English to Kashmiri Transliteration System - A Hybrid Approach

Mir Aadil
Department of Computer Sciences
BGSB University
Rajouri, J&K-India

M. Asger
School of Mathematical Sc. & Eng.
BGSB University
Rajouri, J&K-India

ABSTRACT
Named entities (NE) and Out-Of-Vocabulary (OOV) words in text are treated differently for application like Machine Translation and Information Retrieval. While normal text in the Source Language (SL) is translated on the basis of translation mapping, named entities and out-of-vocabulary are written in the script of the Target Language (TL) without any change to the articulation of the word in both the languages. The process is transliteration. This paper shows the exploitation of Phoneme-Based model for transliteration of English to Kashmiri that can be used in information extraction and machine translation for the language pair. The overall accuracy of the system achieved while tested on medical domain and Wikipedia based English text is 86% (intelligible transliteration).

General Terms
Artificial Intelligence, Natural Language Processing, Machine Translation, Information Extraction.

Keywords
Machine Transliteration, Phoneme-Based Machine Transliteration, English-Kashmiri Machine Transliteration.

1. INTRODUCTION
Machine Transliteration is the automatic method that is employed by an algorithm to transcribe an alphabet or a syllable of a word written in one script or language to some other script. Languages using the same writing system usually do not need transliteration. However for languages with different writing styles and fonts, transliteration is required. Machine transliteration is used in multiple applications like Question-Answering, Information Extraction, Foreign Language learning, Machine Translation, Data Mining, etc. [6]For Named entities (NE) and Out-Of-Vocabulary (OOV) words, transliteration is necessary for the development of a machine translation system for most of the language pairs like English-Kashmiri. English language has a Latin(Roman) script that belongs to the family of Egyptian Hieroglyphs and is written from left to right while as Kashmiri language on the other hand has a Perso-Arabic script that is written from right to left. Kashmiri is also written in Devanagari and Sharada scripts (left to right) but these two scripts are hardly in use nowadays. English language has 26 graphemes while Latin script of Kashmiri has 38 graphemes [1][5]. The work is on the Roman to Perso-Arabic script transliteration not only for its common use but also for its capability to indicate all vowel sounds regularly. The work is a part (sub-module) that is incorporated with a statistical machine translation system of English to Kashmiri. For this purpose a database of English-Kashmiri transliterated word pairs of around 10000 words has been developed that is used with the translation database but only if a word-error is encountered while translating. For Named Entities and still missing words, a dynamic system needed to be incorporated that could transliterate English words based on their phoneme automatically if an OOV or Named Entity is encountered in the text for translation. Carnegie Mellon University Pronunciation Dictionary and The American Heritage Dictionary of English Language has been used to extract ARPAbet syllables of each English word and then transliterate it to Kashmiri word based on a syllable transliteration database [2]. Furthermore, for continuously changing Named Entities and still unknown English words, a different approach is used. The word's grapheme is subject to an automatic phoneme generation based on some rules that results in ARPAbet syllables of the word. The syllables are finally mapped using an English-Kashmiri syllable transliteration database. Since, the syllables are limited; these two approaches are faster, however not as accurate as direct mapping.

2. METHODS USED

2.1 Direct Transliteration
Direct transliteration is based on the direct mapping of an OOV word with the transliterated form in the database of English-Kashmiri transliterated word pairs. This gives more accurate results for the words that make a hit, as all the transliterations are manual and implicitly accurate. The system after normalizing (lowercasing) the word maps it to its transliteration, as shown in Fig 1. However for the words which are not found (a miss) in the database, the method doesn't suffice and is not used further[7].

