AUTOMATIC TRANSLITERATION OF PROPER NAMES SOMALI TO ENGLISH

AHMED MUKTAR OMAR, QU JIAN, SUMETH YUENYONG

Abstract: Transliterating of proper names is the process of converting a word from source natural language (such as Somali) to a target natural language (such as English) while maintaining language pronunciation. Proper names and technical words are challenging in bilingual translation systems and also in Cross-Language Information Retrieval (CLIR) applications, due their absence in the most dictionary. In this paper, we study an automatic transliteration from Somali to English; Somali-English transliteration is an unstudied problem. Our Somali-English transliteration system uses transliteration rules based on the orthographic mapping of the source language characters to the characters of the target language. We also propose an alignment method that maps the Somali characters when there is no direct match character to get accurate transliteration of the target language (English), our novel approach particularly enhances Somali-English transliteration.

Keywords: Somali-English, Somali Transliteration Table, Grapheme-Based

Introduction: Somali is a Cushitic language which belongs to the family of Afro-Asiatic languages (or Hamito-Semitic). The Somali language is similar to Semitic languages such (Arabic and Hebrew). It is a mother tongue for ethnic Somalis in Greater Somalia and is by far well-documented of all Cushitic languages Lecarme & Maury [1].

Somali is the official language of Somalia and Djibouti, as working Language in the Somali regions of Ethiopia and Kenya, Somali uses different writing systems, and the Latin alphabet has been the official writing system in the Federal Republic of Somalia and Djibouti since 1972, Andrzejewski, [2]. It mostly uses the Roman alphabet except for "p,v z" without diacritic signs or special characters, although the " ' " glottal stop stands for the (Arabic Hamza).

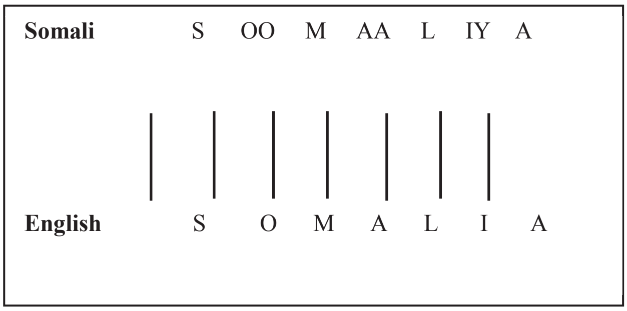
Somali also has three digraph consonants (kh (خ), sh (ش), dh (ط or ظ)) which are based similar Arabic sounds. Somali orthography corresponds mostly to Roman alphabets except where some characters are modified for the usage of Somali characters, where the letters c and x designed to accommodate the voiced and voiceless pharyngeal fricatives, comparable to (h = ح) and (ʕ = ع) Lecarme & Maury [1]. Somali long vowels usually are written by doubling the vowel itself.

The purpose of general transliteration is not to introduce new sounds to the target language which the target does not provide accommodation. However, its purpose is to substitute the original letter to the nearest letter in the target language.

For example, the concept of long and short vowels letters in Somali, long vowels are usually indicated by repeating the vowel itself such as "aa", "ee", "ii", "oo", "uu."

For example, the Somali word Soomaaliya usually transliterated to English as Somalia by omitting the long vowels. Figure 1 demonstrated a basic word transliteration.

Figure 1: Basic Transliteration



The structure of the paper follows. In Section II, we describe the previous study of machine transliteration. In Section III, we describe our character mapping method for Somali-English transliteration; we also discuss our transliteration rules of Somali proper names to English. In Section IV, we detail our experiments, evaluation metrics and the results we obtained; and Section V concludes the paper.

II. Previous work: Transliteration is the process of transliterating a word from a source language to a target language. Many different generative transliteration approaches have been studied in the literature, each of which brings out various processes in different languages. Referable to the many varying features of these methods such as the direction of transliteration, writing systems of different languages, or various applications, classification of these works is not straightforward. Transliteration refers to an orthographical transformation or phonetic change across two languages with different scripts.

