

بسم الله الرحمن الرحيم



**Sudan University of Science and Technology**  
**collage of Graduate Studies**  
**College of Computer science and Information Technology**

## **Term Translation disambiguation in Cross-Language Information Retrieval**

**(Case Study: Translation From Arabic To English)**

إزالة غموض الترجمة باستخدام معامل الارتباط في أنظمة استرجاع المعلومات  
بين اللغات

(دراسة حالة: الترجمة من اللغة العربية إلى اللغة الإنجليزية)

**A Thesis Submitted in Partial Fulfillment of the requirements of M.Sc. in  
computer science**

**Submitted By:**

Ebtihal Mustafa Elamin Mohammed

**Supervised by:**

Dr. Ali Ahmed Al-faki

November, 2015



## Approval Page

The thesis of: Ibtihads Mustafa Elamin Mohamed  
is approved.

Thesis title: .....  
Term Translation Disambiguation in  
cross-language Information Retrieval  
Case Study: translation from Arabic to English  
أزالة الغموض من الترجمة لاستخدام مظاهر الترابط  
في استرجاع المعلومات بين اللغتين  
دراسة حالة: الترجمة من العربية إلى الإنجليزية

### External Examiner

Name: Dr. Albaria Abuoheda

Sign: [Signature] Date: 25. 11. 2015

### 2. Internal Examiner

Name: Mohamed Elhatiz Mustafa Musa

Sign: [Signature] Date: 25. 11. 15

### 1. Supervisor

Name: Ali Ahmed

Sign: [Signature] Date: 25. 11. 15



Sudan University of Science and Technology  
College of Graduate Studies



### Declaration

I, the signing here-under, declare that I'm the sole author of the (M.Sc.) thesis entitled.....

..... Term Translation Disambiguation in cross language  
..... Information Retrieval Case Study: Translation From Arabic to English

which is an original intellectual work. Willingly, I assign the copy-right of this work to the College of Graduate Studies (CGS), Sudan University of Science & Technology (SUST). Accordingly, SUST has all the rights to publish this work for scientific purposes.


Candidate's name: ... Ebtihal Mustafa ELamin Mohamed ...

Candidate's signature: ...  ... Date: ... 26-1-2016 ...

### إقرار

..... أنا الموقع أدناه أقر بأنني المؤلف الوحيد لرسالة الماجستير المعنونة  
..... إزالة غموض الترجمة باستخدام معادلات الارتباط في أنظمة استرجاع المعلومات  
..... بين اللغات - دراسة حالة الترجمة من اللغة العربية إلى اللغة الإنجليزية  
وهي منتج فكري أصيل . وباختياري أعطى حقوق طبع ونشر هذا العمل لكلية الدراسات العليا - جامعه السودان  
للعلوم والتكنولوجيا، عليه يحق للجامعة نشر هذا العمل للأغراض العلمية .

..... اسم الدارس : ...  ...

..... توقيع الدارس : ...  ... التاريخ : ... 26-1-2016 ...

# DEDICATION

To my parents,

to my brother,

to my sisters,

to my friends, and

to my supervisor.



## **ACKNOWLEDGEMENT**

Praise be to Allah, then thanks to my family which provided all necessary help to me. Also all thanks and appreciation to Dr. Ali Ahmed and Dr. Mohammed Mustafa Ali, who have a major role in completion of this research and gave me help and assistance. Thanks to all colleagues who supported me specially Rayan Omer and Mohammed Taha.

# ABSTRACT

Cross-language information retrieval (CLIR), where queries and documents are in different languages, become one of the major topics within the information retrieval community. The important step in CLIR is the translation. This research proposes a term translation disambiguation method based on co-occurrence statistics for translation in Arabic-English CLIR.

There are multiple ways to perform query translations: employing machine translation techniques, using parallel corpora or using bilingual dictionaries. The first two approaches are very labour intensive. Manual hand-coding of linguistic, semantic and pragmatic knowledge is required for a machine translation engine to produce good translations. This can be quite overwhelming when the domain of coverage is wide. A great deal of work is also required for building parallel collections when using the second approach. With the increasing availability of machine-readable bilingual dictionaries, the third approach has become a viable approach to Cross-Language Information Retrieval (CLIR), but in this approach, resolving term ambiguity is a crucial step.

In this research the ambiguity problem was resolved by co-occurrence statistics. Co-occurrence technique based on the hypothesis that correct translations tend to co-occur together in the target language collection. Therefore, the valid translation among a set of possible synonymous candidates of a certain source query term is expected to have high frequency of co-occurrence with the translations of the other terms in the same source query.

After the document set divided to fixed size window to overcome varying in document length problem, the degree of association is calculated using mutual information measure because it simple and produce high correlation between terms even though they not appeared very frequently in document set.

The results of developed method proved that co-occurrence statistics can reduce the ambiguity problem and it works well in case of diacritics and homonymous.

## المستخلص

في أنظمة استرجاع المعلومات قد يكون المستخدم بحاجة لإسترجاع مستندات بلغة محددة باستخدام استعمال مكتوب بلغة أخرى وهذا ما يسمى بأنظمة استرجاع المعلومات متعددة اللغات مثل ان يسترجع مستندات باللغة الانجليزية عن طريق استعمال مكتوب باللغة العربية، في هذه الحالة يجب ترجمة الاستعلام أولاً ثم اجراء عملية الاسترجاع كالمعتاد.

توجد ثلاث طرق لترجمة الاستعلام: الترجمة باستخدام الآلة، والترجمة باستخدام النصوص المتوازية، والترجمة باستخدام القواميس الثنائية. الترجمة الآلية تحتاج لبرمجة جميع قواعد اللغة يدوياً، وهذه عملية صعبة تتطلب قدراً كبيراً من الوقت والجهد. اما الترجمة باستخدام النصوص المتوازية تعتمد على موارد نادرة؛ فمن الصعب الحصول على نفس النص مكتوب بأكثر من لغة، وبناء هذا النوع من النصوص مكلف جداً خاصة اذا كانت الترجمة غير محصورة في مجال محدد. مع توفر القواميس الثنائية التي يمكن قراءتها آلياً اصبحت الطريقة الثالثة هي الاكثر جدوى نسبة لبساطتها وتوفر الموارد التي تعتمد عليها، ولكن بها عيب اساسي وهو الغموض حيث يمدنا القاموس بأكثر من معنى للكلمة الواحدة فتصبح المشكلة في كيفية اختيار احد هذه الترجمات بحيث تؤدي المعنى بدقة اكبر.

في هذا البحث مشكلة الغموض تم حلها عن طريق حساب مدى ارتباط الكلمة مع الكلمات الواردة معها، وهذه الطريقة تعتمد على تلك الفرضية التي تفرض ان الترجمة الصحيحة في اللغة المستهدفة تميل ان تكون كلماتها ذات ارتباط اعلى مع بعضها البعض مقارنة مع الترجمة الخاطئة.

لحساب مدى ارتباط الكلمات يتم تقسيم المستندات لأجزاء متساوية وذلك لتفادي مشكلة الاختلاف في الطول، بعد ذلك يتم حساب الارتباط بين الكلمات باستخدام مقياس يسمى المعلومات المتبادلة لأن حسابه بسيط ويعطى درجة ارتباط عالية بين الكلمات التي ترد مع بعضها البعض حتى وان كان تكرارها في المستندات بسيط.

النتائج التي تم الحصول عليها بعد تطبيق الطريقة المقترحة تشير إلى ان هذه الطريقة قللت من الغموض بصورة ملحوظة، وهي تعمل جيداً اذا كانت الكلمات تحمل نفس الاحرف ومختلفة فقط في التشكيل وايضاً في حالة الكلمات التي يكون عندها اكثر من معنى.

# Table of contents

<b>DEDICATION .....</b>	<b>ii</b>
<b>ACKNOWLEDGEMENT.....</b>	<b>iii</b>
<b>ABSTRACT.....</b>	<b>iv</b>
<b>المستخلص .....</b>	<b>v</b>
<b>Table of contents .....</b>	<b>vi</b>
<b>LIST OF TABLES .....</b>	<b>viii</b>
<b>LIST OF FIGURES .....</b>	<b>ix</b>
<b>LIST OF ABBREVIATION .....</b>	<b>x</b>
<b>CHAPTER ONE: INTRODUCTION.....</b>	<b>1</b>
1.1. Introduction .....	1
1.2. Problem Statement .....	2
1.3. Research Objectives .....	3
1.4. Research Motivation .....	3
1.5. Research Questions .....	4
1.6. Research Scope .....	4
1.7. Research Methodology.....	4
1.8. Research Organization .....	5
<b>CHAPTER TWO: BACKGROUND AND RELATED WORKS .....</b>	<b>6</b>
2.1. Introduction .....	6
2.2. Information retrieval .....	6
2.2.1. Information retrieval processes .....	6
2.2.2. Information Retrieval Models .....	7
2.2.2.1. Boolean Model .....	7
2.2.2.2. Vector space Model.....	8
2.2.2.3. Probabilistic Model .....	9
2.2.3. Text Processing in Information retrieval.....	11
2.2.3.1. Tokenization.....	11
2.2.3.2. Stop-words Removal .....	12
2.2.3.3. Normalization, Stemming and Lemmatization .....	12
2.3. Cross-Language Information Retrieval .....	13
2.3.1. Translation Approaches in CLIR .....	13
2.3.1.1. Dictionary Translation.....	13
2.3.1.2. Machine Translation.....	14
2.3.1.3. Parallel corpora translation.....	15
2.3.2. Resolving Translation problems.....	16
2.3.2.1. Translation Disambiguation .....	16
2.3.2.1.1. Disambiguation using Co-occurrence Statistics.....	16



2.3.2.1.2. Disambiguation using parallel corpora.....	17
2.3.2.1.3. Bidirectional Translation Disambiguation .....	17
2.3.2.2. Lack of Coverage (OOV) Problem Resolution .....	17
2.3.2.2.1. Transliteration .....	18
2.3.2.2.2. Exploiting Web Corpora .....	18
2.4. Evaluation in Information retrieval .....	19
2.4.1. Test Collection/ Corpora .....	19
2.4.1.1. Standard Test collection .....	20
2.4.2. Relevance Judgment.....	21
2.4.3. Evaluation Measures .....	21
2.5. Arabic Language .....	23
2.6. Related Works .....	24
<b>CHAPTER THREE: METHODOLOGY.....</b>	<b>27</b>
3.1. Introduction .....	27
3.2. Previous methods of query translation .....	27
3.3. Proposed method .....	29
<b>CHAPTER FOUR: EXPERIMENT TOOLS AND EVALUATION.....</b>	<b>33</b>
4.1. Introduction .....	33
4.2. Test Collection .....	33
4.3. Tools and techniques .....	35
4.4. Experiments and Results .....	35
<b>CHAPTER FIVE: CONCLUSION AND FUTURE WORK.....</b>	<b>38</b>
5.1. Conclusion.....	38
5.2. Limitation .....	38
5.3. Future Work .....	38
<b>References.....</b>	<b>39</b>