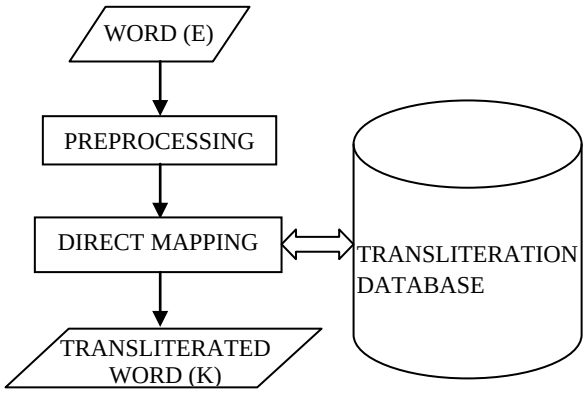

Fig 1: Direct Mapping Based Transliteration

2.2 Phoneme-based Transliteration

2.2.1 Dictionary Look-up
For the words that are not manually transliterated and hence are not found in the transliteration word-pair database, a dictionary-lookup approach based on phoneme of an English word is used. First the word to be transliterated is replaced by its phonemes. The phonemes are then transliterated to the

Kashmiri language script. A database based on Carnegie Mellon University Pronunciation Dictionary (CMD) and The American Heritage Dictionary of English Language (AHD) is looked-up for ARPAbet phonemes substitution of an English word. The ARPAbet phoneme is syllabified to remove inconsistency in the pronunciation dictionaries used. Each phoneme-syllable is transliterated from syllable transliteration database by a simple mapping function. The transliterated syllables are post-processed (un-syllabified) for readability and the final transliterated word is obtained [3]. Fig 2 shows the working of Dictionary-lookup method for English to Kashmiri transliteration.

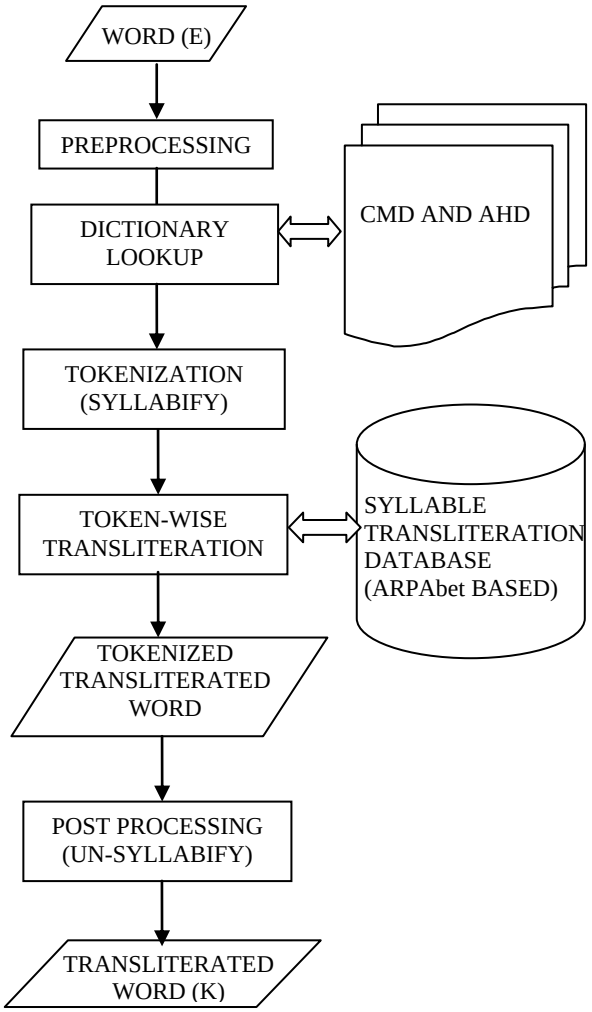

Fig 2: Dictionary Look-up based Transliteration.