Earlier work has been done for Machine Transliteration grapheme-based approaches or phoneme-based approaches; Lee and Choi, [3] proposed a source channel model (SCM) a grapheme-based approach for English-Korean transliteration, they used a direct orthographical mapping from source graphemes to target graphemes. Knight and Graehl, [4] proposed Japanese to English back-transliteration using the similarity of SCM. Wan and Verspoor, [5] modeled a technique to transliterate proper names from English to Chinese using a phonetic procedure. They proposed an algorithm for mapping from English characters to Chinese characters based on heuristics relationships between English spelling and pronunciation, and stable relationships between English phonemes and Chinese characters.

Kang and Kim [6] explored a forward-transliteration and back-transliteration for English-Korean using a direct and pivot method and then they used chunks of phonemes to perform the transliteration and back-transliteration, Kang and Choi [7] also studied an English-Korean back-transliteration using a decision-tree learning. English-Korean word alignment procedure they used similarly as Lee and Choi, [3].

Oh and Choi, [8] also studied model for English-Korean transliteration using pronunciation and contextual rules. Their method composed two phases: alignment and transliteration in their first phase, they aligned English pronunciation units (EPU) taken from a pronunciation phrasebook and aligned it to Korean phonemes to find the probable correspondence between the EPU and phonemes. Virga and Khudanpur, [9] presented English-Chinese transliteration using a phonetic representation of English names into Chinese to support Cross-Lingual Speech and Text Processing Applications.

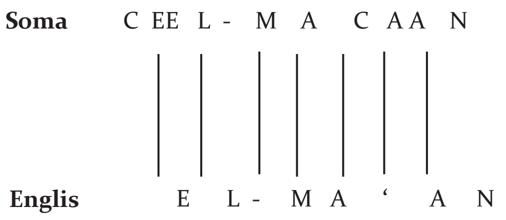
AbdulJaleel and Larkey [10] proposed a generative statistical transliteration model for English-Arabic transliteration using n-gram methods. The n-gram model generates Arabic characters of strings from a string of English characters. Malik, [11] proposed a rule based Punjabi machine transliteration by transliterating a word between two scripts of Punjabi. Most of the transliteration methods have been proposed between English and other languages of the world like Arabic, Chinese, Japanese except Somali, for both transliteration systems and cross-language information retrieval applications.

Grapheme-based transliterations consider transliteration as an orthographic process rather than phonetic process and maps groups of graphemes/ characters in the source language (SL) word "S" directly to groups of graphemes/characters in the target language (TL) word "T" Karimi, [12]. An example of grapheme based transliteration approach is shown in Figure 2.

Figure 2: Spelling-based Transliteration

This approach also is known as (spelling-based or direct methods) as it directly transforms the SL graphemes into the TL graphemes without any phonetic knowledge of the source and target languages. Instantaneously the phoneme-based methods require some steps in the transliteration process. However, most of the Grapheme-based methods directly depend on the information that is attainable from the characters of the words.

Forward transliteration is transliterating a word from a source language such as Somali to a foreign language such as English. For example, forward transliteration of a Somali name "Ceelmacaan" to English is "Elma'an". Backward transliteration or back-transliteration is transliterating a word from its transliterated version back to the language of origin. For example, back-transliteration of "Elma'an" from English to Somali is "Ceelmacaan". An example of forward transliteration is shown in Figure 3.



In this paper, we explore Somali-English forward transliteration using direct orthography mapping with Somali Transliteration rules.

III. Mapping and Transliteration rules: To align Somali/English characters, we use a direct orthographic mapping between the Somali and English characters, a character alignment is, given by a source language word (Somali) and its orthographic equivalent in the target language (English), to find the most probable letters [table 1].

We start by the alignment of the identical letters, in most cases, Somali words are longer than their corresponding English transliterated words. The mapping type is either one-to-one letter or many-to-one letter to avoid null mapping.

