Interactive
Cross Language Information Retrieval
Using Transliterated Names Resolution

Jim Cowie and Ahmed Abdelali

MCCS-04-331

Computing Research Laboraroy
Box 30001
New Mexico State University
Las Cruces, NM 88001

The Computing Research Laboratory was established by the
New Mexico State Legislature
under the Science and Technology Commercialization Commission
as part of the Rio Grande Research Corridor

Contents

Abstract

The English-Arabic component of a cross-language retrieval system is described. This system uses bilingual English-Arabic lexicons to allow the user to construct queries in Arabic. Various methods of browsing the retrieved document set are supported including proper name recognition and word-for-word translation. The system includes an Arabic morphological analyzer, which can be invoked both during the indexing and querying phases.

As proper name recognition and translation has proved very useful for document browsing we focus much of this paper on methods to accurately recognize and transcribe names. Significant differences are encountered in various Arabic sources. These include differences in transliteration of foreign names and words and differences in spelling and terminology associated with national variations from Modern Standard Arabic. Various approaches under investigation to handle these problems are discussed.

Interactive
Cross Language Information Retrieval
Using Transliterated Names Resolution

Jim Cowie and Ahmed Abdelali
Computing Research Laboraroy
Box 30001
New Mexico State University
Las Cruces, NM 88001

Introduction

The aim of Information Retrieval (IR) is to find and retrieve documents relevant to a given query, generally where query and documents are in the same language. Recent research has extended this goal to include document collections in languages different from the language of the query, known as Cross-Language Information Retrieval (CLIR). The toolset we describe here produces a CLIR system with capabilities that will enhance overall performance and compensate for the deficiencies that the standard retrieval process may produce. The system

uses a combination of automatic and user assisted methods to build cross-language queries. It sends the modified queries to language-specific query modules, which retrieve appropriate documents. These are then displayed in a variety of summary forms intended to support the monolingual user of the system.

System Description

The underlying information retrieval system is a standard statistical system. Document collections and indexes are encoded using the Unicode character set (UCS2). This allows mixes of documents in a variety of languages to be retrieved simultaneously. We concentrate here on the front end to the system which allows queries to be constructed in any of three different modes: using a multilingual query, using an English query without user involvement in the formulation of the multilingual queries or using an English query with user involvement in the formulation of the multilingual queries.

The multilingual query mode allows users to access resources in other languages by formulating queries in several languages simultaneously. The user types a Unicode text containing terms in various languages. The documents retrieved by the query will, in general, be in the same languages as the query terms. The most relevant documents retrieved may end up being in one language, reflecting the likelihood of co-occurrence of the terms of the original query. The relevance of the documents retrieved is computed globally over the entire multilingual resource. In the case of cross-language homographs (for instance, up to 40% of the words in English are of Latin or French origin), retrieval would only be slightly affected because the co-occurring words of the query will cause the system to give any documents containing the noisy word a lower score. In other words, even though some of the documents may be mistakenly retrieved on

the basis of the accidental homograph, they will be ranked lower than those retrieved on the basis of the complete query for each language.

The second mode relies on a set of bilingual dictionaries for translating an English query into the different target languages. The resulting query is then used to retrieve documents. With this approach the user has no control over the multilingual query formulation process. Rather, success depends on the quality of the bilingual dictionaries and the size of the index word list of the resources. Ambiguous terms may cause a significant loss in precision by causing the retrieval of non-relevant documents.

Using the third mode, the user becomes involved in making decisions about the terms selected for the various queries in the different languages. Knowledge of the other languages is not essential, but will significantly improve the precision of the query

First, the English query is passed to the bilingual dictionaries, which in turn, produce translations for each query term. The translations are checked against the index word list and only those terms that have a match are kept for further processing. The interface presents these terms to the user along with details about their English correspondences and other information (i.e., part of speech, domain, English gloss if available, and so on). The user is then responsible for choosing the most appropriate terms, based on their meaning, to use for the multilingual queries. Once a query ready, it is used to retrieve documents from the indexed resources and presented to the user.

Reviewing the Retrieved Documents

The options available for viewing the retrieved documents include the highlighting in different colors of the proper names (i.e. of people, organizations and locations) in the text, as well as of the query terms (see Figure 1 for a general view and Figure 2 for the pop-up view of a

particular text). These functionalities allow users to get more information about the text in terms of co-occurrence and other relations between the entities mentioned in the text.

