The Islamic University–Gaza
Research and Postgraduate Affairs
Faculty of Engineering
Master of Computer Engineering

الجامعـــــــة الإســـلاميـة – غـزة
شئون البحث العلمي والدراسات العليا
كـليـــــة الهنـــدســـة
ماجستير هندسة الحاسوب

Developing Interactive Cross Lingual Information Retrieval Tool

بناء أداة تفاعلية متعددة اللغات لاسترجاع المعلومات

Mohammed M S Mortaja

Supervisor

Mohammad A. Mikki

Professor of Computer Engineering

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Engineering

April/2017

إقـــــرار

أنا الموقع أدناه مقدم الرسالة التي تحمل العنوان:

Developing Interactive Cross Lingual Information Retrieval Tool

بناء أداة تفاعلية متعددة اللغات لاسترجاع المعلومات

أقر بأن ما اشـتملت عليه هذه الرسـالة إنما هو نتاج جهدي الخاص، باسـتثناء ما تمت الإشـارة إليه حيثما ورد، وأن هذه الرسالة ككل أو أي جزء منها لم يقدم من قبل الآخرين لنيل درجة أو لقب علمي أو بحثي لدى أي مؤسسـة تعليمية أو بحثية أخرى. وأن حقوق النشر محفوظة للجامعة الإسـلامية – غزة.

Declaration

I hereby certify that this submission is the result of my own work, except where otherwise acknowledged, and that this thesis (or any part of it) has not been submitted for a higher degree or quantification to any other university or institution. All copyrights are reserves to IUG.

Student's name:	محمد محمود مرتجى	اسم الطالب:
Signature:		التوقيع:
Date:	01 – 04 – 2017	التاريخ:

بسم الله الرحمن الرحيم

الجامعة الإسلامية غزة
The Islamic University of Gaza

مكتب نائب الرئيس للبحث العلمي والدراسات العليا
هاتف داخلي 1950

الرقم Ref: ج س غ/35/

التاريخ Date: 2017/04/01

نتيجة الحكم على أطروحة ماجستير

بناءً على موافقة شئون البحث العلمي والدراسات العليا بالجامعة الإسلامية بغزة على تشكيل لجنة الحكم على أطروحة الباحث/ **محمد محمود شكري مرتجى** لنيل درجة الماجستير في كلية *الهندسة* قسم **هندسة الحاسوب** وموضوعها:

بناء أداة تفاعلية متعددة اللغات لاسترجاع المعلومات

Developing interactive cross lingual information retrieval tool

وبعد المناقشة التي تمت اليوم السبت 04 رجب 1437هـ، الموافق 2017/04/01م الساعة الثانية ظهراً، اجتمعت لجنة الحكم على الأطروحة والمكونة من:

أ.د. محمد أمين مكي	مشرفاً و رئيساً	.....
د. أيمن أحمد أبو سمرة	مناقشاً داخلياً	.....
د. إيهاب صلاح زقوت	مناقشاً خارجياً	.....

وبعد المداولة أوصت اللجنة بمنح الباحث درجة الماجستير في كلية *الهندسة*/ قسم: **هندسة الحاسوب**.

واللجنة إذ تمنحه هذه الدرجة فإنها توصيه بتقوى الله ولزوم طاعته وأن يسخر علمه في خدمة دينه ووطنه.

والله ولي التوفيق،،،

نائب الرئيس لشئون البحث العلمي والدراسات العليا

أ.د. عبد الرؤوف علي المناعمة

Abstract

The growing requirement on the Internet have made users access to the information expressed in a language other than their own , which led to Cross lingual information retrieval (CLIR) .CLIR is established as a major topic in Information Retrieval (IR). One approach to CLIR uses different methods of translation to translate queries to documents and indexes in other languages. As queries submitted to search engines suffer lack of untranslatable query keys (i.e., words that the dictionary is missing) and translation ambiguity, which means difficulty in choosing between alternatives of translation. Our approach in this thesis is to build and develop the software tool (**MORTAJA-IR-TOOL**) , a new tool for retrieving information using programming JAVA language with JDK 1.6. This tool has many features, which is develop multiple systematic languages system to be use as a basis for translation when using CLIR, as well as the process of stemming the words entered in the query process as a stage preceding the translation process.

The evaluation of the proposed methodology translator of the query comparing it with the basic translation that uses readable dictionary automatically the percentage of improvement is 8.96%. The evaluation of the impact of the process of stemming the words entered in the query on the quality of the output process in the retrieval of matched data in other process the rate of improvement is 4.14%. Finally the rated output of the merger between the use of stemming methodology proposed and translation process (**MORTAJA-IR-TOOL**) which concluded that the proportion of advanced in the process of improvement in data rate of retrieval is 15.86%.

Keywords:

Cross lingual information retrieval, CLIR, Information Retrieval, IR, Translation, stemming.

الملخص

الاحتياجات المتنامية على شـبكة الإنترنت جعلت المسـتخدمين لهم حق الوصـول إلى المعلومات بلغة غير لغتهم الاصلية، مما يقودنا الى مصطلح عبور اللغات لاسترجاع المعلومات (CLIR). CLIR أنشئت كموضوع رئيسي في "استرجاع المعلومات" (IR).

نهج واحد ل CLIR يستخدم أساليب مختلفة للترجمة ومنها لترجمة الاستعلامات وترجمة الوثائق والفهارس في لغات أخرى. الاستفسارات والاستعلامات المقدمة لمحركات البحث تعاني من عدم وجود ترجمه لمفاتيح الاستعلام (أي أن العبارة مفقودة من القاموس) وايضا تعاني من غموض الترجمة، مما يعني صعوبة في الاختيار بين بدائل الترجمة.

في نهجنا في هذه الاطروحة تم بناء وتطوير الأداة البرمجية (**MORTAJA–IR–TOOL**) أداة جديدة لاسـترجاع المعلومات باسـتخدام لغة البرمجة JAVA مع JDK 1.6، وتمتلك هذه الأداة العديد من الميزات، حيث تم تطوير منظومة منهجية متعددة اللغات لاسـتخدامها كأسـاس للترجمة عند اسـتخدام CLIR، وكذلك عملية تجذير للكلمات المدخلة في عملية الاستعلام كمرحلة تسبق عملية الترجمة.

وتم تقييم الترجمة المنهجية المقترحة للاسـتعلام ومقارنتها مع الترجمة الأسـاسـية التي تسـتخدم قاموس مقروء اليا كأساس للترجمة في تجربة تركز على المستخدم وكانت نسبة التحسين 8.96% ، وكذلك يتم تقييم مدى تأثير عملية تجذير الكلمات المدخلة في عملية الاسـتعلام على جودة المخرجات في عملية اسـترجاع البيانات المتطابقة باللغة الاخرى وكانت نسـبة التحسـين 4.14% ، وفي النهاية تم تقييم ناتج عملية الدمج بين اسـتخدام التجذير والترجمة المنهجية المقترحة (**MORTAJA–IR–TOOL**) والتي خلصـت الى نسـبة متقدمة في عملية التحسـين في نسـبة البيانات المرجعة وكانت 15.86%.

Dedication

To my mother …

To my father …

To my brothers and sisters …

To my dear wife …

To my son and my daughters …

To My Friends…

To all who helped me, I dedicate this work.

Acknowledgment

Firstly, I thank Almighty ALLAH for making this work possible. Then, there are a number of people to whom I am greatly indebted, as without them this thesis might not been written.

To Prof. Mohammad A. Mikki for his guidance, support, and advice.

To my parents for providing me with the opportunity to be where I am. To my dear wife, dear friends Mahmoud El Zaalan & Ali Ali .Without them, none of this would be even possible to do. You have always been around supporting and encouraging me.

To my brothers, sisters, and friends for their encouragement, input and constructive criticism, which are priceless.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Chapter 1. Introduction

The number of Internet users has already exceeded three billions (Internet World Stats, 2016). As a result, Internet has become the major source of knowledge. Nowadays, people have access to various information about products, news, books, movies, public services, science, etc. This unstructured knowledge provided from various sources such as news feeds, encyclopedias, blogs, forums and social networks. However, such a huge data exceeds the human capacity of understanding. Consequently, users need to use Information Retrieval (IR) technologies to find information that is relevant to their needs.

The problem appears not only due to the amount of data, but also users are fluent in different languages face difficulties when they try to retrieve documents that are not written in their mother tongue or if they would like to search documents in all languages they can speak in order to cover more resources with a single query (Andreas, 2012).

For this reason, IR technologies should handle cross-lingual texts. Cross-Lingual Information Retrieval (CLIR) methods can find relevant information in cross-lingual texts. Cross-lingual IR provide new paradigms in searching documents through myriad varieties of languages across the world and it can be the baseline for searching not only among two languages but also in multiple

1.1 Information Retrieval (IR) Methods

In classical IR search engines, both the query and the retrieved documents are in the same language. The classical IR regards the documents in foreign language as the unwanted "noise" (Abusalah, Tait, & Oakes, 2007). These needs to introduce new area of IR, which takes into account all the documents, received regardless of the languages being used. This is where the bilingual, cross-lingual and multilingual IR plays a part. However, to perform these variants of IR, a variety of translation methods are required. These are described in the following sub-sections. Translation can be done to the query, the document or both when any retrieval system involved with many languages. Query translation involves translating the query to the target language. Document translation translates the document into the source language (i.e. the language used for the query).

There are various methods to translate query, document or both. There are three primary tools for translations are dictionaries, machine translation systems and parallel corpora. Query translation, typically, uses either dictionary based or corpus based translation. Document translation, for the most part, only uses machine translation (PothulaSujatha & Dhavachelvan, 2011 , October).

1.1.1 Dictionary Based

A bilingual dictionary is a list of words in the source language and their translation(s) in the target language. Optionally, these dictionaries have translation probabilities assigned that allow for disambiguation and weighting. There are plenty of bilingual dictionaries are available in the literature both in Indian and Foreign languages.

1.1.2 Machine Translation

The Machine translation method simply uses a machine translation system to translate either the document or query. The main drawback of this method is computational expensive. In situations where there is a large collection of documents or when searching for documents on the web, machine translation is impractical.

1.1.3 Parallel Corpora

When compared to dictionary-based corpus based translation typically gives much better performance, as (McNamee, P, & Mayfield, 2002) found. However, the creation of parallel corpora is complicated and quite expensive. It can be extremely difficult to find parallel corpora for certain languages or that are large enough to be of use.

The main problems with both corpus based and dictionary-based translation are coverage and quality. Poor quality corpora and dictionaries can greatly decrease the performance of a system (McNamee, P, & Mayfield, 2002). Coverage relates to out of vocabulary words, or words that are not present in the dictionary or corpus. These words will have no translation, while in some languages that are related this is no problem in other language pairs such as Chinese and English this is a big problem (Zhang & P, 2004). Because of this, there has been considerable research done on automatically or semi-automatically acquiring parallel corpora or bilingual lexicons.

The same methods are used for CLIR and MLIR. These two systems may use translation of all documents into a common language, either automatic translation of the queries or combination of both query and document translations.

1.2 Cross Language Information Retrieval (CLIR):

One area of IR that has seen a great deal of interest and has had many exciting advances made in it, is CLIR. The goal of CLIR is to allow users to make queries in one language and retrieve documents in one or more other languages. The resulting documents can then be translated into the language used for the query to allow the user to get the gist about the information retrieved. For example, a user makes a query in English about "flower arrangement" and receives documents back in Japanese about "Ikebana" which is Japanese flower arrangement. Most systems in CLIR use some type of translation. While there exist non-translation methods, such as cognate matching (Buckley, C., M., & J., 2000), latent semantic indexing (Dumais, S., T., & M., 1997), and relevance models (Lavrenko & V., 2002), here the predominate method is still translation. As such one of the main problems in CLIR is dealing with language translation. What should be translated, how should it be translated, and how to eliminate bad translations are some of the major areas of research in CLIR. In addition, how to acquire large enough amounts of translation data is also an active topic for research.

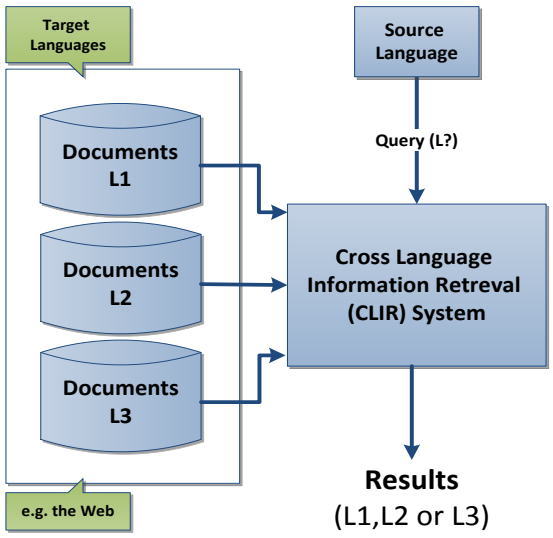

Figure (1.1): Cross-Language Information Retrieval (CLIR) system

1.3 Thesis Contribution

This thesis proposes a new information retrieval tool called **MORTAJA_IR_TOOL**, which include:

- Develop a new stemmer that solve the problem of irregular words, broken plural patterns and blind removal of affixes, using the combination of Table lookup stemmer processes & Affixes Removal stemmer processes.

- Overcome several of previous CLIR approaches limitations, especially the problem of untranslatable query keys, missing words and translation ambiguity
- Develop a new translation methodology, which depend on select the proper translation of bilingual dictionary depending of all words in the query.
- Translation methodology divided to several steps: Normalization, Stemming & Translation of terms.
- The query process by using the stemmer according to the query language, that mean the query will be normalized, stemmed, and then weighted so it convert to normalized words.
- Translation methodology overcome a lot of traditional translation limitations where in cross-language information retrieval, the input is often a combination of a series of keywords rather than a complete sentence, this sequence of query keywords lacks necessary contextual and syntactic semantic information, so they cannot be translated by traditional MT technology in a direct and easy way.
- How to compute the boost value of the translations in the bilingual dictionary of a given query keyword is the focus of the algorithm proposed in this thesis. Our approach based on large-scale bilingual corpora, and the computation is implemented by applying the theories of vector space model (VSM) and lexical mutual information to traditional IR.

1.4 Thesis Organization

The organization of the rest of this thesis is as follows:

Chapter 2 introduces the related work, where Chapter 3 give background about CLIR, Chapter 4 describes the methodology including the proposed stemmer and the new CLIR tool. Chapter 5 will show the results of the work, and Chapter 6 the conclusion of the research, which will summarize the research.

Chapter 2

Literature Review

Chapter 2. Literature Review

Most of the research in the field of information retrieval focused on the English language. There is a large amount of evaluation benchmarks for IR in English. The most basic is the Cranfield collection based on work (Cleverdon & C, 1967). It contains a set of information needs from a database of abstracts. TREC (Simpson, Voorhees, & Hersh, 2014) and CLEF (Suominen, et al., 2014) are the biggest series of evaluation campaigns focused on various tasks of IR. Multi-lingual (Peters, Braschler, & Clough, 2012) and cross-lingual IR are gaining a lot of attention, but most of the current evaluation databases contain just couple of the most commonly used languages such as Chinese or French.

