I J M I
International Journal of Machine Intelligence
ISSN: 0975–2927 & E-ISSN: 0975–9166, Volume 3, Issue 4, 2011, pp-340-345
Available online at http://www.bioinfo.in/contents.php?id=31

ENGLISH TO KANNADA/TELUGU NAME TRANSLITERATION IN CLIR: A STATISTICAL APPROACH

MALLAMMA V REDDY[1*], HANUMANTHAPPA M.[2]
[1,2]Department of Computer Science and Applications,
Bangalore University, Bangalore-560 056, INDIA
*Corresponding Author: Email- [1]mallamma_vreddy@bub.ernet.in , [2]hanu6572@bub.ernet.in

Received: November 06, 2011; Accepted: December 09, 2011

Abstract- Transliteration is mapping of pronunciation and articulation of words written in one script into another script. Transliteration should not be confused with translation, which involves a change in language while preserving meaning. CLIR is the acronym of a great variety of techniques, systems and technologies that associate information retrieval (normally from texts) in multilingual environments. Dictionaries have often been used for query translation in cross language information retrieval (CLIR). However, we are faced with the problem of translating Names and Technical Terms from English to Kannada/Telugu. The most important query words in information retrieval are often proper names. We present a method for automatically learning a transliteration model from a sample of name pairs in two languages.
Key words - Query translation, Bilingual Dictionaries, Transliteration

Introduction

India is a country with 22 official languages and use of computers is fast spreading not only to create employment in the IT sector but also to support productive use of IT in daily life - increase productivity and competitiveness, provide better quality of life, enable inclusiveness and strengthen democracy. Ability of different sections of people to use computers (and increasingly text and data over mobile phones) demand that the Basic Information Processing Kit for Indian languages is constantly upgraded for various hardware and software platforms, new tools added and work promoted with developers, ISVs and System Integrators and application developers to enable/support Indian languages use in different sectors/verticals. And increasingly, Indian language content in Digital form has to be created and supported for applications to be supported and reach a critical mass.

Cross-Language Information Retrieval (CLIR) is a special case of information retrieval in which retrieval is not restricted to the query language itself but extends to all languages supported by the system. CLIR [14] deals with the problem of issuing a query in one language and retrieving relevant information in other languages. It aims to help the user in finding relevant information without being limited by linguistic barriers. Proper names in general are very important in text. Since news stories especially revolve around people, places, or organizations, proper names play a major role in helping one distinguish between a general event (like a war) and a specific event.

Proper names in English are spelled in various ways. Despite the existence of one or more standard forms for someone's name, it is common to find variations in transliterations of that name in different source texts, such as **Goutham** vs. **Goutam**. The problem is more pronounced when dealing with non-English names or when dealing with spellings by non-native speakers.

One possible method to generate transliteration is based on the use of dictionaries, which contains words in source language and their possible transliterated forms in target language. However, this is not a practical solution since proper nouns and technical terms, which are frequently transliterated, usually have rich productivity [1]. This paper discusses another approach based on machine learning to automate the process of machine transliteration.

Transliteration

The Language transliteration is one of the important area in natural language processing. Machine Transliteration is the conversion of a character or word from one language to another without losing its phonological characteristics. It is an orthographical and phonetic converting process. Therefore, both grapheme and phoneme information should be considered. Accurate transliteration of named entities plays an important role in the performance of machine translation and cross-language information retrieval processes. The transliteration model must be design in such a way that the phonetic structure of words should be preserve as closely as possible. The overall machine transliteration system architecture as

