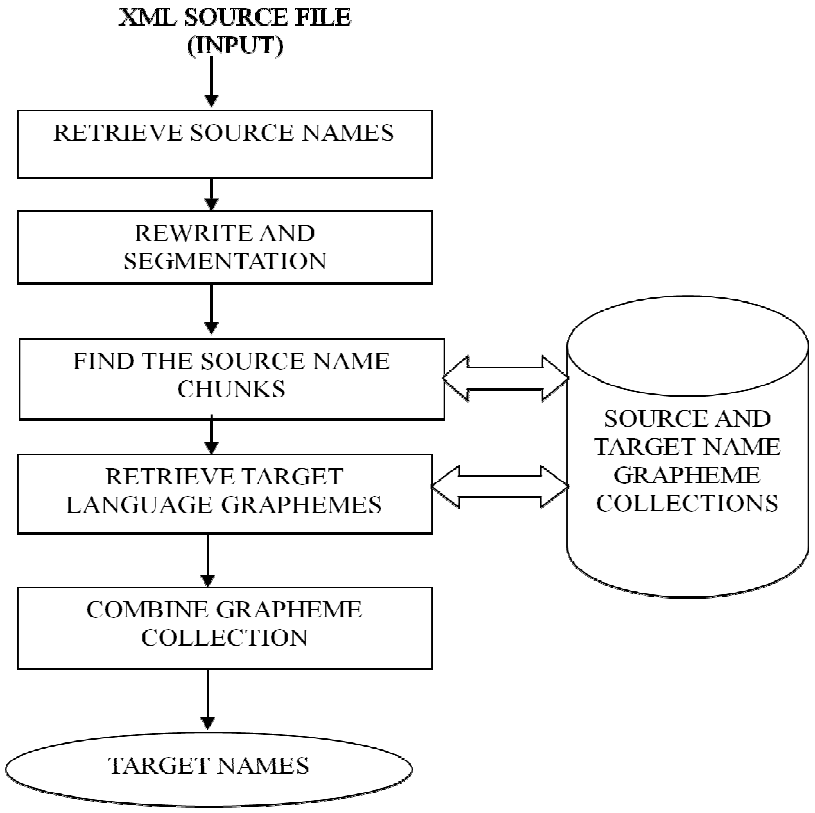

Figure 1: Block Diagram of the Machine Transliteration System

4. Implementation details

The System has been designed and implemented in java using swings for interface that takes various input queries from user and outputs the translated query in Telugu. The internal interaction and working of the system has been implemented using Java. The coding phase aims at translating the design of the system into code. The code has then been executed and tested. The goal is to implement the design in the best possible manner.

Rule-Based method

It requires analysis and representation of the meaning of source language texts and the generation of equivalent target language texts. Representation should be unambiguous lexically and structurally. There are two major approaches:

- The transfer approach in which translation process operates in three stage-analysis into abstract source language representations, transfer into abstract target language representations and generation or synthesis into target language text.
- The two stage 'interlingua' model where analysis into some language-neutral representation starts from this Interlingua representation.

Source Name Retrieval

The input for this system is an xml file. This xml file consists of only names in source language. The xml file is read and the source names are extracted and stored in the array list. Source names are retrieved from the array list one by one for the further process.

Rewrite and Segmentation

There are several rules and methods for rewriting and segmentation. Some of such rules are listed as follows:

- If the second index to the current index of the word is 'a' or 'e' or 'I' or 'o' or 'u' then it is considered as one segment.
- If the second index to the current index of the word is 'h' and the third index to the current index of the word is 'a' or 'e' or 'I' or 'o' or 'u' then it is considered as one segment.
- If the second and third index to the current index of the word is 'a' or 'e' or 'I' or 'o' or 'u' and it is same character i.e. 'aa', 'ee' ,'oo' then is considered as one segment.

- If the second index to the current index of the word the word 'a' , 'o' and the third index to the current index of word is 'e', 'u' then it is considered as one segment.
- If the second and third index to the current index of the word does not satisfy the above four conditions then the current index of the word is considered as one segment.
- After segmentation, the graphemes of source name (English) are compared with the database and target graphemes are collected.

After collecting target graphemes those graphemes merged to generate transliterations in target language (Telugu).

Source Name Chunks

This method was applied with the rule based algorithm. This algorithm is based on translating the linguistic rules into machine readable form. These rules are hand-crafted.

- The first step in the implementation is to rewrite the name.
- This step is used to reduce the unnecessary occurrence of 'h', repeated characters, and replace the characters having the same sound.

Retrieval of Target Language Graphemes

There are several handcrafted rules for rewriting process of named entities. They are:

- The next step in the algorithm is Segmentation.
- The segmentation is also done on the basis of handcrafted rules.

Segmentation is done with the rules as said before. In the segmentation process the names are segmented in to chunks using those rules and are stored in an array list. After segmentation the chunks retrieved from the array list where they are stored one by one for target grapheme retrieval. In the target grapheme collection process the source grapheme is compared with the database and all the relevant graphemes are collected and stored it in the array list. The target graphemes of first grapheme is stored in one array list and that target graphemes of other source graphemes are stored in one array list. The value of second array list is merged with first array list. The value of second array list changed dynamically. After generation of target names from the source names it will is stored in the xml file.

Conclusion

Based on the techniques and methods used to transliterate the named entities from English to Telugu language, we had found that for writing system comprises of the graphemes and phonemes that play major role in transliteration. The writing system for both the Tamil and Telugu languages is same and share common properties during transliteration system development. Thus application of Machine Learning would help in developing a common generator with different production algorithms based on the South Indian Languages like Kannada, Malayalam, Telugu and Tamil.

References:

[1]. Eduard Hovy and Chin-Yew Lin, Automated Text Summarization in SUMMARIST, In Advances in Automatic Text Summarization, 1999.

[2].George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller**,** Introduction to WordNet**,** 1993.

[3]. Harshit Surana and Anil Kumar Singh, A More Discerning and Adaptable Multilingual Transliteration Mechanism for Indian Languages, Proceedings of International Joint Conference on Natural Language Processing, Hyderabad, India, 2008.

[4]. Surya Ganesh, Sree Harsha, Prasad Pingali, and Vasudeva Varma., Statistical Transliteration for Cross Language Information Retrieval using HMM alignment and CRF, Proceedings of International Joint Conference on Natural Language Processing(_CNLP)-2008, NERSSEAL Workshop, Hyderabad, India, 2008.

[5]. The Unicode Standard version 3.0 (http://www.unicode.org)

[6]. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Second edition by Daniel Jurafsky, James H.Martin

[7]. T. Rama and K. Gali, Modeling machine transliteration as a phrase based Statistical Machine Translation Problem, In proceedings of the Named Entities Workshop, ACL–IJCNLP 2009, pp. 124-127, August 2009.

[8]. Nayan , B. R. K. Rao, P. Singh, S. Sanyal, and R. Sanyal, "Named entity recognition for Indian languages," In Proceedings of International Joint Conference on Natural Language Processing (IJCNLP), pp. 97-104, 2008.

[9]. N. A. Jaleel and L. S. Larkey, "Statistical transliteration for english-arabic cross language information retrieval," In Proceedings of the twelfth international conference on Information and knowledge management, November 03-08, 2003, New Orleans, LA, USA.