English/Arabic Cross Language Information Retrieval (CLIR) for Arabic OCR-Degraded Text

Tarek A. Elghazaly, Faculty of Computers & Information, Cairo University, Giza, Egypt, t.elghazaly@fci-cu.edu.eg
Aly A. Fahmy, Faculty of Computers & Information, Cairo University, Giza, Egypt, a.fahmy@fci-cu.edu.eg

Abstract

In this paper, a novel for Query Translation and Expansion for enabling English/Arabic CLIR for both normal and OCR-Degraded Arabic Text model has been proposed, implemented, and tested. First, an English/Arabic Word Collocations Dictionary has been established plus reproducing three English/Arabic Single Words Dictionaries. Second, a modern Arabic Corpus has been built. Third, a model for simulating the Arabic OCR errors has been proposed. Forth, a comprehensive model for Query Translation and expansion is proposed. The model translates the Query from English to Arabic detecting and translating collocations, translating single words and transliterating names. It solves the replacement ambiguity then it expands the Arabic Query to handle the expected Arabic OCR errors. The proposed model gives high accuracy in translating the Queries from English to Arabic solving the translation and transliteration ambiguities and with orthographic query expansion, it gave high degree of accuracy in handling OCR errors.

Keywords: Cross Language Information Retrieval, CLIR, Arabic OCR-Degraded Text, Arabic Corpus.

1. Introduction

The importance of CLIR appears clearly when we consider a case like the Library of Congress [1] which has more than 134 million items and approximately half of the library's book and serial collections are in 460 languages other than English. When people like to retrieve the whole set of documents that represent some interest, they have to repeat search process in each language. Furthermore, as a big number of books and documents are available only in print especially the Arabic ones, they are not ‘full text’ searchable and they need applying the Arabic OCR process whose accuracy is far from perfect [47]. The goal of this paper is to enable users to query in English language against an Arabic OCR-Degraded Text.

The outline of this paper is as follows: The previous work is reviewed in Section 2. The proposed work is presented in the next sections. Arabic words formalization, normalization and stemming are presented in Section 3. Corpus and Dictionaries are presented in Section 4 and 5. In Section 6 & 7 the work done for CLIR through Query Translation and expansion respectively is detailed, followed by the experimental results and the conclusions in Sections 8 & 9.

2. Previous Work

2.1. Arabic Morphological Analysis for Information Retrieval (IR)

Several researches have been done to check the effect of light stemming & root based stemming on IR like in [10] [11] and [12]. Al-Jilayl and Frieder concluded in [48] that light stemmer performs than root based stemmer (using enhanced version of Khoja root based stemmer [49]). The effect of either stemming techniques on Information Retrieval was better than no stemming at all. The same result has also been concluded by Larkey et al. in [50].

2.2. Arabic Corpus

As per Hunston in [28], thhe construction and use of text corpora is continuing to increase [28]. Several research efforts has been done in this field like Kharashi & Evens in [10], Hmeidi et al in [29], Goweder A and De Roeck A. in [30] and Darwish et al. in [45].

2.3. CLIR

In CLIR, either documents or queries are translated. There are three main approaches to CLIR: Machine Translation (MT), Comparable or Parallel Corpus, and Machine Readable Dictionaries.

MT systems seek to translate queries from one human language to another by using context. Disambiguation in MT systems is based on syntactic analysis. Usually, user queries are a sequence of words without proper syntactic structure [14]. Therefore, the performance of current MT systems in general language translations make MT less than satisfactory for CLIR [15].

In corpus-based methods, queries are translated on the basis of the terms that are extracted from parallel or comparable document collections. Dunning and Davis used a Spanish-English parallel corpus and evolutionary programming for query translation [16]. Landauer and Littman [17] introduced a method called Cross Language- Latent Semantic Indexing (CL-LSI), and requires a parallel corpus. Unlike parallel collection, comparable collections are aligned based on a similar theme [18].

Dictionary-based methods perform query translation by looking up terms on a bilingual dictionary and building a target language query by adding some or all of the translations. This technique can be considered the most practical method [19]. Ballesteros and Croft [20] developed several methods using MRDs for Spanish-English CLIR and then improved the effectiveness by many ways including resolving the ambiguity [21],[22],[23]. Pirkola [14] studied the effects of the query structure and setups in the dictionary-based method. Mohammed Aljlay and Ophir Frieder investigated for the Arabic-English CLIR [24] (The opposite direction of this paper). They investigated MT and MRD to Arabic-English CLIR using queries from TREC [25]. They concluded that Query Translation for Arabic/English CLIR through Machine-readable dictionaries is cost effective as compared to the other methods such as parallel corpus, Latent Semantic Indexing (LSI), and MT. Ahmad Hasnah and Martha Evens concluded also in [26] that most comprehensive work is to work with the bilingual MRD with solving the problems of terms translation ambiguity.

2.4. CLIR for Arabic OCR-Degraded Text

For handling OCR-Degraded text in CLIR, Darwish K. investigated in [51] the different methods for query term replacement and he found that Word Term Frequency/Document Frequency (WTF/DF) was the best evaluated approach of the evaluated ones. He proved an approach of producing possible replacements for query terms that could have been generated by OCR proved to be a useful technique for improving retrieval of OCR-degraded text.