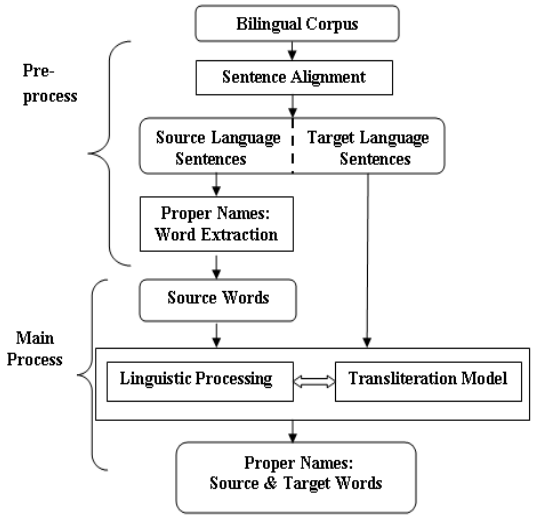


Fig. 1- The overall process for extracting name and transliteration pairs

The proposed transliteration model can be applied to the tasks of the extraction of bilingual name and transliteration pairs. These tasks become more challenging for language pairs with different sound systems, we focus on the extraction of English-Kannada/Telugu name and transliteration pairs. However, the proposed framework is easily extendable to other language pairs.

For the purpose of extracting name and transliteration pairs from parallel corpora, a sentence alignment procedure is applied first to align parallel texts at the sentence level. Then, we use a part of speech tagger to identify proper nouns in the source text. After that, the machine transliteration model is applied to isolate the transliteration in the target text. In general, the proposed transliteration model can be further augmented by linguistic processing, which will be described in more detail in the next subsection. The overall process is summarized in "Fig. (1)".

Preprocessing: This is the first layer of the proposed model which gets an English word as input and simplifies the word by performing some pre-processing steps. At the first step, an English word goes through schwa deletion algorithm.

Transliteration module: The transliteration system is trained from the lists of proper names in English and Kannada by using GIZA++ [11], an extension of GIZA, which determines the translation probability. Training is done with the help of 3800 names in the both English and Kannada in tokenized form.

Post Processing: An input in target language that is Kannada language from last unit is forwarded to the post processing unit where some post processing tasks are applied to it. Post processing tasks further improve the results using various transliteration rules.

comparative Analysis of English And Kannada languages

English is a West Germanic language that arose in the Anglo-Saxon kingdoms of England. It is one of six official languages of the United Nations. India is one of the countries where English is spoken as a second language. There are 21 consonant letters in English. These are B, C, D, F, G, H, J, K, L, M, N, P, Q, R, S, T, V, W, X, and Z. The rest of the letters of the alphabet are called vowels. The vowels are: A, E, I, O, U. As defined in the Constitution of India, English is one of the two official languages of communication (Hindi being the other) for India's federal government and is one of the 22 scheduled languages specified in the Eighth Schedule to the Constitution. A working knowledge of English has become a requirement in a number of fields, occupations and professions such as medicine and computing; as a consequence over a billion people speak English to at least a basic level [4].

Kannada or Canarese is a language spoken in India predominantly in the state of Karnataka, Making it the 25th most spoken language in the world. It has given birth to so many Indian languages like, Tulu, Kodava etc and one of the scheduled languages of India and the official and administrative language of the state of Karnataka [5]. Telugu is also one of the widely spoken languages in India especially in the state of Andhra Pradesh and the district of Yanam. The number of consonants and vowels in Baraha Kannada/Telugu script as given in "Fig. (2)". Both Kannada and Telugu use the "UTF-8" / western windows encode and draw their vocabulary mainly from Sanskrit.