For example as shown in Figure 4, The Somali word Ceel-cadde is usually transliterated into English as El-Adde

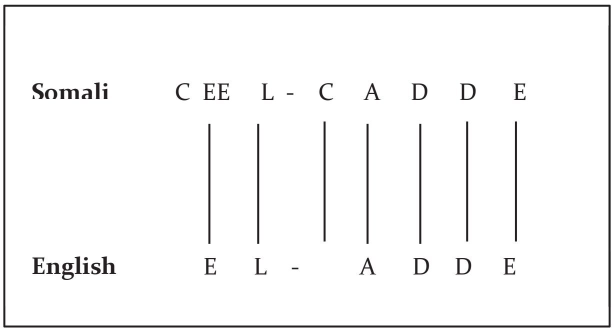


Figure 3: Missing equivalent letters

For example, The Somali word Ceel-cadde is usually transliterated into English as El-Adde.

The drawback of the above mapping of the identical letters only, we have to revise due to the absence of some of the Somali letters and long vowels we have to modify to complete the process of mapping the Somali letters and their corresponding English letters.

3.1 A direct orthographical mapping

Table 1: Somali Transliteration Table

consonant mapping		**vowel mapping**	
Somali	English	Somali	English
‘			
b	[b]	a	a
t	[t]	e	e
j	[j]	i	i
x	[h]	o	o
d	[d]	u	u
r	[r]		
s	[s]		
c	[ʕ] ‘		
g	[g]		
f	[f]		
q	[q]		
k	[k]		
l	[l]		
m	[m]		
n	[n]		
w	[w]		
h	[h]		
y	[y]		
kh	Kh	aa	a
sh	Sh	ee	e
dh	dh	ii	i
		oo	o
		uu	u

The mapping in Table 1 maps only equivalent letters, and it is not enough to get better transliteration, so we developed Somali Transliteration Table (STT) similar to Buckwalter transliteration table Buckwalter, [13], and it allows us to map the Somali letters which are assigned Somali sounds that are either not present or used differently in English. In this case, we would be able to increase the performance of the transliteration.

As can been seem from Table 1, Somali and English both use Roman alphabets, though Somali has 24 letters 19 consonant monographs and five vowel monographs also three consonant digraphs and five long vowels.

Consonant mapping: Consonants can be divided into two constants that have similar phonetic properties and consonants that are either not present in English or pronounced differently, for example, the Somali "b" letter matches English "b" and "p" letters. Consonants which are unique to Somali are (c, x and the " ‘ "glottal stop) for the Arabic Hamza, these consonants frequently occur in words. Somali syllable structure based on Consonant Vowel Consonant (CVC) and clusters of two consonants that do not occur at the beginning or the end of a word, they only happen at syllable boundaries. Somali glottal stop " ’ " or the Arabic Hamza usually is not written unless it happens at the border of a syllable or in mid of the word.

Vowel mapping: Somali has five vowel monographs, Somali vowels have one to one correspondence with English vowels, and Somali be different regarding long vowels it doubles the short vowels, we map the Somali diphthong vowels with double English vowels.

3.2 Challenges in character to character mapping

As shown in Table 1, the total number of letters in Somali and English are not equal. The Somali letters "C", and the " ‘ " glottal stop have no equivalent mapping in English. These letters will never be mapped in Somali to English transliteration using a direct orthographic mapping. Another problem is the use of long vowels in Somali. The basic concept of this regulation is the equivalent as that of the character to character mappings, these rules include consonants that require substitution and long vowels, we only discuss most important rules.

3.3 Dependency rules: Character to character mapping only is not satisfactory, to get a better result for the transliteration. We need to add a particular dependency or appropriate rules for constructing accurate transliteration.

Consonants: Somali consonants are transliterated into their corresponding English consonants, here we discuss the consonants that are unique to Somali and how to transliterate them into English.

Starting with "C" letter called "Ceyn" in Somali, when a "C" occurs at the beginning, and the end of a word than "C" will be omitted.