Over the past few years, research in CLIR has progressed and a many systems have been developed. Some of the prominent systems of CLIR are as follows:

2.1 Mulinex

Mulinex (Capstick, et al., 2000) supports cross-lingual search by giving the users possibilities to formulate, expand and disambiguate queries. Furthermore, the users are able to filter the search results and read the retrieved documents by using only their native language. Mulinex performs the multilingual functionality based on a dictionary-based query translation. Besides the cross lingual functionality, where the query is submitted in one language and the retrieved documents are presented in another language, Mulinex provides the automatic translation of documents and their summaries. In Mulinex, three languages are supported, French, German, and English. In Mulinex, the CLIR process is fully supported by the translation of the queries, documents and their summaries. Hereby, users do not need to have any knowledge about the target language.

2.2 Keizai

The goal of the Keizai project (Oard, He, & Wang, 2008) is to provide a Web-based cross-language text retrieval system that accepts the query in English and searches Japanese and Korean web data. Furthermore, the system displays English summaries of the top ranking retrieved documents. In Keizai, the query terms are translated into Japanese or Korean languages along with their English definitions and thus this feature allows the

user to disambiguate the translations. Based on the English definitions of the translated query terms, the user who does not understand the Japanese or Korean language can select the appropriate translation, out of several possible translations. Once the user selects those translations whose definitions are consistent with the information needed, the search can be performed.

2.3 UCLIR

In UCLIR (Abdelali, Cowie, Farwell, & Ogden, 2003), the Arabic language was included. The system performs its task in any of the following three different modes: the first mode, using a multilingual query (query can consist of terms of different languages), the second mode using an English query without user involvement in the multilingual query formulation, the third mode using an English query with user involvement in the formulation of the multilingual queries. At the end, the user selects the appropriate translation out of the filtered translation list. The selected multilingual terms then can be used to form the multilingual query, which is then submitted to retrieve the relevant documents from the system's entire multilingual resource. After the retrieval process is performed, the relevant retrieved documents can be then translated into English. To perform the document translation, two approaches are used. The first approach is word-level translation, where the user can click on the selected word and this word will be translated using the dictionary and displayed as a popup view to the user with its lexical information. The second approach is a document-level translation, where the whole retrieved document, using a translation system, is translated into English. Similar to Keizai, UCLIR uses "Document Thumbnail Visualizations".

2.4 MIRACLE

In MIRACLE (Mayfield & McNamee, 2004), there are two types of query translations, fully automatic query translation (using machine translation) and user-assisted query translation. In other words, the user submits his query; the system provides him/her with translation alternatives. The interaction between the system and the user, gives the user possibilities to see the effect of his/her decision (selection, deselection of the translation or query refinement) in that the user can cycle the search till it satisfies his/her needs. The

query language is always English, in MIRACLE. However, language resources that are available for English can be leveraged, regardless of the document language. Currently, MIRACLE works with a simple bilingual term list. However, it is designed to readily leverage additional resources when they are available.

2.5 MultiLexExplorer

The goal of the MultiLexExplorer tool (Luca, Hauke, Nurnberger, & Schlechtweg, 2006) is to support multilingual users in performing their web search. Furthermore, the MultiLexExplorer supports the user in disambiguating word meanings by providing the user with information about the distribution of words in the web. The tool allows users to explore combinations of query term translations by visualizing EuroWordNet relations together with search results and search statistics obtained from web search engines. Based on the EuroWordNet, the tool supports the user a lot of functionality.

2.6 Multi Searcher

The "multi Searcher" tool (Ahmed & Nurnberger, 2010) provides users with interactive contextual Information that describes the translation in the user's own language, in order for him/her to have a certain degree of confidence about the translation. In order to consider users as an integral part of the retrieval process, the tool provides the users with possibilities to interact with it, where they can select relevant terms from the contextual information in order to improve the translation and thus improve the CLIR process. "Multi Searcher" deals with two issues concerning the CLIR. Firstly, there is translation ambiguity, where one word in one language can have several meanings in another language. "Multi Searcher" make the use of automatic translation with disambiguation (in the previously mentioned tools, the analyzer, the query terms are analyzed and the senses (possible translations) of the ambiguous query terms are identified. Second, the most likely correct senses of the ambiguous query terms are selected based on co-occurrence statistics.

2.7 Spider (EuroSpider)

The Spider (Braschler, 2004), later a commercialized product EuroSpider, developed at the Suiss Federal Institute of Technology, is a multilingual information retrieval based on thesaurus-based query expansion approach performed over a collection of comparable

multilingual documents.

The Eurospider retrieval system is based on fully automatic indexing (no manual indexing required). The EuroSpider system has been evaluated at TREC-5. It provides many functions of the new generation retrieval systems such as relevance ranking, word normalization, relevance feedback and automatic indexing. Eurospider's architecture allows powerful integration of database management systems and advanced retrieval functions. When adding the Eurospider system to an existing database system, the database applications do not have to be changed. A database system and the Eurospider retrieval system provide complementary functions constituting a good combination.

2.8 TRANSLIB (Tools for Accessing Multilingual Library Catalogues)

Tools for supporting multilingual access to library catalogs have been developed in the framework of the European Project TRANSLIB (TRANSLIB. Advanced Tools for Accessing Multilingual Library Catalogues, 1995). The project is coordinated by the KNOWLEDGE S.A. company in Greece. In this project, existing tools (e.g. machine translation, grammar parser and checker) and corpora (such as terminological databases, thesauri, and electronic dictionaries) have been evaluated.

The developed tools help a user to access OPACs in English, Greek and Spanish. Therefore, it supports three languages (English, Greek and Spanish). The search tools are based on the multilingual thesaurus EUROVOC. The TRANSLIB system has been evaluated by Greek and Spanish librarians and has shown that improvements in user interface are necessary.

2.9 Statistical Transliteration for English-Arabic Cross Language Information Retrieval (selected n-gram model)

Out of vocabulary (OOV) words are problematic for cross language information retrieval. One way to deal with OOV words when the two languages have different alphabets, is to Transliterate the unknown words, that is, to render them in the orthography of the second language. In the present study, it present a simple statistical technique to train an English to Arabic transliteration model from pairs of names. It call this a selected n-gram model because a two-stage training procedure first learns which n-gram segments should be

added to the unigram inventory for the source language, and then a second stage learns the translation model over this inventory.

This technique requires no heuristics or linguistic knowledge of either language. It evaluate the statistically-trained model and a simpler hand-crafted model on a test set of named entities from the Arabic AFP corpus and demonstrate that they perform better than two online translation sources. It also explore the effectiveness of these systems on the TREC 2002 cross language IR task. It find that transliteration either of OOV named entities or of all OOV words is an effective approach for cross language IR.

2.10 Using the Web for Automated Translation Extraction in Cross-Language Information Retrieval

There have been significant advances in Cross-Language Information Retrieval (CLIR) in recent years. One of the major remaining reasons that CLIR does not perform as well as monolingual retrieval is the presence of out of vocabulary (OOV) terms. Previous work has either relied on manual intervention or has only been partially successful in solving this problem. We use a method that extends earlier work in this area by augmenting this with statistical analysis, and corpus-based translation disambiguation to dynamically discover translations of OOV terms. The method can be applied to both Chinese-English and English-Chinese CLIR, correctly extracting translations of OOV terms from the Web automatically, and thus is a significant improvement on earlier work.

2.11 Triangulated Translation

Most approaches to cross language information retrieval assume that resources providing a direct translation between the query and document languages exist. This paper presents research examining the situation where such an assumption is false. Here, an intermediate (or pivot) language provides a means of transitive translation of the query language to that of the document via the pivot, at the cost, however, of introducing much error. The paper reports the novel approach of translating in parallel across multiple intermediate languages and fusing the results. Such a technique removes the error, raising the effectiveness of the tested retrieval system, up to and possibly above the level expected,

had a direct translation route existed. Across a number of retrieval situations and combinations of languages, the approach proves to be highly effective.

2.12 Fast Document Aligner using Word Embedding (FaDA)

FaDA is a free/open-source tool for aligning multilingual documents. It employs a novel cross lingual information retrieval (CLIR)-based document-alignment algorithm involving the distances between embedded word vectors in combination with the word overlap between the source language and the target-language documents. In this approach, it initially construct a pseudo-query from a source-language document. It then represent the target-language documents and the pseudo-query as word vectors to find the average similarity measure between them. This word vector-based similarity measure is combine with the term overlap-based similarity. Its initial experiments show that s standard Statistical Machine Translation (SMT)-based approach is outperform by its CLIR-based approach in finding the correct alignment pairs. In addition to this, subsequent experiments with the word vector-based method show further improvements in the performance of the system.

2.13 Slovak News Information Retrieval

This work proposes an information retrieval evaluation set for the Slovak language. A set of 80 queries written in the natural language is given together with the set of relevant documents. The document set contains 3980 newspaper articles sorted into 6 categories. Each document in the result set is manually annotated for relevancy with its corresponding query. The evaluation set is mostly compatible with the Cranfield test collection using the same methodology for queries and annotation of relevancy. In addition to that it provides annotation for document title, author, publication date and category that can be used for evaluation of automatic document clustering and categorization.

Table (2.1): An Overview of CLIR and MLIR Research

Authors	Languages	Method/Technique	Evaluation Initiatives
Fujii, A., Ishikawa, T. (Fujii & Ishikawa, 2001)	J to E and E to J	Query translation and Document Translation	NTCIR -2 Collection
Jialun Qin, Yilu Zhou, Michael Chau & Hsinchun Chen (Qin, Zhou, Chau, & Chen, 2003)	E to Ch	Dictionary based query translation	TREC Collection
David A. Hull& Gregory Grefenstette (Hull & Grefenstette, 1996)	E to F	Dictionary based query translation	Documents Collection
Chen-Yu Su, Tien-Chien Lin & Shih-Hung Wu (Chen-Yu, Tien-Chien, & Shih-Hung, 2007)	Ch to j and K	Dictionary based query translation	NTCIR -6
Paraic Sheridan& Jean Paul Ballerini (Paraic & Jean, 1996)	G to I	Thesaurus-based query expansion	Documents Collection
Wen-Cheng Lin & Hsin-HsiChen (Wen-Cheng & Hsin-HsiChen, 2003)	J to E and Ch	Query translation	NTCIR -3
Peter A. Chew & Ahmed Abdelali (Peter & AhmedAbdelali, 2008)	E, R, S, F and A	Latent Semantic Analysis	Bible and Quran data
Mizera-Pietraszko J (Mizera-Pietraszko, 2009)	E to F and F to E	Mete data search	Documents Collection
Turdi Tohti, Winira Musajan & Askar Hamdulla, (TurdiTohti, WiniraMusajan, & AskarHamdulla, 2008)	Uyghur, Kazak,Kyrgyz	Query phase reconstruction, character coding	Website data

Marshall Ramsey, Thian-Huat Ong & Hsinchun Chen (Marshall, Thian-HuatOng, & Hsinchun, 1998)	Ch and J	Dictionary-lookup, phonetic, radical, and mnemonic	Training data
Dong-Mo Zhang, Huan-Ye Sheng, Fang Li & Tian-Fang Yao (Dong-Mo, Sheng, Fang, & Tian-Fang, 2002)	E, G and Ch	Case based reasoning and machine learning	Documents
Kazuyuki Yoshinaga, Takao Terano & NingZhong (Kazuyuki, Takao, & NingZhong, 1999)	J and E	Web Information Collector, Document classifier, Ontology generator and Search engine	Web documents
Hsin-Chang Yang & Chung-Hong Lee (Hsin-Chang & Chung-Hong, 2008)	E and Ch	Parallel corpora	Bilingual corpus documents
Chung-hsinLin & Hsinchun Chen (Chung-hsinLin & Hsinchun, 1996)	Ch and E	Indexing and Classification approach	Multilingual Databases
Jeffrey A. Rydberg-Cox, Lara Vetter, Stefan M.Rüger& Daniel Heesch. (Jeffrey, Lara, Stefan, & Daniel, 2004)	Greek, Latin and Old Norse	Query translation	Search engine results
Shuang-Qing Yuan, Fang Li & Huan- Ye Sheng (Shuang-Qing, Fang, & Huan-Ye, 2002)	Ch and E	Novel approach for finding terminology translations from hyperlinks	Website Links (parallel or unparallel corpus)
Akiko Aizawa (Akiko, 2002)	E and J	Evolutionary framework	NTCIR –J1

Table 2.1 describe the foreign languages which are involved in the CLIR/MLIR system, the translation technique/method and finally the evaluation initiatives used in the research work.

Chapter 3

Background

Chapter 3. Background

3.1 Application areas of CLIR

The core field of information retrieval where research on CLIR needed for effective results are (Sanjay & Ganesh, 2016):

3.1.1. Medical application and researches on CLIR

A number of resources available on Web provide the public and healthcare professionals with the most up-to-date findings in medical research, such as PubMed (Chang, Weng, Lin, Hwang, & Oyang, 2006)and MedlinePlus (Shatkay & Hagit, 2005).

Medline Plus is a Web-based consumer health information resource, made available by the National Library of Medicine. PubMed first released in 1996, is a free search engine for accessing the Medline database of life sciences and biomedical topics.

Most of the high level quality resources that are freely available and unlimited for users all around the world are available only in English language. Therefore Non-English users encounter a great language barrier when trying to access medical information from these websites such are also not familiar with medical terminology even in their first language (native language). So there is a big platform for researcher to work on medical information retrieval system, in order solve the problem of language barrier.

3.1.2. Multimedia application and researches on CLIR

Multimedia Information Retrieval (MMIR or MIR) is a hot research discipline whose objective is to extract the semantic information from multimedia data sources such as audio, video, and image. MMIR implies that multiple channels are employed for the understanding of media content, each of these channels are described by media-specific features transformations.

The first version of the Multilingual Multimedia Information Retrieval (MMIR) prototype involves short videos in the domain of news, that are selected from online web TV channels, from UGC portals, or from online news agencies. There has so far been

very little work in the area of Cross-Language Multimedia Information Retrieval (CLMIR). This is an important future research topic as the growth of multilingual and multimedia document collections is likely to lead inevitably to the growth of multilingual multimedia collections.

3.1.3. Mobile Network application and researches on CLIR

This research proposes a Cross-Lingual Information Retrieval approach that is used to search Internet resources for appropriate content and summarize it into another form using the content specification meta-language. This content is then mapped to the target language.

3.1.4. Video Question-Answering system- application and researches on CLIR

Question/Answering on multi-media is a new research issue in recent years. The cross-language QA system have some fundamental problems like video processing, i.e. video Optical Character Recognition (OCR) and video segmentation.

