Research Article

Automatic Transliteration of Proper Names from Somali to English

Ahmed Muktar Omar*, Jian Qu and Sumeth Yuenyong

School of Information Technology, Shinawatra University
99 Moo 10, Bangtoey, Samkhok, Pathum Thani 12160, Thailand

Abstract

Transliterating of proper names is the process of converting words 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 to their absence from most dictionaries. In this paper, we study an automatic transliteration from Somali to English; which is an under-studied 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 matching character to get accurate transliteration in English. Our novel approach particularly enhances Somali-English transliteration.

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

1. 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 as Arabic and Hebrew. It is a mother tongue for ethnic Somalis in Greater Somalia and is by far the most well-documented of all Cushitic languages [1].

Somali is the official language of Somalia and Djibouti and a 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 [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 on similar Arabic sounds. Somali orthography corresponds mostly to Roman alphabets except where some characters are modified for the usage in Somali characters, where the letters c and x designed to accommodate the voiced and voiceless pharyngeal fricatives, comparable to (h = ح) and (ʕ = ع) [1]. Somali long vowels usually are written by doubling of the vowel itself.

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

*__Correspondence__ : muktar7@gmail.com

DOI 10.14456/tijsat.2016.26

the original letter to the nearest letter in the target language. The concept of long and short vowels letters exists in Somali, and 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.

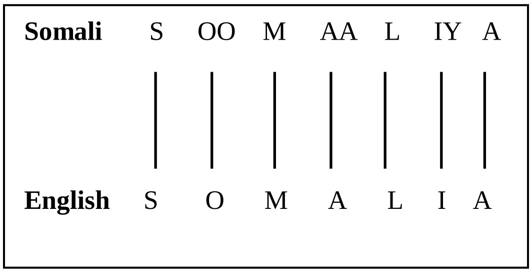

Figure 1. Basic Transliteration.

In this paper we propose a Somali-English transliteration system that 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 matching character to get accurate transliteration in English.

The structure of the paper is as 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.

2. Related Work

Transliteration refers to an orthographical transformation or phonetic change across two languages with different scripts. Many different generative transliteration approaches have been studied in the literature, each of which brings out various processes in different languages. These methods vary by the direction of transliteration, writing systems of different languages, or intended applications. Classification of these works is not straightforward.

Earlier work has been done for Machine Transliteration grapheme-based approaches or phoneme-based approaches. Lee and Choi proposed a source channel model (SCM) a grapheme-based approach for English-Korean transliteration [3]. They used a direct orthographical mapping from source graphemes to target graphemes. Knight and Graehl proposed Japanese to English back-transliteration using the similarity of SCM [4]. Wan and Verspoor modeled a technique to transliterate proper names from English to Chinese using a phonetic procedure [5]. 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 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 [6]. Kang and Choi also studied an English-Korean back-transliteration using a decision-tree learning [7]. The English-Korean word alignment procedure they used is similar to Lee and Choi [3].

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

AbdulJaleel and Larkey proposed a generative statistical transliteration model for English-Arabic transliteration using n-gram methods [10]. The n-gram model generates strings of Arabic characters from a string of English characters. Malik proposed a rule based Punjabi machine transliteration by transliterating a word between two scripts of Punjabi [11].

Grapheme-based transliterations consider transliteration as an orthographic process rather than phonetic process and maps groups of graphemes/characters in the source language word directly to groups of graphemes/characters in the target language word [12]. This approach also is known as (spelling-based or direct methods) as it directly transforms the source language graphemes into the target language 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 as it is written in the 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". This example is shown in Figure 2.

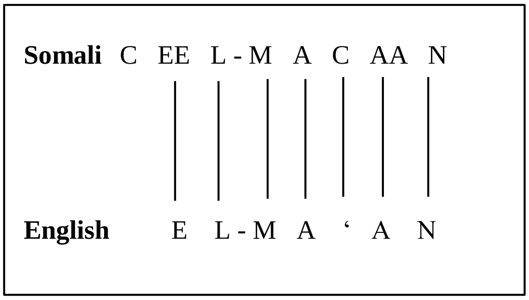

Figure 2. Forward Transliteration.

Most of the transliteration methods have been proposed between English and other common languages such as Arabic, Chinese, or Japanese. Somali, not being a common language, is under-studied for both transliteration systems and cross-language information retrieval applications.

3. 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 in Somali and its orthographic equivalent in English to find the most probable letters.

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 3, the Somali word Ceel-cadde is usually transliterated into English as El-Adde.