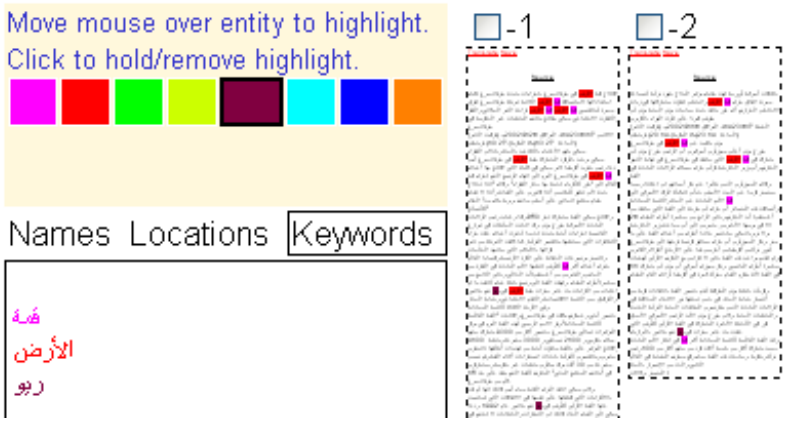

Figure 1. General view of documents retrived

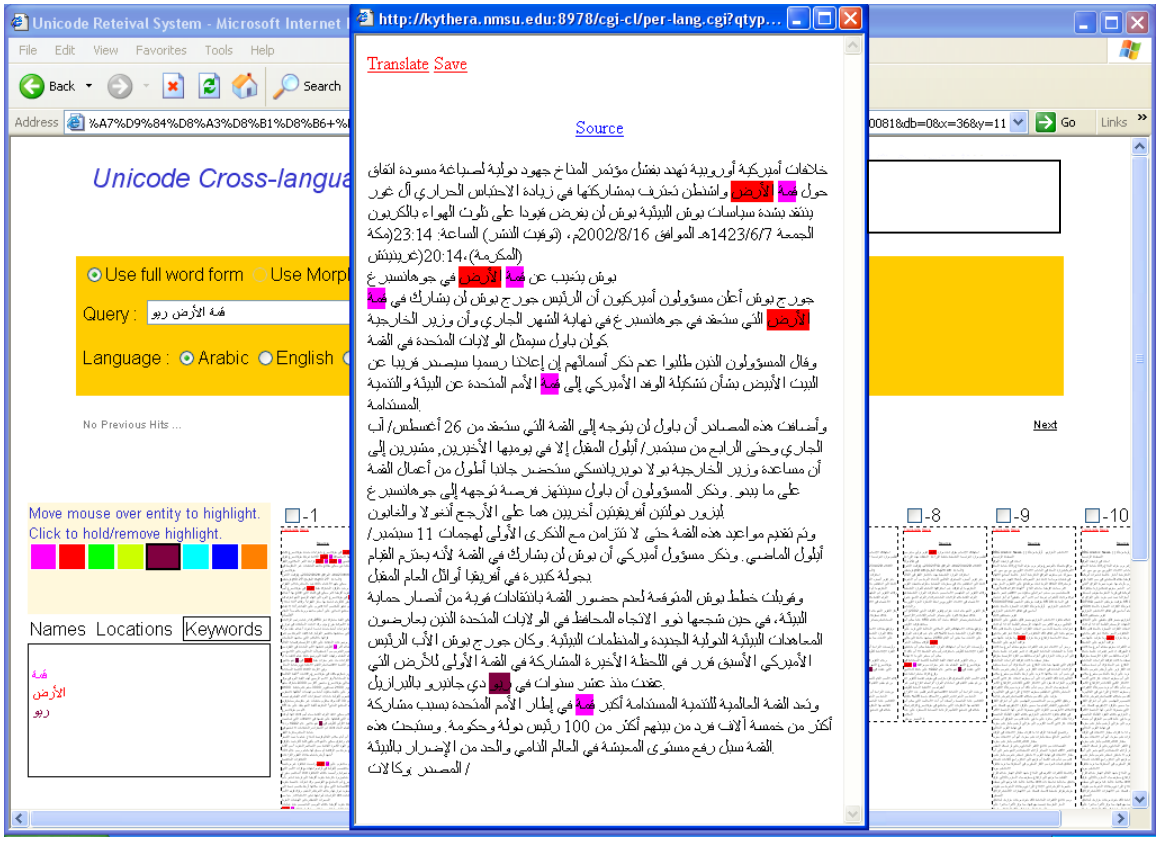

Figure 2. Overview of a retrieved document with query terms markup

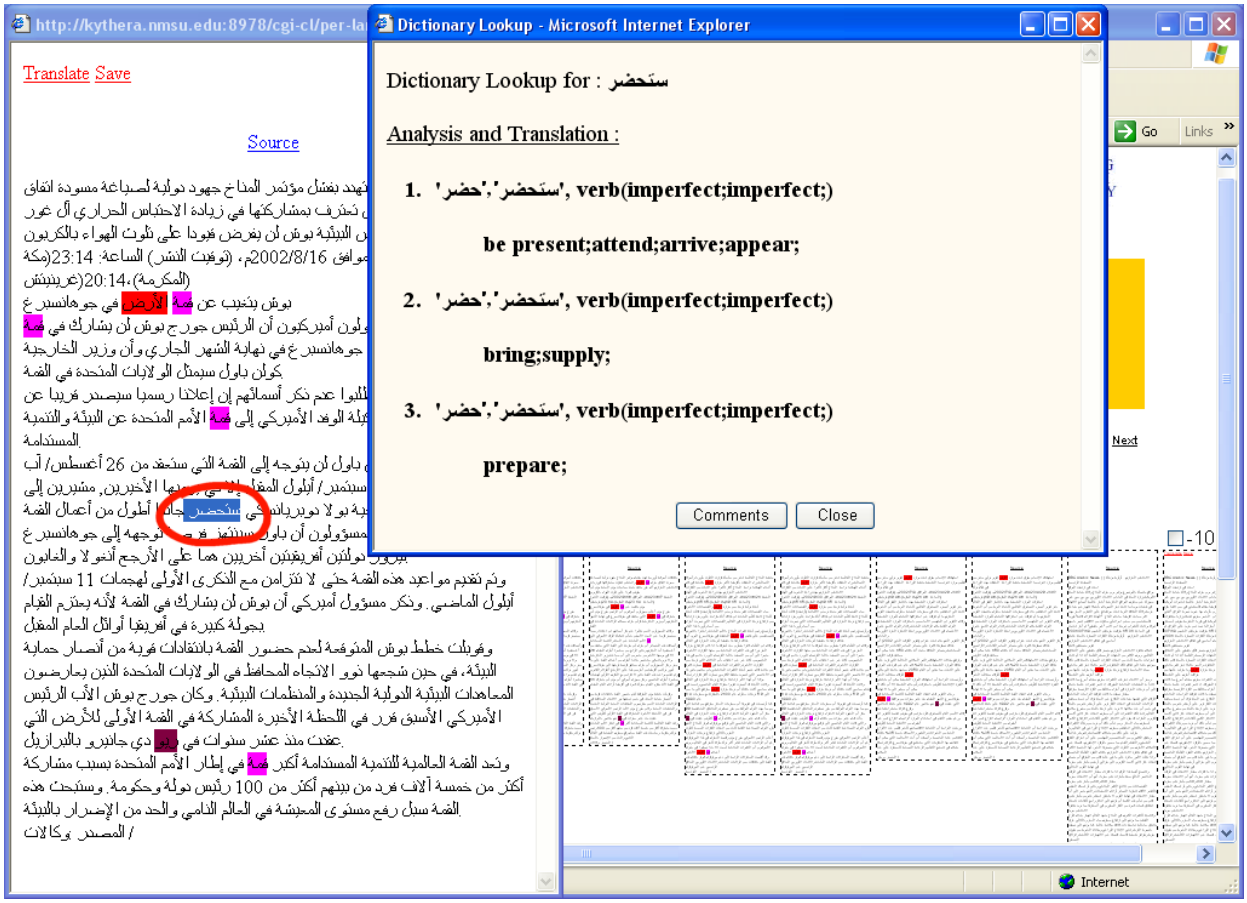

Figure 3. Single word English translation

The documents retrieved in the different languages may be translated into English. This is a useful tool for checking on and evaluating the success of the cross-language retrieval process. To carry out the translation there are two alternative techniques: word-level lookup and document-level translation. Word-level lookup allows the user to see the alternative translations of any single word in the document by clicking on it (see Figure 3). The translation provided includes grammatical information about the word. Document level translation allows the user to translate the full document from its original language into English. The quality and type of translation depends entirely on the translation engines available to handle the task (see Figure 4). In many cases it may only provide sufficient information to decide if further human translation is warranted.

