Morphological Cross Reference method for English to Telugu Transliteration

A. P. Siva kumar[1], Dr. P. Premchand[2] and Dr. A. Govardhan[3]

[1] Department of Computer Science and Engineering, JNTUACE Anantapur, India
sivakumar.ap@gmail.com

[2] Professor, Department of Computer Science Engineering, Osmania University, Hyderabad, India
p.premchand@uceou.edu

[3] Principal & Professor, Department of Computer Science Engineering, JNTUHCE, Nachupalli, India.
govardhan_cse@yahoo.co.in

ABSTRACT

Machine Transliteration is a sub field of Computational linguistics for automatically converting letters in one language to another language, which deals with Grapheme or Phoneme based transliteration approaches. Several methods for Machine Transliteration have been proposed till date based on nature of languages considered, but those methods are having less precision for English to Telugu transliteration when both pronunciation and spelling of the word is considered. Morphological cross reference approach provides user friendly environment for transliteration of English to Telugu text, where both the pronunciation and the spelling of the word is taken into consideration to improve the precision of transliteration system. In addition to alphabet by alphabet transliteration, this paper also deals with whole document transliteration. Our system achieved an correct transliteration with an accuracy of '78%' of Transliteration for Vocabulary words.

KEYWORDS

Transliteration linguistics, grapheme, phoneme.

1. INTRODUCTION

Transliteration is the technique of mapping text written in one language using the orthography of another language by means of a pre-defined mapping. In general, the mapping between the alphabet in one language and the other in a transliteration scheme will be as close as possible to the pronunciation of the word. Depending on various factors like mapping, pronunciation etc., a word in one language can have more than one possible transliteration in another language. This is more frequently seen in the case of transliteration of named entities and vocabulary words. This kind of transliterated text is often referred by the words formed by a combination of English and the language in which transliteration is performed like Telugu, Hindi etc. It is useful when a user knows a language but does not know how to write its script and in case of unavailability of a direct method to input data in a given language. However, English to Telugu transliterated text has found widespread use with the growth of Internet usage, in the form of mails, chats, blogs and other forms of individual online writing.

Telugu is one of the fifteen most spoken languages in the world, the third most spoken language in India which is the official language of Andhra Pradesh. Telugu has 56 alphabets, among them 18 are vowels and 38 are consonants and English has 26 alphabets among them 5 are vowels and 21 are consonants. By using Unicode mapping for phonetic variants of each vowel and

DOI : 10.5121/ijaia.2011.2402

consonant, English text can be transliterated to Telugu. One problem here in Transliteration is text input method [2]. Most of the users of Indian language on the Internet are those who are familiar with typing using an English keyboard. Hence, instead of introducing them to a new Telugu keyboard designed for Indian languages, it is easier to let them type their source language words using Roman script. For Indian Languages, many tools and applications have been designed for text input method [2]. However, Telugu still does not have a user efficient text input method and a user friendly environment, which is widely accepted and used, and an evaluation of the existing methods has not been performed in a structured manner to standardize on an efficient and accurate input method. Another problem with transliteration is, when we consider a word without knowledge of pronunciation, the transliteration (Grapheme) will be different from the transliteration (Phoneme) of the word with knowledge of pronunciation. So in this paper, we try to solve the above problem by combining Grapheme and Phoneme based Transliteration models to form a new Model called Morphological Cross Reference Method which produces correct transliteration for Vocabulary words with knowledge of pronunciation and without knowledge of pronunciation produces same transliteration for Out of Vocabulary words when compared with other transliteration systems.

In Graphemic approach, the source language word is split in to individual sounding elements. For example: bharath is split as bha-ra-th, b(బ్),h(హ్),a(అ) are combined to form bha(భ),r(ర్),a(అ) are combined to form ra(ర),t(ట్),h(హ్) are combined to form th(త్) by using an input mapping Table. The Table contains the phonetically equivalent combination of target language alphabets in terms of source language and its relevant Unicode hexadecimal value of target language alphabets. According to the source input the exact hexadecimal Unicode equivalent of the target language is retrieved and displayed as transliterated text.

