Diacritization And Transliteration Of Proper Nouns From Arabic To English

Hamdy S Mubarak **Mohamed Al Sharqawy** **Esraa Al Masry**

Arabic NLP Dept, Sakhr Software
SAKHR Building, Nasr City, Free Zone, Cairo
Egypt
{hamdys,mas,emasry}@sakhr.com

Abstract

This paper proposes a complete system for the automatic diacritization and transliteration of proper nouns from Arabic to English using a database of name pairs in Arabic and English languages. The system consists of three phases: Correction, Diacritization, and Transliteration. Correction phase corrects the Common Arabic Mistakes (initial Hamza, final Yaa, and final Taa errors) using Normalization and corrects normal concatenation errors. The most frequent transliteration is considered in case of exact match with saved normalized tokens generated from proper names database. The missing diacritics are restored using Sakhr's Morphological Analyzer for analyzed tokens or from the best matching with patterns (for Arabic and Non-Arabic names) and consecutive characters obtained from the diacritized proper names. Transliteration rules are applied for the diacritized proper name to obtain the English equivalent (transliteration).

Our results show an average accuracy of 89% on blind test sets with forced spelling mistakes (and 95% for correct input).

1. INTRODUCTION

Transliteration is the task of transcribing a word or text written in one writing system into another writing system. Person names, locations and organizations as well as words borrowed into the language are the most frequent candidates for transliteration.

In Information Retrieval (IR), the most important query words in are often proper names. In cross language retrieval, a user issues a query in one language to search a collection in a different language. If the two languages use the same alphabet, the same proper names can be found in either language. However, if the two languages use different alphabets, the names must be transliterated or rendered in the other alphabet.

As mentioned in (Al-Onaizan and Knight, 2002), two types of transliteration exist, *forward transliteration* and *backward transliteration*.

Forward Transliteration is the transliteration of a foreign name (in the case of our system, Arabic) into English. Typically, there are several acceptable transliteration candidates. For example, the Arabic name "محمد" (mhmd in Buckwalter encoding) might correctly be transliterated into Mohamed, Mohammed, Mohammad, etc. In fact, the many types of name variation commonly found in databases can be expected.

A recent web search on Google for texts about "Muammar Qaddafi" (spelled in Arabic as معمر القذافي mEmr AlqdAfy in Buckwalter encoding), for example, turned up thousands of relevant pages under the spellings Qathafi, Kaddafi, Qadafi, al-Qaddafi, Al-Qaddafi, Al Qaddafi, etc (and these are only a few of the variants of this name known to occur).

Backward Transliteration is the reverse transliteration process used to obtain the original form of an English name that has already been transliterated into the foreign language. In this case, only one transliteration is retained.

For the transliteration from Arabic to English, some observed problems are:

- Both Arabic and English lack some sounds and letters from the other language. For example, there is no perfect match for " ع" in English and "P" in Arabic. This leads to ambiguities in the process of transliteration.

Another problem associated with Arabic is the omission of diacritics and vowels (fatha, dumma, kasra, shadda,sokoon) in almost all the Arabic writings. The information contained in unvocalized Arabic writing is not sufficient to give a reader, who is unfamiliar with the language, sufficient information for accurate pronunciation. Diacritics are considered to be one of the main causes of ambiguity when dealing with Arabic proper nouns.

Another observed problem in Arabic is the existence of Common Arabic Mistakes (CAM) in which different characters are used interchangeably; like Hamza errors (إ،أ،آ،ا), Yaa errors (ى،ي), and Taa Marbuta errors (ه،ة).

In this paper, we present a complete system for Correction, Diacritization, and Transliteration of Arabic proper nouns using a database of name pairs in Arabic and English languages.

The system searches for the normalized form of user input (in Arabic) in a dictionary of proper names, and if found it returns the most frequent transliteration, otherwise it suggests the appropriate diacritization based on its Morphological Analysis if analyzed or the best matching with patterns obtained from the diacritized proper names. And in case of mismatch with patterns, it uses a probability matrix for diacritization of any consecutive characters. Finally, the system applies transliteration rules to obtain the English equivalent.