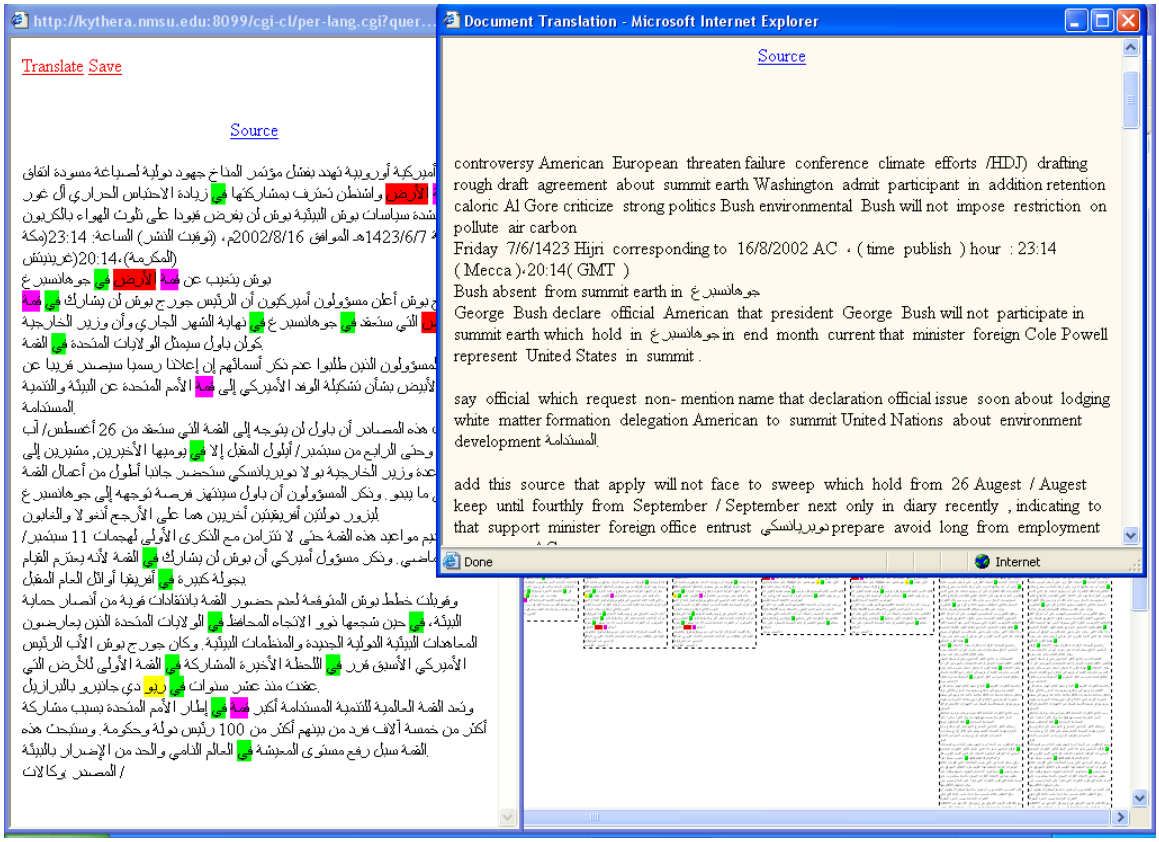

Figure 4. English Translation of the whole document

Proper Names

One of the most useful filters for scanning a set of documents for novel information is the ability to highlight the occurrence of proper names in the thumbnail representations of the document (see figure 5 below). The quality and understandibility of translations is also improved significantly by good proper name recognition and translation. In many cases this process involves transliteration of the proper name. This is complicated significantly by the fact that the original source language of most names is not English and the languages of the authors of the documents is also not English. Take for example the Agence France Presse Arabic Newswire, this provided the document collection used in TREC for Arabic information retrieval evaluation. This news feed is produced by thirteen translators working in Cyprus. Their original