2.5. Comments and Limitations of the Previous Work

As mentioned in [25] and [26], the most cost effective and practical method for CLIR is using MRDs. Darwish K. work in [51] tried in this direction especially in the English/Arabic CLIR supporting also OCR-Degraded Text. But however, it suffered from some limitations. From the Query Translation perspective, it did not provide a solution for the named entities, expressions, and the word collocations in general. For the OCR-Degraded Text handling, it concentrated on the correction of character n-grams (up to 7-gram) but it does not take into consideration neither the higher n-grams nor the position of this character n-gram inside the words. In this paper, we try to find a solid solution for the English/Arabic CLIR for both the normal and the OCR-Degraded Arabic Text overcoming the mentioned limitations.

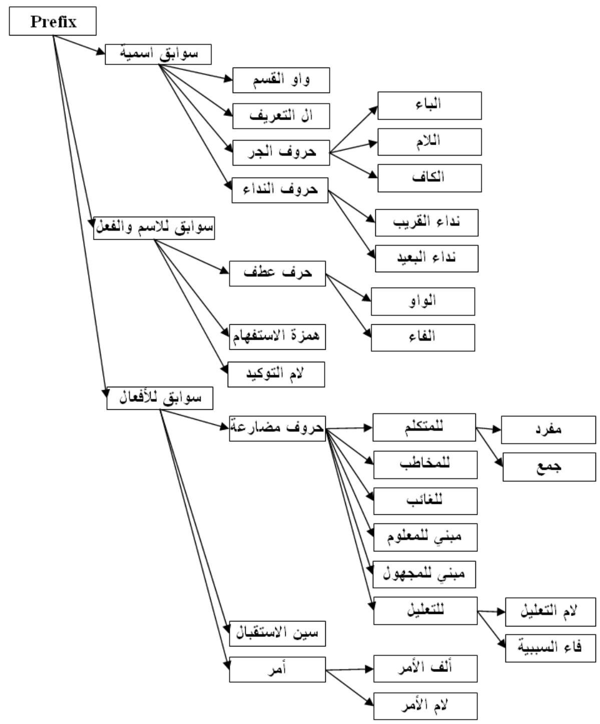

Fig 1: Prefixes in Arabic Language

3. The proposed Light Stemmer for Arabic Information Retrieval

As per the previous work mentioned in section 2, light stemming can be considered as the most effective approach for improving Arabic IR rather than aggressive stemming or the root extraction. In this paper, we propose a light stemmer that normalize then lightly stem Arabic words.

The proposed stemmer works on three steps. First, it normalizes the Arabic word characters that are written differently due to the different writing ways or due to the common writing mistakes. This is to unify ('ي', 'ى') to 'ي', ('أ','إ','ء','آ','ا') to 'ا', and ('ـة'&'ـه') to 'ـه'. The 2nd step is to produce the stems as prefix stripped, suffix stripped, and both prefix and suffix stripped with always considering the longest combinations to form prefixes and same for suffixes. The 3rd step is to discard the produced stems that are not available in the Arabic dictionary. The available affixes are as shown in Fig 1 and Fig 2 [27].

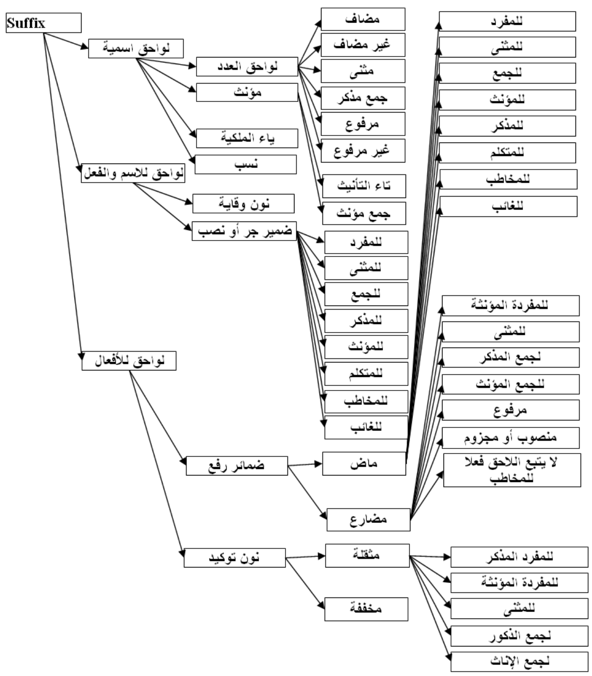

Fig 2: Suffixes in Arabic Language

4. The proposed Modern Arabic Corpus

The proposed Corpus ia a Modern Arabic Corpus that would help in the study of Modern Standard Arabic in general and to use its statistics to solve the Query Translation replacement ambiguity. Moheet portal [31] has been chosen as the main data source as it gets its news articles from 200 sources from different countries, perspectives, and fields. It takes also Arabic articles through translation into Arabic from non-Arabic sources.