"C" also is omitted when it occurs between two different vowels, but it replaced with the ‘glottal stop.

"C" is also transliterated into " ‘ " glottal stop if C appear at the end of mid-syllable and the next syllable is a consonant.

If "C" occurred at the beginning of a word and followed by "U" then "C" is omitted and "U" is transliterated into "O."

"X" letter is assigned to Somali sound, so there is no equivalent English letter, "X" is transliterated into "H" which is the nearest English letter.

"X" always transliterated into "H" no matter the position it occurs.

If "X" occurs in the middle of a word behind "U" letter then "X" is replaced with "H" and "U" is replaced with "O".

Hamza " ' " only shown if it occurs between same vowels, or when it takes place in a single syllable word, but most of the cases are not written.

"Y" letter in Somali is treated as consonant but in English, it is regarded as vowel and consonants.

If "Y" occurred in mid of the word and followed by "I" not between two "I" vowels then "Y" is omitted.

If "Y" occur in mid of the word after "I" not between two "I" vowels then "Y" is omitted.

If "Y" happen in mid of the word after "E" vowel not between two "E" vowels then "Y" is omitted.

Finally, if "Y" occur in mid of the word after "A" vowel and not between two "A" vowels then "Y" is omitted.

Long vowels: Somali long vowels are twice as long as short vowels and are written as double vowels.

The rules of the Somali long vowels transliterated by transforming the long vowels to short vowels.

For example, if long vowel "AA" occurs in a word is substituted with short vowel "A" and the rest of the long vowels follow the same procedure.

Algorithm: Somali transliteration rules

Require Somali dependency rules: Insert string S (Somali word) Search for **pattern** in Regex **Foreach** all dependency rules **If** S contain C letter while C is matched do if ((C) letter occur Init of a word)&& next vowel not "u" omit C; else if (vowel is "u") omit C && replace "u", "o"; else if (C in between same vowels) or ultimate of syllable) Replace (" C " ," ' ") else Omit C; **End while elseif** S contain X letter **While** X matched do if next vowel to X is "u" replace "x" with "h" && "u" with "o"else Replace "x" with "h".**elseif** S contain long vowels while long vowels "aa", "ee" , "ii", "oo" , "uu" is matched do replace all with their short vowels "a" , "e", "i", "o" , "u" **else** if (Y in mid-word vowel before is "i" or vowel after Y is "i" not between two "i")

Omit Y; else if (Y in mid-word vowel before is "e" or vowel after Y is "e" not between two "e")

replace "y" , "i" else if (Y in mid-word vowel before is "a" and next letter is consonant)replace "y" , "i"

Endwhile End if No **pattern** to match **End foreach**

Output S as T (English word)

3.4 Somali transliteration Process

The Somali transliteration architecture and its functionality are discussed in this section, the Somali Transliteration (ST) system figure describes the structure of our proposed transliteration method, the ST uses rules for transliteration, and it is implemented using regular expression pattern matching algorithm [Algorithm1]. A flow chart of the transliteration architecture is shown in Figure 5.

The system takes Somali text as input and search for the letters to align each letter to its correspondence letter in English as a pattern to match in the transliteration table which consists of the character mappings and dependency rules.