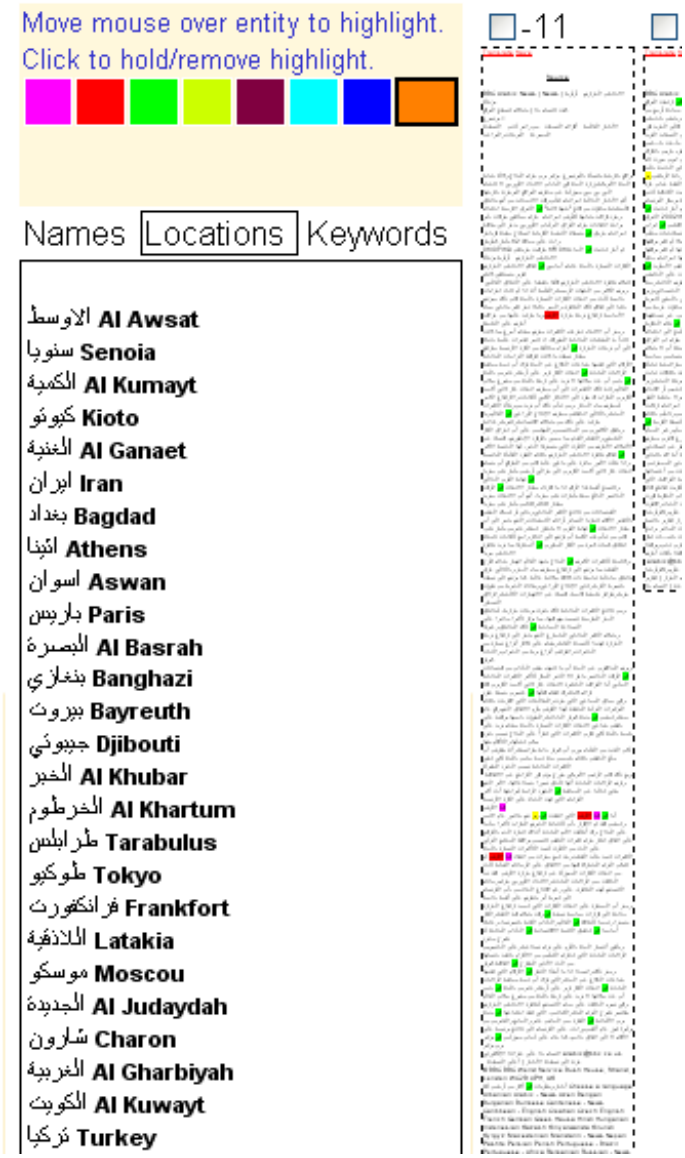

Figure 5. Proper Name Detection

sources are news stories from French and English newswires. Translation/transliteration of proper names is being carried out every day. The assumption is that each translator can handle this task in a sensible and consistent manner. Also variations in the forms of a name do not have a significant impact on a human reader. Unfortunately, however, variations in form in an information retrieval system mean that documents will not be found and names will be incorrectly mapped.

The variations in name formation are not unique to AFP ; documents in Arabic form BBC, CNN, Al-Jazeera and all Arabic online news sources also show variations in the forms used for proper names. Given the importance to our system of proper names we carried out the work described below which attempts to classify the sources of difference and suggests a method for dealing with some of these problems.

Translation of Names

Different transliteration forms for a name

In many cases the names (and also words) imported to Arabic appear in different forms in one source. The following table gives example from the AFP newswire.

Name AFP	**English**	**Occurrences**
لوس انجليس	Los Angeles	21
لوس انجلوس	Los Angeles	23
لوس انجيلس	Los Angeles	2
لوس انجيليس	Los Angeles	34
انجلترا	England	2
انكلتر	England	1
انكلترا	England	1
كارولاينا	Carolina	26
كارولينا	Carolina	14
ويسكونسين	Wisconsin	8
ويسكنسن	Wisconsin	2
ويسكونسن	Wisconsin	16
نيوهامبشير	New Hampshire	15

نيو هامبشر	New Hampshire	9

Table 1. Different spelling for names in Arabic in AFP

The same problems arise when Arabic names are transcribed into English. The academic discussion on rules for translation and Romanization of Arabic words is still ongoing. Searches conducted on three major search engines for names currently in the news illustrate the problem.

Name	**Google**	**Lycos**	**Altavista**
Najib Mahfuz	712	327	432
Nagib Mahfuz	419	345	36
Naguib Mahfuz	920	1114	549
Gamel Abdel Nassir	7	6	3
Gamal Abdel Nasser	10600	23405	4901
Jamal Abdel Nasser	277	192	129
Yasser Arafat	225000	510471	331097
Yasir Arafat	29500	81893	45718

Table 2. Different spelling of Arabic proper names in English

Transliteration from languages other than English

In many cases these names are understandable, but the author of the document has based his rendering of the name on either the original source language form of the name, or his assumption of the phonetics of a source language other than English for English place names. The examples below illustrate this problem.

Word	Wrong Spelling	Correct Spelling
شارلوت	Sharlote	Charlotte
المانيا	Alemania	Germany
اوروبا	Europa	Europe
موسكو	Moscou	Moscow
طرابلس	Tarabulus	Tripoli
الكويت	Al Kuwayt	Kuwait
باولوس	Paulos	Pauls
بروكسل	Brussells	Brussels
برلين	Berlien	Berlin
فلسطين	Palestina	Palestine
بريطانيا	Britania	Britain
بيروت	Bayreuth	Beirut
لاغوس	Lagus	Lagos

Table 3. Spelling based on non-English sources

Different news sources report things differently, based on their sources of information and also on their assumptions about the language of the original location of an event. For example it is common that in French that some characters are not pronounced and for the same word form in English those characters are pronounced.