Kannada Vowels (SwaragaLu)											
ಅ	a	ಆ	aa, A	ಇ	i	ಈ	ee, I	ಉ	u	ಊ	oo, U
ಋ	Ru	ಎ	e	ಏ	ae, Eai	ಐ	ai	ಒ	o	ಓ	oa, O
ಔ	ou	ಅಂ	um	ಅಃ	ah						

Telugu Vowels (Achchulu)											
అ	a	ఆ	aa, A	ఇ	i	ఈ	ee, I	ఉ	u	ఊ	oo, U
ఎ	e	ఏ	ae, E	ఐ	ai	ఒ	o	ఓ	oa, O	ఔ	au
అం	aM	ఆ:	AH								

Kannada Consonants (VyanjanagaLu)											
ಕ	ka	ಖ	kha	ಗ	ga	ಘ	gha	ಙ	Gna	ಚ	cha
ಛ	Cha	ಜ	ja	ಝ	jha	ಞ	ini	ಟ	Ta	ಠ	Tha
ಡ	Da	ಢ	Dha	ಣ	Na	ತ	ta	ಥ	tha	ದ	da
ಧ	dha	ನ	na	ಪ	pa	ಫ	pha	ಬ	ba	ಭ	bha
ಮ	ma	ಯ	ya	ರ	ra	ಲ	la	ವ	va	ಶ	sha
ಷ	Sa	ಸ	sa	ಹ	ha	ಳ	La	ೞ	ksha		

Telugu Consonants (Hallulu)											
క	ka	ఖ	kha	గ	ga	ఘ	gha	ఙ	Gna	చ	cha
ఛ	Cha	జ	ja	ఝ	jha	ఞ	ini	ట	Ta	ఠ	Tha
డ	Da	ఢ	Dha	ణ	Na	త	ta	థ	tha	ద	da
ధ	dha	న	na	ప	pa	ఫ	pha	బ	ba	భ	bha
మ	ma	య	ya	ర	ra	ల	la	వ	va	శ	sa
ష	Sa	స	sha	హ	ha	ళ	La	క్ష	ksha	ఱ	Ra

Fig. 2- English-Kannada/Telugu Character Mapping

Problems In Machine Transliteration From English To Kannada

Transliteration is difficult in both English and Kannada languages due to following reasons:

-**Character Gap**: The number of characters in, both English and Kannada, character sets varies in both the language that makes the transliteration process difficult. The numbers of vowels are 5 and 14 and numbers of consonants are 21 and 35 in both English and Kannada, respectively as explained earlier. So there is character gap in both the languages that leads to problems in transliteration process.

-One-to-Multi mapping Problem: In this problem, single character in one script transform to multiple characters in another script. The Multi-mapping Problem is associated with following characters as shown in Table 1. For example, character'd' in English language can be transliterated into two characters in Kannada 'ದ್' and 'ಡ್'. Some algorithm is required to select the appropriate character at different situations. Like in technical term 'bit', two transliterations are possible 'ಬಿತ' and 'ಬಿಟ' but only second one is correct and first should be discarded by the system.

Table-1 -Multi-mapped Characters

Characters in English	't'	'd'	'c'	'n'	'r'
Transliteration in Kannada	'ತ್' and 'ಟ್'	'ದ್' and 'ಡ್'	'ಚ್' and 'ಛ್'	'ನ್' and 'ಣ್'	'ರ್' and 'ಱ್'
Transliteration in Telugu	'త్' and 'ట్'	'ద్' and 'డ్'	'చ్' and 'ఛ్'	'న్' and 'ణ్'	'ర్' and 'ఱ్'

-Multi-to-One map problem: Here multiple characters in one character set leads to single character in another character set. This problem makes transliteration process difficult. This type of problem is associated with characters shown in Table 2.

For example, in following name, combination of two characters 'ch' in English language forms the single character in Kannada language 'ಚಿತ್' 'Chaitra' 'ಚೈತ್ರ'.

Table-2 -Multi-to-One Map Problems

Multiple characters in English	'ch'	'kh'	'th'	'rh'	'bh'	'sh'
Transliterated characters in Kannada Script	'ಚ್'	'ಖ್'	'ಥ್'	'ರ್ಹ್'	'ಭ್'	'ಶ್'
Transliterated characters in Telugu Script	'చ'	'ఖ'	'థ'	'ర్హ్'	'భ'	'శ'

-'Double occurrence' of certain Characters: This is similar to above explained problem. Here again two characters are clustered and mapped to single character in target language but clusters are made with 'double occurrence' of certain Characters in English language as shown in Table V. For example, 'oo' in word 'cool' are clustered to make one unit because they represent single character in target language. This type of clustering is associated with the following combination of characters.