# LIST OF TABLES

Table 2.1: Distribution of term $t$ over the relevant and non-relevant documents in the collection .....	9
Table 2.2: Set of Arabic Letters .....	24
Table 3.1: sample of correlation matrix .....	31
Table 4.1: Statistic about document set .....	34
Table 4.2: set of queries used in experiment .....	34
Table 4.3: Average DCG values of Google and proposed baselines .....	36

## LIST OF FIGURES

Figure 1.1: incorrect translation to "قرن".....	2
Figure 1.2: incorrect translation to "دار".....	3
Figure 2.1: General IR Processes.....	6
Figure 3.1: Selection of translation equivalent in previous method .....	28
Figure 3.2: steps of proposed method.....	30
Figure 3.3: Flow of proposed translation method .....	31
Figure 4.1: Experiment steps .....	36
Figure 4.1: Average DCG values of two baselines in single graph .....	36

# **LIST OF ABBREVIATION**

IR : Information Retrieval

CLIR : Cross Language Information Retrieval

MT : Machine Translation

DCG : Discounted Cumulative Gain

MSA : Modern Standard Arabic

# Chapter One: Introduction

## 1.1. Introduction

With the growth of Internet rich information become available to all people over the world in different media and languages. User writes a query to retrieve relevant information (usually written documents) to the query. This process is known as Information Retrieval.

Usually the language of retrieved documents is same as query language but sometimes users need to search information in language different from that of the query. For example, one may want to retrieve documents written in English with a query written in Arabic. This has resulted in rise of the Cross-Language Information Retrieval (CLIR), which aims to retrieve information in a language different from the language of the query. Thus cross language information retrieval is typical information retrieval process preceded by translation.

There are two approaches to implement CLIR: a document translation approach and a query translation approach. In the first approach of translating documents can produce accurate translation by using machine translation, because documents have rich context. but it needs to build another indexes to each translation language. Also all web documents must be translated into query language in advance. However, considering the enormous of web documents, this approach is unrealistic. So that second approach of query translation is commonly used, but it suffers from the problem of translation ambiguity, and this problem is amplified due to the limited amount of context in short queries .

There are multiple ways to perform query translations: employing machine translation techniques [1], using parallel corpora [2] or using bilingual dictionaries [3]. The first two approaches are very labour intensive. Manual hand-coding of linguistic, semantic and pragmatic knowledge is required for a machine translation engine to produce good translations. This can be quite overwhelming when the domain of coverage is wide. and a great deal of work is also required for building parallel collections when using the second approach. With the increasing availability of machine-readable bilingual dictionaries, the third approach has become a viable

approach to Cross-Language Information Retrieval (CLIR), but in this approach, resolving term ambiguity is a crucial step, and this is the main objective of research.

## 1.2. Problem Statement

The main problem in dictionary based cross-language information retrieval (CLIR) approaches is the term-sense ambiguity and the difficulty in translating and selecting the accurate translation of query terms. To illustrate the problem suppose you want to translate query "طرق الباب" from Arabic to English, the term "طرق" has multiple translations into English like: rap, tool, knock, percuss, puncture, roads, and enter. And the query term "الباب" also has multiple translations like: the door, chapter, section, gate, subject, and so on. The question now is how to choose the accurate translation of query terms from all alternatives. Another example the term "قرن" can be translated to century, horn, coupling, pairing, and connection. If it comes with some terms like حيوان, خروف, بقر... etc. which mean animal, sheep, and cow, respectively, it must be translated as horn. And if it comes with another terms such as التاسع عشر or العشرين which mean nineteenth and twentieth, respectively, it must be translated as century. Figure (1.1) shows an incorrect translation for term قرن in sentence قرن الحيوان which means animal horn in this sentence. Figure (1.2) shows an incorrect translation for term دار in the sentence دار الحديث حول الثقافة which means talking revolved around culture.

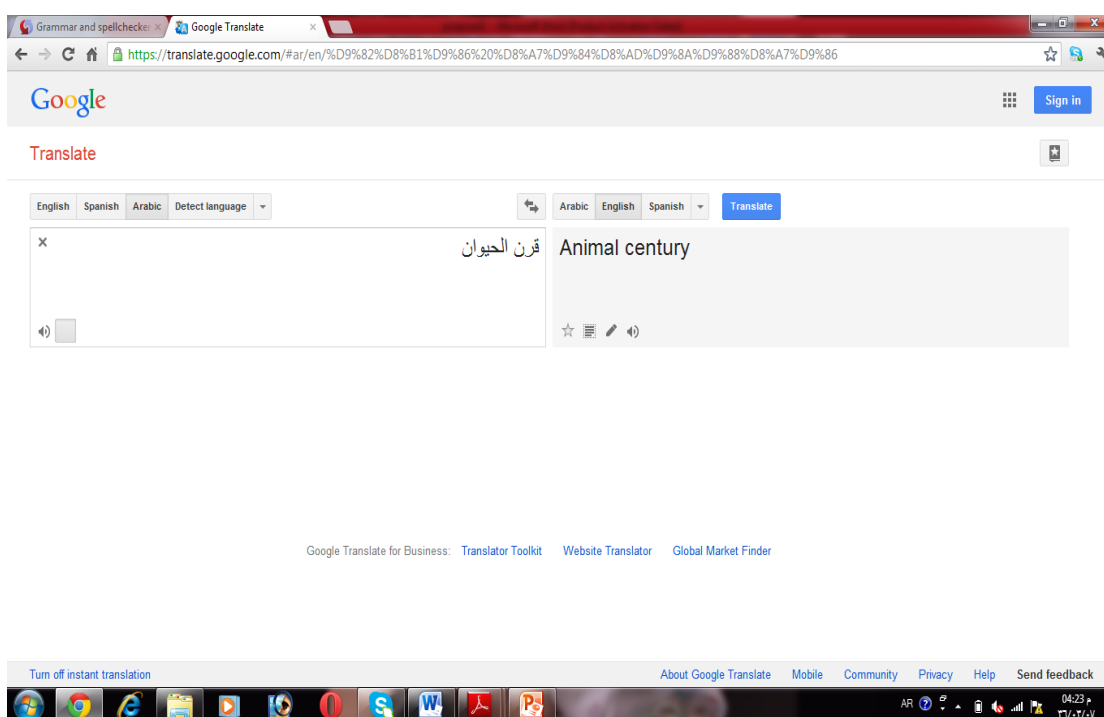


Figure (1.1): incorrect translation to "قرن".

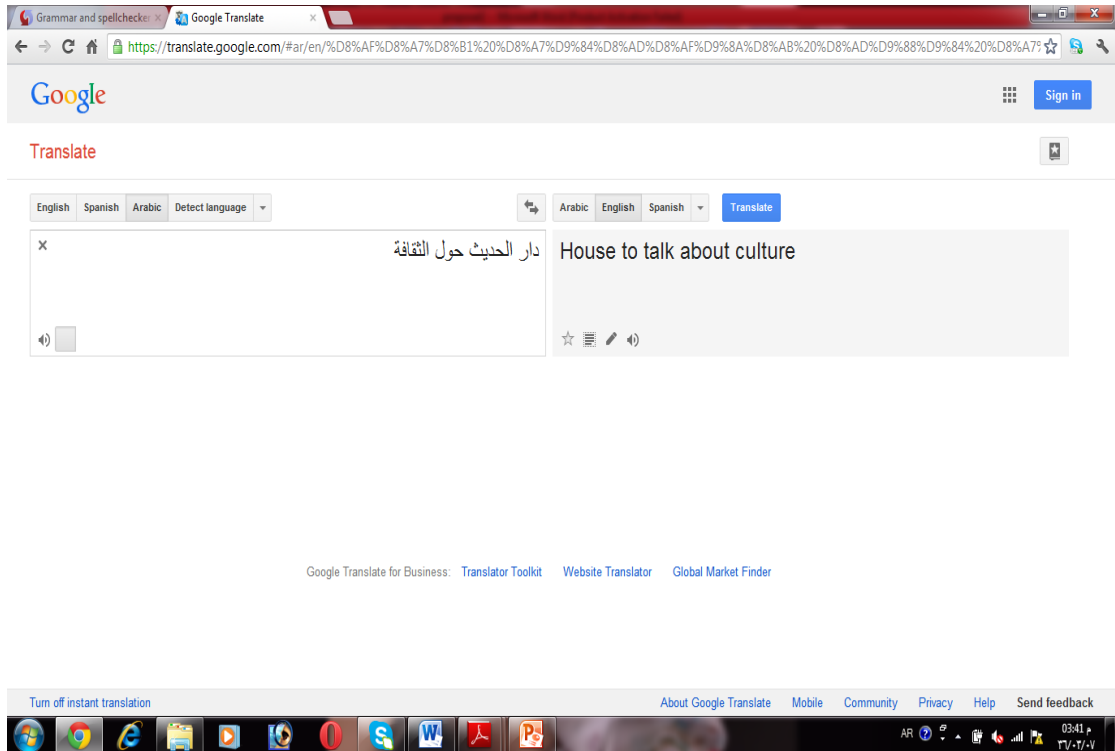


Figure (1.2): incorrect translation to "دار".

### 1.3. Research Objectives

The main objective of research is to develop a term translation method to translate Arabic query into equivalent English query using bilingual dictionaries that select the more accurate sense of query terms. To achieve the above main objective, the following two objectives could be added:

To select the best translation for query terms based on co-occurrence statistics .

To prove that is a dictionary based translation suitable to translate query correctly or not.

### 1.4. Research Motivation

In Cross Language Information Retrieval, translation is most essential process. Researches in this area have many limitations and most of them done in non-Arabic languages. Translating Arabic queries to English queries was selected because English is an international language, and most of resources on the Internet such as books, international and scientific journals, reports, web sites,... etc. are in English language.



## **1.5. Research Questions**

- Is a dictionary based translation suitable to translate query correctly?
- What are techniques that can be used to select the accurate sense of query terms?
- How the proposed method can increase effectiveness of retrieving documents that are written in language differ from query language?

## **1.6. Research Scope**

The scope of research is translating Arabic user query into to equivalent English query, and then uses an off the-shelf IR system for indexing and retrieving the documents.