Arabic AFP	English	Al-jezeera	Other
الغولدن غلوب	Golden globe	جولدن غلوب	القولدن غلوب
اف ب	AFP	إيه إف بي	
رويتر	Reuters	رويترز	

Table 4. Inconsistency between sources

Solutions to the Problem

To help resolve the issue we proposed a tool that will have features, unifying the different translations or transliterations, and finding the closest translations in a target text. The issue is not new, and research has already been carried out in the same context (Siegfried and Bernstein, 1991) (Kwok and Deng, 2003). (Arababi et al., 1994) developed a tool to translate Arabic names to English. (Stalls and Knight, 1998) presented a solution to translate names and technical terms in Arabic to English in order to find the original English words. We approach the problem from another angle to make the task bi-directional from Arabic to English and English to Arabic. The algorithm we describe here not only supports machine translation, information retrieval and information extraction but could also be used for newspaper editing and proofing.

Description of the Algorithm

The process of translation follows two ways.

- Simple Transliteration
- Transliteration with omission and expansion.

The Algorithm:

1- A word - *W* - is presented

2- Pre-Processing consist of normalizing *W* by removing silent consonants

3- Omission

Omit the vowels in *W* → *Wo*

4- Transliteration

Transliterate both *W* and *Wo*

W → *Wt*

Wo → *Wot*

a. Initialize the transliteration table trans[1] = “”;

b. Initialize the transliterations number Ntrans = 1;

c. Load Transliteration Map

d. For every character c in *W'* (*Wot* or *Wt*)

e. Ntrans = Ntrans*Number of mapping(s)