-**Schwa sound:** Schwa deletion is an important issue in grapheme-to-phoneme conversion for Indo-Aryan languages. Schwa is defined as the mid-central vowel that occurs in unstressed syllables [9]. Simple observation of Hindi words [7] provides certain information where schwa is retained and certain contexts where it is deleted without any exception. For example, name 'sapna' is transliterated into 'ಸಪ್ನ' whereby 'a' in between character 's' and 'p' is deleted and 'a' after character 'n' is retained. To find out when to retain schwa sound is not a trivial task.

-**Multiple representation of same word:** There can be multiple ways to write a source language word into target language. For example, for person name 'Rabindranath

Tagore', two representations among various transliterations in target Kannada language are 'ರವಿಂದ್ರನಾಥ ಟ್ಯಾಗೋರ' and 'ರವಿನ್ದ್ರನಾಥ ತಗೊರೆ' . Choosing the correct one is again depends upon the perception of end user.

Experiments

In this section, we focus on the setup for the experiments and a performance evaluation of the proposed model applied to extract bilingual word pairs from parallel corpora.

Experimental Setup

Several corpora were collected to estimate the parameters of the proposed models and to evaluate the performance of the proposed approach. The corpus BUBShabdaSagara-2011[14] for training consisted of 3,500 pairs of English names and transliterations in Kannada and Telugu. The training corpus composed of a bilingual proper name list. The bilingual proper name list consists of first names, last names, and nicknames. For example, (Kalpna, ಕಲ್ಪನ , కల్పనా) and (vishwa విశ్వ , విశ్వా) are first names, (khatriki, ಖತ್ರಿಕ , ఖత్రికి) and (Rao, ರಾವ , రావు) are last names, and (pinko, ಪಿಂಕೊ , పింకో) and (chakul, ಛಕುಲ ,చకుల్) are nicknames, for males and females, respectively. Some first names are also used as last names. For instance, "Reddy" can be either a first name or a last name.

All the experiments carried out here involve the same set of English queries and the same query expansion, translation and retrieval method. The only difference between the experimental conditions is in what dictionaries are used in the query translation. Some samples from the training set BUBShabdaSagara-2011 as shown in Table 3.

In the experiment, we dealt with personal and place names as well as their transliterations from the parallel corpora. The performance of transliteration extraction was evaluated based on the precision rates of transliteration words or characters. For simplicity, we considered each proper name in the source sentence in turn and determined its corresponding transliteration independently.

Software

We perform experiment in Linux environment. The following sections describe briefly the software that was used during the project.

MOSES: Moses is a **statistical machine translation system** that allows to automatically train translation models for any language pair. Only translated texts (parallel corpus) is needed. An efficient search algorithm finds quickly the highest probability translation among the exponential number of choices [13].

GIZA++: *GIZA++* is an extension of the program GIZA which was developed by the Statistical Machine Translation team during the summer workshop in 1999 at the Center for Language and Speech Processing at Johns-Hopkins University [12]. **GIZA++** is a program for aligning words and sequences of words in sentence aligned corpus. We used it to do character alignment of word-aligned pairs.