2.2.2 Grapheme to phoneme

Since it is not practically possible to find all Named Entities and OOVs in the English-Kashmiri transliterated word pair or in the CMD or AH Dictionary of English Language for direct or syllable based transliteration, some words may still count as a word-error [4]. For such words, a slower but fairly effective method is used that consists of changing a grapheme to its phoneme based on a syllable mapping. English word is tokenized into unigram and bigram graphemes (characters) in such a way that almost all token resemble an ARPAbet. For this a huge number of tokenization rules are developed. These tokens are mapped to their equivalent ARPAbet syllables or phonemes. For the tokens which did not resemble ARPAbet equivalents, alphabet (unigram) tokenization is used. The syllables are then transliterated to Kashmiri script tokens. The token may be post processed (un-syllabified) similarly as in Dictionary-lookup method for better readability and to obtain final transliterated word. The elaborated illustration of the method employed is illustrated in Fig 3.

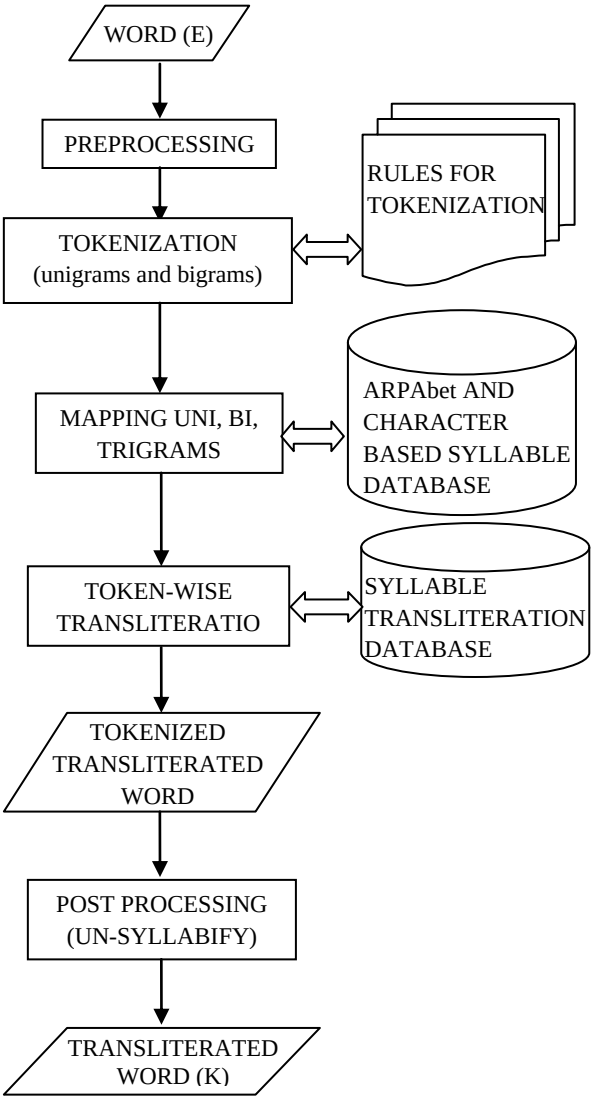

Fig 3: Grapheme to phoneme based transliteration.

3. ALGORITHM

The overall transliteration of English to Kashmiri is summarized in following steps:

3.1 Start by Normalization of first English word (Preprocessing)

Normalize each English word by removing upper-case properties and special characters.

3.2 Search each word in Transliteration Database //Direct Transliteration

If the English word is found in the transliteration database, replace the word by its corresponding transliterated word.

3.3 If not found in Transliteration Database //Phoneme-based Transliteration

3.3.1 Search word in CMU & AHD Database

If the English word is found in CMU & AH Database, perform the following steps:

- Replace the English word by its phoneme in the database
- Tokenize the phoneme into ARPAbet syllables by using black space between the phonemes as delimiter.
- Map each syllable against its transliteration in the ARPAbet based transliteration database and substitute each syllable by its corresponding Kashmiri token.
- Un-syllabify the Kashmiri tokens for better readability to finally get the transliteration of the input English word.