f. Duplicate the translation table

g. For every m_i in mapping(s) c→m_1, m_2, …

Replace the character c with m_i

5- Validation

Check each transliteration for and occurrence in the proper name lexicons

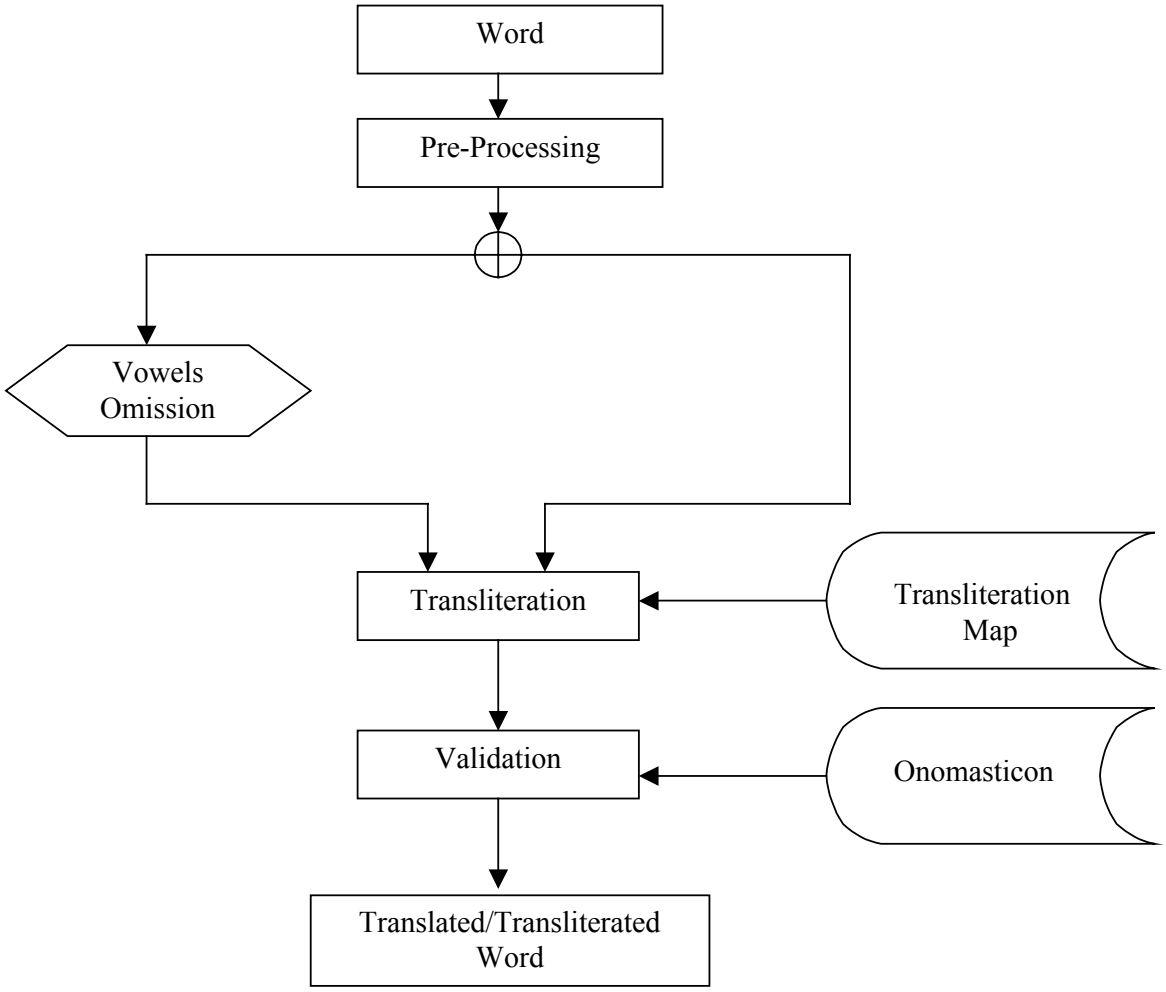

Figure 1. Process flow chart

The core of the process of transliteration is based on the quality of the proper name lexicons and the transliteration maps. The mapping generates all the possible transliterations and using the proper name lexicons we select only the valid names. The current proper name lexicon is a collection of about 180 000 proper name and place names. The size of the proper name lexicon affects the overall performance of the system.

We approached the problem using omission as the strategy to match and find the closest representation for the transliterated word. In modern Arabic the short vowels are omitted, so a

word will have at least three possible pronunciations for every vowel position. The case is very similar in English, the inconsistency in pronunciation of vowels or even some consonant groups has a significant impact on the transliteration process. The spelling of a name in English or Arabic shows vowel change from one speaker to another as the case of "أحمد" transliterated to "Ahmed" and "Ahmad" or "Carolina" transliterated to "كارولاينا" and "كارولينـــــا"

We generate the new compacted form of the word using the omission approach by omitting all the vowels. Exceptions are made for vowels at the start of a word. The next step is expansion, this step is intended to reconstruct all the possible combinations that the word might use. The generated words are then checked against the proper name lexicons to check if the word is a valid combination of characters.

We tuned the transliteration maps to generate the all possible mappings. Some of these mapping may not seem obvious but are often the result of foreign names being used unchanged in English. For example on "Johan" "Johannes" "Johannesburg" the sound of "j" changes although this is a consonant, the same is true for "g" in "Angeles" and "England" and more.

In the following section we illustrate the process by examples from English to Arabic and Arabic to English.

Examples

English to Arabic

Word = 'los angeles'

subWord = 'ls'

subWord = 'angls'

Transliteration LOS	Omission: LS
لؤس	لس
لؤص	لص
لس	
لص	
لعس	
لعص	
7	2

Transliteration ANGELES	Omission: ANGLS
أنإلإس	أنجلس
أنإلإص	أنجلص
أنالاس	أنغلس
أنالاص	أنغلص
أنجإلإس	أنلس
أنجإلإص	أنلص
أنجالاس	انجلس
أنجالاص	انجلص
أنغإلإس	انغلس
أنغإلإص	انغلص
أنغالاس	انلس
أنغالاص	انلص
انإلإس	عنجلس
…	…
37	18

Word	Generated	Omission	Valid Expansion	Correct	Common
Los Angeles	259	36	620	2	1

Table 5. Example of processing English

Arabic to English

Word = لوس انجلس

subWord = لس

subWord = انجلس

Transliteration لوس	Omission: لس
Lws	Ls
1	1

Transliteration انجلس	Omission: انجلس
angls	Angls
Anjls	Anjls
2	2

Word	Generated	Omission	Valid Expansion	Correct	Common
لوس انجلس	2	2	300	1	1

Table 6. Example of processing Arabic

Evaluation

To evaluate the performance of the system we prepared a list of 100 names including personal names, and places name from both native Arabic and non-Arabic.

The results were comparable with those reported by other sources. The results in Tables 7 and 8 show the performance when all names were transliterated.