Generally characters in English and Telugu languages do not adhere to a one-to-one mapping because English has 26 alphabets and Telugu has 56 alphabets. So our system combines Grapheme model with Phoneme based transliteration model in which a parallel corpus is maintained which contains source English words and Telugu phonetically equivalent Romanized text in terms of source language. For example: ‘period’ English word has its relevant Romanized text as ‘piriyad’. If ‘period’ is transliterated using Grapheme based model then the result is ‘పెరిఓడ్ ’’ but by combining Grapheme with Phoneme we can get exact transliteration which is 'పీరియడ్'.

Our system provides an user friendly environment which is platform and browser independent, case insensitive to the vocabulary words which are placed in parallel corpus, case sensitive to the general text, so our transliteration system will work very fast and provides accurate results when compared to the other transliteration systems like Google, Baraha, Quillpad etc.

2. RELATED WORK

There has been a large amount of interesting work in the arena of Transliteration from the past few decades.

Antony P.J, Ajith V.P, Soman K.P [1] proposed the problem of transliterating English to Kannada using SVM kernel which is modelled using sequence labelling method. This framework is based on data driven method and one to one mapping approach which simplifies the development procedure of transliteration system.

V.B. Sowmya, Vasudeva Varma [2] proposed a simple and efficient technique for text input in Telugu in which Levenshtein distance based approach is used. This is because of the relation between the nature of typing Telugu through English and Levenshtein distance.

Chung-chian hsu and chien-hsing chen. Mining [3] identified a critical issue namely the incomplete search-results problem resulting from the lack of a translation standard on foreign names and the existence of synonymous transliterations in searching the Web, to address the issue of using only one of the synonymous transliterations as search keyword will miss the web pages which use other transliterations for the foreign name, they proposed a novel two-stage framework for mining as many synonymous transliterations as possible from Web snippets with respect to a given input transliteration.

Guo Lei, Zhou Mei-ling,Yao Jian-Min, Zhu Qiao-Ming [4] a supervised transliteration person name identification process, which helps to classify the types of query Lexicon and concepts of transliteration characters and transliteration probability of a character.

Roslan Abdul Ghani, Mohamad Shanudin Zakaria, Khairuddin Omar [5], introduced a transliteration approach to semantic languages, easy way and fast process in Jawi to Malay transliteration in which Jawi stemming process was develop to make a word as short as possible but only focus on root word and some prefix and suffix. Vocal filtering and Diphthong filtering methods are also introduced to make a word simpler in Unicode mapping process in which Jawi-Malay rules are also applied to make output more accurate. Other than the above stated method, a dictionary database also provided for checking the words that cannot be found while process occur. This alternative method is used because format writing in Jawi is not remained.

Chun-Jen Lee, Jason S. Chang, Jyh-Shing Roger Jang [6] proposed a new statistical modelling approach to the machine transliteration problem for Chinese language by using the EM algorithm. The parameters of this model are automatically learned from a bilingual proper name list. Moreover, the model is applicable to the extraction of proper names.

Wei Gao, Kam-Fai Wong, and Wai Lam [7] modelled the statistical transliteration problem as a language model for post-adjustment plus a direct phonetic symbol transcription model, which is an efficient algorithm for aligning phoneme chunks as a statistical transliteration method for automatic translation according to pronunciation similarities, i.e. to map phonemes comprising an English name to the phonetic representations of the corresponding Chinese name.

Oi Yee Kwong [8] reported work on approximating phonological context E2C with surface Graphemic features which is based on the observation of graphemic ambiguities and is closely associated with the local contexts of phonological properties of which often determine its expected pronunciation.