## **1.7. Research Methodology**

As mentioned above, there are multiple ways for translation, the selected one is a query translation using dictionary because it is simple, available, and does not require hard coding as in machine translation or scarce resources as in translation by using parallel corpora. the problem is, this method suffer from translation ambiguity.

Co-occurrence used to overcome the ambiguity problems. Co-occurrence technique based on the hypothesis that correct translations tend to co-occur together in the target language collection. Therefore, the valid translation among a set of possible synonymous candidates of a certain source query term is expected to have high frequency of co-occurrence with the translations of the other terms in the same source query. In such cases, the problem becomes how to estimate the strength of association (the degree of similarity) between each paired element in the produced set. This is the co-occurrence problem.

Different similarities measures (degree of association) can be used to measure how frequently two terms co-occur in a predefined window such as mutual information and Chi square.

## **1.8. Research Organization**

The research consists of five chapters, Chapter one contains introduction, research problem and objective, Chapter two represents the background and related work, Chapter three shows the proposed solution, chapter four discusses results of the research, and The last Chapter will contain the Conclusion and Future Work.

# Chapter Two: Background And Related Works

## 2.1. Introduction

Conducting this research requires understanding the basic concepts of information retrieval and cross-language information retrieval. Section 2.1 is an overview of chapter contents, section 2.2 is an introduction to IR processes, IR models, and evaluation of IR systems, section 2.3 about cross language information retrieval concepts and techniques. Section 2.4 is review of previous works in this area.

## 2.2. Information retrieval

Information retrieval is a process of retrieving documents that satisfy users' needs from large collection of unstructured data. This data may be text, images, audio or videos, but in this research we will concerned with textual data.

### 2.2.1. Information retrieval processes

The general processes of IR illustrated in figure 2.1.[4] if a user want to retrieve some information from collection , he describe his information need in form of a query. The process of document representation known as indexing,

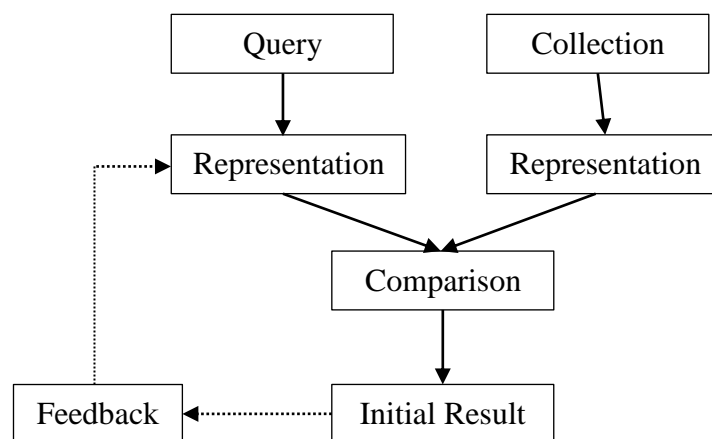


Figure 2.1: General IR Processes

which start with extraction of important keywords by tokenizing documents into words, phrases or N-grams upon on need. This tokens known as terms. Then some preprocessing done on terms like stop-words removing and stemming, this processes will be illustrated in details on next section.

After relevant documents are retrieved the feedback is done either manually by user which known as relevance feedback, or automatically its known as pseudo relevance feedback. In relevance feedback after the system returns an initial set of retrieval results, user marks some returned documents as relevant or non-relevant, then system computes a better representation of the information need based on the user feedback, also when user seen some document may understands more about information need and refine the query. Pseudo relevance feedback, also known as blind relevance feedback, It automates the manual part of relevance feedback, so that the user gets improved retrieval performance without an extended interaction. The method is to do normal retrieval to find an initial set of most relevant documents, to then assume that the top k ranked documents are relevant, and finally to do relevance feedback as before under this assumption [5]. Note that, pseudo relevance feedback can be implemented when IR process based on ranked retrieval model, which will illustrated in details in the next section.

## **2.2.2. Information Retrieval Models**

Information retrieval models describe how documents and query are represented and how relevance score computed to determine which document is relevant to the query. There are multiple retrieval models but the common used are: Boolean model, vector space model, probabilistic model and language model.

### **2.2.2.1. Boolean Model**

Boolean Retrieval Model is the simplest model, documents are represented as a set of terms. Queries formed as a Boolean expression of terms, in which terms are combined with logical set-theoretic operators such as AND, OR and NOT. Retrieval and relevance are considered as binary concepts in this model, i.e. if the terms that represented documents satisfy the Boolean expression that represent the query then this document returned as relevant, so the retrieved elements are an “exact match”

retrieval of relevant documents. There is no notion of ranking of resulting documents that means all retrieved documents are considered equally important. [5, 6]

### 2.2.2.2. Vector space Model

Vector space model is one of ranked retrieval models, in which documents and queries represented as vectors in high dimensional vector space, the dimensions of vector space determined by the number of distinct terms in collection. Each term in vector is represented, it may represented in binary form as in Boolean model, in which, the term take value of 1 if its exist in document or query and take 0 otherwise, but the binary representation don't serve ranking process. so the most commonly used method is based on *tf.idf* weighing schema. *tf* is (term frequency) which means the total number of occurrences of term in specific document, *idf* is (inverse document frequency) is used to determine the importance of specific term. calculated as follow:

$$idf_t = \log \left( \frac{N}{df_t} \right) \quad (2.1)$$

Where *t* is specific term in collection, *N* is total number of documents in collection, *df* is the number of documents contain the term. Note that *idf* is inversely proportional to *df*, that means rare terms is more important than frequent terms. In the standard weighting scheme, the term weight defined as combination between its term frequency and its inverse document frequency, that is:

$$w_{t,d_k} = tf_{t,d_k} \times idf_t \quad (2.2)$$

Where  $w_{t,d_k}$  is weight of term *t* in specific document  $d_k$  and  $tf_{t,d_k}$  is the frequency of term in that document. This standard approach used to assign weight to terms either of document or terms of query[7].

The similarity assessment function that compares two vectors is not inherent to the model, different similarity functions can be used. However, the cosine of the angle between the query and document vector is a commonly used function for similarity assessment. The following formula is typically used:

$$sim(d_j, q) = \cos \theta = \frac{\langle d_j \times q \rangle}{|d_j| \times |q|} = \frac{\sum_{i=1}^{|v|} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{|v|} w_{ij}^2} \times \sqrt{\sum_{i=1}^{|v|} w_{iq}^2}} \quad (2.3)$$

Where  $d_j$  is the document vector,  $q$  is the query vector,  $w_{ij}$  is the weight of term  $i$  in document  $j$ ,  $w_{iq}$  is the weight of term  $i$  in query vector  $q$ , and  $|V|$  is the number of dimensions in the vector that is the total number of important keywords.

As the angle between the vectors decreases, the cosine of the angle approaches one, meaning that the similarity of the query with a document vector increases[6].

### 2.2.2.3. Probabilistic Model

Probabilistic model ranks documents by their estimated probability of relevance with respect to the query and the document. In the probabilistic model, the IR system has to decide whether the documents belong to the relevant set or the non-relevant set for a query. To make this decision, it is assumed that a predefined relevant set and non-relevant set exist for the query, and the task is to calculate the probability that the document belongs to the relevant set and compare that with the probability that the document belongs to the non-relevant set [23].

	relevant (R)	not relevant(R')	Total
Term to present t	rt	st- rt	st
Term to absent t'	R-r	N- st -(R- rt)	N- st
Total	R	N-R	N

Table 2.1: Distribution of term t over the relevant and non-relevant documents in the collection.

N represents the number of documents in the collection,  $r_t$  represents the number of relevant documents containing term t,  $S_t$  represents all documents containing t, and R is the total number of relevant documents.

Consider Table 2.1; the conditional probability that a document R is relevant if it consist a term t is given by

$$p(r|t) = \frac{r_t}{s_t} \quad (2.4)$$

and the probability that a document  $R$  is not relevant if it contains term  $t$  is given by

$$p(R'|t) = \frac{s_t - r_t}{s_t} \quad (2.5)$$

also, the probability that a term  $t$  is present in a relevant document is given by

$$p(t|R) = \frac{r_t}{R} \quad (2.6)$$

and the probability that a term  $t$  is present in a non-relevant document is given by

$$p(t|R') = \frac{f_t - r_t}{N - R} \quad (2.7)$$

with Bayes' theorem, the weight of term  $t$ ,  $W_t$  can be calculated as:

$$W_t = \frac{r_t/R - r_t}{(s_t - r_t)/(N - s_t - (R - r_t))} \quad (2.8)$$

Having calculated the term weight and assuming that terms are independent of each other, the weight for a document  $d$  is calculated by the product of its term weights.

$$W_d = \prod_{t \in d} W_t \quad (2.9)$$

The major purpose is to order documents by estimated relevance according to their weights, not the specific result of the above equation. Therefore, it is often possible to simply express this as a sum of logarithms[24]:

$$W_d = \sum_{t \in d} \log W_t \quad (2.10)$$

The major problem with this model is its dependency on relevance judgments. A similar term weighting can be also used when queries are long. Okapi BM25 measure considers the document frequency ( $f_t$ ), the number of the documents in the collection ( $N$ ), the frequency of a term in the document ( $f_{d,t}$ ) and it normalizes document length. The equation used to compute the similarity between a document  $d$  and a query  $q$  is:

$$BM25(d, q) = \sum_{t \in q} \left[ \log \frac{N - f_t + 0.5}{f_t + 0.5} \right] \cdot \left[ \frac{(k_1 + 1)f_{d,t}}{k_1((1-b) + b \frac{|d|}{avgdl}) + f_{d,t}} \right] \cdot \left[ \frac{(k_3 + 1)f_{q,t}}{k_3 + f_{q,t}} \right] \quad (2.11)$$



In which  $|d|$  is the document length,  $avgdl$  is the average document length in the collection,  $k_3$ ,  $k_1$ , is another parameter to tune term frequency in query  $q$  and other symbols are as defined above [25].

The  $k_1$ , parameter affects the term weight. If it is 0, then the term weight is decreased, meaning that the term weight is not affected by its frequency in the document, and if it is set to a bigger value, the term weight increases as its frequency increases in the document. The tuning constant  $k_3$  affects the number of term instances that participate to the ranking. For example, if  $k_3$  is set to 0, then only one instance of each query term participate to the ranking. The constant  $b$  is used to manage the document length normalization. If it is set to 0, no normalization will take place; if it is set to 1, then normalization is in complete effect. In TREC 6, the value of  $k_3$  was 1.2, the value for  $k_3$  was in the range from 0 to 1000 and the value of the  $b$  parameter was 0.75 [26] .