3.1.5. Enterprise Competition on CLIR

Along with the economic globalization, the information resource in a modern society becomes an important element for modern enterprises competition. CLIR is introduced to the enterprise competitive intelligence collections can effectively resolve the low recall and veracity rate of intelligence collections to some extent and promote the development of CLIR in the enterprise competition intelligence.

3.2 Challenges in CLIR

Queries from users are often too short, which produce more ambiguity in query translation, and reduce the accuracy of the cross language retrieval results. Since the problem of language mismatch in CLIR are more serious than in monolingual IR, it is necessary to exploit techniques for improving the multilingual retrieval performance. In CLIR systems, users often present their query in their native language, and then the system automatically searches documents written in other languages. Therefore, it is a

challenge for CLIR to conquer the barrier between the source language (SL) in query sentences and the target language (TL) in documents to be searched. As discussed in the previous section, most CLIR systems utilize MT technology to resolve this problem. As MT research itself has a number of issues (such as accuracy), the research in CLIR also faces critical issues and challenges that must be addressed (Sanjay & Ganesh, 2016).

3.2.1. Ambiguity in IR

Ambiguity occurs when words have multiple meaning which also referred to as homonymy or polysemy. Ambiguity in IR are semantic and syntactic in nature, where as ambiguity in CLIR are semantic and lexical. So the probability of occurrence of ambiguity in CLIR is higher than normal IR, due to the availability of different languages.

3.2.2. Effective User Feedback about IR

Effective user functionality can be incorporated by the user feedback, about their requirements and information needs. It should also provide readable translations of the retrieved documents to support document selection. System should also provide better support for query formulation and reformulation based on some set of intermediate results.

3.2.3. Complexity in New IR Applications

Question/Answering is relatively a new stream of Information Retrieval. In Question/Answering end-users throw a question in a form of query and retrieve answers for that. However, challenge is to retrieve answers of English questions in different language.

3.2.4. Specialized Terminology and Proper Nouns

Specialized terminology, such as scientific names, is often difficult to translate and is often found in specialized dictionaries or term banks. Specialized terminology tends to be less ambiguous than regular vocabulary although regular vocabulary can have a specialized meaning when used in a certain subject area.

3.3 Query Translation Approach

A major challenge in CLIR is to bridge the language gap between query and documents. Query translation is now serving as a major cross-lingual mechanism in current CLIR systems as shown in figure3.1. CLIR search engines enable users to retrieve content in a language different from language used to formulate the query. Translation of query has the advantage that the computational effort i.e. time and space, is less as compared with other methods (Sanjay & Ganesh, 2016).Query translation has following disadvantages:

(1) Usually a query does not provide enough contexts to automatically find the intended meaning of each term in the query.

(2) Translation errors affect retrieval performance sensibly.

(3) In case of searching a multilingual database, query has to translate into each one of the languages of database.

In CLIR query translation, play an important role that can achieved by following approaches: dictionary based translation approach, corpora based translation approach and machine translation based approach.

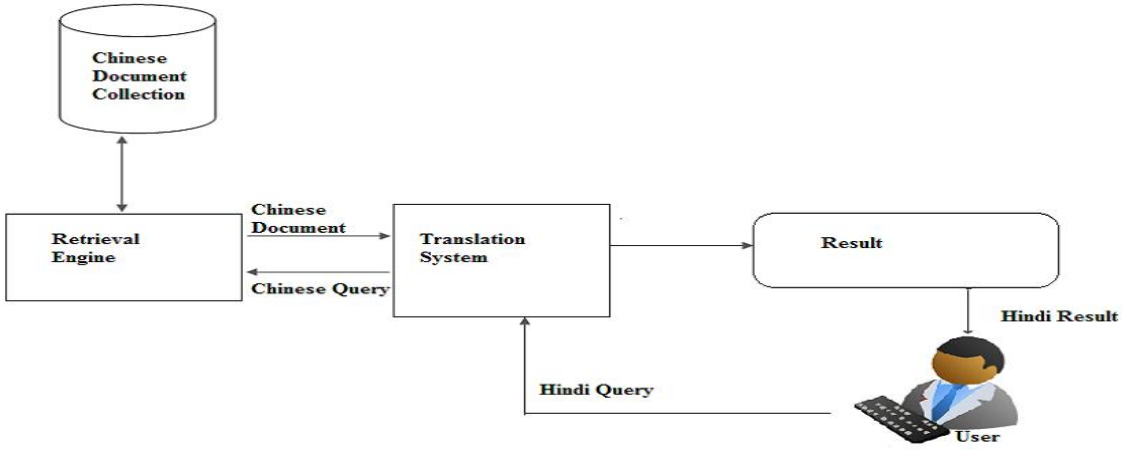

Figure (3.1) : Query Translation Approach

3.3.1. Dictionary Based Translation Approach

In dictionary-based query translation, the query will be processed linguistically and only keywords are translating using Machine Readable Dictionaries (MRD), given in figure 3.2. MRDs are electronic versions of printed dictionaries, either in general domain or specific domain. The use of existing linguistics resources, especially the MRDs, is a natural approach to cross-lingual IR. Translating the query using the dictionaries is much faster and simpler than translating the documents. Some common problems associated with dictionary-based translation are:

(i) Untranslatable words (like new compound words, proper names, spelling variants, and special terms): Not every form of words used in query is always found in dictionary. Sometime problem occurs in translating different compound words (formed by combination of new words) due to the unavailability of their proper translation in dictionary.

(ii) Processing of inflected words: Inflected word forms are usually not found in dictionaries.

(iii) Lexical ambiguity in source and target languages: Relevant forms of lexical meaning for information retrieval are: 1) homonymous and 2) polysemous words. Two words are homonymous; if they have at least two different meanings and senses of words are unrelated e.g. bank (river bank) and bank (financial institution). Polysemous words should have related senses e.g. star in the sky and star. Due to ambiguity in the search keys, matching for retrieving relevant documents may not be successful.

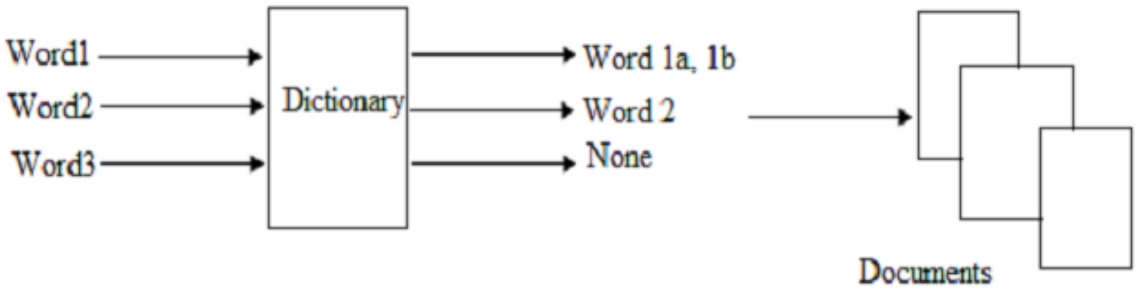

Figure (3.2): Dictionary Based Translation Approach

3.3.2. Corpora Based Translation Approach

Query translation using corpora requires single corpus or many corpuses. Corpora, (plural of corpus) are the systematic collection of naturally occurring language material, such as texts, paragraphs and sentences from one or many languages. In corpus-based methods queries are translated on the basis of multilingual terms extracted from parallel or comparable document collections. A parallel corpus has been used since the early 1990‟s for translation of given word.

A parallel corpus is a collection of texts, each of which is translated into one or more languages other than the original language. Parallel corpora are also used to decide the relationships, such as co-occurrences, between terms of different languages. A parallel corpus is an important kind of source of linguistic meta-knowledge, which forms the basis of techniques such as tokenization, morphological and syntactic analysis.

A comparable corpus is one of the important concepts in corpus-based translation study, introduced by Baker (Fernandes & Lincoln, 2006). Comparable corpora contain text in more than one language. The texts in each language are not translations of each other, but cover the same topic area, and hence contain an equivalent vocabulary. A good example of corpora is the multilingual news feeds produced by news agencies such as Reuters, CNN, BBC, Xinhua News and BERNAMA. Such texts are widely available on the Web for many language pairs and domains. They often contain many sentence pair that are good translations of each other.

3.3.3. Machine Translation Based Approach

Cross-lingual IR with query translation using machine translation seems to be an obvious choice compared to the other two above, as shown in figure3.3. The advantages of using the machine translation is that it saves time while translating large texts. Manning and Schutze (Manning & Schutze, 1999) distinguished four different approaches to deal with machine translation: (a) Word-for-word approach, (b) Syntactic transfer approach, (c) Semantic transfer approach, and (d) Interlingual approach. The ultimate goal of CLIR machine translation (MT) systems is to translate queries from one language to another by using a context. Many factors contribute the difficulties of machine translation, including words with multiple

meanings, sentences with multiple grammatical structures, uncertainty about what a pronoun refers to, and other problems of grammar.

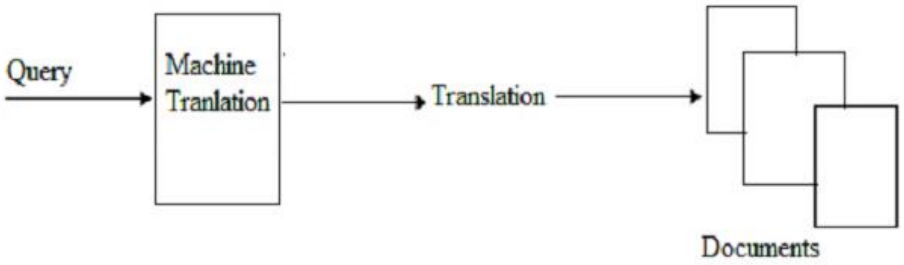

Figure (3.3) : Machine Translation Based Approach

Many researchers criticize MT-based CLIR approach. The reasons behind their criticisms mostly stem from the fact that the current translation quality of MT is poor. Another reason is that MT systems are expensive to develop and their application degrades the retrieval efficiency (run time performance) due to the lengthy processing times associated with linguistic analysis.

MT based approaches seems to be the ideal solution for CLIR. It is mainly because MT systems translate the sentence as a whole, and the translation ambiguity problem is solved during the analysis of the source sentence.

3.4. Document Translation Approach

Document Translation can be the most desirable scenario in CLIR (Sanjay & Ganesh, 2016), if the purpose is to allow the users to search the documents different from their own language and receive results back in user's language, as given in figure (3.4). In this sense, it is truly a better option which does not require a passive knowledge of the foreign language from the user. In document translation approach, all target languages are translated to the source language . The function of this translation is twofold. First, post translation or „as-and-when-needed" or „on-the–fly translation", where documents of any other language being searched by user are translated into user language at query time. IR process mostly uses indexing technique to speed up the searching process of documents. But indexing is not possible in post translation, so this approach is infeasible because it requires more time for translation.

Second, pre translation or „all together before any query is processed" used to browse through a translated version of an original translation in user language or in a language which user can understand figure (3.4).

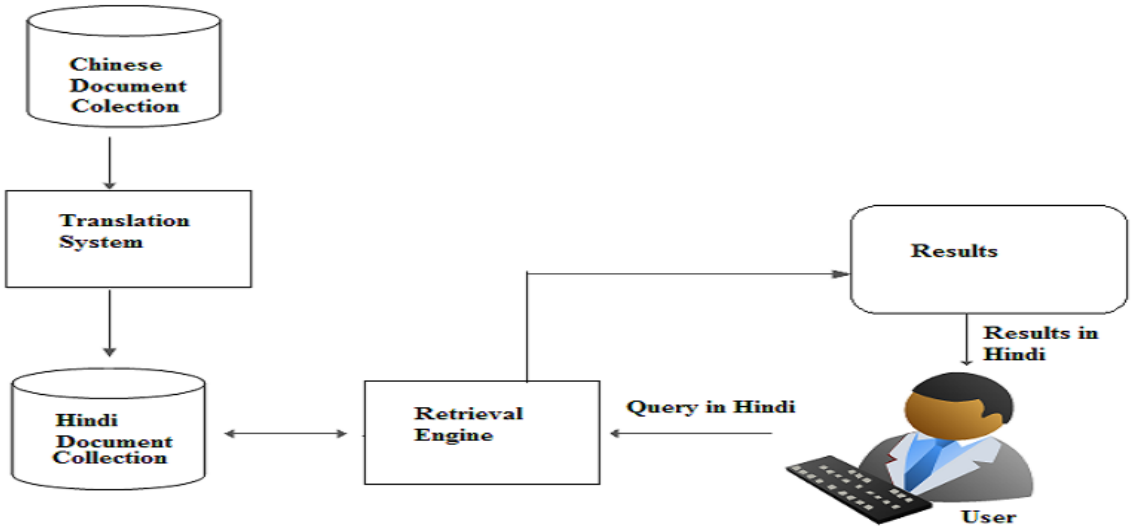

Figure (3.4): Offline Translation

This translation can be called as offline translation. In this approach, documents that are written in different languages are translated to all desired source languages and these documents are indexed before query time. This translation is impossible as a solution for large collection of distributed documents, which are managed by different groups of people, for example internet. Document translation has its own advantages and disadvantages compared to query translation. Some researchers have used it to translate large sets of documents (e.g.,Braschler & Schauble, 2001 ; Franz, Scott McCarley, & Todd Ward, 2000 ; Oard & Hackett, 1998) since more varied context within each document is available for translation, which can improve translation quality. The document translation approach has certain benefit over query translation. These include the following:

(i) A long document provides more contexts to perform translation, so that terms in the target language can be chosen more accurately.

(ii) Translations errors should not harm retrieval too much, as they are weighted against a whole document.

(iii) The translation effort is done at indexing time, thus getting faster retrieval at run time.

However, there are certain issues with document translation as well, such as:

(i) Much more computational effort is needed to index collections.

(ii) Bad scaling performed in case of more than two languages.

3.5 Dual Translation (Both Query and Document Translation Approach)

In this approach – both queries and documents are translated into a common representation figure (3.5) (Sanjay & Ganesh, 2016). This approach requires additional storage space for translated documents but provides scalability when same collection of documents is required in multiple languages. One of the examples of such approach is controlled vocabulary systems . These systems represent all documents using a pre-defined list of language-independent concepts, and enforce queries in the same concept space. This concept space defines the granularity or precision of possible searching.

The major issue of controlled vocabulary systems is that, non-expert users usually require some training and require interfaces to the vocabularies in orderable to generate effective queries.