3. SYSTEM OVERVIEW

The whole model consists of two important phases:

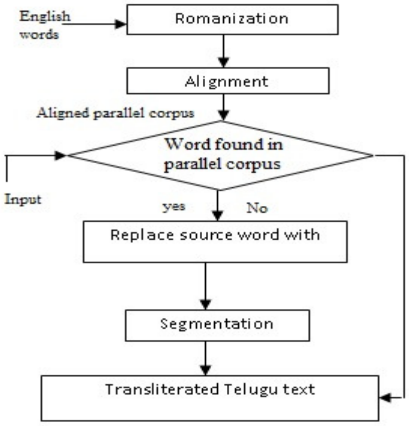


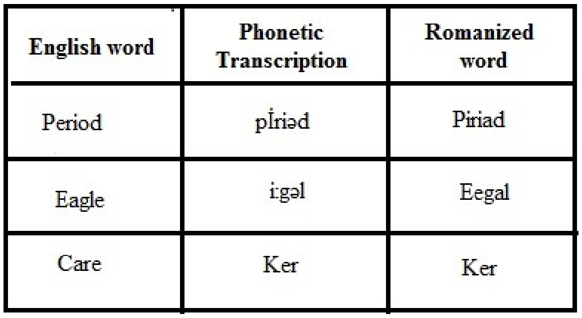
Figure1. Transliteration model

3.1 PRE-PROCESSING PHASE

In pre-processing phase, English vocabulary words for which transliteration will not produce correct results will be Romanized and Aligned in parallel corpus which is used in Transliteration phase to get correct result.

3.2 ROMANIZATION

During this step, the transliteration system is trained for those words which can't be exactly transliterated using either Grapheme or Phoneme individually. During the training step first the words are converted into their phonetics and then according to phonetic symbols, Telugu phonemic equivalent words in terms of English alphabets are generated and maintained as parallel corpus.



English word	Phonetic Transcription	Romanized word
Period	pİriəd	Piriad
Eagle	i:gəl	Eegal
Care	Ker	Ker

Table1.Romanization

3.3 ALIGNMENT

XML is used for storage of parallel corpus in which English words and Romanized words are aligned each other. Our Transliteration system is platform independent one because of using XML for storage purpose and Java script is used for retrieval of Parallel Corpus.

3.4 TRANSLITERATION PHASE

In transliteration phase the user entered English text or given file will be transliterated into Telugu text.

3.5 SEARCHING PARALLEL CORPUS

For each user entered word it will searched in Parallel Corpus, if a word is found in Parallel corpus then the original source word will be replaced with its Romanized equivalent word and it will be sent to Segmentation stage otherwise original source word will be sent for Segmentation stage.

3.6 SEGMENTATION

Based on combination of vowels, consonants the source language text will be segmented. Generally the segmentation unit will end with a vowel. Each segmented unit is called Transliteration unit. There are four rules which are to be followed while segmenting. They are

3.7 RULES

For example: Consider word 'piriad'

i) Consonant followed by vowel pi

ii)	Consonant followed by consonant	ri
iii)	Vowel followed by consonant	ad
iv)	Vowel followed by vowel	ia

During Segmentation two or more alphabets can be phonetically combined only when it had consonant followed by vowel or a consonant followed by consonant but in remaining two, each alphabet in uniquely mapped.

Before Segmentation	After Segmentation
p i r i y a d	pi l ri l a ld

3.8 UNICODE MAPPING

For each alphabet in English there will an hexadecimal unique code is mapped and for transliteration units which are obtained from Segmentation stage these Unicode's are combined to get phonetically equivalent Telugu alphabets. Using this method, we can convert English text into phonetically equivalent ones in Telugu. For Telugu, Unicode range varies from 0C01 - 0C7F.

pi | ri | a | d

పి | రి | అ | డ్

If user enters text in text-area of GUI then the output will be displayed on another text-area which is on the same GUI, otherwise the transliterated text will be saved into another file in the same directory as the source file which is given as input.