Words	Average Correct	Common
All	0.48	0.47
Filtered List	0.62	0.62

Table 7. General performance of the tool for English to Arabic

Number of Words	Average Correct	Common
All	0.30	0.30
Filtered List	0.50	0.50

Table 8. General performance of the tool for Arabic to English

An analysis of the failures was conducted to check the nature of the problems that the system couldn't handle properly. One major category, historical place names, have a translated form rather than a transliterated one. If these names are eliminated before the transliteration step the performance of the algorithm increases to 62% and 50% correct (Table 7,8 Filtered List). Thus we do not convert "Palestine" and "Brussels" which transliterate respectively to "فلســـــطين" and "بروكســـل".

The rest of the problems can be categorized in two major areas:

a- Non standard spelling:

As the case for "Weimar" the transliteration of the word usually is done through the pronunciation of the word from the original language "W" as "V" rather than the regular sound of "W" as in "Washington" or "New Zealand". If the word will be transliterated will end up with "ويمـار" and "ويمـر" but the correct one sound transliterated is "فيمـــر". The same case for "Johannes" which transliterated to "انـسيـوه".

b- Omission of Consonants

Some cases of tansliteration omit consonants, which make the process hard to guess the correct spelling of the word as in "رويـــتر" which transliterate to "Rueters" or "نيوهامشـــير" to "New Hampshire".

Conclusions and Further Work

Allowing multiple possible transliterations and validating the results against proper name lexicons is a useful technique both for retrieval and for filtering/translation. We intend to work further on improving the accuracy of the method.

Our current transliteration table is given below in Appendix A. We intend to further improve this and to explore ways of identifying the sources of foreign names both in Arabic and English, which may give clues to using a subset of the phonetic transliteration, based on the original source language. Experiments carried out previously have shown good levels of success in identifying Spanish proper names in English documents using a trained probabilistic finite state model. We intend to develop source language identifiers for names, which have been mapped into Arabic and English.

We also intend to introduce the method developed Hanks and his team (Hanks 2003) in the development of the Dictionary of American Family Names. Here lists of *guaranteed* forenames were used to predict the original nationality of a surname (e.g. Jacques [French], Hamish [Scottish]) and once this is established either for the surname or forename then the most appropriate phonetic mappings can be chosen.

References

Arbabi, M. Fischthal, S. M. Cheng, V. C. and Bart, E. 1994. Algorithms for Arabic name transliteration. *IBM Journal of Research and Developmen*t, 38(2):183–193.

Hanks (ed.) *Dictionary of American Family Names,* Oxford University Press, 2003

Stalls, B. and Knight, K. 1998. Translating Names and Technical Terms in Arabic Text. *COLING/ACL Workshop on Computational Approaches to Semitic Languages*. Montreal, Quebéc.

Siegfried S. L. and Bernstein J. 1991. Synoname: The Getty's New Approach to Pattern Matching for Personal Names. *Computers and the Humanities*, 25(4): 211-226,

Fraser, A. Xu, J. Weischedel, R. 2002. TREC 2002 Cross-lingual Retrieval at BBN, NIST *Special Publication 500-251: The Eleventh Text REtrieval Conference* (TREC 2002).

Tur, G. Hakkani-Tür, D. Oflazer, K. 2000. Name Tagging Using Lexical, Contextual, and Morphological Information. In *Proceedings of the Workshop on Information Extraction meets Corpus Linguistics*, at LREC-2000, 2nd International Conference on Language Resources and Evaluation, Athens, Greece, 31 May - 2 June, 2000

Kwok K. L. and Deng Q. 2003. GeoName: a system for back-transliterating pinyin place names HTL-NAACL 2003 Workshop: Analysis of Geographic References, pp. 26-30 Edmonton, May-June 2003

Appendix A: English to Arabic and Arabic to English transliteration maps

English	Arabic
oo	ع
ph	ف
sh	ش
ch	ش
gh	غ
tt	ت
a	ا,أ,ع
b	ب
c	ك,س
d	د
e	إ
f	ف
g	غ,ج,
h	ه,ح
i	إ,ي
j	ج
k	ك
l	ل
m	م
n	ن
o	ؤ,ع
p	ب
q	ق,ك
r	ر
s	ص,س
t	ت,ط
u	ؤ,ع
v	ف
w	و
x	اكس,إكس
y	ي
z	ز

Arabic	English
ء	a
آ	a
أ	a
ؤ	o,u
إ	i,e
ئ	i,e
ا	a
ب	b
ة	t
ت	t
ث	th
ج	j,g
ح	h
خ	kh
د	d
ذ	dh
ر	r
ز	z
س	s
ش	sh,ch
ص	s,c
ض	dh
ط	t
ظ	dh
ع	a
غ	gh
ف	f
ق	k,q
ك	k
ل	l
م	m
ن	n
ه	h
و	w
ى	a
ي	y