Sudan University of Science and Technology

College of Computer Science and Information Technology

Department of Computer and Information Systems

Identifying Broken Plural in Arabic Information Retrieval Systems

التعرف على جموع التكسير في نظم استرجاع المعلومات العربية

A thesis submitted in partial fulfillment of the requirement of Bachelor Degree of Computer Science.

Submitted by: Lojain Abdalhakeem Ahmed

Manal Alshazali Osman

Supervised by: Ebtihal Mustafa Alameen.

October, 2016

بسم الله الرحمن الرحيم

Sudan University of Science and Technology

College of Compute Science and Information Technology

Department of Computer and Information Systems

Identifying Broken Plural in Arabic Information Retrieval Systems

التعرف على جموع التكسير في نظم استرجاع المعلومات العربية

A thesis submitted in partial fulfillment of the requirement of Bachelor Degree of Computer Science.

Submitted by: Lojain Abdalhakeem Ahmed

Manal Alshazali Osman

Supervised by: Ebtihal Mustafa Alameen

:Signature by

Date : / October , 2016

Verse:

وَيَسْـَٔلُونَكَ عَنِ ٱلرُّوحِ ۖ قُلِ ٱلرُّوحُ مِنْ أَمْرِ رَبِّى وَمَآ أُوتِيتُم مِّنَ
ٱلْعِلْمِ إِلَّا قَلِيلًا ﴿٨٥﴾

سورة الاسراء الاية 85

DEDICATION:

To Our:

Parents

Sisters and Brothers

Teachers

Friends

ACKNOWLEDGEMENT

We would like to take this opportunity to express my deepest gratitude to all of those who helped us directly and indirectly to successfully finish this research.

Firstly, we would like to express my sincere gratitude to our supervisor **Ebtihal Mustafa Alameen,** Head of the Information Systems Department, for walking with us step by step until the completion of this research, and for helping us understanding Information Retrieval, although she couldn’t be busier.

Beside our supervisor, we would like to thank **Dr. Intesar Ibraheem** for giving us inspiration to develop a better method than we had expected. We also want to show our gratitude for **Mohammed Adany** for helping us during our last steps to finish this research. We couldn’t be more grateful for both of them.

Lastly we are very thankful to our brothers, sisters and colleagues who offered their help and encouraged us to finish this research, and supported us during the hard times.

The thanks firstly and lastly for Allah.

Lojain Abdalhakeem

Manal Alshazali

ABSTRACT

Arabic Language is one of the most widespread languages in the world and it's newly associated with the field of the internet, so information retrieval is one of the most important fields in computer science. It is concerned with operations like indexing, searching and retrieving information, which is required by the user. Search engines are examples of information retrieval system (IRS). IRS faces many challenges when searching with Arabic language, because it is a grammatical language. Plurals in Arabic language are divided to two types Regular Plurals (RP) and Irregular/Broken Plurals (BR), IRS can identify regular plural, because it maintains the basic structure of the word, but it fails to identify BP, because the basic structure of the word changes from singular form to plural form and vice versa and that reflects negativities when applying indexing operation in IRS; because if a user types a query that contains BP, the system retrieves only the documents that contain the plural form while losing the documents that contain the singular form that should also be retrieved. Identifying BP is also one of the challenges that face Arabic IRS and it causes document loss leading to inaccurate results. This study aims at explaining how big of a challenge BP is to Arabic IRS. This study proposes a method to recognize BP and to increase Recall without affecting Precision. Proposed method consists of five stages (pattern recognition – word recognition – singular candidates – selecting the right singular form – expanding the query). This study covers only one pattern of BP patterns which is (فعاليل). Method results were compared with System baseline before and after applying the proposed method. Based on these results this study has successfully identified BP and enhanced retrieval.

المستخلص

اللغة العربية من اللغات الواسعة الانتشار في العالم وارتبطت حديثا بمجال الانترنت ولذلك استرجاع المعلومات هو احد المجالات المهمة في علوم الحاسوب والتي تهتم بعمليات الفهرسة والبحث واسترجاع المعلومات التي يطلبها المستخدم .ومن امثلة نظم استرجاع المعلومات محركات البحث , ونظم استرجاع المعلومات تواجه عدة تحديات عند البحث باللغة العربية لانها لغة نحوية . الجموع في اللغة العربية تنـ قسم الى الجموع السالمة وجموع التكسير .نظم استرجاع المعلومات يمكنها التعرف على الجموع السالمة لانها تحافظ على بنية الكلمة في المفرد والجمع , بينما تفشل في التعرف على جموع التكسير لان بنية الكلمة تتغير في المفرد والجمع والعكس .وهذا يعكس سلبيات عند تطبيق الفهرسة في نظم استرجاع المعلومات ؛ لانه اذا قام المستخدم بكتابة استعلام يحتوي على احدى صيغ جموع التكسير فان النظام يـ قوم باسترجاع المستندات التي تحتوي على الجمع كنتائج ، بينما يتم فـ قد المستندات التي تحتوي على صيغة المفرد والتي ينبغي استرجاعها . التعرف على جموع التكسير ايضا واحدة من التحديات التي تواجه نظم استرجاع المعلومات العربية وتتسبب في فـ قد المستندات وبالتالي عدم د قة النتائج .هذه الدراسة تهدف لتوضيح كيف ان جموع التكسير تمثل تحدي يواجه نظم استرجاع المعلومات العربية وا قترحت طريـ قة للتعرف على جموع التكسير وتحسين الاسترجاع وزيادة الد قة .الطريـ قة المـ قترحة تتكون من خمسة مراحل(التعرف على النمط- تمييز الكلمة - ا قتراح مفردات للكلمة - اختيار المفرد الصحيح – استخدام المفرد الصحيح في توسعة الاستفسار) .غطت الدراسة صيغة واحدة من صيغ جموع التكسير وهي (فعاليل) .تمت المـ قارنة بين النتائج قبل وبعد استخدام الطريـ قة المـ قترحة . بناء على النتائج هذه الدراسة نجحت في التعرف في جموع التكسير وتحسين الاسترجاع وحسنت في د قة النتائج .

LIST OF EXPRESSION

IR	Information Retrieval
IRM	Information Retrieval Models
IRS	Information Retrieval Systems
BM	Boolean Model
VSM	Vector Space Model
GPP	Al Gela Plurals Pattern
KPP	Al Katharh Plurals Pattern
FMJ	Formula of Montahaa Jemoa
BP	Broken Plurals
SNC	Sex Name Collective

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATION

TABLE OF CONTENTS

CHAPTER ONE

INTRODUCTION

CHAPTER ONE:

INTRODUCTION:

1.1 INTRODUCTION:

Information Retrieval has a lot of meanings, but it could be academically defined as: Finding specific information in a pile of unstructured large collection of data that satisfies the needs of the system user [1].

It wasn't a popular field in the past ,but with time and technology development, a lot of people use IR (Information Retrieval) even if they don't know it, closest example is using search engines to search for a popular singer ,another example is to take out your wallet and take out a specific card [1].

1.2 PROBLEM STATEMENT:

Arabic Information Retrieval faces many challenges that don't exist in other languages, because Arabic language is full of derivation (الاشتقاق) and grammatical rules. Another challenge is that Arabic language is a morphological language [2]. In English language, if we want to form the subject from a verb we add (er) to the end of the verb However in Arabic language sometimes it means adding (ون) ,(ين) or (ات) to the end of the word, and sometimes it means changing the whole structure of the word.

.

1.3 RESEARCH OBJECTIVES:

- To increase recall but not to decrease precision.
- To be able to search using both Broken Plural form of the word and the singular form of the word to Improve Arabic Language Information Retrieval Systems by identifying the Arabic Broken Plural and their patterns.
- In this research, it is wanted to develop a method to generate multiple types of the singular form, find the accurate one and expand the query with it.