4. EVALUATION AND RESULTS

The proposed model is trained for 50,000 words containing English vocabulary words. The model is evaluated by taking articles and checking the correctness of transliteration by comparing with Google transliteration system. Accuracy of the system is calculated using the following equation:

$$Accuracy = (C/N) * 100 \quad \text{-> } eq1$$

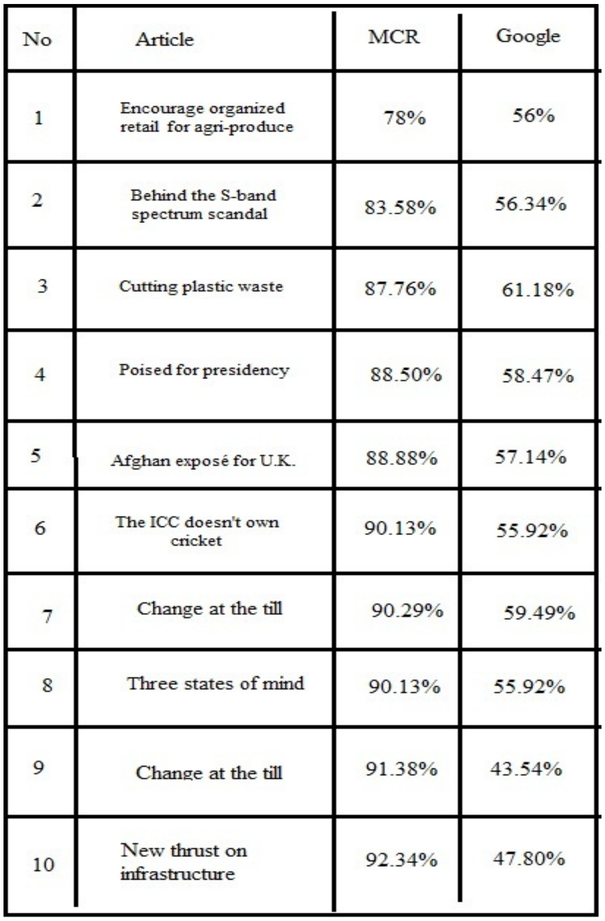
Where C indicates the number of test words with correct transliteration when compared with Google transliteration systems and N indicates the total number of test words.

4.1. COMPARISION WITH GOOGLE TRANSLITERATION SYSTEM

By comparing our Morphological Cross Reference System with publicly available Google Indic Transliteration System the accuracy of the two systems is observed as follows:
The system is evaluated by considering random set of articles and accuracy is experimentally calculated using equation 1.

4.1.1. ACCURACY

Accuracy of transliteration for 10 different articles which are taken from Hindu News Paper for Google and MCR systems is tabulated below.



No	Article	MCR	Google
1	Encourage organized retail for agri-produce	78%	56%
2	Behind the S-band spectrum scandal	83.58%	56.34%
3	Cutting plastic waste	87.76%	61.18%
4	Poised for presidency	88.50%	58.47%
5	Afghan exposé for U.K.	88.88%	57.14%
6	The ICC doesn't own cricket	90.13%	55.92%
7	Change at the till	90.29%	59.49%
8	Three states of mind	90.13%	55.92%
9	Change at the till	91.38%	43.54%
10	New thrust on infrastructure	92.34%	47.80%

Table2. Transliteration Accuracy

From the results it is observed that, when we take sample of articles which consists of both vocabulary and out of vocabulary words, accuracy of Morphological Cross Reference System increases when we go down from the articles containing out of vocabulary words and vocabulary words to the articles containing only vocabulary word. For Google Transliteration systems the accuracy of transliteration for those articles slightly decreases as we go down through the list of articles shown in Table 2.

The results of Table 2 are shown in Figure 2 which displays how the accuracy varies for Google and MCR systems

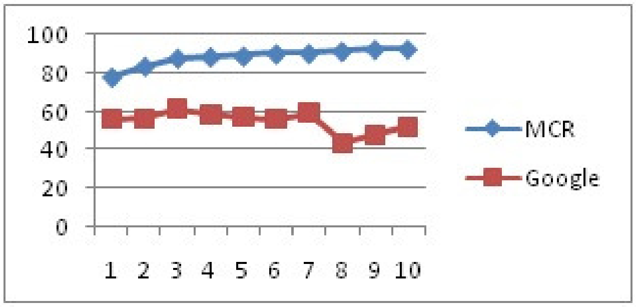
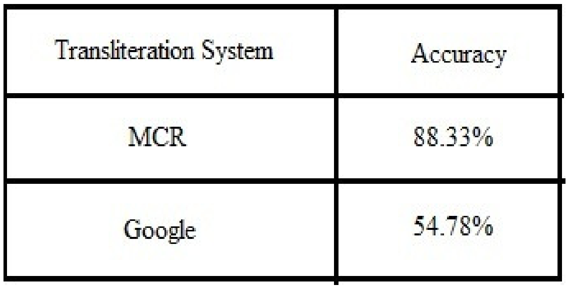