Table-3 -Some samples from the training set BUBShabdaSagara-2011

Source Word	Target Word in Kannada	Target Word in Telugu
Aabheer	ಅಭೀರ	ఆభీర్
Aadarsh	ಆದರ್ಶ	ఆదర్శ
Aadesh	ಆದೇಶ	ఆదేశ్
Baadal	ಬಾದಲ	బాదల్
Baalaark	ಬಾಲಾರ್ಕ	బాలార్క్
chintu	ಚಿಂಟು	చింతు
citragupta	ಚಿತ್ರಗುಪ್ತ	చిత్రగుప్త
damodar	ದಾಮೊದರ	దామోదర్
dakshi	ದಕ್ಷಿ	దక్షి
Ekalinga	ಏಕಲಿಂಗ	ఏకలింగ
mArtanDa	ಮಾರ್ತಂಡ	మార్తాండ
Madhulika	ಮಧುಲಿಕ	మధులిక
Shaila stone	ಷೈಲ ಸ್ಟೊನ	షైల స్టోనే
tapti	ತಪ್ತಿ	తపతి
Ujjayini	ಊಜ್ಜಯಿನಿ	ఉజ్జయిని
Vineeta	ವಿನೀತ	వినీత
Yukti	ಯುಕ್ತಿ	యుక్తి

SRILM: SRILM is a toolkit for building and applying statistical language models (LMs), primarily for use in speech recognition, statistical tagging and segmentation. SRILM is used by Moses to build statistical language models [10].

Evaluation Metric

In the experiment, the performance of transliteration extraction was evaluated based on precision and recall rates at the word and character levels. Since we considered exactly one proper name in the source language and one transliteration in the target language at a time, the word recall rates were same as the word precision rates:

$$Word\ Precision(WP) = \frac{number\ of\ correctly\ extracted\ word}{number\ of\ correct\ words}$$

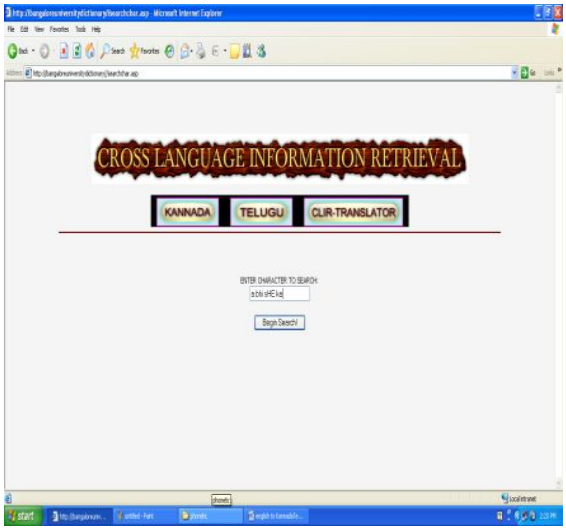
The character level recall and precision rates were defined as follows:

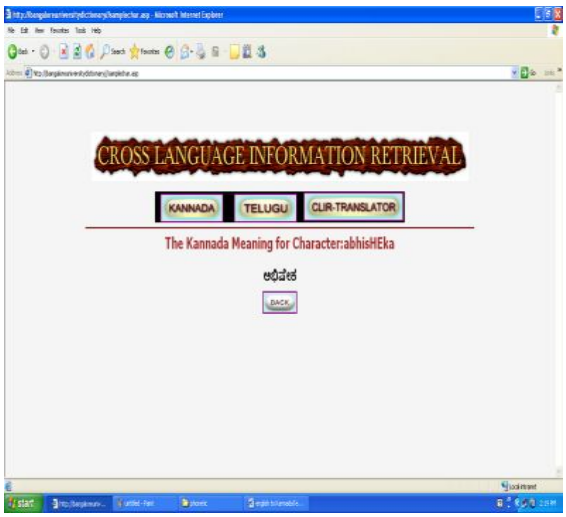
$$Character\ Precision(CP) = \frac{number\ of\ correctly\ extracted\ Characters}{number\ of\ extracted\ \ Characters}$$

$$Character\ Recall(CR) = \frac{number\ of\ correctly\ extracted\ Characters}{number\ of\ correct\ Characters}$$

Results

Cross Language Information Retrieval Tool is built by using the ASP.NET as front end and for a Database the Kannada is encrypted by using the Encoding system. The sample result as shown in "Fig. (5)."