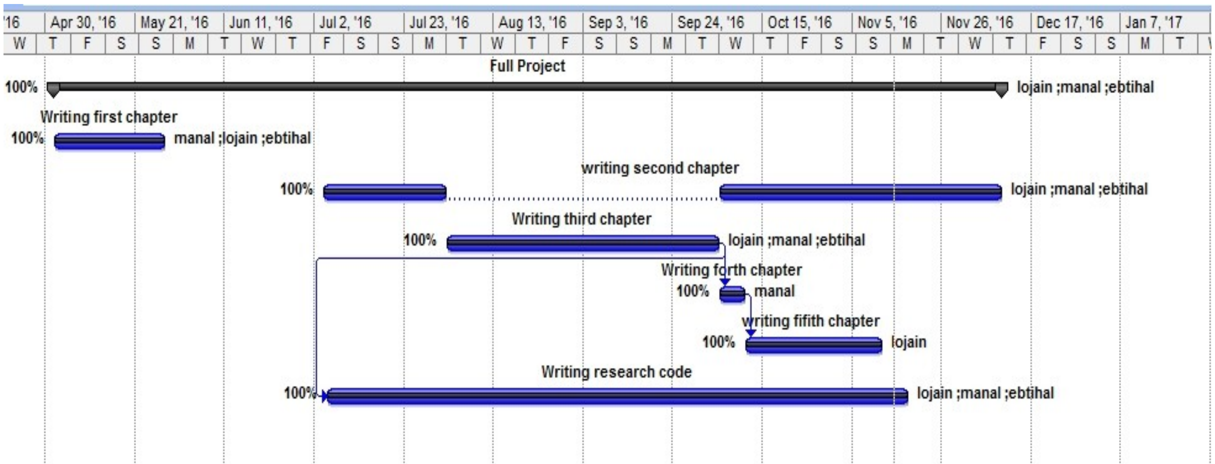
1.4 RESEARCH SCOPE:

This study aims at enhancing IRS by proposing a method to find the right singular form for BP. This method is applied offline to find the right singular form of the pattern (فعاليل), which is a pattern of BP.

1.5 METHODOLOGY:

In this research we used Java Lucene Package as a basic. Lucene is one of IRS, it uses an Arabic Analyzer. Arabic corpus watan-2004 was used in this study, this corpus contains documents about various topics. RapidMiner software was also used to perform some processes on watan-2004 in the first stage of the research.

1.6 GANTT CHART:



1.7 RESEARCH HYPOTHESIS:

IRS (Information Retrieval System) fails to retrieve all documents relevant to a user's query, because IRS fails to deal with Arabic BP (Broken Plurals) properly. This study proposes a method to solve BP identification problem and enhance document retrieval.

1.8 RESEARCH ORGANIZATION:

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

CHAPTER TWO

BACKGROUND AND RELATED WORK

CHAPTER TWO

1 BACKGROUND AND RELATED WORK:

1.9 INTRODUCTION

This Chapter previews basic concepts that are considered the keys of this research and related works in the same field. This chapter is arranged into five sections. Section 2.2 explains the principles of Information Retrieval IR (definition, basic concepts, Models, IR for Arabic language), while section 2.3 explain the Arabic Language and Plurals (Arabic Features Affecting Retrieval, Arabic Number System). Section 2.4 explains the Data set corpus that this research depends on. Section 2.5 presents Evaluation Measures that are used to evaluate the results. Section 2.6 preview Related Works based on Broken Plural identification.

1.10 INFORMATION RETRIEVAL BACKGROUND:

Information retrieval is one of the important areas in Computer Science concerned with indexing and searching operations for web sites to simplify and improve the search process. IR has become one of the most important areas where most of the studies and researches have worked in this area and research efforts are continuing to improve this domain.

Nowadays we frequently think first of web search, but there are another cases such as E-mail search, searching our laptop, corporate Knowledge bases and legal IR.

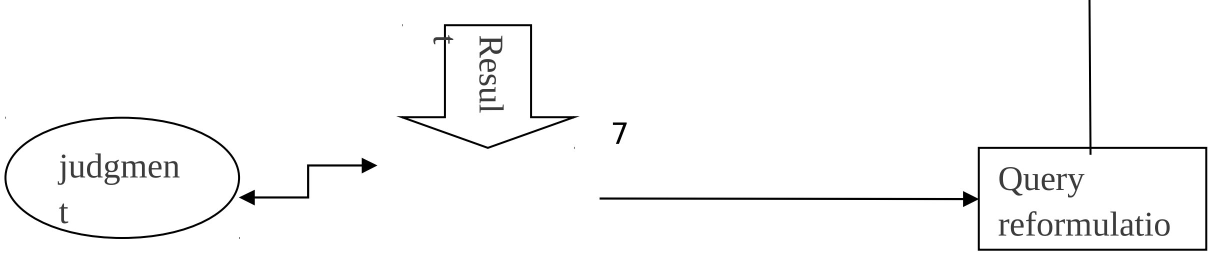
An IRS could be described by different levels: by types of users, types of data, and the types of the information need, along with the size and scale of the information repository it addresses.

1.10.1 INFORMATION RETRIEVAL DEFINITION:

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers) [1].

1.10.2 INFORMATION RETRIEVAL BASIC CONCEPT:

In the beginning of Section 2.2, we mention Information Retrieval consists of two important processes. The first process is (Indexing), and the second process (Searching). Figure 2.1 illustrates the typical Information Retrieval processes, which IRS follows.



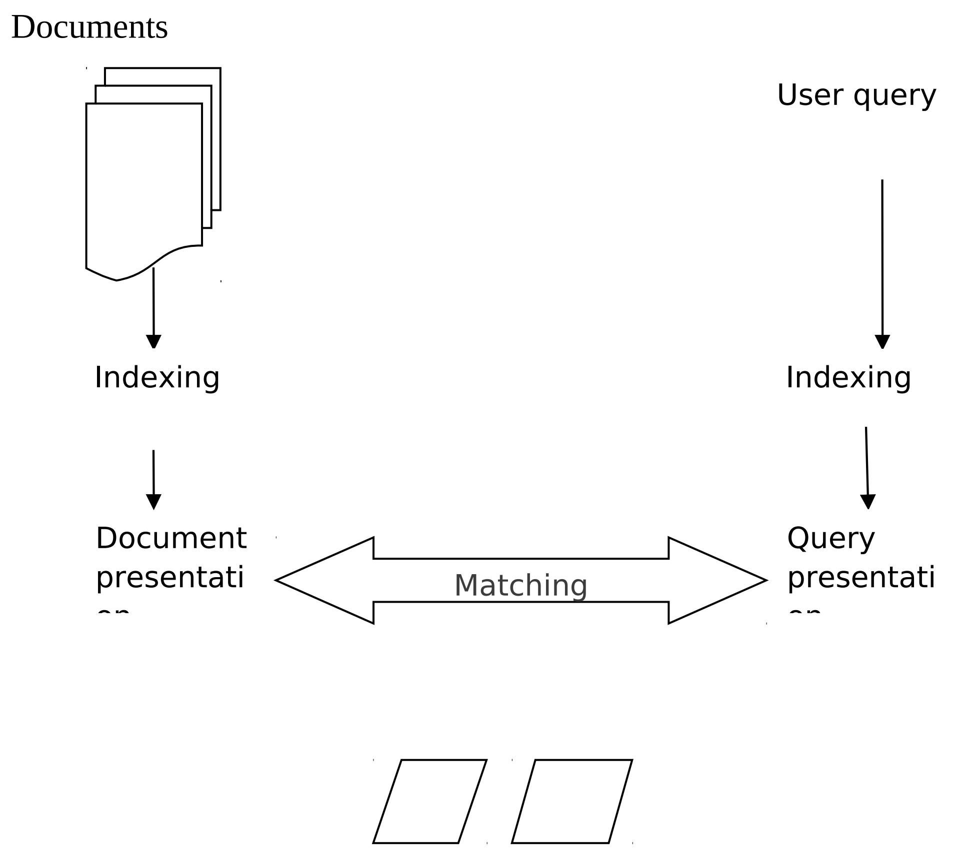


Figure (2.1) Information Retrieval Process

1.10.2.1 INDEXING:

It means storing documents and terms so that we can retrieve documents efficiently, effectively, it requires reasonable space.

In the indexing, there are major steps in inverted index construction:

- Collect the documents to be indexed.
- Tokenize the text.
- Perform linguistic preprocessing of tokens.
- Index the documents that each term occurs in [1].

1.10.2.2 SEARCHING:

In Information Retrieval Systems, we mean by searching how to search about information or documents using IRS and obtain good result from a query [4].

1.10.2.3 PRE-PROCESSING:

Pre-processing is the important stage for Indexing and Searching processes because it makes compatibility within the query statement and indexed data. The importance of this stage is that it allows the user to write query without restricting and makes it compatible with existing index. Pre-processing contains several operations; in the next paragraphs will explain briefly about the most important processes in this stage:

Tokenization: Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called tokens, perhaps at the same time throwing away certain characters [1] :

Input: التعرف على جموع التكسير في نظم استرجاع المعلومات

Output:

المعلومات	استرجاع	نظم	في	التكسير	جموع	على	التعرف

Figure (2.2) Token processes

Lower case: we all know the English language and related for languages (use the same alphabet), it means changing all capital letters to small letters [4].