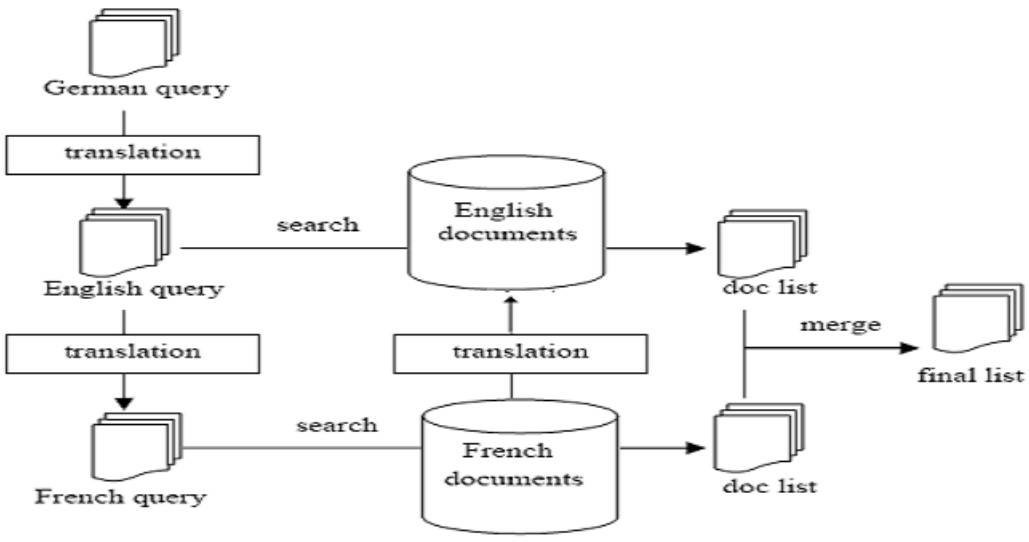

Figure (3.5): Translate both document and query into pivot language

Dual translation approach also called as hybrid translation approach can be performed by pivot language. Direct translation between two languages may not always be possible due to the limitation of translation resources. To perform such type of translation, a resources or a third language is required between these languages, called pivot language. In this process, two types of approaches are possible: either the query or the document is translated first to pivot language, then to the target language; translate both document and query into pivot language as shown in figure (3.5).

3.6 Comparative Study of the Three Approaches

The need for translation has itself been questioned because non-translation based methods of CLIR, such as cognate-matching and cross-language Latent Semantic Indexing have been developed. Document translation into query language or query translation into documents language are the two approaches that couples machines translation and information retrieval.

Query translation and document translation approaches are neither equivalent nor mutually exclusive. They are not equivalent because machine translation is not an invertible operation. Query translation and document translation become equivalent only if each word in one language is translated into a unique word in other languages.

Various researches suggest that document translation should be competitive or superior to query translation. Typical queries are short and may contain key words or phrases only. When these are translated inappropriately, the IR engine has no chance to recover. Translating a long document, MT engine offers the many more opportunities to translate key words and phrases. If some of these are translated inappropriately, the IR engine has at least a chance of matching these to query terms. Query translation approach is flexible and allows for more interactions with the user.

However, query translation often suffers from the problem of translation ambiguity, and this problem is amplified due to the limited amount of context in short queries. From this perspective, document translation seems to be more capable of producing more precise translation due to richer contexts.

One of the critical aspects of document translation approach is that one has to determine in advance to which language each document should be translated and that all the translated versions of the document should be stored. In a multilingual IR environment, one would desire to translate each document to all other languages. This is impracticable because of the multiplication of document versions and the increase in storage requirement. Once a document is pre-translated into the same language as the query, user can directly read and understand the translated version.

Otherwise, a post-retrieval translation is often needed to make the retrieved documents readable by the user (if he/she does not understand the document language).

Query translation and document translation become equivalent only if each word in one language is translated into a unique word in the other languages. Document translation can be performed off-line and on-line but query translation is performed only on-line. Hybrid system that uses both query and document translation are possible because of a trade off between computer resources and quality of translation. Hybrid or dual translation approach provides the relationship between multilingual and the key advantages of these systems are that queries can be expressed and matched unambiguously. In this approach the additional storage space requirement is independent to the number of languages supported. The major problems occurs in this approach are to define the concept space, intermediate representation and conversion of documents into intermediate representation. Differences between two approaches (query translation and documents translation) of CLIR are described in table(3.1). Table (3.2) describes the comparative study of three approaches of CLIR (Sanjay & Ganesh, 2016).

Table(3.1): Difference between Query and Document Translation

Parameter	Query Translation	Document Translation
Size	Small	Large
Language	Prior knowledge of translation language is not required	Prior knowledge of translation language is required
Overhead	Low	High
Recovery	When these are translated inappropriately, the IR engine has no chance to recover	Chance to recover
Ambiguity	Maximum chances of occurring ambiguity	Minimum chances of occurring
		ambiguity
Cost	Low cost	High cost

Table (3.2): Comparison of three Translation Approaches

Parameter	Query Translation	Document Translation	Both Query & Document Translation
Ambiguity	Maximum	Minimum	More than both
Additional Storage Space	Not required	Required	Not required
Translation time	Less	More than query	More than both
Information retrieval	Bilingual	Bilingual	Bilingual and Multilingual
Flexibility	Highly	Less	Less
Working nature	Can provide interface between two language at a time	Can provide interface between two language at a time	Can provide interface between more than two language at a time

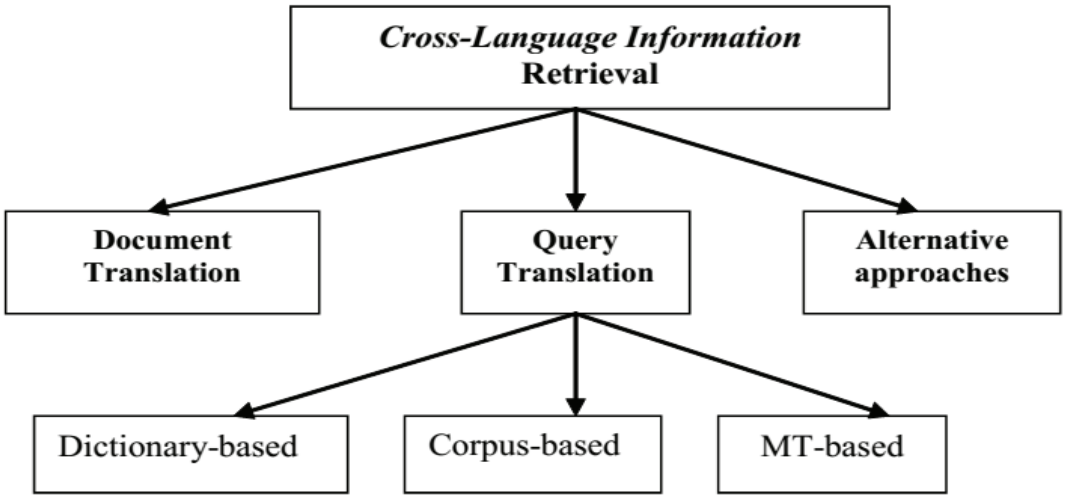

Figure (3.6): CLIR Three Approaches Comparative

3.7 Conclusion and Future of CLIR:

Cross-lingual IR provides new mirror in searching documents through multitude varieties of languages across the world and it can be the baseline for searching not only between two languages but also in multiple languages. Today, most of the cross-lingual researches involved only few famous languages like English, Hindi, Spanish, Chinese and French. Research on languages has increases the development of country. As the world becomes more connected by technology, cross language IR in every language is needed. CLIR is a multidisciplinary area that has been increasingly gaining more attention from the research community. Despite recent advances and new developments, there are still many aspects to be explored .

In Indian context, which is one of the hotspots of linguistic diversity (350 languages) in the globe ،and the fact that a dominant language of one region may be a language of a linguistic minority in other region, cross languages information retrieval systems would play a very important role in allowing the people to go through the documents and literatures of other languages thus breaking the language barrier. We work out here to give a broad overview of the speedy demanding work in the field of CLIR by exploring its aspiration, difficulties, basic tools, major works and future research goals. In reviewing this information, it becomes possible to gain a larger picture of the CLIR field (Sanjay & Ganesh, 2016).

Chapter 4

Methodology and Design

Chapter 4. Methodology and Design

An interactive cross lingual information retrieval tool is developed, which support multi languages, and can uses as computer application.

User can use this tool to retrieve related documents from other languages; the retrieve process can divided into three phases as shown in figure (4.1):

1- Query keywords extraction: we need firstly use a method of document indexing to select keywords from document to search in other documents. Word form normalization and word stemmer is one of the most approaches can be used; so all punctuations, stop words, conjunctions will be removed, and then the stemmer will be used to remove affixes of the word.
2- Keywords translation: most of CLIR tools using multi lingual dictionaries translation to translate keywords, but this approach suffer from many problems such as word inflection, problem of translating word compounds, phrases and social terms.

   The proposed tool depend on Ontology rather than dictionaries as its covers the entire context and relationships which will be helpful for both user and system provider.

   There are several approaches to build Ontology like decision trees or mapping rules but these approaches is either time consuming or complex. Our ontology is build depending on relational database with some additional rules to be use for cross lingual information retrieval.
3- Search for result:

   The result of previous step will be a full-translated query, which will be used in search process to retrieve the relevant document and then these documents ranked and viewed to the user according to the rank.

To search query we need firstly to use a stemmer, many stemmers have developed for Arabic and English language and we use the "porter" stemmer for English language as it performed very well depending on studies and most effective root stemmer "Khoja" (Khoja, 1999) and the most effective light stemmer "Light10" (Froud, 2012) for Arabic language.

Although a number of attempts had been made to develop stemming techniques for the Arabic language, most of those attempts still suffer from many problems such as dealing with irregular words, broken plurals words and the blind removing of affixes that lead to change in meaning of words and reducing the performance of the stemmer.

The next section will discuss a new hybrid-stemming algorithm that solves the above-mentioned problems.

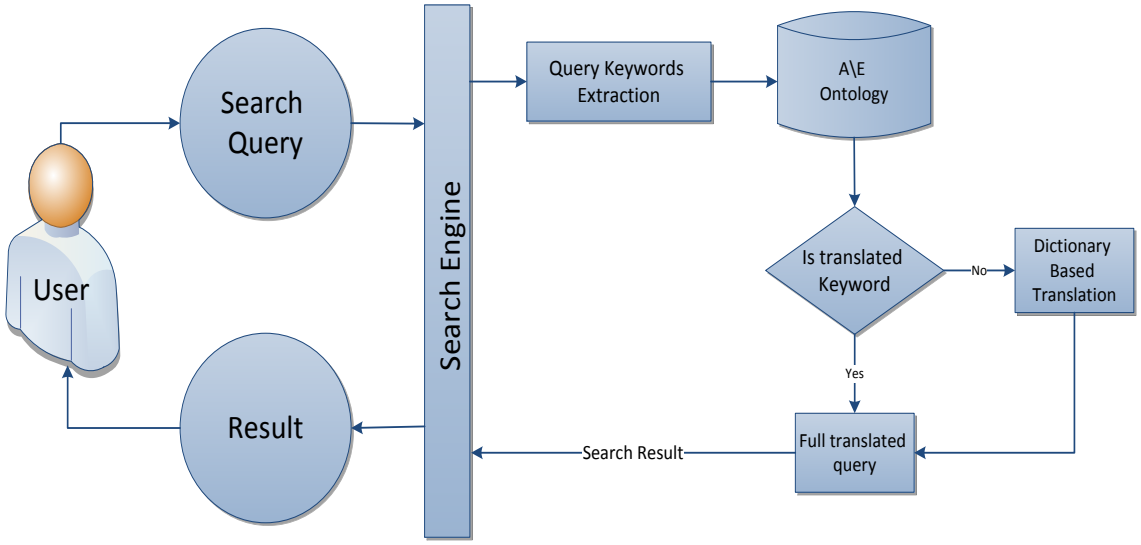

Figure (4.1): The three phases of retrieve process

4.1 Proposed Hybrid Stemmer

In this research, the researcher proposes a new hybrid-stemming algorithm – referred to as "the proposed stemmer" - that integrates between the affixes removal and lookup approaches. The proposed stemmer improves the performance of information retrieval by defining a set of morphological rules that solve many of the ambiguity problems of light stemming like broken plurals and blind removal of the affixes.

The researcher developed an Arabic morphological engine; which takes a set of patterns, affixes, and corpora as input and extracts morphological rules. These rules will be applied into words and then the stem word will be extracted depending on techniques discussed below.

The algorithm divided into two sections; section 4.1.1 describes the main idea of how to extract rules, while section 4.1.2 analyzes the extracted rules and then select the best set of rules to use in the stemmer.

4.1.1 Building Rules – Training

The main goal of this step is to define a set of rules to be used in the stemming algorithm. In this step, Arabic morphological rules built depends on three inputs, which are:

1. Set of Affixes listed in Table 4.1, the affixes include only prefixes and suffixes, as all set of antefixes and postfixes combined with prefixes and suffixes respectively.

Table (4.1): Affixes list

Affixes		
Prefixes	P1	ل – ب – ف – س – و – ي – ت – ن – ا
	P2	ال – لل – فل – ول – وب
	P3	ولل – وال - كال – بال – فال
Suffixes	S1	ة – ه – ي – ك – ت – ا – ن
	S2	ون – ات – ان – ين – تن – كم – هن – نا – يا – ها – تم – كن – ني – وا – ما – هم
	S3	تمل – همل – تان – تين – كمل

2. Predefined lists of patterns. Lists shown in Table 4.2 depends on length.

Table (4.2): Arabic patterns

Length	Patterns
L4	فاعل – افعل – فعّل – تفعل – فعال – مفعل – فعول – فعيل - فعلى – فعلة
L5	افتعل – انفعل – تفاعل – تفعيل – افعال – مفاعل – متفعل – منفعل – مفتعل – مفعول – مفعال – فعلان – فعلاء – فواعل – افاعل – يفتعل – تفتعل – فاعول – فعائل – تفعلل – افعلل – مفعلل – فعالل
L6	استفعل – انفعال – افتعال – افعلال – متفاعل – مستفعل – مفاعيل – يستفعل – افعوعل – متفعلل
L7	استفعال

3. Arabic Corpus - Open Source Arabic Corpus OSAC (Kazem Taghva, 2010).

The main idea of "building rules" step is that the word will be firstly matched against the list of predefined patterns, if there is no pattern match then we will start removing affixes and retry to match it with the predefined patterns after each removal. Figure 4.2 shows the general diagram of the proposed stemmer:

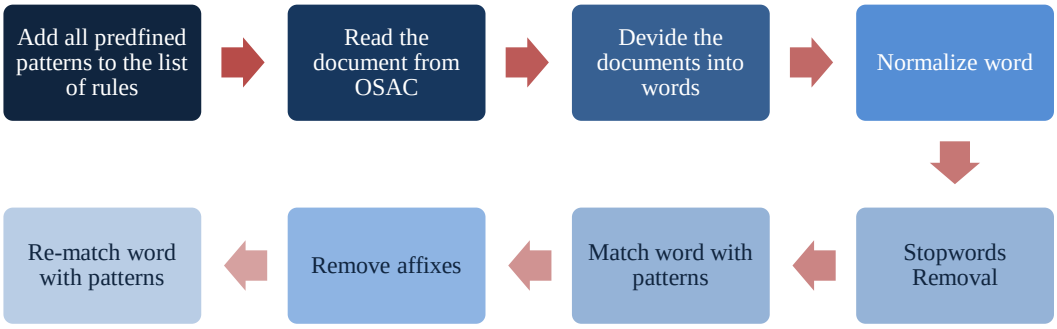

Figure (4.2): Basic steps of building rules process

Flow chart of building rules process - The training phase

As shown in figure 4.4:

1. Matching the word against the Arabic pattern list before removing any affixes, the goal of this step is to solve the problem of blind affixes removal as if there is a word that starts or ends with possible prefix or suffix and the word matches one pattern before removing the affixes, then it is a valid word and the affixes in the word is a part of the original word and must be kept.
   For example, the word "الوان"starts with a possible prefix "ال", but as we see the "ال" is a part of the original word and removing it will lead to have the root "وان"which has no meaning. When applying the match first then the word will match with the pattern of the length five "افعال"and we will return it without change.
   If the match occurs then there is no need to add any additional rule to the list of rules as all predefined patterns added before.