Table 1: Statistics of the established Modern Arabic Corpus

Category	Figures
Corpus Textual Size	6.8 GB
Number of Textual Documents	98,000
Total No. of extracted Arabic words	46,603,112
No. of unique words (exact)	338,335
No of unique words (stemmed)	155,561

Moheet Portal has been crawled for 336 continuous hours. Articles are parsed to extract the plain Arabic words excluding all Latin characters, punctuations, or diacritics. Then the result Text database was analyzed to get the unique exact words and the unique exact stems (after normalization and stemming). The overall figures of the Corpus are illustrated in Table 1.

5. Establishing a Word Collocations Dictionary and Evaluating Single Words Dictionaries

5.1. Establishing the proposed Word Collocation Dictionary

To establish this dictionary, we considered WordNet as the source of English Word Collocations as it is considered as most important resource available to researchers in computational linguistics [37]. Table2 describes some statistics about the WordNet DB.

It has been parsed to extract the collocations (expressions, proper noun, and named entities, with multi-words). The collocations have been translated then reviewed manually [38]. The result Word Collocations Dictionary has the figures in Table 3.

Table 2: WordNet Statistics

POS	Unique Strings	Synsets	POS
Noun	117097	81426	145104
Verb	11488	13650	24890
Adjective	22141	18877	31302
Adverb	4601	3644	5720
Total	155327	117597	207016

Table 3: Detailed figures for the Collocations Dictionary

	2-gram	3-gram	4-gram	5-gram	6-gram	7-gram	8-gram	Total
Noun	51,205	7030	1246	268	73	25	29	59,876
Verb	2,404	297	87	12	5	1	0	2,806
Adv.	437	294	85	12	1	0	0	829
Adj.	426	145	0	2	0	2	0	575
Total	54,472	7,766	1,418	294	79	28	29	64,086

5.2. Single Words Dictionaries

5.2.1. Dictionary1

The main goal of producing this dictionary is to provide a modern dictionary based on a data that are originally from an English/Arabic source and is slimmed to cover the practical Arabic meanings to the English words. The raw data for this dictionary is an English/Arabic dictionary data as one of the outputs of the Arabeyes project [32], [33].

The output Dictionary DB has 87,423 English words and every English word has from one to two Arabic translations.

5.2.2. Dictionary2

The main goal of producing this dictionary is to provide a dictionary based on a data that are originally extracted from an Arabic/English source. The raw data for this dictionary is an Arabic/English Dictionary from Computing Research Laboratory (CRL), New Mexico State University [39].

The output DB has unique 30,389 English words and every English word has from 1 to 248 Arabic translations. The average number of Arabic translations for every English word is 5.

5.2.3. Dictionary3

The main goal of producing this dictionary is to provide a big one based on a data that are originally extracted from an English/Arabic source and is as big as it covers almost the whole set of available English words. The main English words are available from CRL dictionary and WordNet [35]. Every word has been translated through a free industry-known online dictionary [36] and all translations have been collected.

The output DB has around 52,000 English words and Every English word has from one to 205 English translations. Each Arabic word has about 8 corresponding English translations in average.

6. The proposed model for CLIR through Query Translation

In this section we will introduce our proposed model for CLIR through Query Translation. This includes the models for Names Transliteration, Single Words Translation, collocations Translation, solving the ambiguities, and Query Translation.

6.1. Transliteration Model

The model's main idea is to check the longest n-gram character section in the start of the word to be translated directly from the n-gram transliteration table, doing the same for the end, of the words, then the medium sections of the word. As there are often many transliteration probabilities for the same section, all these probabilities are taken into consideration due to the frequency of the corpus and the probability of that section with respect to the transliteration table. Table4 illustrates the transliteration table used [39].

Table 4: Transliteration Table

N-gram	T	P	T	P	T	P	T	P
A	ا	0.6	0	0.2	ي	0.1	ع	0.1
a (End)	ه	0.6	ا	0.4				
ai	ي	0.5	أي	0.5				
alk	وك	0.9	الك	0.1				
au	ا	0.4	او	0.4	و	0.2		
B	ب	1						
bb	ب	1						
a (Start)	ا	0.9	ع	0.1				
au (Start)	ا	0.8	او	0.2				
e (Start)	ا	0.4	0	0.1	ي	0.3	يا	0.3
i (Start)	ا	0.7	أي	0.2	ء	0.1		
mc (Start)	ماك	0.9	مك	0.1				
o (Start)	ا	0.3	او	0.7				
u (Start)	ا	0.8	او	0.2				
wr (Start)	ر	1						
C	ك	0.9	تش	0.1				
cc	ك	1						
ce	س	0.8	سي	0.2				
ch	تش	0.8	ش	0.2				
ci	س	0.2	سي	0.8				
cy	س	0.2	سي	0.8				
ck	ك	1						
D	د	1	0	0				
dd	د	1						
E	0	0.6	ا	0.1	ي	0.3		
ea	ي	0.9	يا	0.1				
e (End)	0	0.9	ـه	0.1				
ee	ي	1						
ey	أي	0.8	ي	0.2				
F	ف	1						
ff	ف	1						
ph	ف	1						
G	غ	0.5	ج	0.4	ق	0.1		
ge	ج	0.8	غ	0.2				