Stop word: It means very common words, which would appear to be of little value helping selected documents matching the user's need [4]. This process removes some word for indexing purpose [4]. Stop word are different from one language to another, for example preposition in English language words like (in, on, if…etc.) denoted as "stop word" and it will be removed [4].

As for Arabic Language (من , الى , في , على) denote as "stop word". Appendix A stated an Arabic *Stop word* [4].

Stop word remove: It means elimination of stop word with the objective of filtering out words with very low discrimination values for retrieval purposes [4].

Stemming: it means removing affixes and allowing the retrieval of documents containing syntactic variations of query terms [4].

1.10.3 INFORMATION RETRIEVAL MODEL (IRM):

The major task in information retrieval is to find relevant documents for a given query [4]. We briefly describe the important models of IR. There are three statistical models firstly Boolean model, secondly vector space model, and thirdly probabilistic model. There is another kind of model called semantic model [5].

1.10.3.1 BOOLEAN MODEL (BM):

In this model documents are represented as a set of ***terms*** [5]**.** The queries formulated as using standard Boolean logic set-theoretic operators such as AND, OR and NOT. Retrieval and relevance are considered as binary concepts in this model, so the retrieved elements are an "exact match" retrieval of relevant documents.

Boolean retrieval models lack sophisticated ranking algorithms and are among the earliest and simplest information retrieval models [5].

1.10.3.2 VECTOR SPACE MODEL (VSM):

This model provides a framework in which term weighting, ranking of retrieved documents, and relevance feedback are possible. Documents are represented as ***features*** and ***weights*** of term features in an *n*-dimensional vector space of terms [5].

The query is also specified as terms vector (vector of features), and this is compared to the document vectors for similarity/relevance assessment. Both queries and documents are specified as lists of terms and mapped into an *n*-dimensional space (where *n* is the number of possible terms). The relevance then depends on the angle between the vectors [5].

In the vector model, the *document term weight wij*(for term *i* in document *j*) is represented based on some variation of the TF (term frequency) or TF-IDF (term frequency-inverse document frequency) scheme (as we will describe below). **TF-IDF** is a statistical weight measure that is used to evaluate the importance of a document word in a collection of documents. The following two formulas are typically used [5]:

Equation (2.1) cosine similarity

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

In the formula given above, we use the following symbols:

- *dj* is the document vector.
- *q* is the query vector.
- *wij* is the weight of term *i* in document *j*.
- *wiq* is the weight of term *i* in query vector *q*.
- |*v*| is the number of dimensions in the vector that is the total number of important keywords (or features).

Equation(2.2) TF-IDF Weighting

$$TF_{ij} = f_{ij} \Big/ \sum_{i=1\ to\ |v|} f_{ij}$$

$$IDF_i = \log(N/n_i)$$

In these formulas, the meaning of the symbols is:

- *TFij* is the normalized term frequency of term *i* in document *Dj*.
- *fij* is the number of occurrences of term *i* in document *Dj*.
- *IDFi* is the inverse document frequency weight for term *i*.
- *N* is the number of documents in the collection.
- *ni* is the number of documents in which term *i* occurs.

1.10.3.3 PROBABILISTIC MODEL:

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

Given the document representation D of a document, estimating the relevance R and non-relevance $\bar{R}$ of that document involves computation of conditional probability $p(R|D)$ and $p(\bar{R}|D)$. The ratio between $p(R|D)$ and $p(\bar{R}|D)$ denoted as Odd, and used as a score to determine the likelihood of the document with representation D belonging to the relevant set. More formally:

$$odd(R|D) = \frac{p(R|D)}{p(\bar{R}|D)}$$

By applying Bayes's theorem:

Equation (2.3) Bayes's theorem

$$odd(R|D) = \frac{p(D|R)p(R)}{p(D|\bar{R})p(\bar{R})} = \frac{\prod_k p(D_k|R)p(R)}{\prod_k p(D_k|\bar{R})p(\bar{R})}$$

Where $p(\bar{R})$ and $p(R)$ are the prior probability of retrieving a relevant document or non-relevant document, respectively, $D_k$denotes thek^{th} term in the document vector [6].

1.10.3.4 SEMANTIC MODEL:

Semantic approaches include different levels of analysis, such as (morphological, syntactic, and semantic analysis), to retrieve documents more effectively. In ***morphological analysis***, roots and affixes are analyzed to determine the parts of speech (nouns, verbs, adjectives, and so on) of the words. Following morphological analysis, ***syntactic analysis*** follows to parse and analyze complete phrases in documents. Finally, the semantic methods have to resolve word ambiguities and/or generate relevant synonyms based on the ***semantic relationships*** between levels of structural entities in documents (words, paragraphs, pages, or entire documents) [5].

1.11 ARABIC LANGUAGE AND PLURAL:

Arabic language is one of the most widespread and commonly used languages, which is used by millions of users in the Internet on a daily basis [7]. Also it is one of the top 10 languages used on the Web in terms of growth during the time period 2000-2011 [8]. To spread Arabic culture must be profiteer the Internet and web site and simplify the search processes for Arabic web sites. Arabic language faces many challenges especially in Information Retrieval System applications such as search engine [7].

1.11.1 ARABIC FEATURES AFFECTING RETRIEVAL:

When we search about information over the internet written in Arabic language we face many challenges like [2]:

In Arabic Language the words are written from ***right-to-left*** and numbers are written from ***left-to –right***, alphabet consists of 28 basic letters all of them change in shape depending on their position in the word, like the ***HAMZA***: (ا, أ, إ, آ, ء, ؤ، ئ), fifteen letters contain dots to differentiate them from other letters like (ب ، ت ، ث ، ج ، خ ، ذ ، ز ، ش ، ض, ظ , غ ، ف ، ق ، ن ، ي), ***Ligature*** (الوصلات) special forms for some character sequences like (ل + ا ≠ لا) ,optional ***diacritics*** -except for holy Quran- (mostly short vowels) like(بُ) and ***Kashida*** symbols that extend t he length of words like (تـــ) [2].

1.11.2 ARABIC NUMBER SYSTEM:

Arabic language has two grammatical genders: feminine and masculine, and three grammatical numbers: singular (المفرد), dual (المثنى), and plural (الجمع), and three grammatical cases: nominative, genitive, and accusative. A noun has the nominative case when it is a subject; accusative when it is the object of a verb, and genitive when it is the object of a preposition. Figure (2.8) shows the Arabic number system hierarchy [9]. The concept of Plural in Arabic differs from English, in English; a plural noun can refer to two or more things, but in Arabic a plural noun refers to three or more things and dual refer to two things [4].

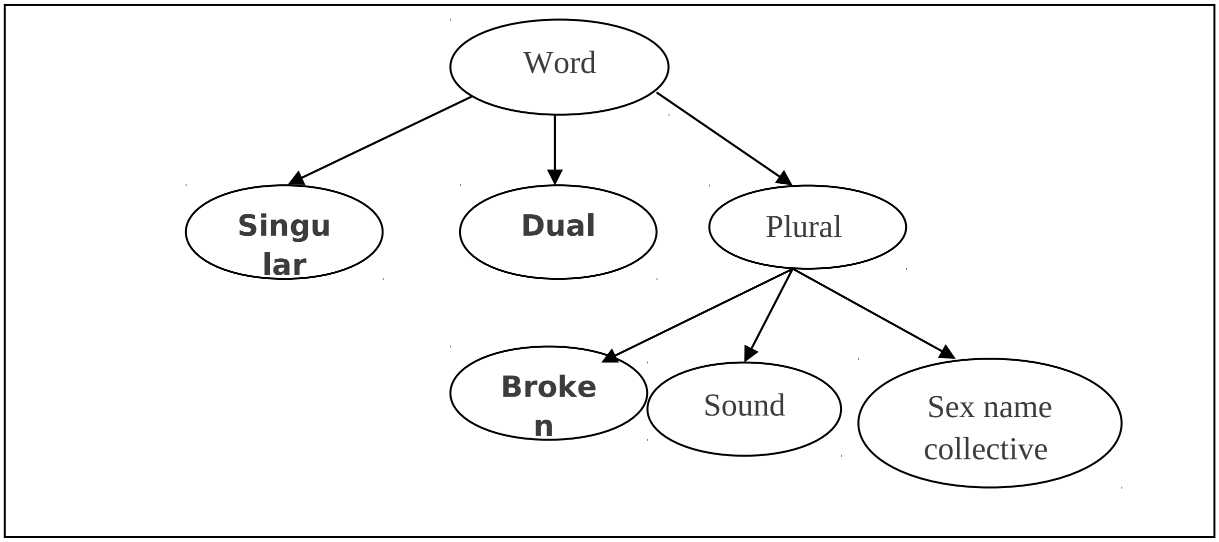


Figure (2.3) Arabic number system hierarchy

1.11.2.1 ARABIC PLURAL:

Plurals in Arabic language are divided into *Sound Plurals* and *Broken Plurals*, the following paragraph will illustrate briefly the *Sound plurals* [10]. There is another type of plurals called *Sex Name Collective (*10] (اسم الجنس الجمعي]. As stated before, this study focuses more on Broken Plurals. Figure 2.9 illustrate the area of BP, which is concerned by this study.