2. If the word does not match any of the predefined patterns, then we need to truncate its prefixes and suffixes to find a new rule, we will start by removing prefixes and suffixes of length three and two respectively.

The removal process must be done depending on some constrains, firstly we start by checking the word length, if it is greater than or equal six then we will remove a prefix of length three, if not then we will check if the length is equal to five and if yes we will remove the prefix of length two.

The same constrains will be checked to remove suffixes of length three and two. The reason of removing prefixes before suffixes will discuss in a special section later in this chapter.

Let us take an example of this step, suppose we have a word like "المنظمات", the length of the word is eight which is greater than six but nothing from the prefixes of length three matches the first three characters "الم", so check the two characters "ال"against the set of prefixes of length two, the prefix will be found and it will be added to a new special prefix list which was established to be used only in the building rules phase and the list will now contain only "ال". The same will be done with suffixes and the special suffix list will be initialized with suffix "ات", and the remaining word will be "منظم".

3. If the remaining word length equal to three then we will stop the process and add the rule to the list of rules; the rule will consist of the special prefix list plus "فعل"plus the special suffix list.

4. If the remaining word length is equal to four then match the word against the list of Arabic patterns of length four, if the match occurs, then add a new rule, if not then try to remove one prefix or suffix according to predefined prefixes and suffixes of length one list and then add a new rule.

   When the word does not match any pattern and also there is no one prefix or suffix matched then the word will be neglected and considered as irregular word.

   By looking into previous example, the word "منظم"will be matched against the set of predefined Arabic patterns of length four, and return the pattern "مفعل, so a new rule will be added to the list of rules.

   Table 4.3 describes the matching process while Figure 4.3 shows the structure of the new rule.

Table (4.3): Matched pattern for word "منظم"

الكلمة	م	ن	ظ	م
النمط المقابل	**م**	**ف**	**ع**	ل

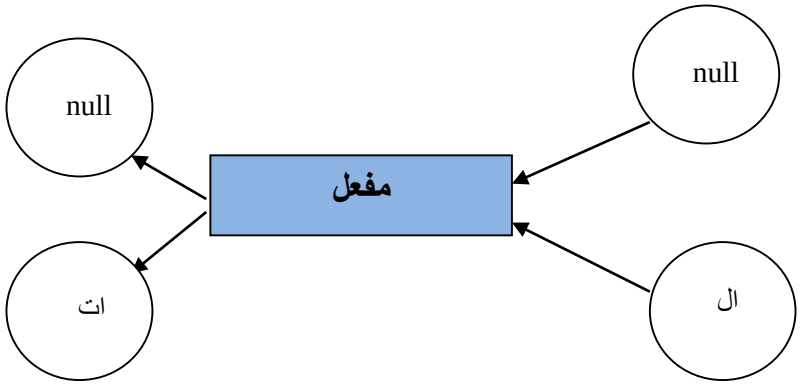

Figure (4.3): Rules tree for pattern "مفعل"

5. Match the word against the list of Arabic patterns of the same remaining length, if the match occur, then add a new rule, if not; try to remove one prefix or suffix according to predefined prefixes and suffixes of length one list, if one of prefixes or suffixes has been removed then reprocess step five with the new word length. After applying the above mentioned seven steps we will have a list of rules that will be used later in the proposed stemmer. Figure 4.3 shows the flowchart of the algorithim .

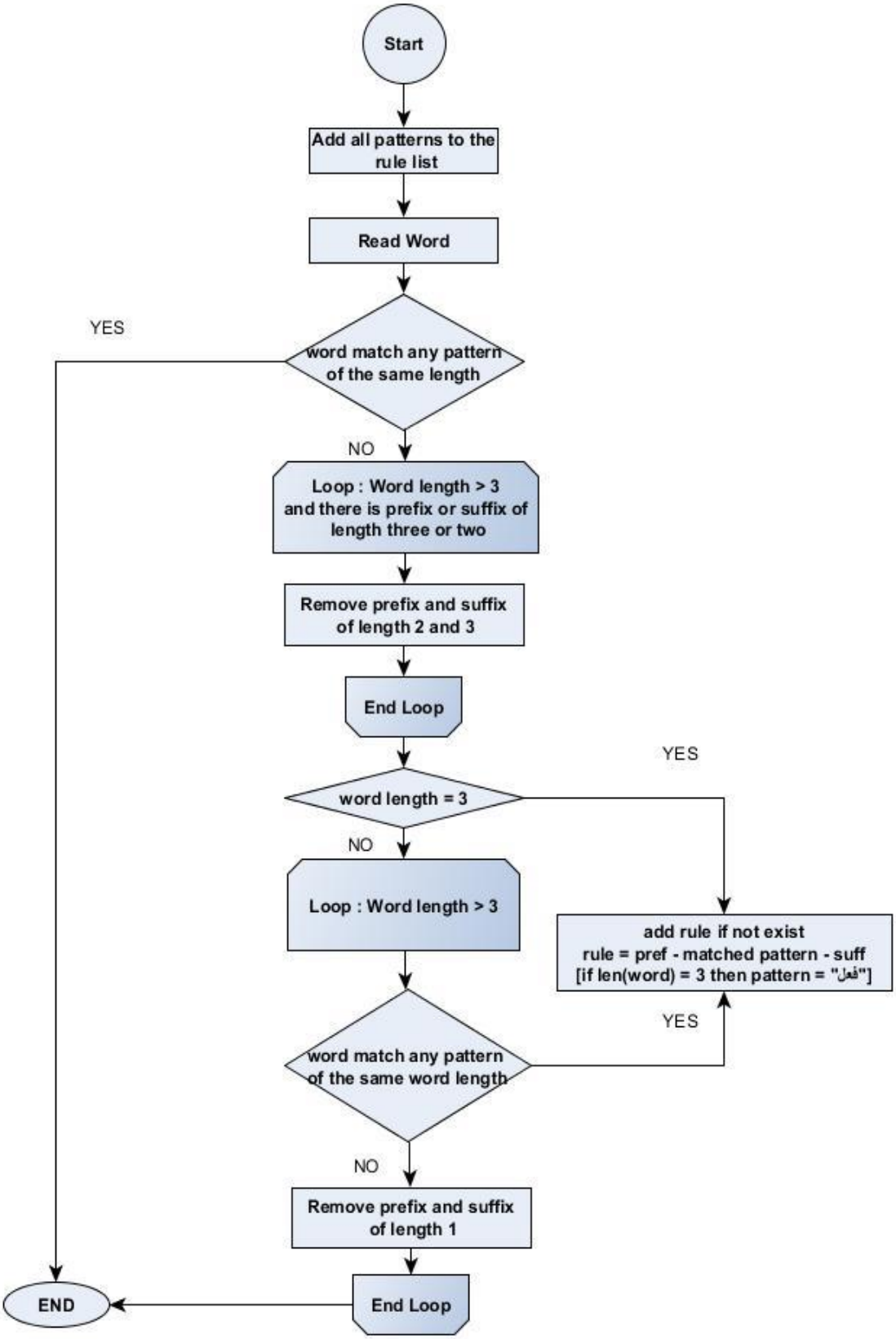

Figure (4.4): Flow chart of building rules process

The morphological structure of Arabic words

The technique used in the previous section -building rules- depends on the morphological structure of the Arabic word, so we need to describe and analyze the structure to determine the reason of removing the prefixes before suffixes and to reorder the predefined patterns. The morphological structure of the Arabic word described in Figure4.5.

Figure (4.5): Morphological Structure of Arabic word

Figure 4.5 describes the structure of the word; firstly add infixes to the root to generate the stem form and then attach the prefixes and suffixes to generate the full word.

A study has been done by the researcher to show the average occurrences of suffixes and prefixes into Arabic words; the study include analyze of about 47,000 words selected randomly from the OSAC corpus. Table 4.4 shows the distribution of suffix and prefix into these words.

Table (4.4): Distribution of prefixes and suffixes into Arabic words

	Number of words	Percent
Only prefixes	15166	32.13%
Only suffixes	12022	25.48%
Has prefixes and suffixes	10169	21.54%
None	9838	20.85%
Total	47195	100%

Table 4.4 shows that more than 20% of the Arabic words do not have any prefixes or suffixes, so blind removal of the affixes will affect those words and will remove original letters from the word. Also it is noticeable from the table that about 33% of the words will have prefixes only and the percent is greater than the percent of words that have suffixes only by about 7%. The researcher also do another study which showed the

distribution of the number of words according to the number of prefixes and suffixes, as exhibited in Figure 4.6.

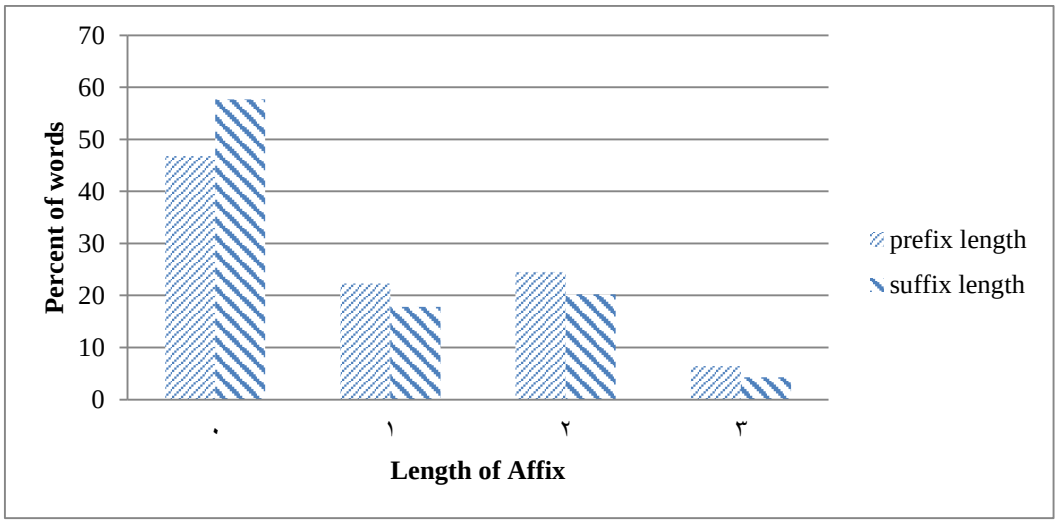

Figure (4.6): Distribution of the number of words according to the number of prefixes and suffixes

From Figure 4.6 it can be confirmed that blind remove of affixes will affect the stemming process, as there is about 45% of words do not have prefixes and about 57% do not have suffixes. Also the percent of words with prefixes is always greater than words with suffixes for all lengths, and as mentioned before, the percent of words with prefixes only is greater than words with suffixes only by about 7%, the researcher decided to remove prefixes before suffixes in the proposed algorithm as it has more popular occurrences.

A study of the frequency of all patterns is conducted; its aim of is to reorder the patterns of the same length according to the number of occurrences. Ordering those patterns will lead to better performance as the word that matches more than one pattern will be matched to the most popular one. Figures [4.7, 4.8, 4.9] show the distribution of words in patterns of the length five, four and six respectively.

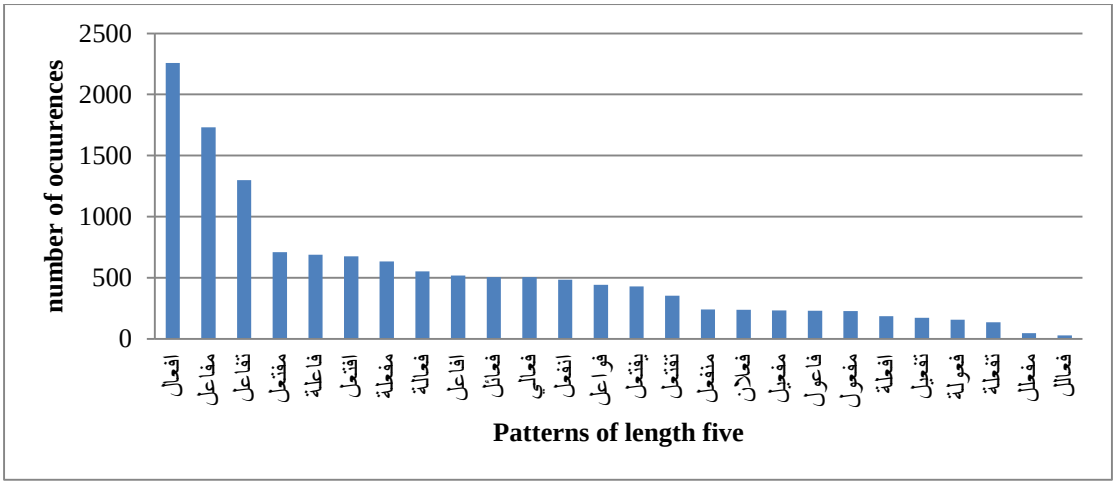

Figure (4.7): The distribution of words in patterns of the length five

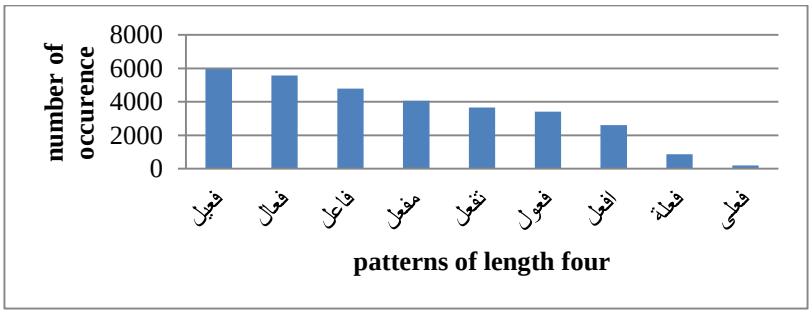

Figure (4.8): The distribution of words in patterns of the length four

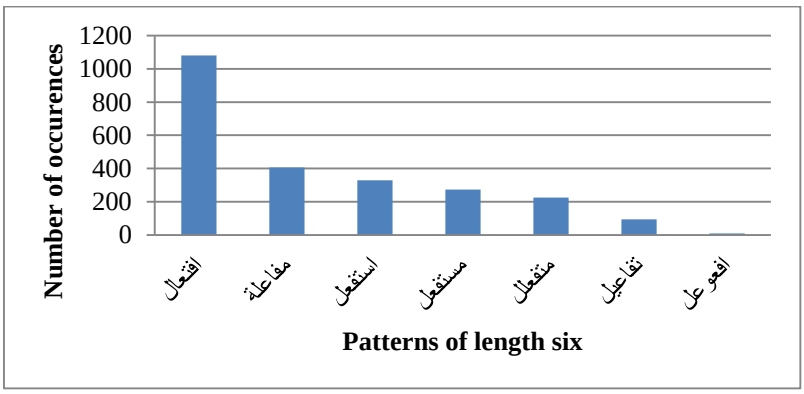

Figure (4.9): The distribution of words in patterns of the length six

Figure 4.7 shows the distribution of words into patterns of length five, the figure shows that the pattern "افعال" is the most popular pattern while the pattern "فعالل"has the least occurrences. The same for patterns with length four as shown in Figure 4.8; "فعيل" is the most popular one for this length while "فعلى" is the least one. For patterns with

length six the pattern "افتعال" has the most number of occurrences while "افعوعل"has the lowest number as shown in Figure 4.9.
Table 4.5 shows the ordered list of the Arabic patterns, depending on the study above.