2. PROPER NAMES DICTIONARY

For training purposes, we needed a list of name pairs, i.e. names written in Arabic and correctly transliterated into English. We used Sakhr Proper Names Database which consists of different types of proper names (Human, Location, Organization, etc). For each proper name, information about Diacritization, Transliteration, Gender, Theme, etc are provided manually.

From the total of 171K Human names (38K Arabic, and 133K non Arabic), we used only 51K names for Transliteration (13K Arabic, and 38K non Arabic) which contain the transliteration information, and use the entire list for Correction and Diacritization processes.

Examples for Arabic Human names: آمنة بنت وهب، أبو الأسود الدؤلي، الملك عبد الله الثاني, and for Non Arabic Human names: آبل غومبا، آدم سميث، جورج بوش.

In the case of Forward Transliteration (where we want to convert a name originally from Arabic into English), there is usually more than one acceptable corresponding name in English. For example, the name " طارق" has 4 different equivalents in our database: "Tarek", "Tareq", "Tarik", and "Tariq".

The distribution of names with different number of alternatives is summarized in table 1.

# Alternatives	%	Examples
1	76	Medhat مدحت, Islam إسلام
2	16	Ahmed, Ahmad أحمد
3	4	Osama, Usama, Ossama أسامة
4	2	Elias, Elyas, Ilyas, Alias إلياس
5+	2	Mohamed, Mohamad… محمد

Table 1: Number of Alternative Names

It's also observed that the maximum number of alternatives was 11 (ex: the name "يوسف" has these equivalents: Youssef, Yusuf, Yousef, Yusef, Youssif, etc.)

3. CORRECTION

From a random sample of 1000 Queries to Sakhr Arabic Search Engine on the Web (http://johaina.sakhr.com/), about 70% of these queries are Proper Names, and 13% of them have spelling mistakes in Hamza, Yaa, and Taa Marbuta. To solve this problem, we use Text Normalization for both the input and the Proper Names Dictionary.

3.1 Text Normalization

Text Normalization is a process by which text is transformed in some way to make it consistent in a way which might not have been before, and it's often performed before comparison. For Arabic, all shapes of Hamza (ا،أ،إ،آ) are converted to Plain Alef (ا), Yaa (ي) is converted to Alef Maqsura (ى), and Taa Marbuta (ة) is converted to Haa (هـ). This makes both اسامه and أسامة match.

3.2 Concatenation Errors

Another type of spelling errors is the concatenation of two or more tokens in the user input which leads to the need for splitting mechanism to obtain correct tokens. Tokens end with any letter that doesn't cause visual ambiguity - i.e. has isolated and final forms only (no initial or middle forms) like the letters (د،ذ،ر،ز،و،ا) – when concatenated to other tokens, they need a Simple Splitting; otherwise a Complex Splitting is needed.

Examples: محمداحمدعبدالله (Simple) and منالحسنمحمد (Complex).

4. DIACRITIZATION

Diacritization of proper names is used in many applications like Address Book, Text to Speech, and also as an intermediate step in Transliteration which is used in Machine Translation Systems (MT), and Multilingual Personal Information (Banking, ID Cards, Passports, etc.)

In Arabic, words consist of prefixes, suffixes, and stem; the stem can be determined by a root and a morphological pattern pair. The root represents the stem original letters while the morphological pattern decides how the stem will be pronounced.

Because short vowels (diacritics) are commonly not presented in Arabic orthography, this creates a problem in transliterating unknown proper names (Out Of Vocabulary) since these missing diacritics should be deduced before transliteration to obtain the appropriate pronunciation.

4.1 Proper Names Patterns

To diacritize unknown proper name, we use a set of rules to deduce its "pattern", and search in the proper names database to retrieve a list of proper names with the same candidate pattern. To select the best match (minimum number of character substitutions), the "Hamming distance" values are used in addition to statistical information for patterns frequencies. Suggested diacritics are the same as this favored pattern.

While deducing proper name pattern, some issues should be taken into consideration for Arabic names only (for non Arabic names, only Long vowels (ا،و،ي) are preserved) like:

1. Long vowels (ا،و،ي), Ending characters (ة،ى), different Hamza shapes (أ،إ،آ،ء،ؤ،ئ) are preserved.
2. Definite article (ال), some initial characters (بو،م), and some ending characters or suffixes (..اوي،اني،ني،ان) are also preserved.
3. Other characters (consonants) are replaceable and can match with any other consonant.
4. Special cases for preserving some consonants (ت،م) are applied.
5. The "Sun Letters" (other than the letters in string " أبغ حجك وخف عقيمه") should have an extra Shadda.