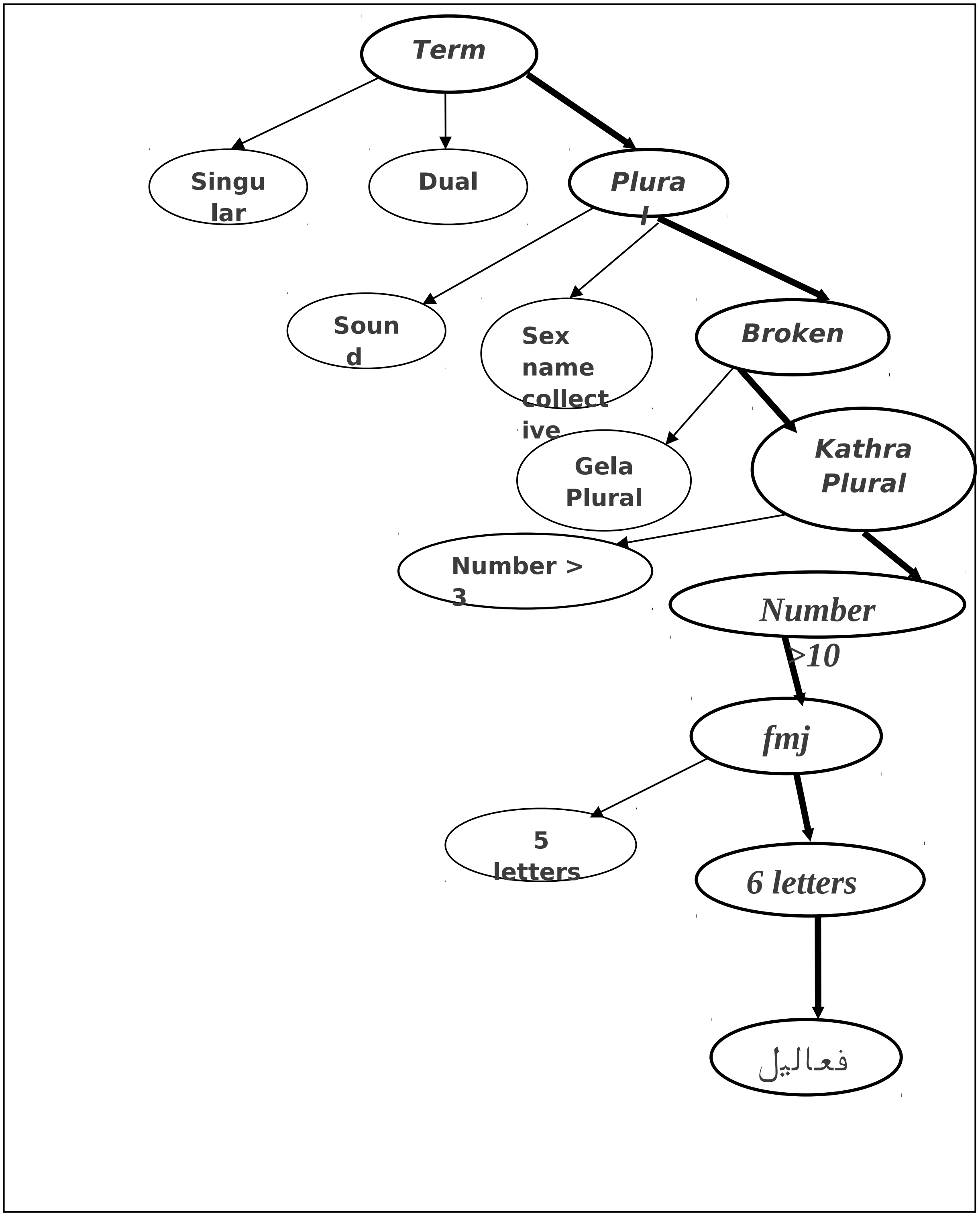


Figure (2.4) The area of BP which concerned by this study

- **Sound plural:**

Sound plural is a type of plural in Arabic language; it is divided to two types ***masculine plural*** and ***feminine plural [10].***

- ***masculine*** **plural : "**it is a Sound plural which is used for more than two, remained unscathed changed single, called this plural of a male or a recipe for same masculine by adding suffix (ين,ون) to singular nouns" [4], for example the word (مسلم ,"Muslim") pluralized to (مسلمون ,"Muslims") added suffix (ون)in the nominative case [10] . Also same word pluralized to (مسلمين ," Muslims")added suffix (ين) in the genitive and accusative cases [10] .

- ***Feminine plural*** : "it is Sound plural which is used for more than two, remained unscathed changed single, called this plurals of a female or a recipe for same Feminine by adding suffix (ات) to singular noun"[4].for example the word (طالبة, student) pluralized to (طالبات ,"students")in the nominative , genitive and accusative cases [10].

1.11.2.2 SOUND PLURAL AND INFORMATION RETRIEVAL SYSTEM:

Information retrieval system such as search engines can handle this type of plurals because sound plural has a particular rule. In addition the structure of singular remain unchanged ,all these features made this type of plural (sound plural) easy to identify and facilitate the understanding and it is applicable to Information Retrieval Systems, but another type of plurals which is called Broken Plural has no particular rules. and the structure of the singular will change, and the Information Retrieval Systems cannot handle this type of plural, this study shows how to recognize Broken plurals, and discuss in detail in the following paragraphs about this type of plural [4].

- **Broken Plural (BP) :**

The Broken Plurals(جموع التكسير) is another type of plurals in Arabic language, in this type of plurals there are no a specific rules governed, each word treated quite differently based on some patterns as we will show in this section, where it is difficult to identify them. Broken plurals similar to irregular nouns in English (e.g.: foot/feet), but this type of plurals are very common in Arabic, it represents more than 40% of the plurals in Modern Standard Arabic, while the remaining percentage 60% is assigned to the other types of plurals, sound masculine and feminine plurals[4].

Although BP does not have a fixed rule it depends on, but it comes in a set of patterns where these patterns were divided into two categories, namely (**Al Gela Plurals** جموع ال قلة) and (**Al Katharh Plurals** 11] (جموع الكثرة].

- **Al Gela Plurals Pattern (GPP) (انماط جموع ال قلة).**

  One of Broken Plurals patterns, this category has several patterns, which can be used for plurals from three to ten. This category has four patterns as in the following table [11].

Table (2.1) Al Gela Plurals pattern (انماط جموع ال قلة)

NO	***Pattern***	***Example/ plural***	***Singular***	***Plurals***
1	أَفْعُل	أنفس	نفس	**3<=10**
2	أَفْعَال	أحمال	حمل	**3<=10**
3	أَفْعِلَة	أحمرة	حمار	**3<=10**
4	فِعْلَة	صبية	صبي	**3<=10**

- **Al katharh plurals pattern (KPP)** انماط جموع الكثرة :

This category has several patterns which can be used for plurals more than ten or more than three except **Formula of Montahaa Jemoa (FMJ) (** صيغ منتهى الجموع**)** which has to be used for more than ten **[10].**This category have twenty three patterns, seven of these patterns are syntax of montahaa jemoa [14] in the Table(2.2) some of **al katharh plural pattern [10]** and table (2.3) shows **FMJ. [4]**

Table (2.2) Al katharh plurals pattern(انماط جموع الكثرة **)**

NO	***Pattern***	***Example/ plural***	***Singular***	***Plurals***
1	فُعْل	حمر	حمراء	**> 10**
2	فُعَل	غرف	غرفة	**> 10**
3	فَعَلَة	سحرة	ساحر	**> 10**
4	فُعُول	قرود	قرد	**> 10**
5	فِعلان	صبيان	صبي	**> 10**
6	أَفْعِلَاء	أصدقاء	صديق	**> 10**

Table(2.3) Syntax of montahaa jemoa. (صيغ منتهى الجموع)

NO	*Pattern*	*Examples/plural*	*Singular*
1	فواعل	لوائح	لائحة
2	فعائل	رسائل	رسالة
3	فعالل	دراهم	درهم
4	مفاعل	مسارح	مسرح
5	افاعل	اصابع	اصبع
6	مفاعيل	مناشير	منشار
7	فعاليل	تقارير	تقرير

This research concerns analyzing this category of patterns(صيغ منتهى الجموع), it investigates how to identify patterns and make analysis to extract rules to help to distinguish is the word represent a plural, singular or not . Then get all singular forms of word if it represents plurals. In addition, expand the correct singular form. Chapter 3 discusses the following pattern faaleel (فعاليل).