3.3.2 If not found in CMU & AHD Database

If the word is not still found, automatic syllable generation is required to get transliterated word as summarized in following steps:

- Tokenize the English words in unigrams, bigrams and trigrams trying to match the articulation with ARPAbet syllables as much as possible.
- Substitute each token by its corresponding syllable by mapping it in ARPAbet and character based syllable database.
- Transliterate each syllable from syllable transliteration database to obtain a tokenized transliterated word in Kashmiri script.
- Un-syllabify the Kashmiri Tokenized word for better readability to finally get a transliterated output of the input English word.

3.4 Repeat the algorithm for next word.

The tokenization of grapheme of English words is achieved by following steps:

1. Traverse the whole word string and traverse backwards until a vowel is encountered.
2. If the next character or grapheme is a consonant, set the syllable delimiter prior to that consonant.
3. Else put the syllable delimiter preceding that vowel.
4. Repeat from step two until all string is consumed.
5. Treat multiple consonants that are not bound in syllable boundaries as bi-gram graphemes if it is one of the ARPAbet Phonemes
6. Else token the remaining as unigram graphemes.
7. Repeat for all boundless graphemes.

Table 1 shows the vowel Orthography based on the position of occurrence in the English word to be transliterated while Table 2 shows the same for Consonants [5]. Both the tables also show the International Phonetic Alphabet (IPA) pronunciations for each ARPAbet.[8]

Table 1. English Vowels mapped to Kashmiri.

ARPAbet	IPA	Kashmiri (Perso-Arabic)		
		Initial	Middle	Final
AA	a	آ	◌َا	◌َا
AE	æ	اَی	◌َے	ے
AY	aɪ	آئ	◌َائِ◌	◌َائ
AW	aʊ	آؤ	◌َاؤ	◌َاؤ
AO	ɔ	اَو	◌َو	◌َو
OY	ɔɪ	اَئِ◌	ائِ◌	ائے
EH	ɛ	اَے	◌َے	ے
ER	ɝ	اَر	◌َر	◌َر
EY	eɪ	ای	ی	ی
IH	ɪ	اِ	◌ِ	◌ِ
IY	i	اِی	◌ِی	◌ِی
OW	oʊ	او	◌ُو	◌ُو
UH	ʊ	اُ	◌ُ	◌ُ
AH	ʌ	آ	◌َ	◌َا

Table 2. English Consonants mapped to Kashmiri.

ARPAbet	IPA	Kashmiri (Perso-Arabic)		
		Initial	Middle	Final
B	B	بـ	ـبـ	ـب
CH	tʃ	چـ	ـچـ	ـچ
D	d	دـ	ـدـ	ـد
DH	ð or ḍ	ڈـ	ـڈـ	ـڈ
F	f	فـ	ـفـ	ـف
G	g	گـ	ـگـ	ـگ
HH	h	حـ	ـحـ	ـح
JH	dʒ	جـ	ـجـ	ـج
K	k	کـ	ـکـ	ـک
L	ɬ	لـ	ـلـ	ـل
M	m	مـ	ـمـ	ـم
N	n	نـ	ـنـ	ـن
P	p	پـ	ـپـ	ـپ
R	r or ɹ	ر	ـر	ـر
S	s	سـ	ـسـ	ـس
SH	ʃ	شـ	ـشـ	ـش
T	t	تـ	ـتـ	ـت
TH	θ	تھـ	ـتھـ	ـتھ
UW	u	اُو	◌ُو	◌ُو
V	v	وـ	ـوـ	ـو
W	w	وـ	ـوـ	ـو
Y	j	یـ	ـیـ	ـی
Z	z	زـ	ـزـ	ـز
ZH	ʒ	ژـ	ـژـ	ـژ

4. PERFORMANCE EVALUATION

The transliteration system gives fairly good results in every domain; however to check its accuracy with maximum word-errors, it has been tested on quite difficult text sets from medical sciences that contain difficult terminology and from Wikipedia that usually includes foreign transliterated words in English, Named Entities and acronyms. The evaluation metrics used is Word Accuracy Rate that is used to compute the percentage rate of accurate transliterations out of total transliterations generated. The system has been tested for more than 15000 words and has achieved an intelligible transliteration accuracy of 86 %. One of the test sets in English along with its transliteration result is as shown in Table 3.