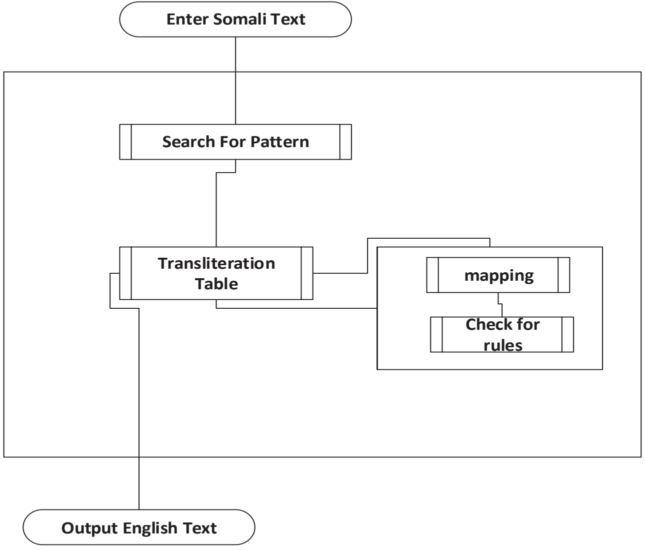


Figure 4: Transliteration architecture

If there are no similar letters in the alignment, it looks for the dependency rules to check if there is a rule to use then it applies and send to the transliteration unit. The transliteration unit replaces the matched pattern to transliterate and then outputs the word as an English word.

IV. Experiments: In this section, we describe the experimental data which includes the data used for testing as well the data used for validation purpose, the transliteration rules that are measured and metrics used to estimate performance.

4.1 **Data Set:** We obtained 600 Somali words and English transliteration from SomaliNames website [14] which contains bilingual of the most frequently used Somali names with their English transliterations, the data were divided randomly into two sets 400 names were used for experimenting the rules and the remaining 200 words were chosen as testing for the transliteration rules. The 600 English transliterated words were used as a reference to verify the correctly transliterated Somali names.

4.2 **Evaluation metrics:** The results of Somali transliteration to English were measured by the number of the transliterated words that correctly matched the transliterated words obtained from the Somali website divided by the total number of the phrase in the validation set.

Word accuracy (WA), also known as transliteration accuracy, measures the proportion of transliterations that are correct.

$$WA = \frac{\textbf{Number of correctly transliterated words}}{\textbf{Total Number of Reference words}}$$

The transliteration accuracy or word accuracy is measured the percentage of transliterated Somali words to English.

4.3 Results: After the selected input Somali texts, they are transliterated into English texts by using the Somali Transliteration tool. Then the transliterated English texts are verified for mistakes and accuracy. Measurement is accomplished with the help of the transliterated words retrieved from SomaliNames of Somali and English [14].

Table 2: Results of Somali transliteration

Type	**Accuracy**
DOM	36.13%
STT	62.69%
STT with DR	96.64%

The experimental results is shown in Table 2, it is clear that the Somali transliteration (STT) with Dependence rules that we have developed gives more than 96.64% accuracy where the Table STT indicates better outcomes 62.69% than the direct orthographical mapping (DOM) which gave 36.13% accuracy on the Somali names list. So our transliteration system accomplishes the requirement of transliteration across Somali to English.

Conclusion: In this paper, we explored the automatic forward transliteration for Somali proper names to English. This language-pair has not been studied before in any automatic transliteration work. All our transliteration rules and alignments procedure we explored here based on a direct orthographic transformation, by avoiding the intermediary phonetic interpretation of the phoneme-based methods the transliteration error rate is reduced.

Spelling-based approaches are considered to be easier regarding implementation and show better performance than the phonetic-based approaches because they do not depend on pronunciation dictionaries which may not consist of the pronunciations of all words. Last but not least, we have chosen transliteration rules rather than machine learning methods due to the unavailability of a large corpus containing Somali-English paired-words, In addition to that, the Somali transliteration table along with the dependency rules improved accuracy significantly.

References:

1. Jacqueline Lecarme and Carole Maury., A software tool for research in linguistics and lexicography: Application to Somali. Computers and Translation (Paradigm Press) 2, 1987.
2. Andrzejewski, B. W. The Introduction of a National Orthography for Somali. London: School of Oriental and African Studies, 1974.
3. Lee, J. S., and Choi, K. S. English to Korean statistical transliteration for information retrieval 1998.
4. K. Amrutha Varshini, K. Padma Vasavi, An Area Efficient Vlsi Architecture For Da-Based Reconfigurable Fir Filter.; Engineering Sciences international Research Journal : ISSN 2320-4338 Volume 3 Issue 2 (2015), Pg 45-49
5. Knight, K., and Graehl, J. Machine transliteration. In Proceedings of the 8th Conference on European Chapter of the Association for Computational Linguistics, pages 128–135 1998.
6. Wan, S, and Verspoor C.M. Automatic English-Chinese name transliteration for development of multilingual resources. In Proceedings of the 17th international conference on Computational linguistics-Volume 2, pp. 1352-1356., 1998.
7. Kang, I.H. and Kim, G. English-to-Korean transliteration using multiple unbounded overlapping phoneme chunks. In Proceedings of the 18th conference on Computational linguistics-Volume 1 (pp. 418-424). 2000 ACL
8. Kang, B. J., and Choi, K. S. Two approaches for the resolution of word mismatch problem caused by English words and foreign words in Korean information retrieval. International Journal of Computer Processing of Oriental Languages, 14(02), 109-1312001.
9. A. Uma Maheswari, On A Class Of Quasi Hyperbolic Kac Moody Algebras Of Rank 11.; Engineering Sciences international Research Journal : ISSN 2320-4338 Volume 3 Issue 2 (2015), Pg 50-56
10. Oh, J. H., and Choi, K. S. An English-Korean transliteration model using pronunciation and contextual rules. In Proceedings of the 19th international conference on Computational linguistics-Volume 1 (pp. 1-7). 2002. ACL.
11. Virga, P., and Khudanpur, S. Transliteration of proper names in cross-lingual information retrieval. In Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition-Volume 15 (pp. 57-64). 2003. ACL
12. AbdulJaleel, N., and Larkey, L. S. Statistical transliteration for English-Arabic cross-language information retrieval. In Proceedings of the twelfth international conference on Information and knowledge management (pp. 139-146). 2003 ACM.
13. Anwar Bhasha Pattan, Makkena Madhavi Latha , FPGA Implementation Of A 64 Point Radix-2

Single Path Delay Feedback FFT Architecture.; Engineering Sciences international Research Journal : ISSN 2320-4338 Volume 3 Issue 2 (2015), Pg 57-60

14. Peter Malik, M. G. Punjabi machine transliteration. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of s (pp. 1137-1144)., Ed. New York: McGraw-Hill, pp. 15–64. 2006.
15. Karimi, S. 2008. Machine transliteration of proper names between English and Persian (Doctoral dissertation, RMIT University, Melbourne).
16. Chetna, Dr. Hardeep Singh, Representation Of Numbers Divisible By A Sufficiently Large Square.; Engineering Sciences international Research Journal : ISSN 2320-4338 Volume 3 Issue 2 (2015), Pg 73-76
17. Buckwalter,Timothy A. "Lexicographic notation of Arabic noun pattern morphemes and their inflectional features." In Proceedings of the Second Cambridge Conference on Bilingual Computing in Arabic and English, pp. 5-7. 1990.
18. Somali names list with their English transliteration retrieved Somalinames. http://www.somalinames.com/index.php/somali-names/soma-names.csv March 12 2014.
19. Dr.Ca Jyothirmayee, Chemical And Microbial Analysis Of Potable Water In Public – Water Supplywithin Five Mandals In The Upland Area Of West Godavari District, Andhra Pradesh, India.; Engineering Sciences international Research Journal : ISSN 2320-4338 Volume 3 Issue 2 (2015), Pg 65-72
20. Dr. Dhananjaya Reddy, Role Of Mathematics Teacher In Using Technology.; Engineering Sciences international Research Journal : ISSN 2320-4338 Volume 3 Issue 2 (2015), Pg 61-64

Ahmed Muktar Omar
School of Information Technology, Shinawatra University, main campus 99 Moo 10,
Bangtoey, Samkhok Pathum Thani 12160
Qu Jian
Program Director, Lecturer, School of Information Technology, Shinawatra University,
main campus 99 Moo 10, Bangtoey, Samkhok Pathum Thani 12160
Sumeth Yuenyong
Lecturer, School of Information Technology
Shinawatra University, main campus 99 Moo 10, Bangtoey, Samkhok Pathum Thani 12160