- Sex Name Collective (SNC) :

  It is used to indicate plural and also indicates sex [10].example (ثمر ," fruits") and its singular form (ثمرة ,“ fruit”) [10] .

1.11.2.3 BROKEN PLURALS CHALLENGE FOR INFORMATION RETRIEVAL SYSTEM:

The broken plurals identification represents a problem especially for Information Retrieval Applications. It is difficult to deal with Arabic broken plurals and reduce them to their associated singulars, because there is no specific rule governed, each word treated quite differently, where it is difficult to identify them, and no obvious rules exist, also there are no standard stemming algorithms that can deal with this type of plural [4] .

Some patterns are analyzed to extract some rules which are used to identify and determine if a word represents a plural or not, and another patterns that were difficult to be analyzed. To denote the importance of this study, there are some examples derived to explain some of these challenges in the following section [4].

Challenge 1: Some patterns are difficult to be identified due to their plural structure have three letters. The root of words in Arabic language is (Fa-al, فُعل ") and that makes identifying the word that represents a plural from a word that represents a verb a difficult process for IRS. For instance the word (جلس) is singular, and the word (حُمر) represent plurals.

Lucene is an example of IRS, it uses Arabic Analyzer to preprocess the input words first. Then, it stripes and return the origins of the word without suffix, prefix or special characters (التشكيل) supplemented to the word [4].

Challenge 2: stemming process deletes some of the original characters for some words as a suffix, therefore the meaning of the word is missing [4] .

Challenge 3: Some words that match a plural form but they are not. For instance the suitable pattern for word (احتواء) is (افعلاء), but this word does not represent a plural. For more explanation in contact, the word (ا قرباء) represent a plural [3].

Challenge 4 : Preprocessing may remove some letters from the word as suffix, but these letters are in fact original letters, then it is made difficult to identify because the word become vague ,for example word (4] (قوانين].

This study mainly focuses on the third challenge and the fourth challenge .Next two sections explore simply the problem statement this research, which is based on challenge 3. When reviewing some State-of-art works, current Arabic based IRS failed to recognize the BP. For instance, when write the term (تحاليل), documents that contain (تحليل) are not retrieved, so the precision of these search engine are low. Therefore, this study is proposed to increase recall without decreasing precision. See the example below:

D1: “ اجريت العديد من التحاليل على مرضى السرطان لايجاد العلاج المناسب لهذا المرض ”.

D2: “ يتم تحليل العينات المؤخوذة من مرضى السرطان بمختبر تتوفر فيه عدد من الشروط ”.

Assuming that was indexed these documents using Arabic Analyzer, if the user write query Q1: “ مجموعة التحاليل التي اجريت على مرضى السرطان ” and then press the search button, the system will retrieve the first document D1 only, because it contains the word (تحاليل , ” analysis”). However, the system will not retrieve the second document D2 although it contains the word (تحليل ,” analysis”) which represents the singular of the word (تحاليل). Although (تحليل) should have been retrieved but it won’t be retrieved because the system fails to recognize that the word represents the singular of the word (تحاليل).

As for challenge 4, if the user types a query that contains the word (قوانين), then any Arabic Analyzer like Lucene doesn’t recognize its right singular form and that is because the Arabic Analyzer removes (ين) from (قوانين) and it becomes (قوان). However, the word (قوان) is meaningless. We will be addressing these problem and try to solve them.

1.12 INFORMATION RETRIEVAL DATA SET:

The data set is a very important component for information Retrieval systems, it contains a large number of documents and files, and our analysis depends on words, which are extracted from this data set. This research depends on Watan-2004 corpus, we will address this dataset more in the next chapter “**Research Methodology**”.

1.13 EVALUATION MEASURE:

The performance of an IR system can be measured in different ways, depending on retrieval task and relevance judgment used. If the binary relevance judgments are employed for assessing documents, then ***precision*** and ***recall*** measures can be used [7].

1.14 RELATED WORK:

There are only few studies addressing the problem of broken plurals, they differ from one another. Some of these studies work on deriving broken plurals from their singulars or roots, while others aimed at extracting singulars from plural forms [3]. To the best of our knowledge, there are two studies proposed approaches to identify Arabic Broken plurals, and some studies used these approaches for other topics such as translation as shown in the following paragraph.

Study [3] proposed three approaches to identify Broken Plurals (BP). The first approach is the ***Simple Broken Plurals Matching Method***. The basic idea is to get a word, use light-stems to produce morphological information such as stemming prefix and suffix, then returns TRUE if the word match one BP patterns in the list or FALSE. This method identifies plurals by match the word with broken plurals patterns by checking some letters in the word with equivalent letters on same positions in pattern. Although it is simple method that made it easy to implement but the main problem with the simple BP matching approach is that the BP patterns are too general to achieve a good performance, which means there are several words have the same pattern but they do not represent BP. The results showed that the ***Simple Broken Plural Matching*** approach has low precision (13.73%) - on a test set of about 187,000 words.

To improve the performance of identification the same authors proposed ***Restricted Broken Plural matching Method***. In this approach, the development of Simple

Broken Plural method it increase the precision which obtain more specific BP patterns by restricting the original one. The main idea is to allow only a subset of the alphabet to be used in the meta characters (ع) ,(ف), and (ل) positions of the patterns was the restricted matching method, in which the broken plural patterns are used to detect broken plurals according to sets of rules that govern their applicability. This method makes some tests to identify a word that represent plural or not. Those steps are summarized as the following. First check the word with BP patterns, if it matches one of these patterns, then checks the position of character based on rule that obtained from analyzed patterns. The results of this approach showed an increase in the precision reaching about 75%. The third approach for identifying broken plural was built on the top of the previous. This approach used a dictionary which lists broken plural stems. This dictionary was constructed automatically by extracting all instances of broken plural stems that match broken plural patterns. Next, sets of rules, as in the previous approach, were extracted. Results showed that a significant improvement in precision, reaching 92%, compared to other two approaches [3].

Study [4] **Proposed Broken Plurals Processing Method for enhancing the performance of Arabic Information Retrival Systems.** Information Retrival System (IRS) faces a fundamental challenge in some languages especially Arabic language because it considered a morphological language. A plural in the Arabic language is divided into two types Sound Plurals (SP) and Broken plurals (BP). IRS can identify the sound plurals simply because it keeps the structure of the words singular and plural form. Whereas IRS fails to recognize the BP form of the word is derived from its plural form and vice –verse. In addition, this is reflected negatively when implementing indexing in Arabic IR. For instance, if a user typed a query contains plural form, system can retrieve all documents contain plurals form the result, while system misses documents which contain singular form for the same word which should be retrieved. BP identification represent one of challenges faces Arabic IRS and causes loss of relevant documents ; this is therefore lead to reduce Arabic IRS accuracy as a result . This study aims at exploring how Arabic BP represent challenge faces Arabic IRS, and suggests a methodology based on the analysis words to resolve

BP identification problem and retrieval. The proposed consists three stages which are: ***preprocess***, **BP *identification*** ,***query expansion*** . That study covered three pattern of formula of montaha Jemoa (FMJ) which are (Tfaaeel تفاعيل – faaeel فعاعيل - fyaeel فياعيل). Method results were compared with (System baseline) after applying the proposed method . His research findings, that study successfully able to identify broken plural word and enhance retrival and precision [4].

1.15 SUMMARY:

In this chapter, researchers reviewed an Information Retrieval concept, Models, Applications. Also gave a brief about Arabic language and challenges which faces Arabic language for Information Retrieval System. Additionally reviewed an Arabic number system, which is, explained the types of plurals in Arabic Sound plurals and Broken Plurals. Challenges, which face Broken Plurals identification, are also discussed. This chapter reviewed the selected Data set, the evaluation measures, and reviewed the related work to this research.

The next chapter explains the proposed methodology and its stages, which are followed to solve the research problem.