### **2.2.3. Text Processing in Information retrieval**

This section review the commonly used text preprocessing techniques that are part of the document and query representation task in Figure 2.1.

#### **2.2.3.1. Tokenization**

To represent documents and build the index tokenization is important step in which text is divided into small pieces such as words, phrases, symbols, N-grams or other meaningful elements called tokens, this tokens then used to build the index. In addition to dividing text tokenization perform some other operations such as: removing punctuation and converting uppercase letters to lowercase[8].

Tokenization is not easy task as it seem because it depends on language, for example, in Arabic and English languages can tokenize on whitespace but this method is not feasible on some others language such as Japanese and Chinese languages. Also in whitespaces separated languages some error can occurs, when token is acronym or hyphenation-separated words, different possible segmentations can take place and thus invalid segment may occur[7]. For example, if the token 'flip-flop' appears in text, one might consider hyphenation as punctuation and remove it, then the token become 'flipflop', while another may use the hyphenation as a delimiter for the word

end and the token divided to two parts 'flip' and 'flop'. this errors can increase ambiguity in information retrieval.

### **2.2.3.2. Stop-words Removal**

Stop-words are very commonly used words in a language that play a major role in the formation of a sentence but which seldom contribute to the meaning of that sentence such as prepositions (i.e. by, of, to), articles (i.e. a, an), pronouns (i.e. it, he, which) .etc. These words are expected to occur in 80 percent or more of the documents in a collection, that is, it has high document frequencies so it cannot distinguished between documents. Stop-words are usually eliminated from both query and documents. Is seem that is good to eliminated stop-words because it has a little importance and its removal results in elimination of possible spurious indexes, thereby reducing the size of an index structure by about 40 percent or more. However, doing so could impact the recall if the stop-word is an integral part of a query, for example, a search for the phrase 'To be or not to be,' where removal of stop-words makes the query inappropriate, as all the words in the phrase are stop-words. Many search engines do not employ query stop-word removal for this reason[6].

### **2.2.3.3. Normalization, Stemming and Lemmatization**

Normalization is the process of produce canonical form of tokens so that matches occur despite of differences in the character sequences of the tokens in order to maximize matching between a query token and document collection tokens. There are multiple approaches to normalize tokens. One common approach is to remove a certain symbol from a token such as hyphen or any non-character symbol, for example, tokens 'pre-processes' and 'preprocesses' are both mapped onto the term 'preprocesses', in both the document text and queries, then searches for one term will retrieve documents that contain either. Another approach is to convert all text into single case, e.g. lower case, for example, covert a sentence 'Green Tree' to 'green tree' this approach known as case folding.

Stemming is a process of reducing inflected forms of a word to their base form by trimming the suffix and prefix of an original word. For example, computer, computing, and computation have one stem word is comput. Stemming is language-

dependent process, In English, the most famous stemmer is Porter stemmer, and Light10 stemmer for Arabic. Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma. The main differences between stemming and lemmatization is that stemmers just cut the suffix and prefix to produce a stem, whereas lemmatizer use morphological analysis to produce lemma, for example, if we enter the word 'having' to both stemmer and lemmatizer, stemmer will produce 'hav', and lemmatizer will produce 'have'.

The main goal of normalization, stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form that is increase matching between query and documents and increase recall, possibly at the cost of precision. [5, 7]

## **2.3. Cross-Language Information Retrieval**

Traditionally, when user issues query, the retrieved documents language is same as language of the query. But sometimes user might need to retrieve documents in language differ from query language where information written in a language other than the user language is rich. To satisfy this needs research of Cross-Language of information retrieval have been active in recent years.

### **2.3.1. Translation Approaches in CLIR**

Translation in cross-language information can be done either by translate queries to documents language or translate documents into language of a query. because full documents translation of large collection is impractical, researches focus on the second alternative of query translation. There are several methods for translation but the commonly used approaches are machine translation, parallel corpora and dictionary translation. The next subsections contains illustration of each method.

#### **2.3.1.1. Dictionary Translation**

Machine-readable dictionaries have become increasingly available and are used in the translation of CLIR. Translation using Machine-readable dictionaries typically

performed via simple dictionary lookup, so this approach to translation is relatively simple when compared to the previous alternatives but suffers from two weaknesses ambiguity and lack of coverage.

Ambiguity is major problem affects systems employing dictionary translation, because bilingual dictionary provide multiple translations for each query term. Choosing the accurate translation from set of alternative is nontrivial task. Early systems addressed the problem of ambiguity in primitive way by simply selecting the first translation offered by dictionary. This method exploits the fact that in bilingual dictionaries, the most commonly used translation is listed first. This basic disambiguation strategy has obvious shortcomings and was soon replaced by more sophisticated techniques exploiting term co-occurrence statistics. this approach should be able to determine the most likely translation for a given query by examining the pattern of term co-occurrence within some representative text collection or a monolingual corpora[14].

Some words such as newly coined terms, technical terms, compound words, proper names, acronyms and abbreviations are not match any entry in dictionary, such terms known as out-of-vocabulary (OOV). Out-of-vocabulary (OOV) is the second problem when using dictionary and it might degrade the effectiveness of the retrieval system. Early solution to this coverage problem is using of domain specific bilingual dictionaries. These dictionaries delivered access to uncommon vocabularies and technical terms[15]. Another solution is to stem the terms, this solution partially addressed this problem but still incomplete. For this reason, research changed from domain specific resources to transliteration[16].

### **2.3.1.2. Machine Translation**

Machine translation (MT) is the process by which computer software is used to translate a text from one natural language (such as English) to another (such as Arabic). Translation is not just word-for-word substitution, the meaning of a text in the source language must be fully restored in the target language. So a translator must interpret and analyze all of the elements in the text and know how each word may influence another.

There are two basic types of machine translation systems are: the rule-based MT system and the statistical MT system, and the third type is hybrid systems that is combine basic types. Rule-based MT systems use manually constructed rules and resources such as lexical, phrasal, syntactic, semantic, syntactic, etc. Statistical MT systems tries to generate translations using statistical methods based on large set of texts and their translations in a different language (parallel texts).

Machine translation had several disadvantages. MT systems do not exist for many languages pairs, and its development require significant amount of time and resources. even if a system works well for one pair of language, each new languages require significant efforts. MT needs large training data, if the training data is insufficient OOV problem will occurs. MT system is not good choice when we want to translate queries because it needs more context than is in a query. Usually queries are short and not well formed [7, 16].

### **2.3.1.3. Parallel corpora translation**

Parallel corpora is a large collection of documents and their translation in one or more other languages, Analysis of these paired documents can be used to infer the most suitable translations of terms between languages in the corpus. Parallel corpora are rich resources that contain translation relations between texts, sentences, phrases, and words. It can be acquired from different sources such as International organizations which publish a huge volume of parallel documentation every year in a several languages like United Nations, World Wide Web is also rich source for parallel corpora because most of organizations web sites contents are provided in different languages. Or can translated by human manually or using machine translation systems.

To use parallel corpora in translation, must align original and translated text into sentence or paragraph level. These sentence-aligned pairs then used to training statistical translation model.

Using parallel corpora in translation in CLIR have several advantages, the first one it provide good translations for new terms, technology, proper names and slang terms, especially when they are obtained from the Web, it beneficial sources for extracting linguistic knowledge such as morphological analysis. Parallel corpora are also used

to disambiguate translations when several alternatives are available for source terms[7]. the key disadvantages of the corpus-based approach to query translation is the difficulty of obtaining suitable document collections. Parallel corpora can be extremely time-consuming to produce, even when restricted to specific information domains[7,14].

## **2.3.2. Resolving Translation problems**

The main goal of CLIR is to generate an approximate translation for users queries. the previous section illustrated the commonly used translation techniques, their advantages and limitations. Now, we can summarize that the major problems can faced translation techniques are sense ambiguity problem and lack of coverage or out of vocabulary (OOV) problem. The following sub sections illustrate some proposed solutions to address this problems.

### **2.3.2.1. Translation Disambiguation**

When the query terms can be translated into different meanings in the target language, the various translations can introduce noise to the retrieval process, and decrease the precision of the results. Disambiguation techniques are typically employed to reduce translation errors introduced during query translation in cross-lingual information retrieval. Previous work has proposed several techniques to address this problem. Next sections discuss some of these techniques.

#### **2.3.2.1.1. Disambiguation using Co-occurrence Statistics**

We can disambiguate translation of terms with their frequently co-occur neighbors, based on the hypothesis that correct translations tend to co-occur together in the target language collection[7]. For example, assume that we know only two senses of the word bank, repository for money, and a pile of earth on the edge of a river. We can expect the first sense of bank if it co-occur frequently with words such as money and loan, and expect the second sense if it's associate with words such as river, bridge, and earth[20]. In this strategy, estimation of degree of co-occurrence or statistical similarity between terms is essential step.

different statistical similarities measures can be used to measure the degree of similarity or association between two terms in set of document such as Mutual Information, Dice Coefficient, Log Likelihood Ratio or Chi-Square Test.

### **2.3.2.1.2. Disambiguation using parallel corpora**

Parallel corpora contain a set of documents and their translations in one or more other languages. Analysis of these paired documents can be used to infer the most suitable translations of terms between languages in the corpus.

Parallel corpora are often used to determine the relationships, such as co-occurrences, between terms of different languages, and can be employed to train a statistic translation model [10].

As we mentioned above parallel corpora can use to the resolve both OOV and translation ambiguity problem. but the key disadvantages of the corpus-based approach to query translation is the difficulty of obtaining suitable document collections. Parallel corpora can be extremely time-consuming to produce, even when restricted to specific information domains.

### **2.3.2.1.3. Bidirectional Translation Disambiguation**

In bidirectional translations, after getting all possible translations of the sentences in target language, these candidates are retranslated back into source language to get more suitable one. That means translations are executed in both directions from a source language to a target language and vice versa.

The hypothesis here is that if the set of equivalent senses for a source term is backward-translated term by term into the source language, using dictionaries for example, the preferred translation is then the target word, whose set of equivalent translations into the source language contain the original source term[7].

### **2.3.2.2. Lack of Coverage (OOV) Problem Resolution**

OOV is a widespread problem in CLIR, it is arises from the fact that some terms such as newly coined terms, technical terms, compound words, proper names, acronyms and abbreviations during translation may not found in translation resource, and may degrades performance of CLIR systems. Researchers propose



multiple solutions to address coverage problem, next two sections illustrate some of them.