Figure 2. Comparison of accuracy for different articles for Google and MCR systems

On average the accuracy for MCR and Google Transliteration System is show in Table 3



Transliteration System	Accuracy
MCR	88.33%
Google	54.78%

Table 3. Average Accuracy

From the results it is observed that Morphological Cross Reference System gives an accuracy of '88.33%' which '33.55%' more than Google Transliteration systems which gives an accuracy of 54.78%'.

The results of Table 3 are shown in Figure 3 in terms of bar diagram.

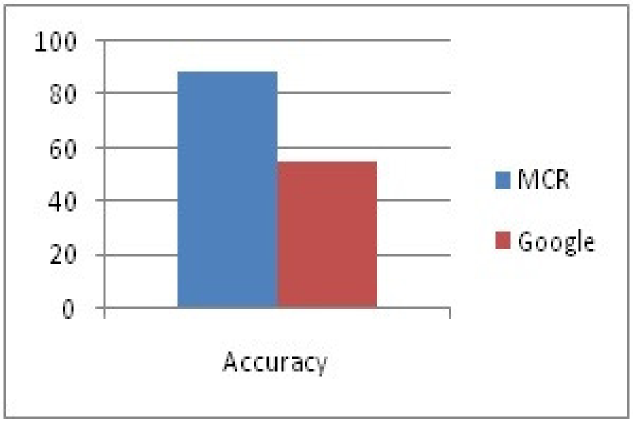
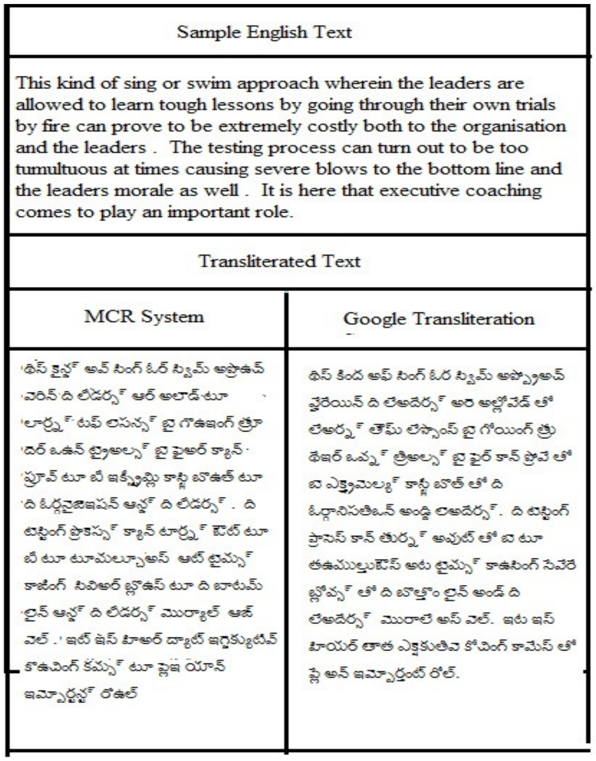


Figure 3. Comparison of overall accuracy for Google and MCR Systems

As per the graph the variation among the best Transliteration System and our system is proved high.

4.1.2. SAMPLE TEST DATA

A sample Test Data used to compare the results of our model with that of Google Transliteration System for Vocabulary words excluding silent words which are taken from Leader ship article from Hindu News Paper is shown in Table 4.