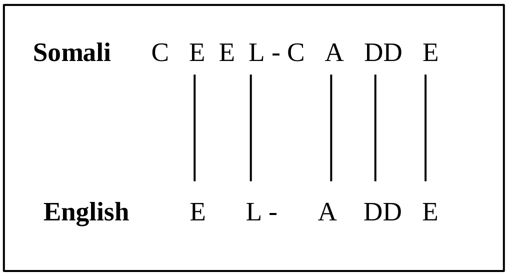

Figure 3. Missing equivalent letters.

The drawback of the above direct mapping is the absence of some Somali letters and long vowels in English. This is the problem addressed by our method.

3.1 Somali Transliteration Table and its Problems

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

The mapping in Table 1 maps only equivalent letters, and it is not enough to get good transliteration, so we developed Somali Transliteration Table (STT) similar to Buckwalter transliteration table [13]. It allows us to map the Somali letters with 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.

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) and for the Arabic Hamza, these consonants frequently occur in words. Somali syllable structure is based on Consonant Vowel Consonant (CVC) and clusters of two consonants that do not occur at the beginning or the end of a word, but 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 the middle of the word.

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

Vowel mapping: Somali has five vowel monographs; Somali vowels have one to one correspondence with English vowels. However, Somali is different regarding long vowels, which are short vowels repeated twice. We map the Somali diphthong vowels with double English vowels.

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. We came up with novel dependency rules which address the problem of no direct equivalent letter in Somali to English transliteration.

3.2 Dependency rules

Character to character mapping only is not satisfactory to get an acceptable 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 then "C" will be omitted.

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

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

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

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

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

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

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

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

If "Y" occurs in the middle of the word and is followed by "I" but not between two "I" vowels, then "Y" is omitted.

If "Y" occurs in the middle of the word after "I", but not between two "I" vowels, then "Y" is omitted.

If "Y" happens in the middle of the word after an "E" vowel, but not between two "E" vowels, then "Y" is omitted.

Finally, if "Y" occurs in the middle of the word after and "A" vowel, but 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 it is substituted with short vowel "A" and the rest of the long vowels follow the same procedure.

The exact order of application of these dependency rules can be seen in Algorithm 1.