Table (4.5) : The ordered list of the Arabic patterns

Length	Patterns
L4	فعيل – فعال – فاعل – مفعل – تفعل – فعول –افعل –فعلة – فعلى
L5	افعال – مفاعل – تفاعل – مفتعل – فاعلة – افتعل – مفعلة – فعالة – افاعل – فعائل – فعالي – انفعل – فواعل – يفتعل – تفتعل – منفعل – فاعول – مفعول – افعلة – تفعيل – فعولة – تفعلة – مفعلل – فعالل
L6	افتعال – مفاعلة – استفعل – مستفعل – متفعلل – تفاعيل – افعوعل – افعلال – مفاعيل
L7	استفعال

4.1.2 Rule-based Stemmer

The proposed stemmer has been developed depending on the rule list that has been derived in the previous section. The rule list will consist of all possible combinations of prefixes and suffixes that can be added to the predefined list of Arabic patterns to create a new form of the patterns. Figure 4.3 is an example of the generated rules for the pattern "مفعل".

For a new input word, the word is string matched to the list of rules, matching is done only with the rules that have the same length of the word, and by matching prefixes and suffixes the pattern will be extracted.

The list of rules that has been generated needs to be reviewed before use, as any rule that appears only from one word, it will be considered as irregular word and needs to be removed from the list.

Any word that does not match rule from the remaining rules will be considered as irregular and will be returned without stemming. This process will also manage the problem of missing space between words.

The proposed stemmer Algorithm

To start describing the proposed stemmer steps, first we define a set of diacritical marks, punctuations and a list of stopwords as shown in Table 4.6.

Table (4.6): A set of diacritical marks, punctuations and a list of stopwords

Diacritical marks	ً - ُ – ّ- َ- ِ- ْ									
Punctuations	.	÷	~	.	;	{				
	,	×	=	+	/	}				
	?	؟	'	%	\	)				
	:	>	!	^	?	(				
	"	<	@	&	$	_				
	'	|	#	*	'	-				
Stopwords	بعد	كذلك	فقط	ومن	هؤلاء	هو	وقد	اضحى	او	ضمن	فكان
	ضد	التي	ثم	لا	فإن	عنها	كانت	ظل	و	اول	متكون
	يلي	وبين	هذه	ليسب	فيه	منه	لذلك	مابرح	ما	وله	مما
	الى	فيها	أنه	وكانت	ذلك	بها	أمام	مافتئ	لا	ذات	أبو
	في	عليها	تكون	أي	لو	وفي	هناك	ماانفك	الي	اي	بإن
	من	إن	قد	ما	عند	فهو	قبل	بات	إلي	بدلا	الذي
	حتى	وعلى	بين	عنه	اللذين	تحت	معه	صار	مازال	اليها	اليه
	وهو	لكن	جدا	حول	كل	لها	يوم	ليس	لازال	انه	يمكن
	يكون	عن	لن	دون	بد	أو	منها	إن	لايزال	الذين	بهذا
	به	مساء	نحو	مع	لدى	إذ	إلى	كأن	مايزال	فانه	لدي
	وليس	ليس	كان	لكنه	وثي	علي	إذا	ليت	اصبح	وان	وأن
	أحد	منذ	لهم	ولكن	أن	عليه	هل	لعل	أصبح	والذي	وهي
	على	الذي	لأن	له	ومع	كما	حيث	لاسيما	أمسى	وهذا	وأبو
	وكان	أما	اليوم	هذا	فقد	كيف	هي	ولايزال	امسى	لهذا	آل

The stemming process will proceed by the following steps:

- Tokenization: This is a necessary and meaningful step in natural language processing. The function of a tokenizer is to break down a text stream into segments so that they can be introduced into a morphological sensor or a position tagger. The tokenizer is responsible for defining boundaries of a word; it is based mainly on the white spaces and punctuation marks as delimiters between words or major segments.

In our algorithm the following separators are used " \r\n\t.,;:'\"()?!" which depend on new lines, white spaces, tabs and some punctuation marks.

- Normalization: In the proposed algorithm, normalization include the following steps:
  - Removing the shadda and the diacritics – Table 4.6.
  - Removing punctuations – Table4.6.
  - Removing all numbers.
  - Replacing "أ", "إ" and "آ" by "ا"
  - Replacing "ي" by "ى" at the end of the words.
  - Replacing the sequence "يء" by "ئ".
  - Removing the tatweel character "-".
  - Remove stopwords – Table 4.6.

After preprocessing steps, each word is matched against a list of predefined rules which has been generated previously. If the word matched a rule then remove all affixes depending on that rule and take the remaining word as stem word, if not then return the word itself.

Returning the stemmed words is not enough; as some words may match broken plural pattern, and then it needs to be converted to the singular form of that pattern. So our predefined patterns need to be classified to two parts: broken plurals patterns and normal patterns.

The broken plurals patterns are important for Arabic text as it form around 10% of any Arabic content; the word takes different morphological form than the singular one so it needs to be reformed to the singular form. The researcher defines a dictionary of these patterns and their singular forms, and after getting the stemmed word from the previous

step, then it will be matched against this dictionary, if the pattern is one of broken plurals patterns then the word is converted to its singular form, and if not it will be returned without change. Table 4.7 describes the broken plurals patterns and all singular forms of it.

Table (4.7): Broken plurals and its singular form(s)

Broken Pattern	Singular Form	Plural Example	Singular Form
مفاعيل	مفعول	مجانين	مجنون
أفعال	فعل	أصوات	صوت
فعلاء	فعل – فعال – فاعل –فعيل	سمحاء – جبناء – عقلاء - اطباء	سمح – جبان – عاقل – طبيب
فواعل	فاعل - فوعل	شوارع - مواسم	شارع – موسم
فعائل	فعيل	ضمائر	ضمير
فعايا	فعيه	خلايا	خليه
فعاليل	فعويل - فعليل	ملايين - براميل	مليون – برميل
افعياء	فعي	اغبياء	غبي
فواعيل	فاعول	طوابير	طابور

From the table above we can notice that some broken plurals have more than one singular form, so when the word is matched against one of them, it is needed to return all singular forms of that pattern.

4.2 Translation process

System architecture for CLIR can generally be classified as query translation, document translation, or interlingual methods Query translation is the most frequent option, because the texts are shorter than the documents, and the computational cost of the translation is lower.

However, a high number of researchers claim it is particularly difficult to solve problems of ambiguity during the translation process, because the queries are too short and do not offer a relevant context, although user interaction might enhance results.

A new translation methodology has developed which depend on select the proper translation of bilingual dictionary depending of all words in the query, approach can be divided to several steps:

1 – Normalization and stemming:

A query word may have different inflected forms without significantly changing its core meaning. This presented a potential issue for a resource. If such a database were to include these multiple forms, it would significantly increase its size without adding much value.

So the query should be processed by using the stemmer according to the query language, that mean the query will be normalized, stemmed, and then weighted so it will be converted to normalized words.

2- Translation of terms

In cross-language information retrieval, the input is often a combination of a series of keywords rather than a complete sentence. This sequence of query keywords lacks necessary contextual and syntactic semantic information, so they can not be translated by traditional MT technology in a direct and easy way. Neither can the translation problem be resolved by simply looking up the bilingual dictionary. For example, the English word "bank" corresponds to two Arabic meanings in a typical English-Arabic dictionary: "بنك" and "ضـــفة". Then the problem arises: should the word "bank" be translated into "بنك" or "ضفة"?

In common sense, when the user inputs {bank; credit} as a query, we may well conclude that he is most likely looking for information about "بنك" and "ائتمان", although we cannot completely exclude the small possibility of interpreting "bank" as the meaning of "ضـفة" in this context. Therefore, the retrieval algorithm should rank among all the retrieval results the documents containing the keyword "بنك" before those containing the keyword "ضفه".

The translation and transform of query sentence in CLIR is formalized as follows:

Suppose there is a SL keyword query sentence:

$$Query_S = W_{S1}\ W_{S2} \cdots W_{Si} \cdots W_{SN} \quad (4.1)$$

where W_{Si} represents i^{th} the keyword in the query sentence, for example, the keywords "bank" and "credit" in the query instance "bank credit".

The SL keyword query sentence is translated and transformed into a TL keyword query sentence, which is represented as follows:

$$\begin{aligned} Query_T = & (W_{T11} \wedge boost_{T11}\ \ W_{T12} \wedge boost_{T12} \cdots W_{T1N} \wedge boost_{T1N}) \\ & \cdots (W_{Ti1} \wedge boost_{Ti1}\ W_{Ti2} \wedge boost_{Ti2} \cdots W_{TiN} \wedge boost_{TiN}) \cdots \\ & (W_{TN1} \wedge boost_{TN1}\ W_{TN2} \wedge boost_{TN2} \cdots W_{TNN} \wedge boost_{TNN}) \end{aligned} \quad (4.2)$$

where $W_{Ti1}\ \ W_{Ti2} \cdots W_{TiN}$ are the translations in the bilingual dictionary of the keyword W_{Si} of the source language keyword query sentence, and $boost_{Ti1}\ \ boost_{Ti2} \cdots boost_{TiN}$ are the weight of each translation of the keyword W_{Si} in the bilingual dictionary, which is referred to as boost value.

4.3 Basic Theory of Information Retrieval

For an IR system, the kernel problem is ranking, i.e., the so-called relevance computing. Based on vector space model, a relevance computation formula is defined as follows:

$$score(q, d_i) = \frac{\sum_{t \in q} tf(t \text{ in } d_i)\, idf(t)\, boost(t)}{\max(score(q, d_i))} \quad (4.3)$$

where q represents query sentence, d_i represents the i th document and t represents a query keyword. The denominator max(score(q,d_i)) is a normal factor which does not affect the ranking result but makes comparable the relevance values of different queries. The addition of the normal factor is for the convenience of computing the boost value when query keywords are translated and transformed. In addition, the idf(t) in (4.3) is defined as

$$idf(t) = -\log p(t) = -\log\left(\frac{N}{M}\right) \quad (4.4)$$

where M represents the total number of the documents, and N represents the number of the documents which contain the keyword t.

4.3.1 Boost Value Computation

How to compute the boost value of the translations in the bilingual dictionary of a given query keyword is the focus of the algorithm proposed in this thesis. Our approach based on large-scale bilingual corpora, and the computation is implemented by applying the theories of vector space model (VSM) and lexical mutual information to traditional IR.

Supporting Knowledge Base. In addition to the bilingual dictionary, the translation and transform of the query sentence also make use of a large-scale bilingual corpus of aligned sentence pairs. It is meant to take the bilingual sentence pair as the basic retrieval unit. The corpus used in the experiment contains altogether 1,162,918 English-Arabic sentence pairs, which amount to approximately 18 million English and Arabic words. This aligned bilingual corpus is used as the retrieval source for boost value computation.

Boost Value Computation. Suppose there is a sequence of SL query keywords:

$$W_{S1}\ W_{S2}\ \cdots W_{Si}\ \cdots W_{SN}$$

and the corresponding sequence of the TL translations of the SL query keywords according to their respective SL entries in the bilingual dictionary:

$$(W_{T11}\ W_{T12}\ \cdots W_{T1N})\cdots(W_{Ti1}\ W_{Ti2}\ \cdots W_{TiN})\cdots(W_{TN1}\ W_{TN2}\ \cdots W_{TNN})$$

Three query sentences are designed for the purpose of boost value computation:

$$Query_1 = (W_{Si}\ AND\ W_{Tij})\ AND\ (W_{S1}\ AND\ W_{S2}\cdots AND\ W_{SN}) \quad (4.5)$$

$$Query_2 = (W_{Si}\ AND\ W_{Tij})\ AND\ (W_{S1}\ OR\ W_{S2}\cdots OR\ W_{SN}) \quad (4.6)$$

$$Query_3 = (W_{Si}\ AND\ W_{Tij})\ OR\ (W_{S1}\ OR\ W_{S2}\cdots OR\ W_{SN}) \quad (4.7)$$

Hence the formula for the computation of the boost value of W_{Tij} :

$$boost(W_{Tij}) = 2^{\alpha} \times mean(score(q, d_i)) + \beta \quad (4.8)$$

where mean(score(q,d_i)) represents the average relativity of the retrieval results. The relativity computation formula is defined as (equation 4.3) . Also in formula (4.8), β is equivalent to the datum value of transform and is set to be 0.5; α is the weight coefficient of the query sentence and is determined in the following way:

$$\alpha = \begin{cases} 3, & if\ q == Query_1 \\ 2, & else\ if\ q == Query_2 \\ 1, & else\ if\ q == Query_3 \\ 0, & else \end{cases} \quad (4.9)$$

4.4 Similarity measurement

After text preprocessing each document from the collection of documents [Doc_1, Doc_2… Doc_n] is represented as a vector d. Each dimension in the vector d stands for a distinct term (word) in the term space of the document collection [$Term_{11}$, $Term_{12}$…. $Term_{1t}$]. Then the collection can be represented in a matrix form as shown in Figure(4.10):

	$\boldsymbol{Term_1}$	$\boldsymbol{Term_2}$	…	$\boldsymbol{Term_t}$
$\boldsymbol{Doc_1}$	t_{11}	t_{12}	…	t_{1t}
$\boldsymbol{Doc_2}$	t_{21}	t_{22}	…	t_{2t}
…	…	…	…	…
$\boldsymbol{Doc_n}$	t_{n1}	t_{n2}	…	t_{nt}

Figure (4.10): Weight Matrix of Vector Space Model

The term T vector will consist of all unique words that appear in each document of the collection, so the matrix will be sparse matrix as every word does not normally appear in each document.