<table>
<tr><th colspan="2">Sample English Text</th></tr>
<tr><td colspan="2">This kind of sing or swim approach wherein the leaders are allowed to learn tough lessons by going through their own trials by fire can prove to be extremely costly both to the organisation and the leaders . The testing process can turn out to be too tumultuous at times causing severe blows to the bottom line and the leaders morale as well . It is here that executive coaching comes to play an important role.</td></tr>
<tr><th colspan="2">Transliterated Text</th></tr>
<tr><th>MCR System</th><th>Google Transliteration</th></tr>
<tr><td>ధిస్ కైన్డ్ అవ్ సింగ్ ఓర్ స్విమ్ అప్రోఉచ్ వెరిన్ ది లీడర్స్ ఆర్ అలౌడ్ టూ లార్న్ టఫ్ లెసన్స్ బై గొఉఇంగ్ త్రూ దెర్ ఒఉన్ ట్రైఅల్స్ బై ఫైఅర్ క్యాన్ ప్రూవ్ టూ బీ ఇక్స్ట్రీమ్లి కాస్ట్లి బొఉత్ టూ ది ఓర్గనైజైషన్ అన్డ్ ది లీడర్స్ . ది టెస్టింగ్ ప్రొకెస్స్ క్యాన్ టూర్న్ ఔట్ టూ బీ టూ టూమల్ట్యుఅస్ ఆట్ టైమ్స్ కాజింగ్ సివిఅర్ బ్లొఉస్ టూ ది బాటమ్ లైన్ అన్డ్ ది లీడర్స్ మొర్యాల్ ఆజ్ వెల్ .' ఇట్ ఇస్ హిఅర్ ద్యాట్ ఇగ్జెక్యుటివ్ కొఉచింగ్ కమ్స్ టూ ప్లెఇ యాన్ ఇమ్పార్టన్ట్ రొఉల్</td><td>ధిస్ కింద అఫ్ సింగ్ ఓర స్విమ్ అప్ప్రోఅచ్ వ్హేరెయిన్ ది లేఅదేర్స్ అరె అల్లోవేడ్ తో లేఅర్న్ తౌఘ్ లేస్సొంస్ బై గోయింగ్ త్రు థెఇర్ ఒవ్న్ త్రిఅల్స్ బై ఫైర్ కాన్ ప్రొవె తో బె ఎక్స్త్రెమెల్య్ కాస్ట్ల బొత్ తో ది ఒర్గానిసతిఒన్ అండ్ద లేఅదేర్స్. ది టెస్టింగ్ ప్రొసెస్ కాన్ తుర్న్ అవుట్ తో బె టూ తుముల్తుఒస్ అట టైమ్స్ కౌసింగ్ సేవేరె బ్లోవ్స్ తో ది బొత్తొం లైన్ అండ్ ది లేఅదేర్స్ మొరాలె అస్ వెల్. ఇట ఇస్ హియర్ తాత ఎక్షకుతివ కోచింగ్ కామెస్ తో ప్లే అన్ ఇమ్పోర్తంట్ రోల్.</td></tr>
</table>

Table 4. Transliterated text for MCR and Google system

4.1. 3. ERROR RATE

Error Rate is defined as the ratio of Wrongly Transliterated words to Total number of Test words. Error Rate of MCR System can be calculated using the following equation:

$$Error\ Rate = (W/N) \quad -> \quad eq2$$

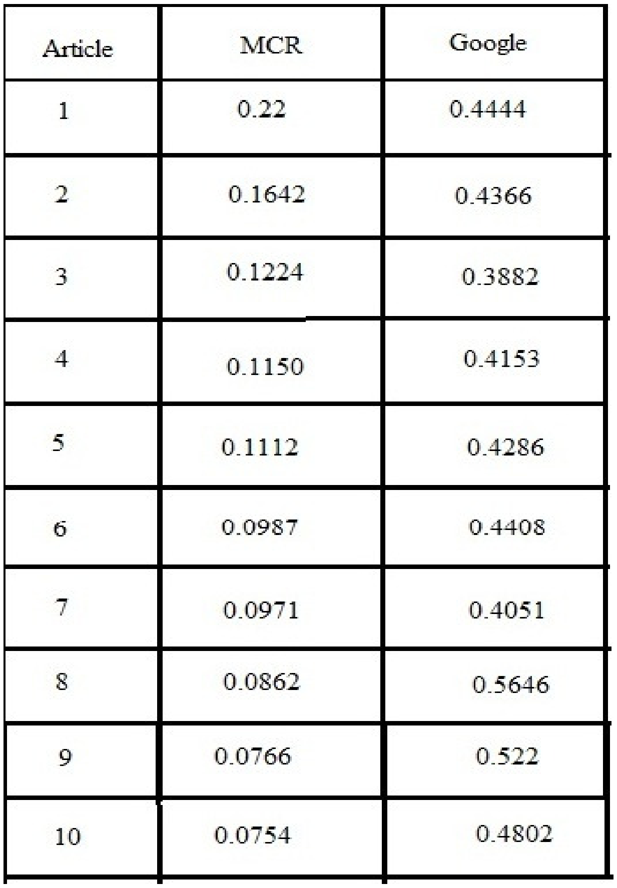
Where W indicates number of wrongly Transliterated words when compared with Google Transliteration System, N is total number of test words.

The error rate for MCR system is more when we consider out of vocabulary words like Ghulam Nabi Azad (ఝలమ్ ణబి ఆజడ్), Manmohan Singh (మన్మోహన్ సింగ్), and for abbreviations like GSM (జిఎస్ఎం), ICC(ఐసిసి) but error rate is less for vocabulary words lecture(లెక్చర్), inflation(ఇన్ఫ్లెఇషన్).

The error rate for Google system is more when we consider vocabulary words like lecture (లెచ్తురె ,(ఇంఫ్లతిఒన్ (ఇంఫ్లతిఒన్)and also for abbreviations like GSM (గ్సం,(ICC(ఇచ్క్) but error rate is less for out of vocabulary words like Ghulam Nabi Azad (ఘులం నబి అజాద్), Manmohan Singh (మన్మోహన్ సింగ్).

The system is evaluated by considering set of articles and the error rates which are obtained by using equation 2 are shown in Table 6. For MCR system Error Rate is calculated using above formula and found as '0.22' and that of Google is '0.40' for Vocabulary words excluding Silent words. For Vocabulary words including Silent words error rate on MCR system is found as 0.19 and that of Google is found as 0.51.

The error rates for our System and already existing system are tabulated below.



Article	MCR	Google
1	0.22	0.4444
2	0.1642	0.4366
3	0.1224	0.3882
4	0.1150	0.4153
5	0.1112	0.4286
6	0.0987	0.4408
7	0.0971	0.4051
8	0.0862	0.5646
9	0.0766	0.522
10	0.0754	0.4802

Table 5. Comparison of Error Rates

The results of error rates in Table 5 are shown in Figure 4 which displays how the error rate varies for MCR and Google systems.

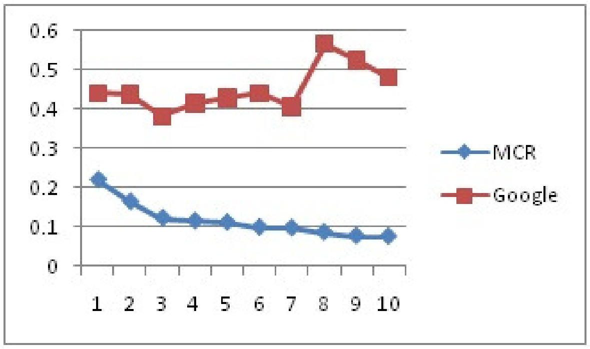
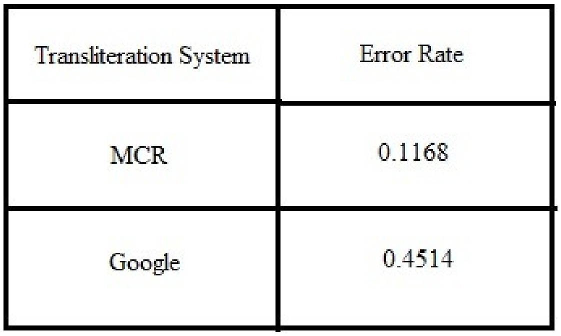