The result statistical Transliteration system from English to Kannada/ Telugu for The corpus BUBShabdaSagara-2011 trained data set in Table 4.

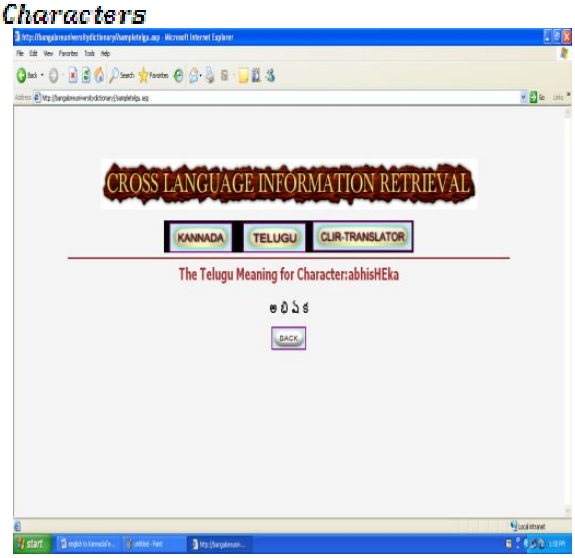


Fig 5: sample result for Kannada/Telugu

Table-4 -The average rates of transliterated word extraction for overall corpora

Run Desc	WP	CP	CR
KANNADA	86.0%	94.4%	96.3%
TELUGU	85.09%	94.8%	93.3%

Conclusion

We presented our English-Kannada and English-Telugu CLIR system developed for the Ad-Hoc bilingual Task. Our approach is based on query Translation using bi-lingual dictionaries. A new statistical modeling approach to the machine transliteration problem has been presented in this paper. The parameters of the model are automatically learned from a bilingual proper name list. Moreover, the model is applicable to the extraction of proper names and transliterations. The proposed method can be easily extended to other language pairs that have different sound systems without the assistance of pronunciation dictionaries. Experimental results indicate that high precision and recall rates can be achieved by the proposed method.

References

[1] Jong-Hoon Oh Key-Sun Choi (2005) *IEICE TRANS.INF. & SYST.*, VOL.E88-D, NO.7,pp 1737-1748.

[2] Jasleen kaur, Gurpreet Singh Josan (2011) *International Journal on Computer Science and Engineering (IJCSE)* ISSN: 0975-3397 Vol. 3 No., 1518.

[3] Hall P.A.V. and G.R. Dowling (1980) 12, 381-402.

[4] Karen Kukich (1992) *ACM Comput. Surv*. 24(4): 377-439.

[5] Virga and S. Khudanpur (2003) *Proceedings of the ACL Workshop on Multi-lingual Named Entity Recognition 2003*.

[6] English Language Accessed from *"http://en.wikipedia.org/wiki/English_language"*, on Jan2010.

[7] The Karnataka Official Language Act, Government of Karnataka. Retrieved 2007-06-29.

[8] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico,Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer,Ondrej Bojar, Alexandra Constantin, Evan Herbst (2007) *Moses: Open Source Toolkit for Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic.*

[9] Article SRILM accessed from *http://www.speech.sri.com/projects/srilm/* on 2011.

[10] Mallamma V. Reddy, Hanumanthappa M. *http://bangaloreuniversitydictionary//menu.asp*

[11] GIZA++ accessed from *"http://www.fjoch.com/GIZA++.html"* on 2011.

[12] Article GIZA++ accessed from *http://wiki.apertium.org/wiki/Using_GIZA%2B%2B* on 2011.

[13] Moses article accessed from *http://www.statmt.org/moses/manual/manual.pdf* on 2011.

[14] Mallamma V. Reddy, Hanumanthappa M. (2011) *International Journal of Computer Science and Information Technologies*, Vol. 2 (5), page-1876-1880. IISN: 0975-9646.