Algorithm1: 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
initial 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
If 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".
If 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”, and
“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
Endif No pattern to match
End foreach
Output S as T (English word)
```

3.3 Somali transliteration Process

The Somali transliteration architecture and its functionality are discussed in this section. Figure 4 describes the structure of our proposed transliteration method, the rules for transliteration, and its implementation using regular expression pattern matching algorithm (Algorithm1).

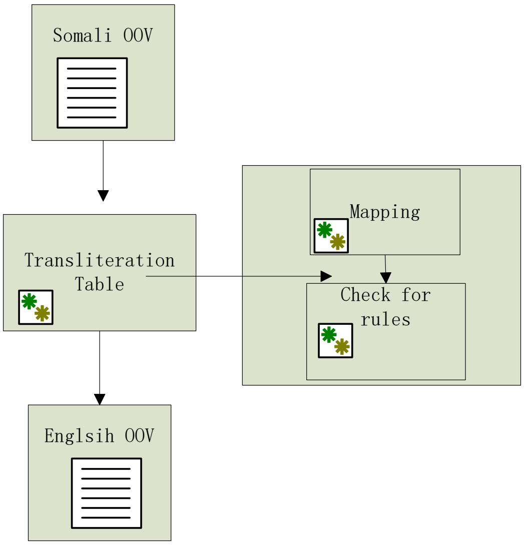

Figure 4. Somali Transliteration Architecture.

The system takes Somali text as input and searches for the letters to align each letter to its corresponding letter in English as a pattern to match in the transliteration table which consists of the character mappings and dependency rules. If there are no similar letters in the alignment, it looks for the dependency rules to check if there is an applicable rule, which it then applies and sends to the transliteration unit. The transliteration unit replaces the matched pattern to transliterate and then outputs the word as an English text.

4. Experiments

In this section, we describe the experimental data which includes the data used for testing as well as the data used for validation purpose.

4.1 Data Set

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

4.2Evaluation 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.

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

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

BLEU (Bilingual Evaluation Understudy) allowances for multiple reference translations, it is used to evaluate many term to term bilingual translation researches [14].

The BLEU technique offers a score between 0 and 1, which is a scale showing how alike the target language word is to the reference word; 0 is the least score and 1 is the best score.

We evaluate the BLEU score using the Interactive BLEU tool developed by Madnani, N [15]. The system metric measures the n-gram (n=1 to 4) precisions of the hypothesis against the reference.

4.3 Results

After the selected input Somali texts they are transliterated into English texts by using the Somali Transliteration tool, 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 [16].

Table 2. Results of Somali transliteration.

Type	Accuracy
DOM	34.62%
STT	64.73%
STT with DR	96.53%

From the experimental results shown in Table 2, it is clear that the Somali transliteration (STT) with Dependence rules that we have developed gives more than 96.53% accuracy. Also, the Table STT indicates better outcomes at 64.73% than the direct orthographical mapping (DOM) which gave 34.62% accuracy on the Somali names list. So our transliteration system accomplishes the requirement of transliteration across Somali to English.

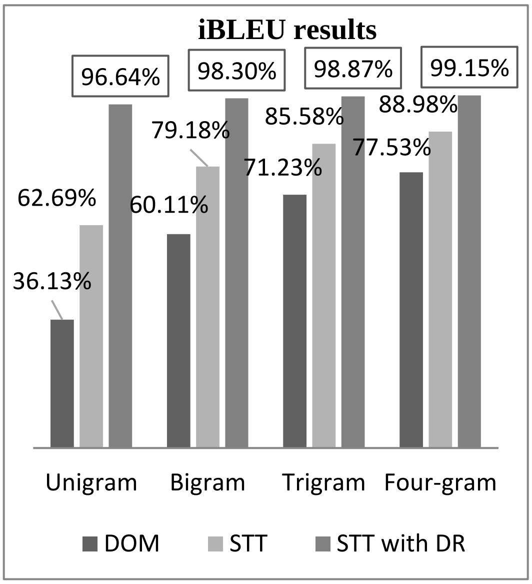

Figure 5. iBLEU Score.

Figure 5 demonstrates the results given by the BLEU interactive tool to compare the DOM and our proposed STT and dependency rules over the 600 proper names from the Somali Names data. Using a unigram, bigram 3-gram and 4-gram, DOM and STT DR score for all grams shows above 0.9664 (almost 1) as the BLEU score is between 0 to 1. The STT alone scored 0.6269, nearly twice as high as the DOM.

We see that the higher the N-gram, the higher the result, with the DOM scores of 0.3613, 0.6011, 0.7123 and 0.7753 for Unigram Bigram, Trigram and four-gram, respectively.

5. Conclusion

In this paper, we explored the automatic forward transliteration for Somali proper names to English. This language-pair is not well studied in any automatic transliteration work. All our transliteration rules and alignments procedure presented here are based on a direct orthographic transformation. Our method avoids the intermediate phonetic interpretation used in phoneme-based methods and reduces the transliteration error rate.

Spelling-based approaches are considered to be easier to implement 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. The transliteration table along with the dependency rules proposed in this paper improved accuracy of transliteration significantly.

6. References

[1] Lecarme, J. and Maury, C., A Software Tool for Research in Linguistics and Lexicography: Application to Somali, *Computers and Translation*, Vol.2 No, 1, pp.21-36, 1987.

[2] Andrzejewski, B.W., *The Introduction of a National Orthography for Somali*, School of Oriental and African Studies, 1974.

[3] Lee, J.S. and Choi, K.S., English to Korean Statistical Transliteration for Information Retrieval, *Computer Processing of Oriental Languages*, Vol. 12, No. 1, pp.17-37, 1998.

[4] Knight, K. and Graehl, J., Machine Transliteration, *Computational Linguistics*, Vol.24, No.4, pp.599-612, 1998.

[5] 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.

[6] 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.

[7] 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*, *Vol.14*, No.2, pp.109-131, 2001.

[8] 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.

[9] 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.

[10] 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.

[11] Malik, M.G., Punjabi Machine Transliteration, In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pp. 1137-1144, 2006.

[12] Karimi, S., 2008, *Machine Transliteration of Proper Names between English and Persian*, Ph.D. Thesis, RMIT University, Melbourne, 216p.

[13] Buckwalter, T.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.

[14] Papineni, K., Roukos, S., Ward, T. and Zhu, W.J., BLEU: A Method for Automatic Evaluation of Machine Translation, In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 311-318, 2002.

[15] Madnani, N., iBLEU: Interactively Debugging and Scoring Statistical Machine Translation Systems, In *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference*, pp. 213-214, 2011.

[16] Somali Names List with Their English Transliteration Retrieved Somalinames Directory. Available Source: http://www.somalinames.com/index.php/somali-names/soma-names.csv