CHAPTER THREE

RESEARCH METHODOLOGY

CHAPTER THREE:

RESEARCH METHODOLOGY

1.16 INTRODUCTION:

As mentioned before Arabic language is full of derivation and grammatical rules and is a very morphological language.

There are two kinds of plurals in Arabic language, Regular plural (الجمع السالم) and Irregular/Broken plural (جمع التكسير). The Regular plural is easy to work on during the stemming process, but the BP on the other hand might cause problems with the current stemmers.

Currently search engines uses Light stemming to process the Arabic language. The Light stemmers merely remove the initials (و , وال , كال , لل , فال , ال , بال), and also remove suffixes (ي, ة , ه , ية , يه , ين , ون , ات , ان , ها). Which is effective in the case of Regular plurals, however it is not effective in the case of BP, because the BP changes the structure of the word.

This chapter shows how we can fix that by improving the identification of Broken plural.

1.17 TOOLS AND TECHNIQUES:

1.17.1 INFORMATION RETRIEVAL DATA SET (ARABIC CORPUS):

Watan-2004 is an Arabic corpus, which contains 20291 documents organized in 6 categories (Culture, Economy, International, Local, Religion, Sports) [12].

1.17.2 RAPIDMINER :

RapidMiner Studio makes predictive analytics lightning-fast, radically reducing the time to unearth opportunities and risks. Our cutting-edge approach brings together all the necessary tools for accelerating the creation, delivery, and maintenance of high-value predictive analytics. [13].

1.17.3LUCENE PACKAGE:

Lucene is a java full-text search library that is used to smoothly add search functionality to a website or an application. The Lucene package makes searching easy by adding content to a full-text index. It then allows you to perform queries on this index, returning results ranked by the relevance to the query.

Lucene is able to achieve fast search responses because, instead of searching the text directly, it searches an index instead [14].

1.18 RESEARCH FRAMEWORK:

Figure (3.1) shows the general framework of the proposed method. In next sections, we are going to describe each component found in the framework.

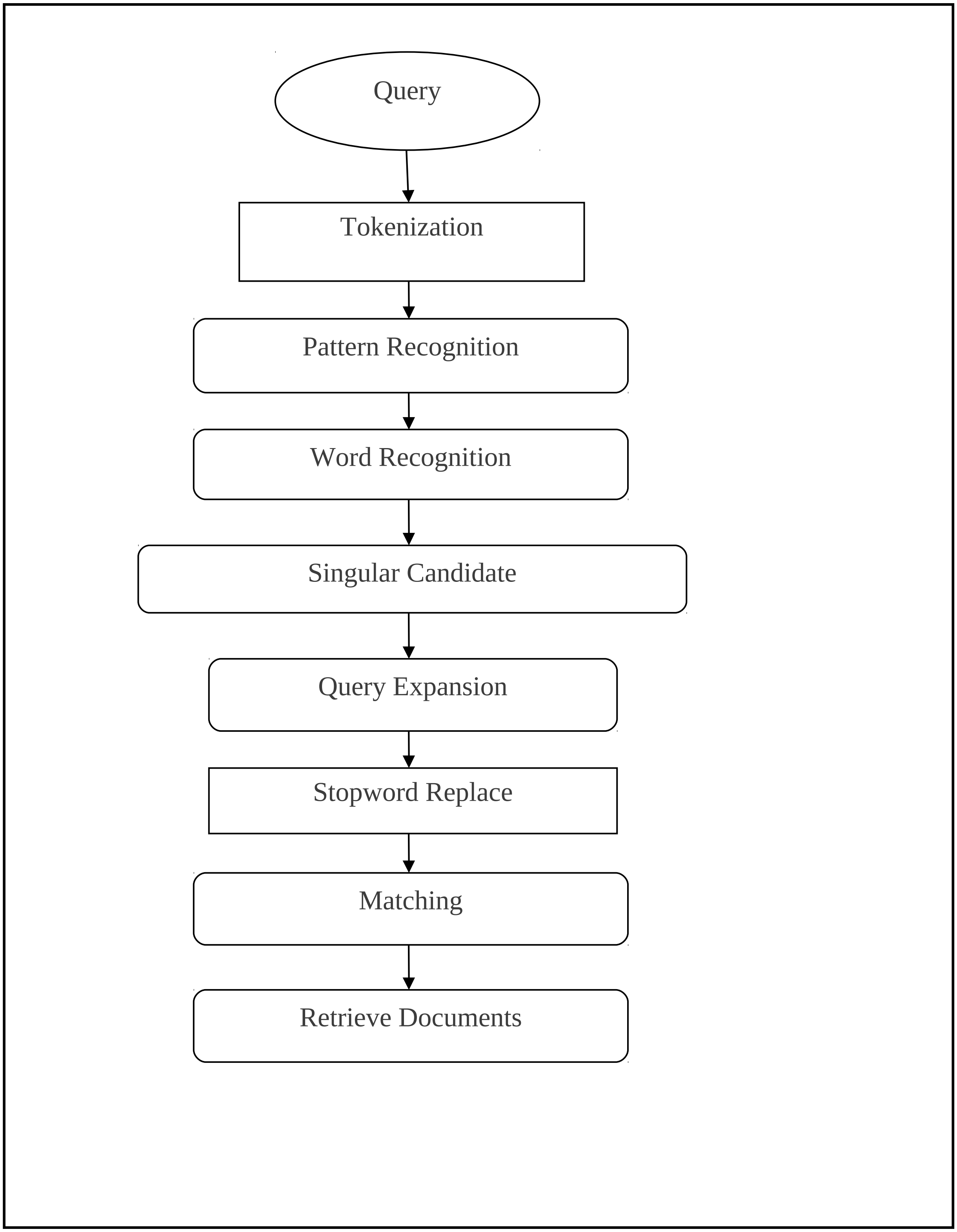


Figure (3.5) general framework of the proposed method

1.18.1 QUERY STATEMENT:

The query the user types to get the documents about information that he/she needs.

1.18.2 BROKEN PLURAL IDENTIFICATION:

1.18.2.1 PATTERN RECOGNITION:

This research works on one pattern of BP patterns, which is (فعاليل), a practical way of recognition was developed to identify this pattern:

- First must check if the length of the word = six.
- Then check if the following rules apply to this pattern.

These two steps will be illustrated in the table (3.1) and table (3.2).

Table (3.4) Pattern recognition (check if the length of the word = 6)

5	4	3	2	1	0
ل	ي	ل	ا	ع	ف

Table (3.5)check if these rules apply to this pattern

Letter No.	Rule
2	Must be (ا)
4	Must be (ي)
3,5	Must be the same

After removing duplicates using RapidMiner the following table contains all the words with faleel (فعاليل) pattern.

1.18.2.2 WORD RECOGNITION:

In this stage we analyzed all the words that has a (فعاليل) pattern to find the words that represents a BP. It was found that 42 words had a (فعاليل) pattern, but only 21 of these words actually represent BP. In the following table, words written in bold are the words that have a (فعاليل) pattern and represent a BP.

Table (3.6) All the words with faleel (فعاليل) pattern.

تـ قاريـر	قوانين	أنابيب	فنانين	بيانين	انابيب
تصاميم	احاسيس	اساريـر	تحاليل	محاليل	صناديد
نواهيه	جواسيس	الانين	ثمانين	خزانين	قواريـر
دياويو	فلاسيس	أساريـر	يضاهيه	مدانين	أحاسيس
مكانين	يعانين	كلاكيك	تعانين	بتاسيس	رحاميم
مداليل	أكاليل	تلا فيف	رهانين	لتاسيس	مجاريـر
كيانين	أخاديد	اكاليل	مشاكيك	تلا بيب	وتاسيس

1.18.2.3 SINGULAR CANDIDATE:

After identifying the BP in the query, at this stage the system generates six different words that serve as potential singular forms, one of them is the right singular form for the BR. For example, let us assume that a user types a query such as

Q: "تـ قارير بيت الاستثمار العالمي"