### **2.3.2.2.1. Transliteration**

Transliteration is a process in which words in one alphabet are represented in another alphabet. For example, Roman alphabet name 'محمد' is transliterated to English as Muhammad.

Transliteration can be done by identifying similarities in the orthographic structures of two languages. These similarities are subsequently used to generate rules specifying how sub-strings written in one language are spelled in another. But this approach only really works when the languages share a similar alphabet, such as English and French. Transliteration between languages with dissimilar alphabets such as Arabic and English requires an additional intermediate process known as phonetic mapping. Phonetic mappers generate rules representing the phonetic presentation of a language. During the mapping process, all proper nouns are transformed into a corresponding phonetic sequence. This phonetic sequence is matched with a phonetic sequence in the target language, then transformed into a final translation[7, 14].

The transliteration technique that described above known as Transliteration Generation, another technique known as Transliteration Mining technique attempts to mine the transliterations of out-of-vocabulary query terms from the document collection. You can get additional details about it from Saravanan et al. in [18].

### **2.3.2.2.2. Exploiting Web Corpora**

As more data is being put on the Web every day, there is a great potential to exploit the Web as the corpus to automatically find effective translations for unknown query terms.

Pu-Jen Cheng et al. in [19] proposed an online system to deal with the translation of unknown queries. They note that for some language pairs, such as Chinese and English, as well as Japanese and English, the Web consists of rich texts in a mixture of multiple languages. Many of them contain bilingual translations of proper nouns, such as company names and personal names. They search for English terms only for

pages in a certain language, e.g., Chinese or Japanese, which are normally returned in a long ordered list of snippets of summaries (including titles and page descriptions) to help users locate interesting documents. Then they mine query translations from these dynamically retrieved bilingual search result pages to extract semantically close translation.

## **2.4. Evaluation in Information retrieval**

To measure effectiveness of information retrieval system in standard way, we must have an evaluation corpus which consisting of three type of components, are: document collection, set of test queries represent the information needs, and set of relevance judgments, standardly a binary assessment of either relevant or non-relevant for each query-document pair[5]. The next sub sections illustrate these components.

### **2.4.1. Test Collection/ Corpora**

Test corpora is set of documents used to practically test and assess new proposed algorithms. There are three approach to gather text corpora from the Web. These are: automatic crawling and harvesting based on a pre-defined list of URLs, automatic or manual downloading based on submission of queries to search engines, and manual collection of documents[7].

Test collection can be categorized to different types depending specific criteria. In terms of languages test collections can be categorized into two types: single language corpora and multilingual corpora. The single language corpora also known as monolingual corpora, all documents are written in a single language. An example of a monolingual collection is the Arabic Agence France Presse (AFP), which is an Arabic newswire collection acquired from articles taken from the AFP Arabic newswire and created by the Linguistic Data Consortium (LDC). In multilingual corpora, documents are written in several monolingual languages or consist of several monolingual corpora. Such types of multilingual corpora highlight language-specific, typological or cultural features and they are mostly collected from both newspapers and newswire sources. Parallel and comparable corpora can be also considered as multilingual corpora.

In terms of vocabulary types, test collections can be classified as general corpora or specialized corpora. A general corpus usually contains different genres and domains such as regional and national newspapers, legal documents, encyclopedias and periodicals. Whereas specialized corpus/ test collection contains terminology in a specific domain[7].

#### **2.4.1.1. Standard Test collection**

Here is a list of the most standard test collections used for ad hoc information retrieval system evaluation.

**CRANFIELD:** The Cranfield collection. This was the pioneering test collection in allowing precise quantitative measures of information retrieval effectiveness, but is nowadays too small for anything but the most elementary pilot experiments. Collected in the United Kingdom starting in the late 1950s, it contains 1398 abstracts of aerodynamics journal articles, a set of 225 queries, and exhaustive relevance judgments of all (query, document) pairs[5].

**TREC:** Text Retrieval Conference (TREC). The U.S. National Institute of Standards and Technology (NIST) has run a large IR test bed evaluation series since 1992. Within this framework, there have been many tracks over a range of different test collections, but the best known test collections are the ones used for the TREC Ad Hoc track during the first 8 TREC evaluations between 1992 and 1999. In total, these test collections comprise 6 CDs containing 1.89 million documents (mainly, but not exclusively, newswire articles) and relevance judgments for 450 information needs, which are called topics and specified in detailed text passages. Individual test collections are defined over different subsets of this data. This is probably the best sub collection to use in future work, because it is the largest and the topics are more consistent. Because the test document collections are so large, there are no exhaustive relevance judgments.

**REUTERS:** The most used test collection has been the Reuters-21578 collection of 21578 newswire articles. More recently, Reuters released the much larger Reuters Corpus Volume 1 (RCV1), consisting of 806,791 documents. Its scale and rich annotation makes it a better basis for future research.

20 NEWSGROUPS: collected by Ken Lang. It consists of 1000 articles from each of 20 Usenet newsgroups (the newsgroup name being regarded as the category). After the removal of duplicate articles, as it is usually used, it contains 18941 articles.

### **2.4.2. Relevance Judgment**

With respect to a user information need, a document in the test collection is given a binary classification as either relevant or nonrelevant. A document is relevant if it addresses the stated information need, not because it just happens to contain all the words in the query. This distinction is often misunderstood in practice, because the information need is not overt. So this task is done by human to determine total number of relevant documents for each topic. this approach is infeasible, especially for large test collections, due to the large effort needed. Instead, a sample of documents for each topic is only assessed. This approach is known as pooling[21]. In the pooling technique, the top k documents, e.g., 100 retrieved by each participating retrieval algorithm are collected and all these selected documents are pooled together into a single pool. Documents that were not selected in the unified pool are often considered as irrelevant. Duplicates in the pool are removed and documents are presented to assessors in a random order without any information about which document was returned by which algorithm or what rank a document obtains. Although, the pooling method has been questioned since documents not in the pool are handled as irrelevant, even if they are relevant, the analysis of Buckley and Voorhees in [22] showed that the technique is stable and sufficient to acquire accurate comparisons and it is useful in measuring effectiveness of IR systems[7].

### **2.4.3. Evaluation Measures**

The basic measures to evaluate information retrieval systems effectiveness based on binary relevance judge are recall and precision. Recall is the ratio of relevant documents retrieved by a search over the total number of existing relevant documents in collection.

$$Recall = \frac{\text{number of relevant retrieved documents}}{\text{number of relevant documents in collection}} \quad (2.12)$$

Precision is the ratio of relevant documents retrieved by a search over the total number of documents retrieved by that search.

$$precision = \frac{\text{number of relevant retrieved documents}}{\text{number of retrieved documents}} \quad (2.13)$$

Recall measures the ability of a system to retrieve all relevant documents, and precision measures the ability of a system to retrieve only relevant documents. Suppose that a document collection contains 100 documents, a user issues a query, and the number of relevant documents to the user query is 10. If the system returns all documents in the collection, it has 100% recall, and precision is 10%. If the system returns one relevant document, the precision of the system is 100% and recall is 10%. That means high precision is achieved almost always at the expense of recall and vice versa. A new single measure that combines precision and recall is the F-score, which provides the harmonic mean of precision and recall. Formally computed as:

$$f = \frac{2}{\frac{1}{p} + \frac{1}{r}} \quad (2.14)$$

Where  $p$  is precision and  $r$  is recall. One of the properties of the harmonic mean is that it tends to be closer to the smaller value. Thus the F-score is automatically biased toward the smaller of the precision and recall values. Therefore, for a high F-score, both precision and recall must be high [6].

Precision, recall, and the F-measure are set-based measures. They are computed using unordered sets of documents without considering relevance ranking as in Boolean models. But most IR systems don't return a set of documents; instead, they return a list of documents ranked by their probability of relevance to the query. If we are to evaluate such ranked retrieval results, we need to extend these measures.

Recall and precision can be defined in a ranked retrieval setting. The Recall at rank position  $i$  for document  $d_i$  is the fraction of relevant documents from  $d_1$  to  $d_i$  in the result set. The Precision at rank position  $i$  or document  $d_i$  is the fraction of documents from  $d_1$  to  $d_i$  in the result set that are relevant. Average precision is computed based on the precision at each relevant document in the ranking. This measure is useful for computing a single precision value to compare different retrieval algorithms on a query  $q$  [6].

When graded relevance is used, the Discounted Cumulative Gain (DCG) can be used. The DCG is becoming an increasingly popular measure for evaluating performance. The assumption in this measure is that lower ranked documents (documents with greater ranks) are less valuable for users and less likely to be tested by them. In that perspective, the most relevant documents (highly relevant) are more valuable than those documents with marginal relevance. Thus, if a graded relevance scale is used to judge the relevance of documents, then it can be employed by the DCG as a measure the value level or gain from testing a document. Thus, from the top of the list the gain begins to accumulate and it may be reduced or discounted as other documents are examined. DCG at a particular rank  $p$  ( $DCG_p$ ) is defined as follows:

$$DCG_p = R_1 + \sum_{i=2}^p \frac{R_i}{\log_2 i} \quad (2.15)$$

Where  $R_i$  is the graded relevance level of the document retrieved at rank  $i$ . The denominator  $\log_2 i$  is the discount of the gain. For example, if we have 10 ranked documents judged on 0-3 relevance scale as follow: 3, 2, 3, 0, 0, 1, 2, 2, 3, 0. To calculate DCG at point 10 we implement the following equation:  $r_1 + \frac{r_2}{\log_2 2} + \frac{r_3}{\log_2 3} + \dots + \frac{r_{10}}{\log_2 10}$

$$DCG = 3 + \frac{2}{1} + \frac{3}{1.59} + \frac{0}{2} + \frac{0}{2.32} + \frac{1}{2.58} + \frac{2}{2.81} + \frac{2}{3} + \frac{3}{3.17} + \frac{0}{3.32}$$

Hence, DCG at point 10 ( $DCG_{10}$ ) = 9.61.

## 2.5. Arabic Language

Arabic is one of the oldest languages that originated in the Arabian peninsula in pre-Islamic times. It is Semitic languages, which also includes Hebrew, Aramaic and Amharic and its first documented inscription was found around 328 C.E [7].

Script of Arabic consists of two types of symbols: these are the letters and the diacritics (known also as short vowels), which are certain orthographic symbols that are usually added to disambiguate Arabic words. For instances, SEEN (س) is a letter equivalent to "S" in English, whereas سُ is a diacritized letter with the sound 'su', like in the word Sudan. Short vowels are always omitted in written MSA texts as Arabic

speakers could distinguish easily between words with similar forms from the context in which they occur.