N-gram	T	P	T	P	T	P
ie	ي	0.7	أي	0.3		
j	ج	0.9	ي	0.1		
k	ك	1				
kk	ك	1				
kh	خ	1				
l	ل	1				
ll	ل	0.8	لل	0.1	ي	0.1
m	م	1				
mm	م	1				
n	ن	1				
nn	ن	1				
o	و	0.7	ا	0.1	0	0.2
ois	وا	0.8	وس	0.1	بوس	0.1
oo	و	1				
ou	و	0.6	او	0.4		
ough	او	0.4	وف	0.2	و	0.4
ough (End)	ه	0.8	وف	0.2		
p	ب	1				
pp	ب	1				
q	ك	0.5	ق	0.5		
qu	كو	0.6	ك	0.3	ق	0.1
r	ر	1				
rr	ر	1				
s	س	0.6	ز	0.2	ص	0.2
sch	ش	0.8	سش	0.2		
s (End)	س	0.6	ز	0.4		
sh	ش	1				
ss	س	0.8	ص	0.2		
t	ت	0.7	ط	0.3		
th	ت	0.3	ث	0.4	ذ	0.3
tio	ش	0.8	شيو	0.2		
tt	ت	0.9	ط	0.1		
u	و	0.8	0	0.2		
ue (End)	0	0.8	و	0.2		

6.2. Single Words Translation Model

The main idea of the proposed model is to get the input as phrase which is not a collocation or a multi-words expression, tokenize that phrase, remove stop words, and get the Arabic equivalent for each word. If the English word does not have an Arabic equivalent word, then the word will be transliterated through the transliteration mode.

6.3. The Model for checking Collocations parts

The main idea of this model is to check if the current part of the search sentence is a part of collocation. Continuous checking for that purpose will lead to get the longest collocation in the search sentence. For example both "United Nations" and "United Nations Children's Fund" are collocations. This continuous checking will succeed in finding the correct collocation which is the second one (the longest). The main benefit of considering the longest collocation is getting the most accurate translation as described in the next section in details.

The model checks the entered query words in the Word Collocations Dictionary either exact or stemmed (through the WordNet rues) taking into consideration that only base forms of words even those comprising collocations, are stored in WordNet [40].

6.4. Solving Translation / Transliteration Ambiguity

Every single English word –that are available in dictionaries- has one or more Arabic Equivalents up to 248 ones (as in Dictionary2). Also, a word that is not available in dictionaries and has to be transliterated has many probabilities as every character has one or many probability. A word Like Lincoln will have 18 Arabic transliterations. As the query may have many words, the ambiguity will be very high. In this section we propose several methodologies for solving the ambiguity of translation and transliteration trough collocations dictionary, using corpus, and using transliteration probabilities.

6.4.1. Word Collocations Dictionary

If the query has the phrase "United Nations Children's Fund", the direct translation will be for every words respectively (20, 6, 14, 19). This means that only this English phrase would have 20*6*14*19=31,920 Arabic translations which is totally unpractical especially that the mentioned English phrase has only one Arabic translation which is " صندوق الأمم المتحدة لرعاية الطفولة ". Using the proposed collocation dictionary solves this problem and gives the correct translation accurately and directly.

If the word is not a part of a word collocation, the next two methods (transliteration probabilities and Corpus) are used.

6.4.2. N-gram Transliteration probabilities

This method used in case that the word is not a part of a collocation and is not available in the dictionary. IT proposes Arabic word which is the result of concatenating the Arabic character(s) which have the highest transliteration probability to each English character(s), with respect to the transliteration table made by Nasreen AbdulJaleel and Leah S. Larkey after their statistical study [39].

6.4.3. Corpus

This handles both cases for translation or even transliteration.It is working by always sorting the transliterations/translation of every word in the query descending according to their frequencies in the corpus. The resulting Arabic query will have the most used Arabic translation/transliteration for every English word.

6.5. The Segmentation & Query Translation Model

The proposed English to Arabic query translation model works with all the proposed models to produce an accurate Query Translation. . Fig 3 describes the model in the "Query Translation" part.

7. Improving Arabic OCR-Degraded Text Retrieval

In this section introduces the final step of the proposed model which is handling the Arabic OCR errors. It starts with defining the OCR accuracy, presenting the model for simulating Arabic OCR errors, and establishing the real training and test sets.

7.1. Defining the OCR accuracy

Scientists and OCR commercial providers usually consider the OCR accuracy from character point of view as the below definition considering the error as character insertion, substation, or deletion:.

$$\text{Character Accuracy} = \frac{\text{n - Number of Errors}}{\text{n}}$$

Where n is the total number of characters in the correct text ("groundtruth") [52].

However, this definition is considered sometimes misleading from many of the OCR commercial consumers who prefer to count the OCR accuracy from word point of view. Considering a sample page of 200 words that contain 100 character in total, assuming the OCR output of this page having only 10 character errors each in a separate word, this means character accuracy of 98% where it means word accuracy of 90%. This difference is one of the reasons of considering complete words in modeling the OCR Errors in this paper.

7.2. Modeling the OCR errors

Generally the current models for simulating OCR-Errors are mainly depending on a 1-gram and sometimes n-gram character replacement algorithm. However, in Arabic, as character shape defers up to its position in the words (begin, middle, end, Isolated). So, it is too difficult to include all these variables (7-gram character for example) plus the character position in one model.