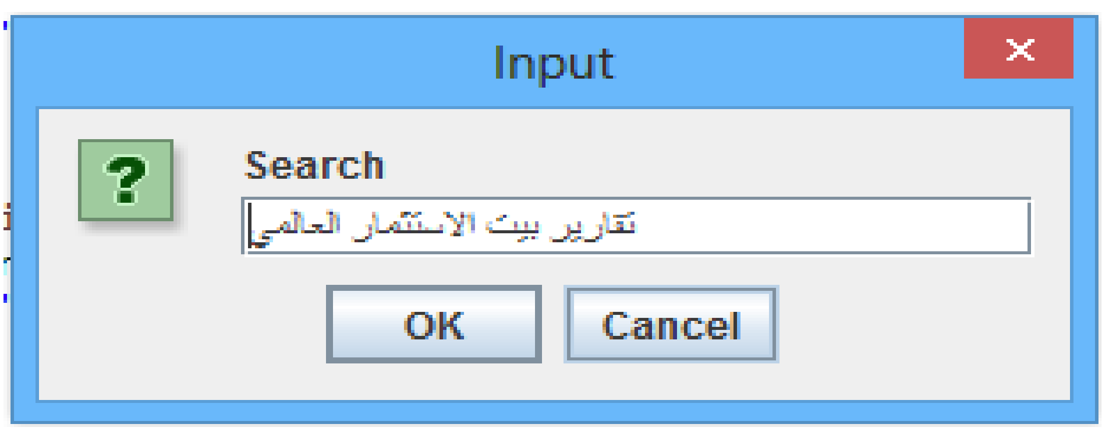


Figure (3.6) User query

The system then identifies the Broken Plural in this query and then generates six potential singular forms:

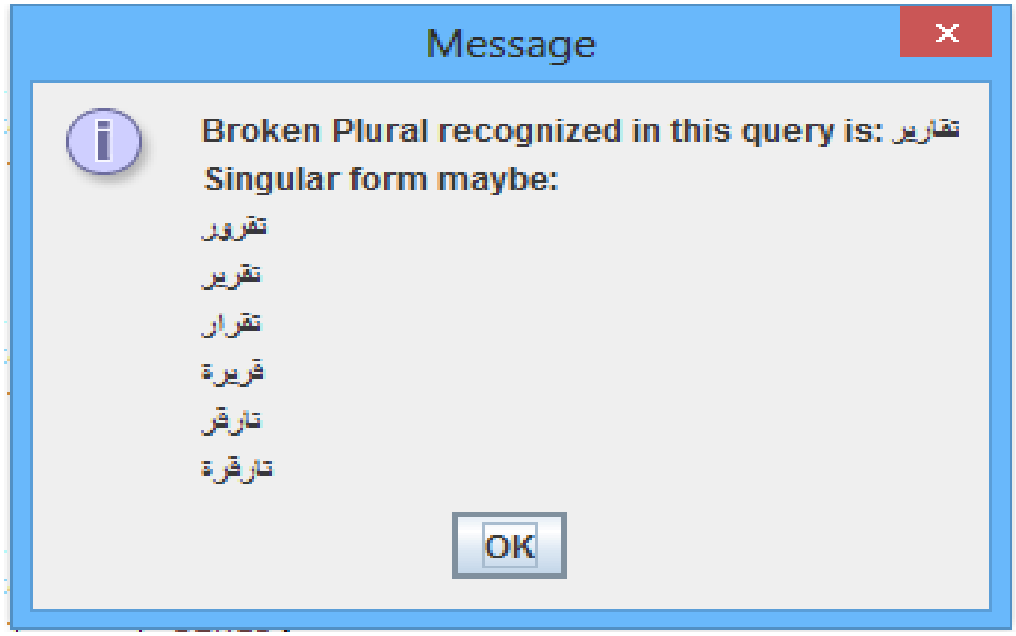


Figure (3.7) Singular candidate

1.18.2.4 SINGULAR FORM RECOGNITION:

After identifying the BP in the query statement, we must then find its singular form. Words of this pattern have six different possible singular forms as illustrated in the next following six figures.

5	4	3	2	1	0
ل	ي	ل	ا	ع	ف
ب	ي	ب	ا	ن	أ

5	و	3	1	0
ب	و	ب	ن	أ
ل	و	ل	ع	ف

Figure (3.8) faalol (فعلول) singular form example

The first singular form is (faalol فعلول), another example for this form is (اخاديد) and its singular (اخدود).

5	4	3	2	1	0
ل	ي	ل	ا	ع	ف
ن	ي	ن	ا	و	ق

5	1	3	2	0
ن	و	ن	ا	ق
ل	ع	ل	ا	ف

Figure (3.9) falaal (فالعل) singular form example

The second form is (falaal فالعل). Another example for this form is (جواسيس) and the singular form is (جاسوس).

5	4	3	2	1	0

ل	ي	ل	ا	ع	ف
ر	ي	ر	ا	و	ق

ة	5	1	3	2	0
ة	ر	و	ر	ا	ق
ة	ل	ع	ل	ا	ف

Figure (3.10) falela (فالعلة) singular form example

The third form is (falela فالعلة). There is no other example for this form.

5	4	3	2	1	0
ل	ي	ل	ا	ع	ف
ر	ي	ر	ا	ق	ت

5	4	3	1	0
ر	ي	ر	ق	ت
ل	ي	ل	ع	ف

Figure (3.11) feleel (فعليل) singular form example

The forth form is (feleel فعليل). Examples for this form are ((اكاليل) ,(تصميم) (تصاميم (تلبيب) (تلابيب) ,(صنديد) (صناديد) ,(اكليل)).

5	4	3	2	1	0
ل	ي	ل	ا	ع	ف

س	ي	س	ا	ح	ا

5	2	3	1	0
س	ا	س	ح	ا
ل	ا	ل	ع	ف

Figure(3.12) felal (فعلال) singular form example

The fifth form is (felal فعلال). This form has no other examples.

5	4	3	2	1	0
ل	ي	ل	ا	ع	ف
ر	ي	ر	ا	س	ا

5	3	1
ر	ر	س
ل	ل	ع

Figure (3.13) elal (علل) singular form example

The sixth form is (elal علل). This form also doesn’t have other examples.

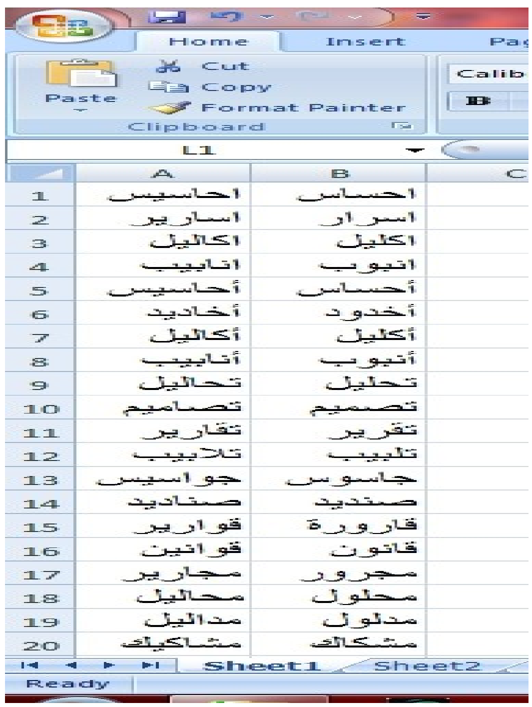
All these six singular forms that we mentioned and used are discretionary forms, and does not exist in the Arabic language.

For example if the user types the Query: "ت قارير بيت الاستثمار العالمي" the system must then identify the right singular form for the BP, Right Singular Form for the previous query is: ت قرير

1.18.2.5 *OFFLINE WORK:*

The BR and their right singular forms were stored in an excel file. The file consists of two rows; the first row contains all words that are BP, the second row contains the right singular form of the BP (first row).

After the system identify the BP in the user's query, it then compares each BP word in the excel file with the word in the user's query until it finds a match, after finding a match, the system takes the singular form from the second row and expands the query with it.



	A	B
1	احاسيس	احساس
2	اسارير	اسرار
3	اكاليل	اكليل
4	انابيب	انبوب
5	أحاسيس	أحساس
6	أخاديد	أخدود
7	أكاليل	أكليل
8	أنابيب	أنبوب
9	تحاليل	تحليل
10	تصاميم	تصميم
11	تقارير	تقرير
12	تلابيب	تلبيب
13	جواسيس	جاسوس
14	صناديد	صنديد
15	قوارير	قارورة
16	قوانين	قانون
17	مجارير	مجرور
18	محاليل	محلول
19	مداليل	مدلول
20	مشاكيك	مشكاك

Figure (3.14) Offline work

1.18.2.6 *QUERY EXPANSION:*

Is the process of adding search terms to user's search query to improve retrieval process and get the documents the user needs.

This is done to increase documents retrieval by adding the right singular form of the BR to the existing query. For example:

If the user types the following query "ت قارير بيت الاستثمار العالمي", the system adds the right singular form "ت قرير" to the query making it "ت قارير بيت الاستثمار العالمي ت قرير". If the system doesn't make the expansion then the system will not retrieve the documents that contain "ت قرير", and that will affect the result.