Table 3. Wikipedia text transliterated.

Input Text:
The Baba Ghulam Shah Badshah University is a university in India which came into existence by the Act of the Jammu and Kashmir Legislative Assembly in 2002. The university, named after saint Baba Ghulam Shah Badshah, focuses on post graduate training and undertakes research in fields such as management, environment, biodiversity, biotechnology, computer sciences, information technology, and applied mathematics. Javed Musarrat is new vice chancellor of this university
Output Text:
تھے بابا غلام شاہ بادشاہ یونیورسٹی اِس آ یونیورسٹی ان انڈیا وچ کیہ اِن ٹو ایکز ٹیٹس بائے تھے ایکٹ آف تھے جموں اینڈ کشمیر لیجسلاٹیو اسمبلی ان 2002 . تھے یونیورسٹی ، نیمڈ آفٹر سینٹ بابا غلام شاہ بادشاہ ، فوکوسیس آن پوسٹ گریجویٹ ٹریننگ اینڈ انڈر ٹیکز ریسرچ ان فیلڈز سچ اینِ مینجمنٹ ، یسیٹی ، بیٹیکنولوجی ، کمپیوٹر سائنسز انوائرومینٹ ، بیوی . ، انفارمیشن ٹیکنالوجی ، اینڈ اپلائڈ متھ میٹکس جاوید مسرت اِس ہی وائس چانسلر آف دس یونیورسٹی

5. CONCLUSION

This paper elaborates the hybrid method used to develop English to Kashmiri transliteration system mainly based on Phonemes. Besides direct mapping (word to transliteration), phonemes are extracted from phonetic dictionaries available and rules are devised to transliterate the phonemes. The grapheme of an English word is exploited to extract its phonemes. Also, character to character transliteration is used for words whose transliteration is not attained until the last stage of the transliteration.

However, the system is not highly accurate because of multiple mappings of same phoneme and due to the lack of a good word processor for Kashmiri language that can be dependently used to un-syllabify the tokenized transliteration word. Even after these challenges, the system is able to transliterate Roman script to Latin Script quite well and almost all the transliterations are intelligible and understandable.

6. REFERENCES

[1] Bhalla, D. and Joshi, N. 2013 Rule Based Transliteration Scheme For English To Punjabi. International Journal on Natural Language Computing. Vol. 2 (No. 2), 67-73. .

[2] CMU. The CMU Pronunciation Dictionary. www.speech.cs.cmu.edu/cgi-bin/cmudict. School of Computer Science, Carnegie Mellon University, Pittsburgh, USA, 2006.

[3] Joshi, H., Bhatt, A. and Patel, H. 2013. Transliterated Search using Syllabification Approach. Forum for Information Retrieval Evaluation, Delhi, India.

[4] Oh, J. and Choi, K. 2002. An English-Korean Transliteration Model Using Pronunciation and Contextual Rules. In proceedings of the 19th International Conference on Computational Linguistics, Taipei, Taiwan, 758-764.

[5] https://en.wikipedia.org/wiki/Kashmiri_language#Perso-Arabic_alphabet

[6] Deep, K. and Goyal, V. (2011). Development of a Punjabi to English Transliteration System. International Journal of Computer Science and Communication. Vol. 2 (No. 2), 521-526.

[7] Josan, G. and Kaur, J. 2011. Punjabi To Hindi Statistical Machine Transliteration. International Journal of Information Technology and Knowledge Management, Vol. 4(No. 2), 459-463.

[8] Abbas R. A. and Madiha I. English to Urdu Transliteration. Proceedings of the Conference on Language & Technology, NUC and ES, Lahore, Pakistan.