The proposed model is a word based noisy channel model. It will be trained and tested on the complete words from both the training and test sets.

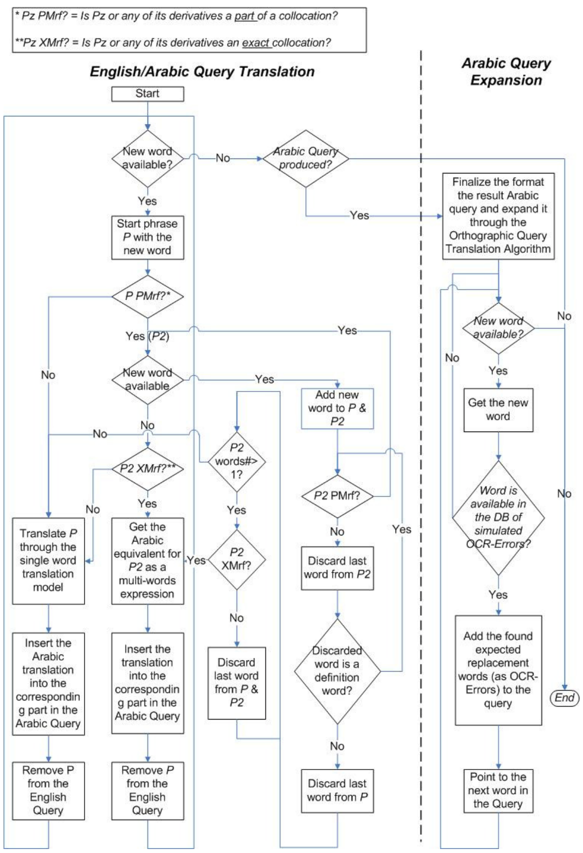

Fig 3: The Query Translation and Expansion Model

7.3. Improving IR for Arabic OCR-Degraded Text through Orthographic Query Expansion

The proposed Orthographic Query Expansion model attempts to find different mis-recognized versions of a query word in the text being searched. It starts by checking every word in the Arabic Query against the word DB resulted from training the model of simulating the Arabic OCR errors on the established Training Set. Then the query is expanded by every word found as a probable mistaken word provided by the OCR. Fig 3 in the "Arabic Query Expansion" part illustrate the model and Fig 4 illustrates an illustrative example.

Word	Individual words Translations/ Transliterations Probabilities	After Solving the ambiguity using Collocations part in the model	Default Translation After applying the proposed model completely	Expanded Arabic Query to include the expected OCR Errors
Lincoln	18	18	1	2
United	20	1		1
States	12			2
Civil	18			2
War	8			1
Stories	6	6		2
Total	3,732,480	108	1	16

Fig 4: Example of Query Translation and Orthographic Expansion

Orthographic Query expansion depends directly on the simulation of the OCR-Errors and so we can consider its accuracy as the accuracy of the OCR-Errors simulation model.

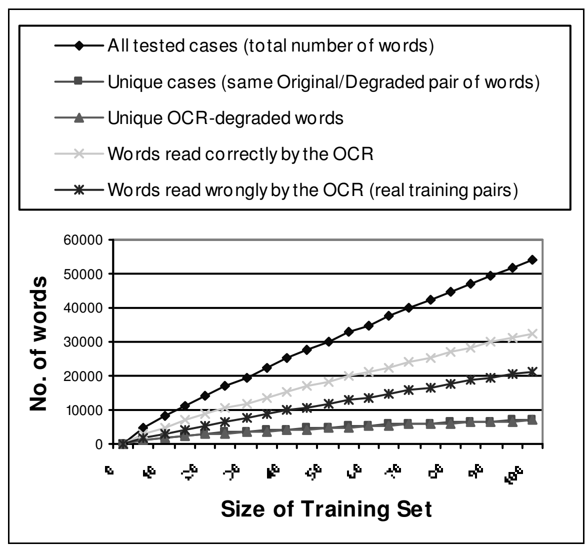

Fig 5: Training Set Statistics

7.4. Establishing the Training and Test Sets

For establishing the Training and Test Sets, a pool of 150 long documents from the corpus have been reformatted to have one word per line, converted to image based document (PDF) using "Adobe Acrobat Professional" [53], then the Arabic OCR process (through Sakhr OCR [41]) has been applied on them to have as a result 150 long documents (30 pages each) with their original text and the corresponding OCR-Degraded Text.

Table 5: Training Set Statistics

Trn. Set size (no. of docs)	No. of all tested cases (total number of words)	No. of unique cases (same Org. / Deg. pair of words)	No. of unique OCR-degraded words	Number of words read correctly by the OCR	No. of words read wrongly by the OCR (real training pairs)	Max no. of n-gram chars i.e. longest word (Deg. / Org)
5	4429	1367	1321	2720	1709	15/14
10	8003	1992	1927	4851	3152	15/14
15	11266	2557	2468	6907	4359	15/14
20	14043	2943	2836	8588	5455	15/14
25	16946	3296	3174	10376	6570	15/14
30	19485	3712	3577	11906	7579	15/14
35	22106	3960	3818	13514	8592	15/14
40	25344	4335	4181	15473	9871	15/14
45	27832	4546	4373	16970	10862	15/14
50	30114	4799	4617	18295	11819	15/14
55	32660	5169	4980	19795	12865	15/14
60	34932	5381	5186	21116	13816	15/14
65	37374	5592	5382	22538	14836	15/14
70	39975	5952	5728	24109	15866	15/14
75	42275	6158	5929	25511	16764	15/14
80	44621	6366	6131	26912	17709	16/14
85	47002	6572	6333	28378	18624	16/14
90	49298	6730	6481	29724	19574	16/14
95	51659	6901	6646	31151	20508	16/14
100	53910	7184	6921	32527	21383	16/14