Chapter 5

Experimental Results

Chapter 5. Experimental Results

This chapter introduces the new proposed Tool **(MORTAJA-IR-TOOL)** and show how the stemmer has solved the problem of irregular words, broken plural patterns and blind removal of affixes.

MORTAJA-IR-TOOL will help to overcome several of previous CLIR approaches limitations, especially the problem of untranslatable query keys, missing words and translation ambiguity

MORTAJA-IR-TOOL used to compare between the four phases:

Phase1 traditional phase without stemming & just traditional translation. Phase2 without stemming but with **MORTAJA-IR-TOOL** modified translation, Phase3 with **MORTAJA-IR-TOOL** stemming & just traditional translation, Finally Phase4 the proposed one with **MORTAJA-IR-TOOL** stemming & with **MORTAJA-IR-TOOL** Modified translation.

The comparison will be according to number of hit files, and the effect of using the stemmer over text classification & the effect of using the modified translation.

All results are generated using OSAC corpora as data set on 64-bit machine with 8GB RAM and core i7 processor. **MORTAJA-IR-TOOL** is developed using java programming language with JDK 1.6.

5.1 Datasets specifications

This section describes and identifies the specifications of datasets used in all experiments over all algorithms. Datasets used are: the open source corpus OSAC collected by Saaed and Ashour (Saad , 2010) (available free for IR researchers).

5.1.1 Open Source Arabic Corpus - OSAC

OSAC corpus is collected from different web sites and includes 22429 text documents including CNN and BBC documents, each text document belongs to one of the ten classes (Economics, History, Education & Family, Religious and Fatwas, Sports,

Health, Astronomy, Low, Stories, Cooking Recipes). The corpus contains about 18000000 words. Table 5.1 describes the distribution of text documents over the ten classes.

Table (5.1): Distribution of text documents over the ten classes of OSAC Corpus

#	Category	Number of documents
1	Economics	3102
2	History	3233
3	Education & Family	3608
4	Religious and Fatwas	3171
5	Sports	2419
6	Health	2296
7	Astronomy	557
8	Low	944
9	Stories	726
10	Cooking Recipes	2373
Total		22429

5.2 Stemming Effects

Firstly, the comparison will be between the effects of stemming process using the traditional translation process.

Table (5.2): Effect of Stemming and normalization – were the translation is traditional

Phases	Processes	Total Files	Hit files	Hit Rate
Phase1	Without stemming & Traditional translation	22429	16329	**72.80%**
Phase3	With stemming & Traditional translation		17256	**76.94%**

Table 5.2 show with numbers the effect of stemming process using the traditional translation process , as shown in this comparison the effect of stemming process is positive and the value is : 76.94 – 72.80 = **4.14%** .

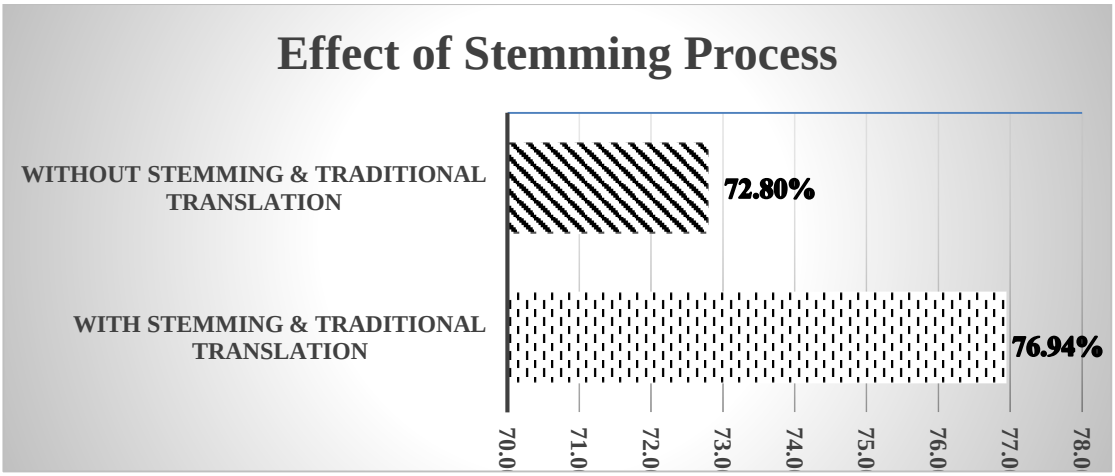

Figure (5.1): Effects of Stemming and normalization – were the translation is traditional

Figure 5.1 introduces the effect of stemming process using the traditional translation process in OSAC corpus, the figure shows that using the new process stemmer will increase the average number of files hit rate about 4%.

5.3 Translation effects

Secondly, the comparison will be between the effects of modified translation process without using the stemming process.

Table (5.3): Effects of translation without using the stemming process

Phases	Processes	Total Files	Hit files	Hit Rate
Phase1	Before stemming - just Traditional translation	22429	16329	72.80%
Phase2	Before stemming - Modified translation		18338	81.76%

Table 5.3 show with numbers the effect of translation without using the stemming process , as shown in this comparison the effect of translation process was positive were the value is : 81.76 – 72.80 = **8.96%** .

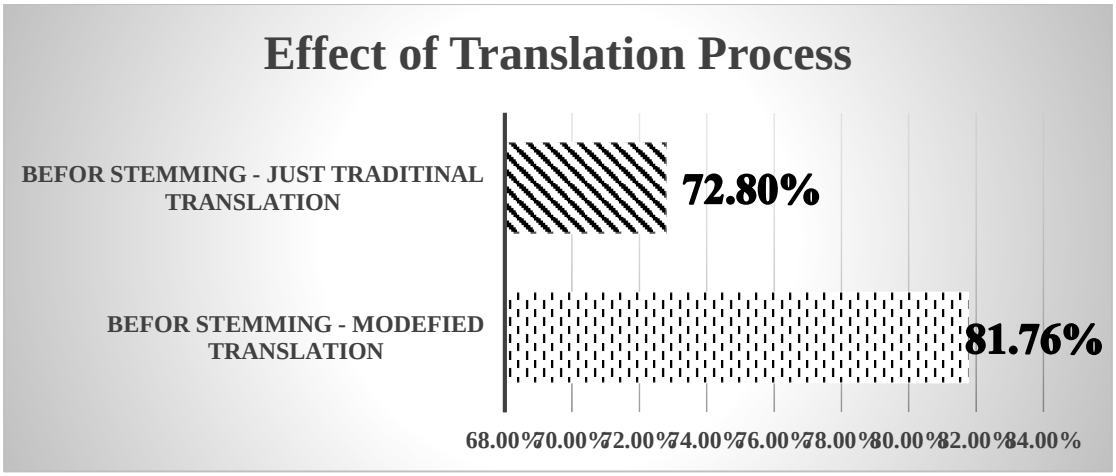

Figure (5.2): Effects of modified translation process – without stemming process

Figure 5.2 introduces the effect of modified translation process without using the stemming process in OSAC corpus, the figure shows that using the new translation process will increase the average number of files hit rate about 9%.

5.4 Translation & Stemming effects

Finally, The merger between the use of stemming methodology proposed and translation process (**MORTAJA-IR-TOOL**) which concluded that the proportion of advanced in the process of improvement in data rate of return was 15.86% as shown in table 5.4.

Table (5.4): All phases comparison

Phases	Processes	Total Files	Hit files	Hit Rate
Phase1	Before stemming - just Traditional translation	22429	16329	72.80%
Phase2	Before stemming - Modified translation		18338	81.76%
Phase3	stemming - just Traditional translation		17256	76.94%
Phase4	stemming & Modified translation (**MORTAJA-IR-TOOL**)		19885	88.66%

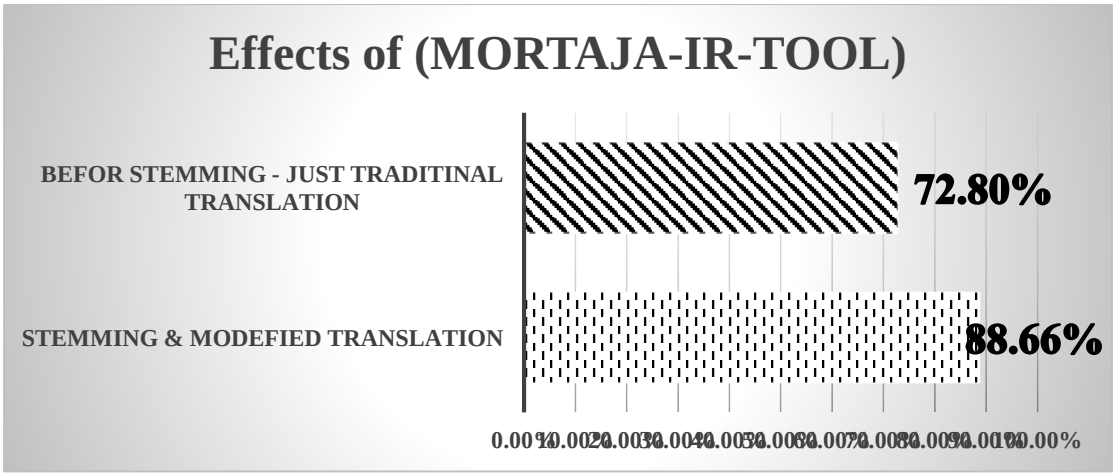

Figure (5.3): Effects of (MORTAJA-IR-TOOL)

Figure 5.3 introduces the effect of modified translation process with the using of the stemming process (MORTAJA-IR-TOOL) in OSAC corpus, the figure shows that using the new translation process and the stemming process will increase the average number of files hit rate about 16%.

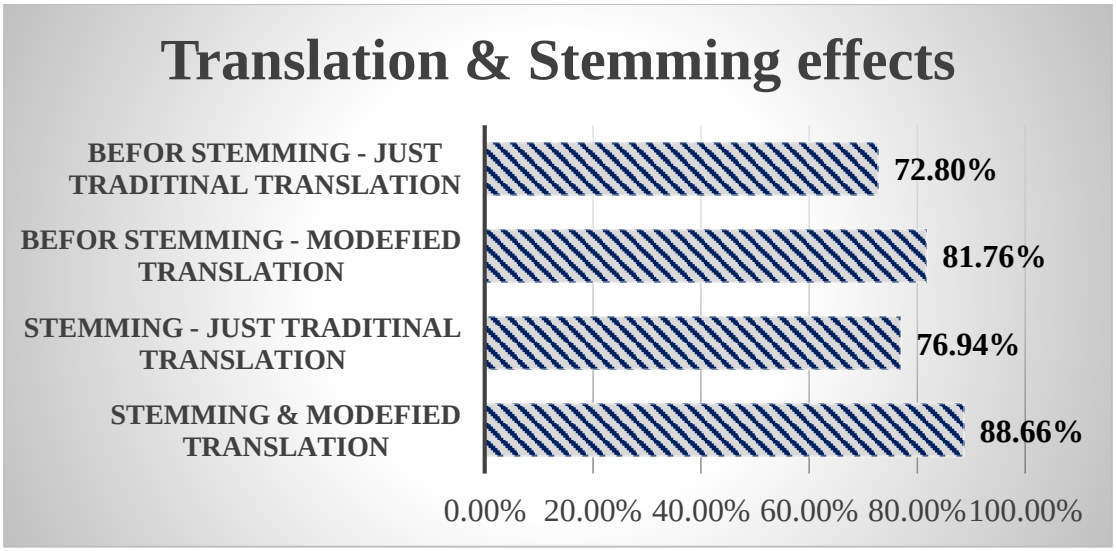

Figure (5.4): Translation & Stemming effects

Figure 5.4 introduce the effect of all 4 phases as shown in the figure Translator percentage of improvement was **8.96%**, as well as the stemming improvement was **4.14**%. Finally, the rated output of Translation & Stemming was **15.86**% , comparing with the base phase (phase1).

Chapter 6

Conclusion and Future Work

Chapter 6

Conclusion and Future Work

6.1 Conclusion

CLIR established as a major topic in Information Retrieval (IR). As queries submitted to search engines suffer lack of untranslatable query keys (i.e., words that the dictionary is missing) and translation ambiguity, which means difficulty in choosing between alternatives of translation. (**MORTAJA-IR-TOOL**) a new tool for retrieving information using programming JAVA language with JDK 1.6, this tool has many features for translation and stemming the words entered in the query process.

Translator percentage of improvement was 8.96%, as well as the stemming improvement was 4.14%. Finally, the rated output of (**MORTAJA-IR-TOOL**) was 15.86%.

6.2 Future Work

In the future work, we shall work on extending the new IR tool to include more stemmers for more languages, weighting techniques and classification techniques that allow researchers to make more accurate decisions when analyzing and comparing techniques.

We shall define a dataset that can be used for testing any stemmer for any language by defining queries, so the stemmers can be compared together according to the results of those queries.

Finally we shall work on building an ontology for translating languages and to solve the ambiguity for languages translation.

The Reference List

The Reference List

A., E.-H. (2008). A Comparative Study on Arabic Text Classification. *Egyptian Computer Science Journal*.

Abdelali, A., Cowie, J., Farwell, D., & Ogden, W. (2003). Uclir: A Multilingual Information Retrieval Tool. *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Arti ficial, 8(22)*, pp. 103-110.

Abusalah, M., Tait, J., & Oakes, M. (2007). Literature Review of Cross Language Information Retrieval. *World Academy of Science, Engineering and Technology*, (pp. 669-671).

Adamson, B. a. (1974, Augest). The Use of an Association Measure Based on Character Structure to Identify Semantically Related Pairs of Words and Document Titles. *Information Storage and Retrieval, 10*(7).

Ahmed, F., & Nurnberger, A. (2010). Can we Support People to Get Information From Text They Can't Read or Understand? *in Proceedings of the 33rd Annual ACM SIGIR conference in Research and Development in Information Retrieval*, (pp. 837-838). Switzerland.

Akiko, A. (2002). A co-evolutionary framework for clustering in information retrieval systems. *National Institute of Informatics.*

Al-Ameed k. Hayder, A.-K. O.-K.-S. (2005). Arabic Light Stemmer: A new Enhanced Approach. *The second international conference on innovations technology*.

Al-Ansary, S. N. (2008). *Towards analyzing the International Corpus of Arabic: Progress of Morphological Stage* (3rd ed.). Bibliotheca Alexandrina.

Al-Shalabi R., K. G. (2006). Arabic text categorization using KNN algorithm. *computer science and information technology CSIT06*.

Anas Boubas, L. L. (2011). A Novel Approach for an Arabic Stemmer Using Genetic Algorithms. (pp. 77-82). Dubai: international Conference on Innovations in Information Technology.

Andreas, F. A. (2012). Literature Review of Interactive Cross Language Information Retrieval Tools. (pp. 479-486). international arab journal of information technology.