Figure 4. Comparison of Error Rates

On an average the error rates for MCR and Google Transliteration System are shown in Table 6



Transliteration System	Error Rate
MCR	0.1168
Google	0.4514

Table 6. Comparison of Error Rates

The Average error rate for MCR and Google Transliteration system are graphically represented in Figure 5

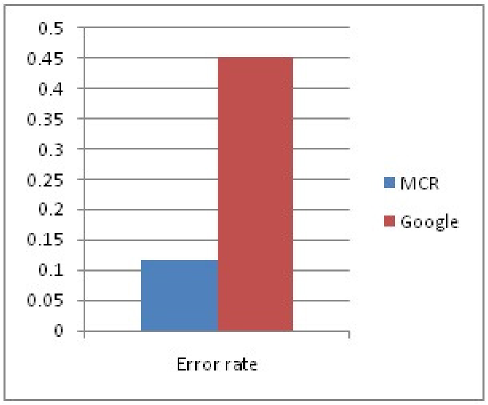


Figure 5. Graphical Representation of Comparison of Error Rates

Hence by observing the above graph we can say that our MCR system will produce less error rates than that of Google Transliteration System.

5. FUTURE WORK AND CONCLUSION

In this paper we addressed the problem of transliterating English to Telugu language using Morphological Cross Reference System. This framework based on data driven method and one to one mapping approach simplifies the development procedure of transliteration system and facilitates better improvement in transliteration accuracy when compared with that of Google Transliteration System. The model is trained on English Vocabulary words that don't have exact transliteration by considering Phonemic or Graphemic Transliteration models individually. The system is evaluated by considering set of English articles and comparing the Telugu transliterated text with Google Transliteration system. From the experiment we found that transliteration result increase the overall transliteration performance to a great extent. The model will work efficiently for English vocabulary words but in future it will be extended to work accurately for named entities and proper nouns also. We hope this will be very useful in natural language applications like creating blogs, chatting, sending emails and in many areas.

REFERENCES

[1] Antony P.J, Ajith V.P, Soman K.P, “Kernel Method for English to Kannada Transliteration”, International Conference on Recent Trends in Information, Telecommunication and Computing, 2010.

[2] V.B.Sowmya, Vasudeva Varma, “Transliteration Based Text Inpu Methods for Telugu”, Springer-Verlag Berlin Heidelberg, 2009.

[3] Chung-chian hsu and chien-hsing chen. Mining, “Synonymous Transliterations from the World Wide Web”, ACM Transactions on Asian Language Information Processing, Vol. 9, No. 1, Article 1, March 2010.

[4] Guo Lei, Zhou Mei-ling,Yao Jian-Min, Zhu Qiao-Ming, “A Supervised Method for Transliterated Person Name Identification”, Second International Symposium on Electronic Commerce and Security, 2009.

[5] Roslan Abdul Ghani, Mohamad Shanudin Zakaria, Khairuddin Omar, “Jawi-Malay Transliteration”, International Conference on Electrical Engineering and Informatics 5-7 August 2009, Selangor, Malaysia.

[6] Chun-Jen Lee, Jason S.Chang, Jyh-Shing Roger Jang, “Extraction of Transliteration Pairs from Parallel Corpora Using a Statistical Transliteration Model”.

[7] Wei Gao, Kam-Fai Wong, and Wai Lam, “Phoneme-Based Transliteration of Foreign Names for OOV Problem”, Springer-Verlag Berlin Heidelberg, 2005.

[8] Oi Yee Kwong, “Graphemic Approximation of Phonological Context for English-Chinese Transliteration”, Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, Suntec, Singapore, 2009.

[9] Shih-Hung Wu and Yu-Te Li, “Curate a Transliteration Corpus from Transliteration/Translation Pairs”, IEEE IRI2008, July 13-15, 2008, Las Vegas, Nevada, USA.

[10] Ranbeer, M., Nikita, P., Prasad, P., Vasudeva, V.: Experiments in Cross-lingual IR among Indian Languages, In: International Workshop on Cross Language Information Processing (CLIP 2007).