7.5. Training Set Statistics

To be able to examine the accuracy of modeling the OCR errors against different sizes of Training Set, different sets of documents have been selected from the Training and Test Set Pool. These sets have the range from 5 to 100 relatively long documents (550-900). Statistics about the Training Set are illustrated in Table 5 and FIG 5.

7.6. Defining the Model Accuracy

The following definition has been considered to define the accuracy for modeling the OCR errors:

$$Accuracy_{TSSn} = \frac{AccurateRplacements_{TSSn}}{TotalOCR\_DegradedWords}$$

i.e. $Accuracy_{TSSn}$ (for certain Training Set Size) equals the no of accurate replacements (with respect to the training set size) divided by the total number of OCR-Degraded words.

In other words, if the mistaken OCR-degraded word is available in the training set with the correct original word, then this will be considered as accurate replacement. Otherwise, it will be considered as not accurate. The accuracy for a specified training set size is the number of accurate replacements that are available in this Training-Set, divided by the total number of the OCR-Degraded words.

8. Experiments

Experiments have been performed to test every option in the model (drawn in Fig 3). This is including the Query Translation accuracy with the different parameters and options for translation, transliteration, and solving ambiguities. Then, the Orthographic Query Expansion part has been tested against different training set sizes.

Table 6: Testing results of the Query Translation Model – Collocations Coverage and Accuracy

Coverage for the fed collocations	95%
Accuracy for the collocations translation	95%

Table 7: Testing results of the Query Translation Model– Transliteration Coverage and Solving Ambiguity

	Corpus	Transliteration Probabilities
The real accurate translation is one of the produced transliteration	90%	
1st hit is the best one of the produced translation	70%	63%
1st hit is the real best translation / translation	60%	54%

100 queries have been fed to the Query Translation Model with. Every query has from one to 9 English words including proper names and queries about different fields (political, history, shopping, events,

tourism, and miscellaneous). The experiments analyzed the effect of the proposed models on solving the translation ambiguity using the collocations dictionary, the different single words dictionaries, and the proposed corpus. The experiments also examined the transliteration model efficiency and solving the transliteration ambiguity through both the corpus statistics and the characters transliteration probabilities. Table 6, Table 7, and Table 8 summarize the results.

Table 8: Summary Testing results of the Query Translation Model – Single Words Dictionaries Coverage and Solving Ambiguity

	Dict. 1	Dict. 2	Dict. 3
The real accurate translation is one of the produced translation	94%	82%	94%
1st hit is the best one of the produced translation	98%	95%	80%
1st hit is the real best translation	92%	78%	75

For the Orthographic Query expansion part, which depends directly on the OCR-Errors Simulation model, 50 documents from the Training and Test Set pool have been selected as the Test set. There is no intersection between the Training and Test sets. TABLE 9 illustrates the statistics of the Test set. Fig 6 illustrates the model accuracy across different training set sizes.

Table 9: Test Set Statistics for the OCR-Errors Simulation Model

Category	**Statistical Number**
No of documents (long documents)	50
No of unique pairs (Original word-degraded word)	4208
Total no. of words	26,579
No of words read correctly	15, 823
No of words read wrong	10,756
No of words read wrong and (and not read as NULL)	10,175

9. Conclusions

The most important contribution was proposing the Query Translation and Expansion model which covers the collocation, transliteration and the normal single English words inside the Query and expands the Arabic query to handle the expected Arabic OCR Errors.

The collocation detection and translation model, supported by the well-introduced collocation dictionary, gives high accuracy in detecting and translating collocations even when they are written in non exact way (derivations). The only non-detected collocations are those which are local one like 'African Cup' i.e. as the collocation dictionary size increases, the collocation detection, translation, and so the overall query translation accuracy will also be enhanced.

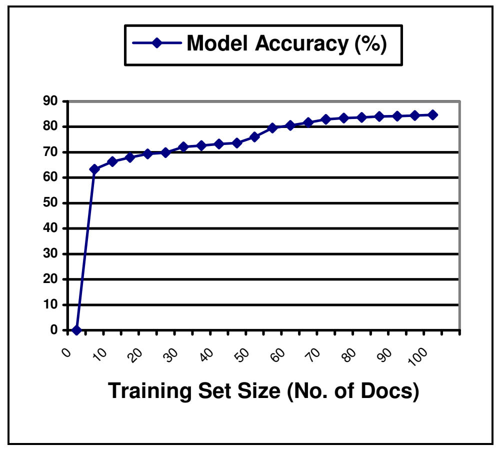

Fig 6: Accuracy of the OCR-Errors Simulation model

Solving the transliteration ambiguity is effective through either the corpus or the n-gram character transliteration probabilities. However, the corpus option gave better results.