For 171K diacritized Human names (38K Arabic, and 133K non Arabic), we have generated about 3K and 10K patterns for Arabic and non Arabic respectively. Examples for these patterns are shown in Table 2 and Table3:

Pattern	Diac1	%	Ex.	Diac2	%	Ex.
ـ ـ اء	ـَـَاء	61	عَلَاء، بَهَاء	ـِـَاء	28	فِدَاء، لِقَاء
ال ـ ـ ـ و ـ ي	اَلْـ ـ ـ ـُـ و ـِ يّ	70	اَلْبَرْغُوثِيّ	اَلْـ ـ ـ ـُـ و ـِ يّ	10	اَلْبَرْمُوسِيّ

Table 2: Arabic Proper Names Patterns

Pattern	Diac1	%	Ex.	Diac2	%	Ex.
-ي-و--	ـِي ـُو ـْـ	69	ريمُوتْد	ـِي ـُو ـَـ	14	نِيكُولَس
-و---ي	ـُو ـِـْـِي	44	شُومِسْكِي	ـُو ـْـَـِي	28	كُورْدَري

Table 3: Non Arabic Proper Names Patterns

For the unknown Arabic name "سماء" which matches the 1st Arabic pattern, the system suggests its correct diacritization "سَمَاء" after getting the nearest known name "سناء". Similarly, it diacritizes the name "البعقوبي" as "اَلْبَعْقُوبِيّ".

When the input pattern doesn't match any of the existing patterns, the system recursively splits it into smaller patterns and searches for them. The overall diacritization is the concatenation of all these diacritized sub patterns after applying some concatenation rules.

Example: The name "كرومازوف" is splitted into 2 parts "كروما" and "زوف" which match the patterns "-ـ و ـ ا" and "-ـ و ـ", which leads to the final diacritization "كُرُومَازُوف".

4.1 Bigram Diacritization Matrix

After pattern splitting, if the name pattern doesn't match any of existing patterns, the system uses a Bigram Diacritization Matrix to diacritize the name. This matrix has all the consecutive characters associated with the diacritics probabilities obtained from Proper Names database. The names "اَلْعُوِيرَان" and "كُونْدِلِيزَا" are sample outputs using this matrix.

4.2 Morphological Analysis

Morphological Analyzer is also used for diacritizing normal words that exist in Arabic proper names, examples: هاني الضَّابط، أحمد عَقْل، صلاح فُولاذ

The decision of being Arabic or non Arabic name is important for diacritization and hence transliteration, and it's done through these surface rules:

1. If the name exists in Proper Names database, the Arabic and non Arabic probabilities are taken into consideration, i.e. some names tend to be an Arabic name (ex: محمد with probability of 75%), some names tend to be non Arabic name (ex: جون with probability of 98%), while others are neutral (ex: آدم with probability of 50%).
2. If the name contains one of the following characters "ح، ظ، ع، ص، ط، ض، ق", it tends to be an Arabic name. These characters are obtained statistically from Proper Names database (ex: Probability of character 'ض' in Arabic name is 84% and 16% for non Arabic name, like: حامد قرضاي).
3. If the name matches an Arabic pattern and doesn't match any of the non Arabic patterns, it's considered an Arabic name (and vice versa) and gets diacritized consequently.
4. Unless user input is provided, the name is considered non Arabic name and diacritized therefore.

5. TRANSLITERATION

Transliteration is a mapping from one system of writing (alphabet) into another, word by word. Transliteration attempts to be exact, so that an informed reader should be able to reconstruct the original spelling of unknown transliterated words. To achieve this objective, transliteration may define complex conventions for dealing with letters in a source script which do not correspond with letters in a goal script.

Transliteration is opposed to transcription, which specifically maps the sounds of one language to the best matching script of another language. Also, transliteration should not be confused with translation, which involves a change in language while preserving meaning.