Basically, the Arabic alphabet has 28 letters and, unlike English, there is no lower and upper case for letters in Arabic. An additional character, which is the HAMZA (ء), has been also added, but, usually it is not classified as the 29th letter. Table 2.2 illustrates the complete set of the Arabic alphabet. Each of the letters in the set can be extended using short vowels, resulting in approximately 90 elements. For example, the letter SEEN can have the sound 'sa' (written in Arabic as سَ), 'su' (written as سُ) and 'si' (written as سِ). The diacritics can change the meaning of the word. For example the word "سَمَك" means fish, and word "سُمك" means sickness, both words have same letters "س", "م" and "ك" but they are differ in diacritics.

ص	ش	س	ز	ر	ذ	د	خ	ح	ج	ث	ت	ب	أ
ي	و	هـ	ن	م	ل	ك	ق	ف	غ	ع	ظ	ط	ض

Table 2.2: Set of Arabic Letters

## 2.6. Related Works

Ahmed and Nürnberger [9] they describe the implementation and evaluation of an Arabic/English word translation disambiguation approach that is based on exploiting a large bilingual corpus and statistical co-occurrence to find the correct sense for the query translations terms. The correct word translations of the given query term are determined based on their cohesion with words in the training corpus and a special similarity score measure.

Gao et. al [10] explore several methods to improve query translation for English-Chinese CLIR. First, they try to identify noun phrases (NP) in a query and translate them as units. Phrases usually have fewer senses, thus the translation of a multi-word concept as a phrase is more precise. In addition to the NPs stored in the dictionary, new multi-word NPs are identified automatically using a statistical model. Second, to deal with the translation ambiguity problem, they propose a method based on statistics of co-occurrences. The method tries to select the best translation according to its coherence with the other translation words. Finally, to increase the coverage of the bilingual dictionary, additional words and translations are automatically generated from a parallel bilingual corpus. We tested our methods using TREC

Chinese documents. their results show that each of the methods can bring significant improvement over simple dictionary approaches. A combination of the methods achieves even better retrieval performance than a high-quality machine translation (MT) system.

Akira et al [11], proposed a disambiguation method for dictionary-based query translation, achieving adequate retrieval effectiveness by utilizing Web documents as a corpus and using co-occurrence information between terms within that corpus. They select translation by the following steps:

1. Obtain the number of retrieved documents for each term in the query from the Web search engine.
2. Obtain numbers of retrieved documents for all possible combinations of each pair of translation candidates, whose occurrence frequency for each term exceed the threshold, from the Web search engine (using an AND operator).
3. Calculate the average of co-occurrence for all possible combinations of the translation-candidate pairs.
4. The term sets whose co-occurrence exceed the threshold value are selected as the target language query.

In the experiments, their method achieved 97% of manual translation case in terms of the average precision.

Mirna Adriani [12] proposed a sense disambiguation technique based on a term similarity measure for selecting the right translation sense of a query term. In addition, she apply a query expansion technique which is also based on the term similarity measure to improve the effectiveness of the translation queries. The results of her Indonesian to English and English to Indonesian CLIR experiments demonstrate the effectiveness of the sense disambiguation technique. As for the query expansion technique, it is shown to be effective as long as the term ambiguity in the queries has been resolved. In the effort to solve the term ambiguity problem, they discovered that differences in the pattern of word-formation between the two languages render query translations from one language to the other difficult.

Lisa Ballesteros and W. Bruce Croft [13] First they present a technique based on co-occurrence statistics from unlinked corpora which can be used to reduce the



ambiguity associated with phrasal and term translation based on the concept that correct translations of query terms should co-occur in the text and incorrect translations should not. The translations are first filtered with part-of-speech tags for reducing ambiguity. Each translation candidate of a query term is then paired up with a translation candidate for another query term. Each pair's pattern of co-occurrence is calculated, and the ones with the highest co-occurrence values are chosen as the query translation. achieve more than 90% monolingual effectiveness. Finally, they compare the co-occurrence method with parallel corpus and machine translation techniques and show that good retrieval effectiveness can be achieved without complex resources.

Aljlayl, et al in [17] proposed method of translation in two directions from a source language to a target language and vice versa, The hypothesis here is that if the set of equivalent senses for a source term is backward-translated term by term into the source language, using dictionaries for example, the preferred translation is then the target word, whose set of equivalent translations into the source language contain the original source term.

# Chapter Three: Methodology

## 3.1. Introduction

Cross Language Information Retrieval is a process of retrieving documents in language differ from language of query. With increasing number of machine readable texts in various languages accessible via the World Wide Web, this attract users to interest in retrieving information in across languages. People use CLIR because they might able to read documents in foreign languages, but have difficulty formulating foreign queries, or they know foreign keywords or phrases, and want to read documents associated with them, in their native language.

As mentioned previously, translating queries using a bilingual dictionary gives rise to a number of problems, namely, the ambiguity problem and out of vocabulary problems with unidentified acronyms, names or proper nouns. In our work, we concentrate on solving the ambiguity problem by choosing the correct sense for each translated term.

## 3.2. Previous methods of query translation

As we mention in section 2.3.1.1 the early systems addressed the problem of ambiguity in primitive way by simply selecting the first translation offered by dictionary. This method depend on hypothesis that in bilingual dictionaries, the most commonly used translation is listed first. Another approach not select one translation for each query term, instead it substitute each query term with it's all possible translation candidates. In this approach the query term that has translations candidates more than other terms will gain weight more than other terms in a query. hence, terms with more possible translation candidates have more effect on retrieved list over those with a few number of translations. This approach known as the unbalanced query.

Both approaches have obvious shortcomings and was replaced by more sophisticated techniques using term co-occurrence statistics. these query translation techniques carry out the query translation task through a repeated selection process among the possible translation equivalents of each query term. The most appropriate translation

equivalent of each query term is selected in sequence to form a target. For example, Gao et al. in [10] and Adriani in [12] select the most likely translation to query terms according to the highest association scores between terms.

This query translation method also has a problem. Let us consider an example in Figure (3.1), where the user query consists of four query terms ( $s_1, s_2, s_3, s_4$ ), the translation equivalents of  $s_1$  are  $e_{11}, e_{12}, e_{13}$  and  $e_{14}$ , and the solid lines connect the  $e_{23}$  to the translation equivalents with the highest association score among the translation equivalents of each user query term. However, the target query is composed of the translation equivalents in the dark circle ( $e_{13}, e_{23}, e_{32}, e_{43}$ ). The translation equivalent  $e_{23}$  is chosen not by how strongly it is associated with those actually selected as target query terms,  $e_{13}, e_{32},$  and  $e_{43}$  but by the fact that it is most strongly associated with other terms such as  $e_{12}, e_{31},$  and  $e_{42}$  that are in fact not qualified to be final target query terms.

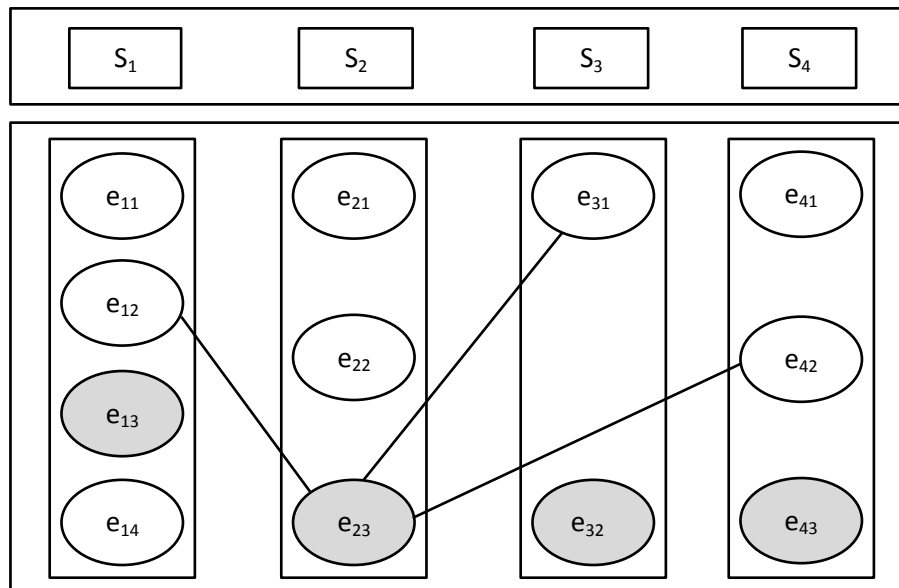


Figure 3.1: Selection of translation equivalent in previous method.

That is caused because the query translation method focuses on selecting a translation equivalent of each query term, not constructing an entire target query. Therefore, other terms that are not chosen as part of the target query may affect the selection of translation equivalents, which may not have strong relations with each other in the target query.

In bidirectional translation method which proposed by Aljlayl et al. in [17], the translation is selected term by term regardless of other query term. This method is

not accurate specially in case of homonymous (same term has different meaning). For example, if query is قرن الحيوان (animal horn), if we translate قرن as century, and then retranslate century back to Arabic will produce the original term قرن.

In our proposed method we combine terms co-occurrence with bidirectional translation and we avoid the limitation of these techniques. The next section illustrate the proposed method in details.

### 3.3. Proposed method

We propose a term-sense disambiguation technique for selecting the best translation sense of a word from the possible senses given by a bilingual dictionary, this technique based on hypothesis that the correct senses tend to co-occurred frequently. Basically, given a set of original query terms, we select for each term the best sense such that the resulting set of selected senses contains senses that are closely related or statistically similar with one another. The steps of proposed method is shown as follow, summarized and illustrated in Figure (3.2) :

1. Prepare two documents corpus one for source language (Arabic) and other for target language (English).
2. Calculate the co-occurrence matrix between each term-pair in each corpus, this co-occurrence used to determine degree of correlation between terms. We obtain the degree of similarity or association relation between terms using a term association measure, called mutual information. The mutual information measures how frequently two terms co-occur in a predefined window. The term association value between two terms is calculated using the formula described below:

*mutual information* =

$$\frac{N_{11}}{N} \log_2 \frac{NN_{11}}{N_1.N_1} + \frac{N_{01}}{N} \log_2 \frac{NN_{01}}{N_0.N_1} + \frac{N_{10}}{N} \log_2 \frac{NN_{10}}{N_1.N_0} + \frac{N_{00}}{N} \log_2 \frac{NN_{00}}{N_0.N_0} \quad (3.1)$$

Where:

$N$ : Total number of documents windows.

$N_{11}$ : Number of windows contain both terms.

$N_{10}$ : Number of windows contain first term and not second term.