The three single words dictionaries gave different results, yet compared with the other two dictionaries, Dictionary 1, which is based on "ArabeYes" project data, gave a significant accuracy although it is much smaller. This highlights the importance for the dictionary for Query Translation to be modern and practical. The 2 other dictionaries gave many possible Arabic equivalents, even if they are rarely used or are likely to mislead in query translation. The corpus may give those non-relevant translations a weight, not because it is the correct translation in this case, but may be because the term is frequently used in general, but indicating another meaning rather than what is meant in the query.
Orthographic Query expansion based on the proposed model for simulating the OCR errors starts with giving intermediate accuracy with very limited training set then high accuracy after increasing the size of the training set.

References

[1] The official web site of the Library of Congress, Retrieved Dec 4, 2006 from: http://www.loc.gov/about/facts.html

[2] The official web site of the United Nations, Retrieved Dec 4, 2006 from: http://www.un.org

[3] Tayli, M., and Al-Salamah. A. "Building Bilingual Microcomputer Systems," *in Communications of the ACM*, Vol. 33, No.5, Pages 495-505, 1990. http://www.africa.upenn.edu/Software/Biling_Comp.html

[4] John McCarthy. "A Prosodic Theory of Non-concatenative Morphology," *In Linguistic Inquiry* 12, 373-418, 1981.

[5] John McCarthy. "Formal Problems in Semitic Phonology and Morphology," *A Doctoral Dissertation*, MIT 1979, Cambridge, Massachusetts. Reproduced by the Indiana University Linguistics Club. Indiana: 1982.

[6] McCarthy, John. "Template Form in Prosodic Morphology," *in Papers from the Third Annual Meeting of the Formal Linguistics Society of Midamerica,* Laurel Stvan et al. (eds.), Bloomington: Indiana University Linguistics Club, pp. 187 -218, 1992.

[7] Kay, M. "Non-concatenative Finite-State Morphology," *in Proceedings of the Third Conference of the European Chapter of the Association for Computational Linguistics*, Copenhagen, pp. 2-10, 1987.

[8] Beesley, K., Lauri Karttunen.. *A Short History of Two-Level Morphology*, Xerox Palo Alto Research Center 2001.

[9] Kiraz, George. "Multi-tape Two-level Morphology: A Case Study in Semitic Non-linear Morphology," i*n Proceedings of Coling 94*, pp. 180, 186, 1994.

[10] Al-Kharashi, I. A. and Evans, M. W. "Comparing words, stems, and roots as index terms in an Arabic information retrieval system," *Journal of the American Society for Information Science (JASIS)* 5(8), pp 548-560, 1994.

[11] Abu-Salem, H., Al-Omari, M., and Evens, M. "Stemming Methodologies over Individual Query Words for an Arabic Information Retrieval System," *JASIS* 50(6): 524-529, 1999.

[12] Beesley, K. "Arabic Morphological Analysis on the Internet,' *in proceedings of the 6th International Conference and Exhibition on Multi-lingual Computing*, Cambridge, 1998.

[13] Retrieved Dec 4, 2008 from: http://www.glue.umd.edu/~dlrg/clir/arabic.html

[14] Pirkola, A. "The Effects of Query Structure and Dictionary Setups in a Dictionary-based Cross-Language Information Retrieval," *SIGIR* 1998, Melbourne, Australia.

[15] Oard D. "A Comparative Study of Query and Document Translation for Cross-language Information Retrieval," *in proceedings of the 3rd Conference of the Association for Machine Translation in the Americas*, 472-83, 1998.

[16] Mark W. Davis and Ted E. Dunning. "Query Translation Using Evolutionary Programming for Multi-lingual Information Retrieval," *in proceedings of the Fifth Annual Conference on Evolutionary Programming*, 1995

[17] Landauer, T. K., Dumais S.T, and Littman, M. L. "Full Automatic Cross-Language Document Retrieval using Latent Semantic Indexing," 1996, *update of the original paper on the 6th Conf. of UW center for New OED and Text Research*, pp. 31-38, 1990

[18] Sheridan, P. and Ballerini, J.P. "Experiments in Multilingual Information Retrieval using the SPIDER System," the *19th Annual International ACM SIGIR* 1996, 58-65.

[19] Adriani, M., and Croft, W. "The Effectiveness of a Dictionary-Based Technique for Indonesian-English Cross-Language Text Retrieval," *CLIR Technical Report IR-170*, University of Massachusetts, Amherst, 1997

[20] Ballesteros, L., and Croft, B. "Dictionary Methods for Cross-Lingual Information Retrieval," *7th DEXA Conf. on Database and Expert Systems Applications*, Pages 791-801, 1996.

[21] Ballesteros, L., and Croft, B. "Phrasal Translation and Query Expansion Techniques for Cross-language Information Retrieval," *SIGIR* 1997, 84-91.

[22] Xu, J. and Croft, W. B. "Query Expansion using Local and Global Document Analysis," *the 19th Annual International ACM SIGIR* 1996, Zurich, Switzerland, Pages 4-11.