5.1 Transliteration Standards

Although there is no Universal Transliteration system from Arabic to English, there are some common systems like ISO 233, Qalam, Buckwalter, etc. Any transliteration system should consider the following issues:

1. Transliteration ignores assimilation of the article before the "sun letters", and may be easily misread by non-Arabs. For instance the proper name word "An-nour" would be more correctly transliterated to "Al Nour".
2. A transliteration is ideally fully reversible: a machine must be able to transliterate it into Arabic and back.
3. Rendering several Arabic phonemes with an identical transliteration, or digraphs for a single phoneme (such as sh) may be confused with two adjacent phonemes.
4. ASCII transliterations using capital letters to disambiguate phonemes are easy to type but may be considered un-aesthetic.

Examples of character mapping in different systems are shown in Table 4.

Letter	Name	ISO	Qalam	Buckwalter
ث	Thaa	t̲	Th	v
خ	Khaa	h̲	Kh	x
س	Seen	s	S	s

Table 4: Transliteration Examples

We used a manual character mapping similar to Qalam, with some enhancements, to preserve the spelling rather than the pronunciation. The system has these additional features:

1. A customized mapping of letters. For example the mapping of letter 'ج' may be to 'g' in Egypt or 'j' in Gulf.
2. Restoring the single character abbreviation (like transliterating "دبليو" to "W.").
3. Fine tuning of the Transliteration based on statistics obtained from the database (for consecutive vowels, or special patterns and conditions).

6. EVALUATION

As described before, in case of forward transliteration, there is more than one acceptable transliteration. Ideally, our gold standard should maintain a set of equivalent English names for each Arabic entry, but it is not possible to gather all the possible transliterations for all the Arabic names. So, we evaluated the system accuracy manually through experienced linguists.

From a random sample of 1000 proper names (Arabic and non Arabic) with a total of 2200 tokens, we have normalized all the inputs (i.e. forcing Common Arabic Mistakes CAM) and evaluated the accuracy of Correction, Diacritization, and Transliteration. Manual assessment shows accuracies of 96%, 90%, and 89% for Correction, Diacritization, and Transliteration in order (and raise to 97% in Diacritization, and 95% in Transliteration if the input is correct.)
Results are shown in Table 5:

Functionality	Proper Names (+CAM errors) Accuracy %	Proper Names (-CAM errors) Accuracy %
Correction	96	-
Diacritization	90	97
Transliteration	89	95

Table 5: Correction, Diacritization, and Transliteration Accuracy

6.1 Transliteration Sample Output

Transliteration sample outputs of blind inputs are shown in Table 6:

Proper Name (+CAM errors)	Transliteration
ابوالوكل مؤنس بن فرحان الرويلي	Abu Al Wakl Monis Bin Farhan Al Rwiely
وضحه مساعد عبدالرحمن ابراهيم الجليبي	Wadha Musaed Abdul Rahman Ibrahim Al Joliby
شاديه علي وصل مرشود الحازمي	Shadia Ali Wasl Marshod Al Hazmi
هينار كي آر هولتيت	Hinar Ki R. Holtit
شانتال ميلون دلسول	Shantal Meillon Dalsol

Table 6: Sample Outputs

7. CONCLUSION

We have introduced a complete system for Correction, Diacritization, and Transliteration of names from Arabic to English with an accuracy of 89% on blind test-data. The system uses bilingual training data, along with morphological analysis (Sakhr's Morphological Analyzer), some heuristic rules and observations to achieve these results in combination with traditional statistical language processing and machine learning algorithms.

8. REFERENCES

AbdulJaleel, Nasreen and Leah S. Larkey (2003). Statistical Transliteration for English-Arabic Cross Language Information Retrieval, CIKM 2003: Proceedings of the Twelfth International Conference on Information and Knowledge Management, New Orleans, LA.

Al-Onaizan, Yaser and Kevin Knight (2002). Machine Transliteration of Names in Arabic Text, ACL Workshop on Computational Approaches to Semitic Languages.

Mehdi M. Kashani, Fred Popowich, and Fatiha Sadat (2005). Automatic Transliteration of Proper Nouns from Arabic to English, School of Computing Science, Simon Fraser University, National Research Council of Canada.