- $N_{01}$ : Number of windows contain second term and not first term.
- $N_{00}$ : Number of windows not contain neither first nor second term.
- $N_{11}$ : Number of windows contain first term regardless of second term.
- $N_{10}$ : Number of windows contain second term regardless of first term.
- $N_{01}$ : Number of windows not contain first term regardless of second term.
- $N_{00}$ : Number of windows not contain second term regardless of first term.

The documents divided to fixed size windows to handle problem of ununiformed distribution which can be arise if documents are varying length.

3. For each query term retrieve set of senses from dictionary, and formulate all translation candidates for the query. Then get the average correlation for each candidate.
4. Get the three top candidates with maximum correlation and retranslate it back to source language, the one that produce the original query is selected as correct translation for a query.

Flow of query translation processes was shown in figure (3.3).

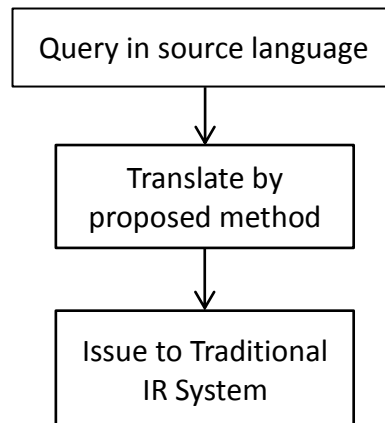


Figure 3.2: steps of proposed method

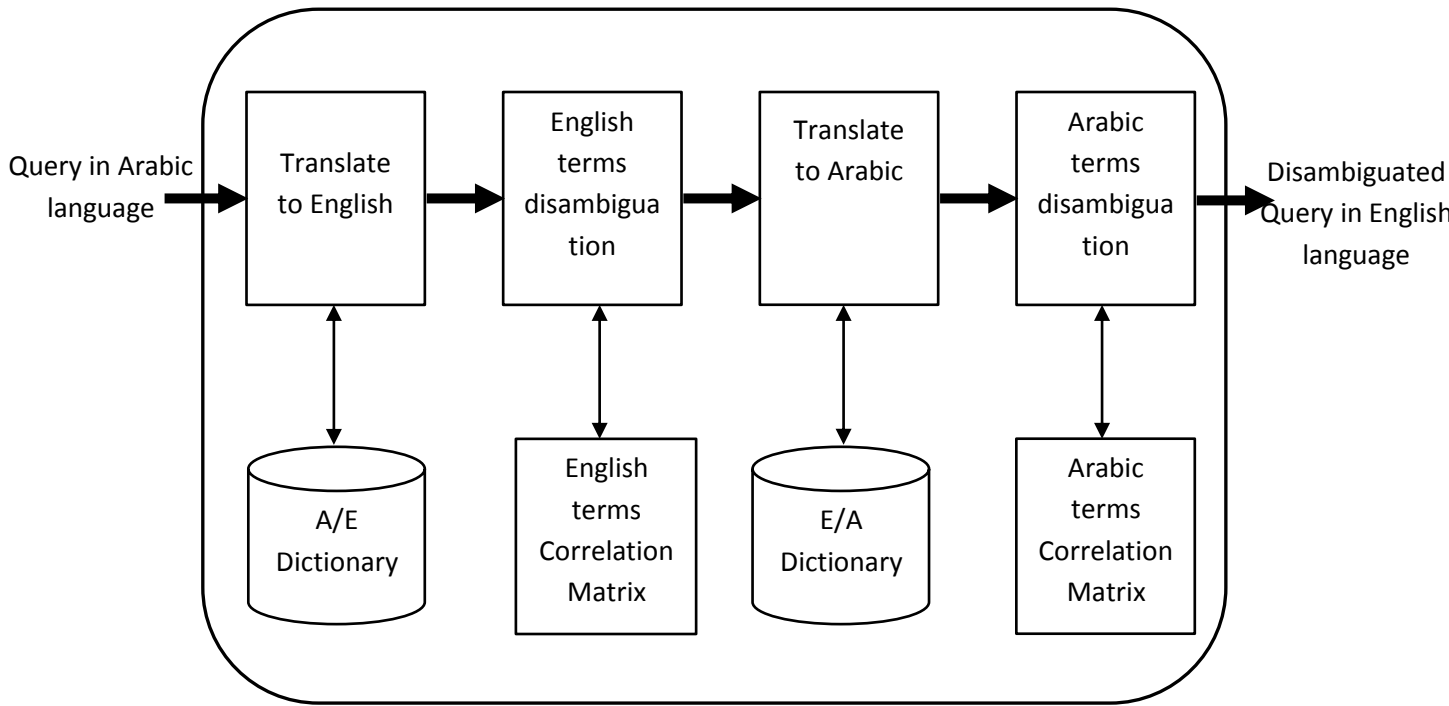


Figure 3.3: Flow of proposed translation method

Now, we can illustrate the previous steps with example. Suppose we have a query "بدأت في القرن التاسع عشر", the term "بدأت" can have these translation in English (started, began), and the translations of term "في" is (in, at, on etc.), "القرن" it can translated as (horn) of animals, or (century) means 100 years, and "التاسع عشر" is found in dictionary as phrase and translated as nineteenth.

And suppose correlation matrix as follow:

	began	Started	horn	century	Nineteenth
began	-	0.0	0.0	0.0	0.017
started	0.0	-	0.721	0.170	0.170
horn	0.0	0.721	-	0.530	0.830
century	0.0	0.170	0.530	-	0.970
nineteenth	0.017	0.170	0.830	0.970	-

Table 3.1: sample of correlation matrix

Suppose we have the following translation candidates and their average correlation: (began in nineteenth horn), (began in nineteenth century), (started in nineteenth horn), and (started in nineteenth century) have 0.4325, 0.4953, 0.50, and 0.57 correlation values respectively.

The three maximum correlated sentences (started in nineteenth century), (started in nineteenth horn), and (began in nineteenth century) are retranslated back to Arabic and the sentence that produce the original query is selected as translation. If more than one sentence produce the original query we select the highest correlated one.

You can note that the term "in" dose not appears in correlation matrix because such terms are frequently appear in documents, thus they have strong correlation with each term and then don't used to distinguished between the translation candidates.

Some Arabic phrases translated to single English term. for example, numbers from eleven to nineteenth, such as "التاسع عشر" is translated to nineteenth and "الخامس عشر" is translated to fifteenth, another phrases such as "وحيد القرن" is translated to rhino, and so on. such phrases and its translations are existed in dictionaries.

In our proposed algorithm this fact was taken into account. We search for two terms as phrase, if we found the phrase in dictionary we take the corresponding meaning, elsewhere we search for each term independently and then calculate correlation value between them.

Implementing this process has two benefits, the first one is reduce number of translation candidates for each query contain such phrases and then reduce computation. The second benefit is: by translating phrases we get more accurate senses than translating each term independently.

# **Chapter Four: Experiment Tools And Evaluation**

## **4.1. Introduction**

This chapter illustrates the tools and measures that selected to build and evaluate the proposed algorithm. Section 4.2 describe the test collection, section 4.3 explain the tool that used to build correlation matrix, and the IR system that used to index and retrieve documents, section 4.4 discuss the experiments and results.

## **4.2. Test Collection**

As mentioned in section 2.2.4, To measure effectiveness of information retrieval system in standard way, we must have an evaluation corpus which consisting of document collection, set of test queries represent the information needs, and set of relevance judgments, standardly a binary assessment of either relevant or nonrelevant for each query-document pair. In the next three subsection we discuss the three test collection components that used in this research.

### **4.1.1.Document Set**

In this experiment two document sets is used one for English and another for Arabic. The first one is used to build correlation matrix between the English terms, and evaluation the effectiveness of proposed algorithm. The Arabic document set used to build correlation matrix between Arabic terms which is just used in translation in the opposite direction.

The English data set that used in this experiment was taken from 20 newsgroup standard corpus in addition to some documents collected manually for some queries not covered in the standard corpus. Arabic document set was taken from watan document set in addition to manually collected document to satisfy each query not covered in document set.

To prepare the document set, two processes was done on it. The first process is to remove all stopwords from document set because it appears frequently in each document thus it cannot distinguishes between candidates. The second process is



split documents into fixed size windows to overcome the problem of inaccurate results that can occur if the document have varying length. Table (4.1) describe statistics about documents set.

Description	Details	Numbers
Number of documents	Arabic documents	499
	English documents	2,132
Number of words	Arabic documents	312,358
	English documents	748,543
Number of distinct words	Arabic documents	34,422
	English documents	28,998

Table 4.1: Statistic about document set

#### 4.1.2. Query Set

Query set is second component of test collection, we select number of queries contain same terms with more than one totally different meaning. Set of queries are listed in the table (4.2).

#	Query	Equivalent in English
Q01	بدأت في القرن التاسع عشر	Started in nineteenth century
Q02	قرن الحيوان	Animal horn
Q03	وحيد القرن	Rhino
Q04	العصر الحديث	Modern era
Q05	سمك القرش	Shark fish
Q06	سمك الخشب	Wood thickness
Q07	ضفة النهر	River bank
Q08	اخذ قرض من البنك	Take a loan from the bank
Q09	مجلس الامن	Security council
Q10	جناح الطائرة	The plane wing
Q11	جناح الفندق فخم	Luxury hotel suite
Q12	اليد اليمنى	Right hand
Q13	حق المواطن	Citizen right
Q14	دار الرعاية الاجتماعية	Social care home
Q15	دار القمر حول الارض	The moon revolved around the earth

Table 4.2: set of queries used in experiment

### **4.3. Tools and techniques**

To test and assess proposed algorithm we use RapidMiner. RapidMiner is a free open source integrated environment written in java. It provides machine learning, data mining and text analysis procedures including data loading and transformation, data preprocessing and visualization, modeling, evaluation and deployment.

### **4.4. Experiments and Results**

This section describes the experiments that were carried out to evaluate the developed method for translation in a cross language information retrieval. It also compare this method to Google translator as lower baseline.

The experiment is done by translating the query set twice, one by Google translator, and another by proposed method. Then the translated queries are issued to the same retrieval model Google search engine, figure (4.1) illustrate experiment steps. The retrieved list is assessed and the following results in table (4.3) and figure (4.2) was recorded.