[23] Ballesteros, L., and Croft, B. "Resolving Ambiguity for Cross-Language Retrieval," *SIGIR* 1998, 64-71

[24] Mohammed Aljlayl, Ophir Frieder. "Effective Arabic-English Cross-Language Information Retrieval Via Machine-Readable Dictionaries and Machine Translation," *Information Retrieval Laboratory*, Illinois Institute of Technology, 2002

[25] The Text REtrieval Conference (TREC), co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense, http://trec.nist.gov/

[26] Ahmad Hasnah and Martha Evens. "Arabic/English Cross Language Information Retrieval Using a Bilingual

Dictionary," Department of Computer Science University-Qatar, and Illinois Institute of Technology.

[27] Mohamed Attia M. Elaraby Ahmed. "A Large-Scale Computational Processor of the Arabic Morphology, and Applications," *M.Sc. Thesis*, Cairo University, Faculty of Engineering, 2000.

[28] Hunston, S. *Corpora in applied linguistics*. Cambridge University Press May 2002.

[29] Abdelali, A., Cowie, J., Soliman. H. "Building A Modern Standard Arabic Corpus," *Workshop on Computational Modeling of Lexical Acquisition,* The Split Meeting. Croatia, 25th to 28th of July 2005.

[30] Goweder, A. and De Roeck, A. "Assessment of a significant Arabic corpus," *Presented at the Arabic NLP Workshop at ACL/EACL* 2001, Toulouse, France, 2001.

[31] Retrieved Dec 4, 2006 from: Moheet web site http:// www.moheet.com

[32] Retrieved Dec 4, 2008 from: http://www.arabeyes.org

[33] Last time visited April 15, 2007 from: http://sourceforge.net/project/showfiles.php?group_id=34866&package_id=93898

[34] Last time visited April 15, 2007 from: http://crl.nmsu.edu/Resources/lang_res/arabic.html

[35] Last time visited April 15, 2007 from: http://wordnet.princeton.edu/

[36] Last time visited April 15, 2007 from Sakhr Online Dictionary, http://dictionary.Sakhr.com/

[37] Christiane Fellbaum. *WordNet, An Electronic Lexical Database*, MIT Press, 1998.

[38] Abir Adly, Senior Translation Consultant.

[39] Nasreen AbdulJaleel and Leah S. Larkey. "Statistical Transliteration for English-Arabic Cross Language Information Retrieval," in proceedings of the twelfth international conference on Information and knowledge management table of contents, New Orleans, LA, USA, 2003.

[40] *WordNet documentations, MORHY (7N)*, Princeton University, Cognitive Science Laboratory, http://wordnet.princeton.edu/, January 2005.

[41] Sakhr Software, www.Sakhr.com.

[42] Tim Buckwalter. *Buckwalter Arabic Morphological Analyzer Version 2.0*. LDC Catalog No. LDC2004L02, ISBN: 1-58563-324-0, Linguistic Data Consortium.

[43] Nizar Habash and Owen Rambow. "Arabic Diacritization through Full Morphological Tagging", Center for Computational Learning Systems. Columbia University, *Proceeding of NAACL HLT*, Companion Volume, pages 53-56. Rochester, NY, April 2007, Association for Computation Linguistics.

[44] Andreas Zallmann, Ashish Venugopal, and Stephan Vogel. "Bridging the Inflection Morphology Gap for Arabic Statistical Machine Translation. School of Computer Science," Carnegie Mellon University, in p*roceeding the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, June 4-9, 2006, New York, New York, USA. The Association for Computational Linguistics 2006.

[45] Darwish K, Doermann D, Jones R, Oard D & Rautiainen M. "TREC-10 experiments at University of Maryland CLIR and video," *Text RE-trieval Conference TREC10 Proceedings*, Gaithersburg, MD, pp 549-562, 2001.

[46] Microsoft Word-breaker white paper, Microsoft Corporation, 2004.

[47] Tapas Kanungo, Gregory A. Marton, Osama Bulbul. "OmniPage vs. Sakhr: Paired model evaluation of two Arabic OCR products," in Proc. of SPIE Conf. on Document Recognition and Retrieval, 1999

[48] Mohammed Aljlayl, Ophir Frieder. “On Arabic Search: Improving the Retrieval Effectiveness Via Light Stemming Approach,” *in proceeding the 11th ACM International Conference on Information and Knowledge Management*, Illions Institute of Technology (pp.340-347). New York: ACM Press.

[49] Khoja, S., Garside R. “Stemming Arabic Text”. Computing Department, Lancaster University, UK, Retrieved April 2007
from: http://zeus.cs.pacificu.edu/shereen/research.htm

[50] L. S. Larkey, M.E. Connell.. “Arabic Information Retrieval at Umass in TREC-10", *Text REtrieval Conference*, 2001.

[51] Kareem Darwish. "Probabilistic Methods for Searching OCR-Degraded Arabic Text," *A PhD Dissertation*, University of Maryland, College Park, 2003.

[52] S. V. Rice, J. Kanai, and T. A. Nartker, "The 3rd Annual Test of OCR Accuracy", *TR 94-03*, ISRI, University of Nevada, Las Vegas, April, 1994.

[53] Adobe Company, www.adobe.com.