Attia Nehar, D. Z. (2012). An efficient stemming for Arabic Text Classification. *Innovations in Information Technology (IIT)* (pp. 328-332). Abu Dhabi: International Conference on Innovations in Information Technology (IIT).

Aziz, B. A.-S. (2011). Statistical Bayesian Learning for Automatic Arabic Text Categorization. *Journal of Computer Science, 7*(1).

Aziz, Y. A. (2011, Septemper). The Enhancement of Arabic Stemming by Using Light Stemming and Dictionary-Based Stemming. *Computer Science & Communications* , *4*(9).

Bawab, M. (2004). Arabic language processing in information systems.

Braschler, M. (2004). Combination Approaches for Multilingual Text Retrieval Eurospider Information Technology. *Information Retrieval, 7*, pp. 183–204.

Buckley, C., M., M., & J., A. W. (2000). Buckley, C., M. Mitra, J. A. Walz and C. Cardie. *TREC 6, Information Processing and Management 36 (2000)*, (pp. 109–131).

Capstick, J., Diagne, A., Erbach, G., Uszkoreit, H., Leisenberg, A., & Leisenberg, M. (2000). A System for Supporting Cross-Lingual Information Retrieval. *International Journal of Information Processing and Management*, (pp. 275-289).

Chang, D. T., Weng, Y. Z., Lin, J. H., Hwang, M. J., & Oyang, Y. J. (2006). Protemot: prediction of protein binding sites with automatically extracted geometrical templates. *Nucleic acids research, 34(2)*, pp. W303-W309.

Chen-Yu, S., Tien-Chien, L., & Shih-Hung, W. (2007). Using Wikipedia to Translate OOV Terms on MLIR. *Proceedings of NTCIR-6 Workshop Meeting*, (pp. 15-18). Tokyo, Japan.

Christopher D. Manning, P. R. (2009). *Introduction to Information Retrieval* (2nd ed.). Cambridge: Cambridge University Press.

Chung-hsinLin, & Hsinchun, C. (1996). An Automatic Indexing and Neural Network Approach to Concept Retrieval and Classification of Multilingual (Chinese-English) Documents. *IEEE Transactions on Systems, Man, And Cybernetics-Part B: Cybernetics*, *26(1)*.

Cleverdon, & C. (1967). The Cranfield tests on index language devices. *Aslib Proceedings*, (pp. 173–194).

Darwish, K. (2003). Probabilistic methods for searching OCR-degraded Arabic text. *Doctoral Dissertation- Unpublished Ph.D. Thesis*.

De Roeck, A. N.-F. (2000). A morphologically sensitive clustering algorithm for identifying Arabic roots. *Proceedings*.

Dilekh T., B. A. (2012). Implementation of a New Hybrid Method for Stemming of Arabic Text. *46*(8).

Dong-Mo, Z., Sheng, H.-Y., Fang, L., & Tian-Fang, Y. (2002). The model and design of a casebased reasoning multilingual natural language interface for database. *Proceedings of the first international conference on machine learning and cybermetics.*

Dumais, S., T., T., A. L., & M., L. ,. (1997). Automatic cross-language retrieval using latent semantic indexing. *AAAI Spring Symposuim on Cross-Language Text and Speech Retrieval.*

El-Disooqi, A. a. (2009). Stemming techniques of Arabic Language: Comparative Study from. *ISSR Cairo University*.

El-Khair, I. A. (2006). Effects of stop words elimination for Arabic information retrieval: a comparative study. *International Journal of Computing & Information Sciences, 4*(3), 119-133.

El-Kourdi M., B. A. (2004). Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm. *Computational Linguistics*.

Evens, R. A.-S. (1998). A computational morphology system for Arabic. *Association for Computational Linguistics*, 66-72.

F Ahmed, E. D. (2009). Revised n-gram based automatic spelling correction tool to improve retrieval effectiveness. In *Information Retrieval and Natural Language Processing* (Vol. 40, pp. 39-48). Computer Science and Computer Engineering with Applications.

F., C. A. (2011). Building an Arabic stemmer for information retrieval. *School of Information Management and Systems. University of California at Berkeley*.

Feldman R., S. J. (2007). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. *Cambridge University Press*.

Fernandes, & Lincoln, P. (2006). Corpora in Translation Studies: revisiting Baker's tipology. *Fragmentos: Revista de Língua e Literatura Estrangeiras*, p. 30.

Frieder, A. a. (2002). Improving the retrieval effectiveness via light stemming approach. *Information and knowledge management*, 340-347.

Froud, L. a. (2012). Stemming Versus Light Stemming for Measuring the Simitilarity between Arabic Words with Latent Semantic Analysis Model.

Fujii, A., & Ishikawa, T. (2001). Evaluating Multi-lingual Information Retrieval and Clustering at ULIS. *Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*, (pp. 144-148). Japan.

Gerard Salton, C. Y. (1975). A vector-space model for automatic indexing. *Communications of the ACM, 18*(1).

Hmeidi, Q. Y. (2014). Extracting the roots of Arabic Words without removing affixes. *Information Science & Library Science*.

Hsin-Chang, Y., & Chung-Hong, L. (2008). Multilingual Information Retrieval Using GHSOM. *ISDA*, (pp. 225-228).

Hull, D. (1996). Stemming algorithm – a case study for detailed evaluation. *Journal of the American Society for Information Science, 47*(1), 70-84.

Hull, D. A., & Grefenstette, G. (1996). Querying across languages. *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, (pp. 49-57). doi:10.1145/243199.243212

Internet World Stats. (2016, June). Retrieved from http://www.internetworldstats.com/stats.htm

Jane Wightwick, M. G. (2007). *Arabic verbs and essentials of grammars* (2nd ed.). McGraw-Hill.

Jeffrey, A. ,.-C., Lara, V., Stefan, M. ,., & Daniel, H. (2004). Cross-lingual searching and visualization for greek and latin and old norse texts. *JCDL*, (p. 383).

K JAWAHAR BABU, V. H. (2012). The Role Of Information Retrieval In Knowledge. *International Journal of Social Science & Interdisciplinary Research, 1*(10).

Kanaan G., A.-S. R. (2009). A comparison of text-classification techniques applied to Arabic text. *Journal of the American Society for Information Science and Technology*.

Kanaan, D. R. (2014, January). AN IMPROVED ALGORITHM FOR THE EXTRACTION OF TRILITERAL ARABIC ROOTS. *European Scientific Journal, 10*(3).

Kazem Taghva, R. E. (2010). Arabic Stemming Without A Root Dictionary. *Information Technology: Coding and Computing, 2005. ITCC 2005. International Conference. 1*, pp. 152-157. Las Vegas: Information Science Research InstituteUniversity of Nevada.

Kazuyuki, Y., Takao, T., & NingZhong. (1999). Multi-lingual Intelligent Information Retriever with Automated Ontology Generator. *Third International Conference on Knowledge based Intelligent Information Engineering Systems.*

Khoja, S. a. (1999). R. Stemming Arabic Text. *Computing Department, Lancaster University, Lanca*, 1-7.

Kiraz, N. H. (2005). Morphological Analysis and Generation for Arabic Dialects. *Semitic '05 Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages* (pp. 17-24). Columbia: Center for Computational Learning Systems.

Kjersti Aas, a. L. (1999). Text Categorisation: A Survey. *Technical report, Norwegian Computing Center*.

Kraaij, W. a. (1996). Viewing stemming as recall enhancement. *SIGIR '96 Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 40-48). Proceedings of ACM SIGIR96.

Krovetz, R. (1993). Viewing morphology as an inference process.

Larkey, B. a. (2002). Improving stemming for Arabicinformation retrieval: Light stemming and co-occurrence analysis. *ACM 2002 Article. Bibliometrics Data Bibliometrics.* Tampere, Finland: Proceedings o fthe 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'02).

Lavrenko, & V., M. C. (2002). Cross-lingual relevance models. *SIGIR '02 Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, (pp. 175–182).

Lennon, P. T. (1988). An evaluation of some conflation algorithms for information retrieval. *Taylor Graham Publishing London, 3*(177-183), 230-236.

Luca, E., Hauke, S., Nurnberger, A., & Schlechtweg, S. (2006). MultiLexExplorer-Combining Multilingual Web Search with Multilingual Lexical Resources. *the Combined Workshop on Language-Enabled Educational Technology and Development and Evaluation of Robust Spoken Dialogue Systems*, (pp. 17-21). Germany.

M. R. Al-Maimani, A. N. (2011). Arabic information retrieval: techniques, tools and challenges. (pp. 541 - 544). Dubai : GCC Conference and Exhibition- IEEE.

M., D. (2003). *Data mining: Introductory and advanced topics* (1st ed.). Pearson Education.

Manning, D. C., & Schutze, H. (1999). Foundations of statistical natural language processing. *MIT*.

Marshall, R., Thian-HuatOng, & Hsinchun, C. (1998). Multilingual Input System for the Web -An Open Multimedia Approach of Keyboard and Handwriting Recognition for Chinese and Japanese. *ADL*, (pp. 188-194).

Mayfield, J., & McNamee, P. (2004). Triangulation Without Translation. *The 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 490-491). USA.

McNamee, P, & Mayfield, J. (2002). Comparing cross-language query expansion techniques by degrading translation resources. *SIGIR '02 Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (2002), pp. 159–166.* , (pp. 159–166).

Mizera-Pietraszko, J. (2009). Interactive Document Retrieval from Multilingual Digital Repositories. *IEEE Xplore Digital Library, IEEE Computer Society Press, str. ICADIWT*, (pp. 423-428).

Mustafa, S. H. (2012). Word Stemming for Arabic Information Retrieval: The Case for Simple Light Stemming. *ABHATH AL-YARMOUK: "Basic Sci. & Eng." , 21*(1), 123-144.

Nwesri A., S. T. (2005). Stemming Arabic Conjunctions and Prepositions. *String Processing and Information Retrieval, Lecture Notes in Computer Science, 3772*, 206-2017.

Nwesri A., S.M.M. Tahaghoghi and Falk Scholer. (2007). Arabic Text Processing for Indexing and Retrieval. *String Processing and Information Retrieval*.

Oard, D., He, D., & Wang, J. (2008). User-Assisted Query Translation for Interactive Cross-Language Information Retrieval. *International Journal of Information Processing and Management, 44(1)*, pp. 181-211.

Paraic, S., & Jean, P. B. (1996). Experiments in Multilingual Information Retrieval Using the SPIDER System. *SIGIR*, (pp. 58-65).

Peter, A. C., & AhmedAbdelali. (2008). The Effects of Language Relatedness on Multilingual Information Re-trieval: A Case Study With Indo-European and Semitic Languages. *The second International workshop on cross lingual information access.*

Peters, C., Braschler, M., & Clough, P. (2012). Multilingual information retrieval: From research to practice. *Multilingual Information Retrieval: From Research to Practice*.

PothulaSujatha, & Dhavachelvan, p. (2011 , October). A Review on the Cross and Multilingual Information Retrieval. *International Journal of Web & Semantic Technology (IJWesT) 2(4),* (pp. 115-124).

Qin, J., Zhou, Y., Chau, M., & Chen, H. (2003). Supporting Multilingual Information Retrieval in Web Applications: An English-Chinese Web Portal Experiment. *ICADL*, (pp. 149-152).

Qusay Walid Bsoul, M. M. (2011). Effect of ISRI Stemming on Similarity Measure for Arabic Document Clustering. *7th Asia conference on Information Retrieval Technology* , 584-593.

R. Al-Shalabi, G. K.-S. (2003). New approach for extracting Arabic roots. *International Arab Conference on Information Technology* , 42-59.

R., D. (2006). Machine Learning for Arabic text Categorization. *Journal of the American Society for Information Science and Technology*, 1005-1010.

Rijsbergen, C. a. (1994). An evaluation method for stemming algorithms. *SIGIR '94 Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 42-50). London: Proceedings of the Annual International ACM-SIGIR Conference on Research and Devlopment in Information Retrieval (UK).

Saad, M. K. (2010). *Open Source Arabic Language and Text Mining Tools.* Retrieved Augest 2010, from http://sourceforge.net/projects/ar-text-mining

Said, D. W. (2009). A Study of Text Processing Tools for Arabic Text Categorization. *Electronics Research Institute, Cairo, Egypt* .

Salton, G. (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer.* Boston, MA, USA: Addison-Wesley Longman Publishing Co.

Sanjay, K., & Ganesh, C. (2016, February). A SURVEY ON CROSS LANGUAGE INFORMATION RETRIEVAL. *International Journal on Cybernetics & Informatics (IJCI), 1*.

Sawalha, M. A. (2013, May). Comparative evaluation of Arabic language morphological analyzers and stemmers. *International Journal of Managing Information Techn, 5*(2).

Shatkay, & Hagit. (2005). Hairpins in bookstacks: information retrieval from biomedical text. *Briefings in Bioinformatics, 6(3),* pp. 222-238.

Shuang-Qing, Y., Fang, L., & Huan-Ye, S. (2002). Finding terminology translations from hyperlinks on the internet. *Proceedings of the first international conference on machine learning and cybermetics.*

Simpson, M. S., Voorhees, E. M., & Hersh, W. (2014). *Overview of the TREC 2014 Clinical Decision Support Track.*

Suominen, Hanna, Schreck, Tobias, Leroy, Gondy, . . . Keim, D. (2014, Sep 29). Task 1 of the CLEF eHealth evaluation lab 2014 visual-interactive search and exploration of eHealth data. *CLEF 2014 Working Notes*, pp. 1-30.

Swarup, S. K. (2011). Particle swarm optimization based K-means clustering approach for security assessment in power systems. *30*.

Tashaphyne. (2010). *Arabic light stemme*. (Tashaphyne) Retrieved from http://tashaphyne.sourceforge.net

Thabtah F., H. W.-s. (2008). VSMs with K-Nearest Neighbour to Categorize Arabic Text Data.

(1995). *TRANSLIB. Advanced Tools for Accessing Multilingual Library Catalogues.* ARRB Transport Research.

TurdiTohti, WiniraMusajan, & AskarHamdulla. (2008). Character Code Conversion and Misspelled Word Processing in Uyghur, Kazak,Kyrgyz Multilingual Information Retrieval System. *ALPIT* , (pp. 139-144).

Wen-Cheng, L., & Hsin-HsiChen. (2003). Description of NTU Approach to NTCIR3 Multilingual Information Retrieval. *Proceedings of the third NTCIR workshop on research in information retrieval automatic text summarization and question answering*. Tokyo.

Wikipedia. (2016). Retrieved from Wikipedia: http://en.wikibooks.org/wiki/Arabic

Youssef Kadri, J.-Y. N. (2006). Effective Stemming for Arabic Information Retrieval. *THE CHALLENGE OF ARABIC FOR NLP/MT*.

Zhang, Y., & P, V. (2004). Using the web for automated translation extraction in cross-language information retrieval. *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, (pp. 162–169).