The following table, table (4.3) shows the performance obtained from Google translation and the proposed translation. Values in the table are presented in an average DCG at document cut-off levels from 1 to 10 and the first top 10 documents retrieved are used for the final performance evaluation. The use of all points (from 1..10) are provided for drawing the curves for each point. Figure (4.1) plots the results of these two baseline runs together in a single graph

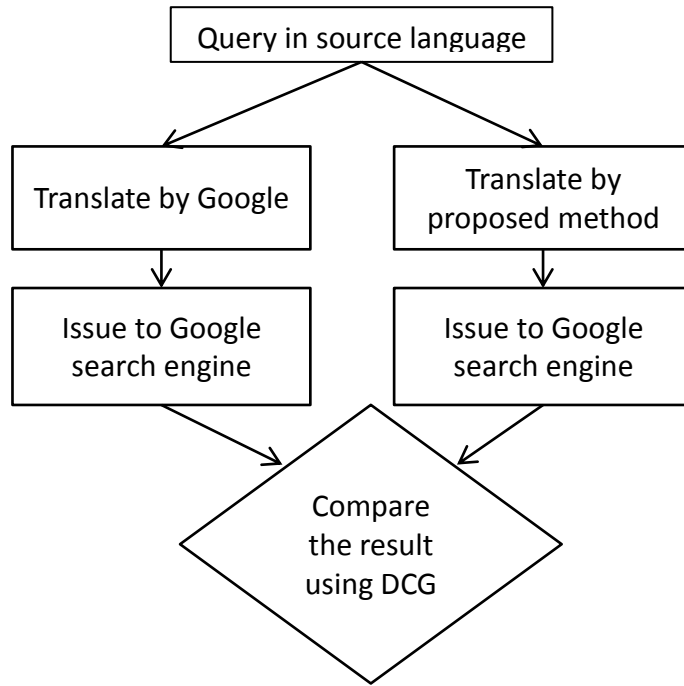


Figure 4.1: Experiment steps.

Measure	Average DCG									
Points	1	2	3	4	5	6	7	8	9	10
Google	2.33	4.53	5.84	6.67	7.47	8.22	8.93	9.58	10.04	10.60
Proposed	2.73	5.33	6.97	8.04	9.07	9.98	10.83	11.52	12.11	12.71

Table 4.3: Average DCG values of Google and proposed baselines.

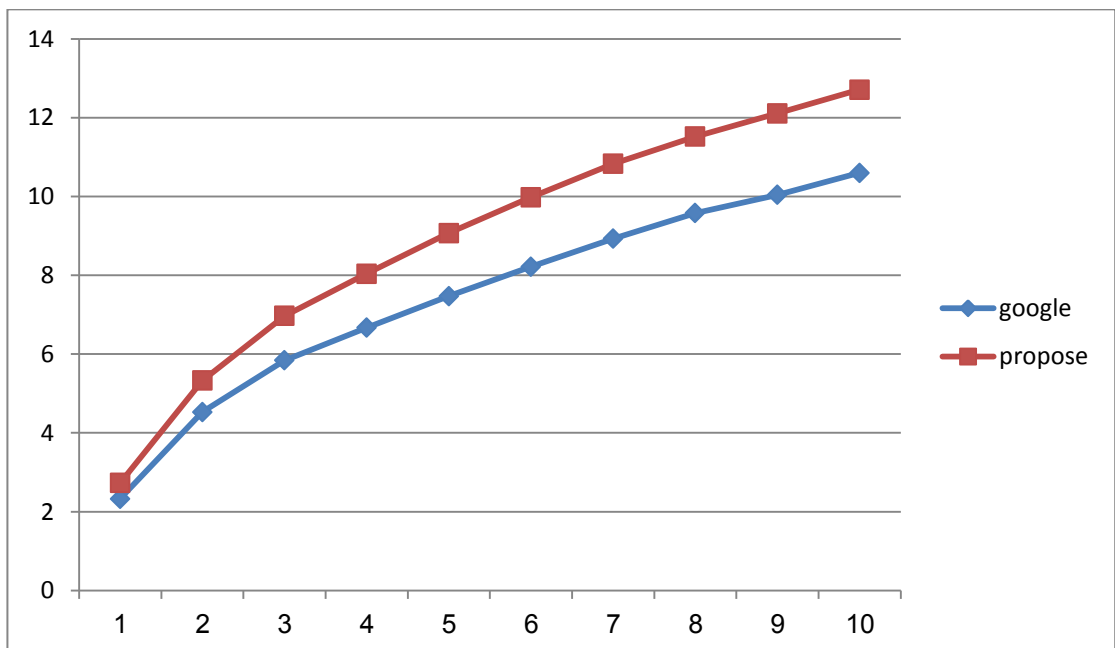


Figure 4.2: Average DCG values of two baselines in single graph.

The DCG values across all the queries were averaged and the statistical Student's t-test measure was used to compare significance of differences among the conducted experiments. The difference in retrieval scores is statistically significant (p-value < 0.000000835) at top 10 documents, using the Student t-test significance measure.

# Chapter Five: Conclusion And Future Work

## 5.1. Conclusion

Users sometimes need to retrieve information in language differ from query language. Hence, the query must be translated firstly and then passed to retrieval system. Translation using dictionaries is simple but it suffer from translation ambiguity because the dictionary provide more than one sense for a term.

In this research the ambiguity problem was resolved by co-occurrence statistics, and some improvement was gained as showed in section 4.4. the proposed method also works well in case of diacritics and homonymous.

## 5.2. Limitation

Although the proposed approach showed significant improvement for translation, there are some major limitation. The proposed query translation method selects the best translation among all translation candidates. However, for queries with many query terms, it would be too expensive to construct all possible candidate target queries. if a query consists of  $n$  terms and each term has  $m$  translation equivalents on average, there are  $m^n$  candidate target queries. For example, if a query consists of 20 terms and each term has 10 translation equivalents on average, there are  $10^{20}$  candidate target queries.

## 5.3. Future Work

Due to scope and time constraints of this research, they are many issues can be added in the future researches:

1. Decrease number of translation candidates for each term by determining the position of term in the sentences. for example, is a word is verb or noun. This can reduce number of translation candidates up to half.
2. Resolve the problem of out of vocabulary terms by transliteration or any technique used to this purpose.

## References

- [1] Oard, Douglas W. and Hackett, Paul. (1997). Document Translation for Cross-Language Text Retrieval at the University of Maryland. In Proceeding of the Sixth Text Retrieval Conference (TREC-6). Gaithersburg, MD: NIST.
- [2] Sheridan, P., and Ballerini, J. P. (1996). Experiments in Multilingual Information Retrieval using the SPIDER System. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zürich, Switzerland: ACM Press.
- [3] Adriani, Mirna and Croft, W. Bruce. (1997). The Effectiveness of a Dictionary-Based Technique for Indonesian-English Cross-Language Text Retrieval. CIIR Technical Report IR-170, University of Massachusetts, Amherst.
- [4] Nie, J. Y. (2010). Cross-language information retrieval. Synthesis Lectures on Human Language Technologies
- [5] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze (2009). An Introduction to Information Retrieval. Cambridge University Press, Cambridge, England.
- [6] Ramez Elmasri, Shamkant B. Navathe.(2010). Fundamentals of Database Systems, sixth edition. The University of Texas at Arlington, Georgia Institute of Technology.
- [7] Mohammed Mustafa Ali. (2013). Mixed-Language Arabic- English Information Retrieval. University of Cape Town. Cape town.
- [8] Stefan Buettcher, Charles L. A. Clarke, Gordon V. Cormack. (2010) Information Retrieval: Implementing and Evaluating Search Engines. The MIT Press, Cambridge Massachusetts, London, England.
- [9] Ahmed, F. and A. Nürnberger (2008). Arabic/English word translation disambiguation using parallel corpora and matching schemes. Proceedings of EAMT.
- [10] Gao, J.Y. Nie, E. Xun, J. Zhang, M. Zhou and C. Huang (2001). Improving query translation for cross-language information retrieval using statistical models. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM

- [11] Akira Maeda, Fatiha Sadat, Masatoshi Yoshikawa, and Shunsuke Uemura.(2000) Query term disambiguation for Web cross-language information retrieval using a search engine. in Proceedings of the fifth international workshop on Information retrieval with Asian languages. ACM.
- [12] Mirna Adriani. (1999). Using Statistical Term Similarity for Sense Disambiguation in Cross-Language Information Retrieval. *Information retrieval* 2(1): 71-82.
- [13] Lisa Ballesteros, W. Bruce Croft (1998). Resolving ambiguity for cross-language retrieval. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, ACM.
- [14] Zhou, D., M. Truran, T. Brailsford, V. Wade and H. Ashman (2012). "Translation techniques in cross-language information retrieval." *ACM Computing Surveys (CSUR)* 45(1): 1.
- [15] Pirkola, A. (1998). The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, ACM.
- [16] Ballesteros, L. A. (2001). Resolving ambiguity for cross-language information retrieval: A dictionary approach, University of Massachusetts Amherst.
- [17] Aljlayl, M. & Frieder, O. & Grossman, D. (2002). On bidirectional English–Arabic search. *Journal of the American Society for Information Science and Technology*, 53, 1139-1151.
- [18] Saravanan, K., R. Udupa and A. Kumaran (2013). Improving Cross-Language Information Retrieval by Transliteration Mining and Generation. *Multilingual Information Access in South Asian Languages*, Springer: 310-333.
- [19] Cheng, P.-J., J.-W. Teng, R.-C. Chen, J.-H. Wang, W.-H. Lu and L.-F. Chien (2004). Translating unknown queries with web corpora for cross-language information retrieval. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, ACM.
- [20] Guthrie, J. A., L. Guthrie, Y. Wilks and H. Aidinejad (1991). Subject-dependent co-occurrence and word sense disambiguation. Proceedings of the 29th annual meeting on Association for Computational Linguistics, Association for Computational Linguistics.

- [21] Van rijbergen, C. & Sparck jones, K. (1975). Report on the need for and provision of an "ideal" information retrieval collection.
- [22] Buckley, C. 2000. The TREC-9 Query Track. eds. Voorhees, E. M. and Harman, D., In: Proceedings of the 9th Text REtrieval Conference (TREC-9). 2000. 81-85.
- [23] ROBERTSON, S. E. & JONES, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information science*, 27, 129-146.
- [24] WITTEN, I. H., MOFFAT, A. & BELL, T. C. (1999). *Managing gigabytes: compressing and indexing documents and images*, Morgan Kaufmann.
- [25] SPARCK JONES, K., WALKER, S. & ROBERTSON, S. E. 2000. A probabilistic model of information retrieval: development and comparative experiments: Part 1. *Information processing & management*, 36, 779-808.
- [26] WALKER, S., ROBERTSON, S. E., BOUGHANEM, M., JONES, G. J. & JONES, K. S. Year. Okapi at TREC-6 Automatic ad hoc, VLC, routing, filtering and QSDR. In: TREC, 1997. Citeseer, 125-136.