Query before Expansion:

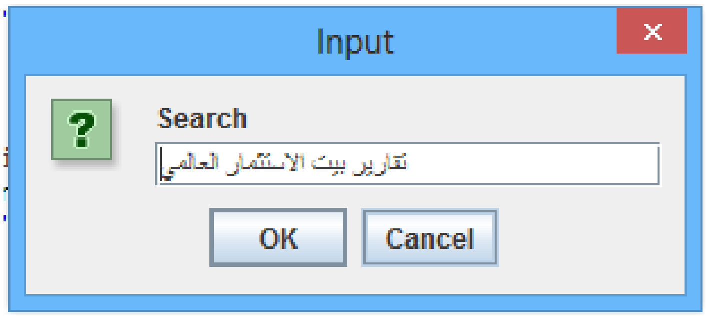


Figure (3.15) Query before expansion

Query after Expansion:

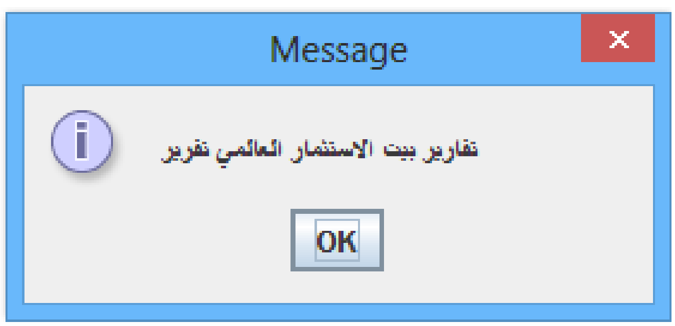


Figure (3.16) Query after Expansion

1.18.2.7 PRE-PROCESSING:

In this stage the system performs a set of processes on the query depending on the language, examples:

- **Tokenization:**

Is the process of dividing the query into a set of words.

ت قرير	العالمي	الاستثمار	بيت	ت قارير

Figure (3.17(token process

- **Stop word Removing:**

The process of removing stop word from the query. Stop word like "من","على". Stop word differ from a language to another.

- **Stemming**:

Removing suffix and prefix from the query.

Query Before Stemming:

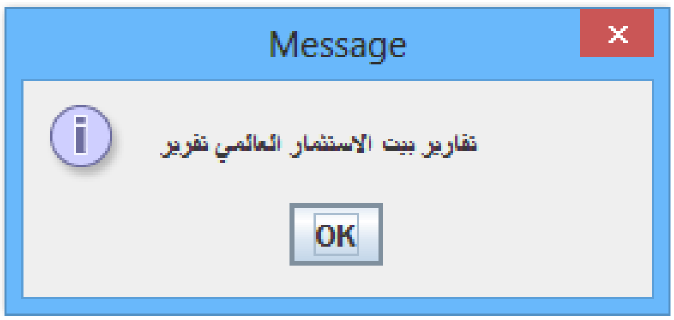
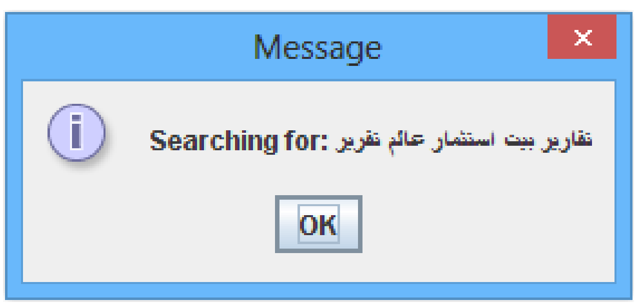


Figure (3.18) Query Before Stemming

Query After Stemming:



Figure(3.19) Query After Stemming

1.18.2.8 MATCHING:

This is the phase were the system match between the query and the indexed documents to retrieve the documents that the user needs.

1.19 SUMMARY:

In this chapter we viewed in details the method used to solve the problem this research is trying to solve, how to identify a BR and how to find its right singular form.

In the next chapter we will discuss the result and effectiveness of this method and compare its results to the Base Line system (Lucene) result.

CHAPTER FOUR

RESULTS AND DISCUSSION

CHAPTER 4:

RESULTS AND DISCUSSION

1.20INTRODUCTION:

As mentioned in the second chapter there were previous studies that identified BR in IRS. This search propose a new method to identify BP in IR systems, this method is innovated from a problem in a study [4].

We applied this methodology on a sample collection of documents, the sample contains 60 documents, some of these documents contain the BP form of the words, some of them contain the singular form of the word and some contain both BR and singular form of the word. Queries are formulated and relevant documents are determined and so results are calculated using Recall, Precision and F-measure.

The results were calculated before applying the methodology (Based Line system, also known as Lucene) and calculated again after applying the proposed methodology. Documents in sample collection were selected from watan-2004 Arabic corpus.

The next chapter discusses the result calculated before and after applying this methodology on a sample collection data selected from watan-2004 corpus, queries that were used contained BR form. Results were compared between Based Line System and the proposed methodology.

1.21EVALUATION:

The performance of IR methods can be evaluated in different ways, the most common ones are Recall, Precision and F-measure.

- **Recall**:

Recall is defined as is the fraction of relevant documents that are retrieved.

Equation (3.4) Evaluation method recall

$$Recall = \frac{number\ of\ relevant\ documents\ retrieved}{number\ of\ relevant\ documents\ in\ the\ collection}$$

- **Precision**:

Precision is defined as is the fraction of retrieved documents that are relevant.

Equation (3.5) Evaluation method precision

$$Precision\ = \frac{number\ of\ relevant\ documents\ retrieved}{number\ of\ retrieved\ documents}$$

- **F-measure:**

F-measure is used to balance between “Recall” and “Precision” on the system performance.

Equation (3.6) Evaluation method F – measure

$$F = \frac{2\ \times Precision\ \times Recall}{Precision + Recall}$$

1.22 RESULTS:

The next chapter review the results calculated from Based Line System and the proposed method, then compares between them.

1.22.1PATTERN RESULTS:

1.22.1.1 COMPARISON BETWEEN BASE LINE METHOD AND PROPOSED METHOD:

A sample collection was structured, it contains 30 documents, some of them contain the word (تقارير) which is the BR based on (فعاليل) pattern, some other documents contain the word (تقرير) which is the right singular form of (تقارير). Some queries were formulated to test the proposed methodology.

Query (1): “تقارير اخبارية”.

Table (4.7) Query (1) Baseline result

No. of Documents	Relevant in the Collection	Retrieved	Retrieved Relevant	Recall	Precision	F-measure
All	18	20	10	0.555	0.5	0.523
Top 15		15	10	0.555	0.666	0.605

As shown in table above, 20 documents are retrieved. The relevant documents in the sample collection equal (18), the Base Line system retrieved (20) documents, only (10) of them are relevant to the query. This lowered the precision, recall and F-measured as a result. Overall recall value equal (0.555), precision value equal (0.5), and so F-measure value became (0.523). Table below shows a comparison between the Base Line method and the proposed method for Query (1).

Table (4.8) Query (1) Proposed method result

No. of Documents	Relevant in the	Retrieved	Retrieved Relevant	Recall	Precision	F-measure

	Collection					
All	18	29	18	1	0.62	0.765
Top 15		15	11	0.611	0.733	0.666

As shown in the table above, retrieved documents after applying the proposed methodology are (29). The relevant documents in the collection are (18) and the relevant documents retrieved after the proposed method is applied are (18). As a result, the measures increased making Recall value (1), and that means that all the relevant documents in the collection were retrieved. It is also noted that the precision value increased to (0.62) from (0.523(, F-measure value enhanced a little as it equals (0.666).

Sample Collection for another query was structured, it also contains 30 documents, some of them contain the word (قوانين) which is the BR based on (فعاليل) pattern; other documents contain the word (قانون) which is the right singular form of (قوانين). Some queries were formulated to test the proposed methodology.

Query (2): قوانين العمل

Such query was chosen because some documents are relevant to this query. The following table illustrates the result for Base Line method for Query (2):

Table(4.9) Query (2) Baseline result

No. of Documents	Relevant in the Collection	Retrieved	Retrieved Relevant	Recall	Precision	F-measure
All	18	20	13	0.72	0.65	0.683

Top 15		15	10	0.55	0.667	0.602

As shown in the table above it is noticed that only (20) documents were retrieved from the collection. Total relevant documents in the collection are (18), Base Line system retrieved only (13) relevant documents, making Recall value equals (0.72) and Precision value equals to (0.65), and so the F-measure value equals (0.683). Table below shows a comparison between the Base Line method and the proposed method for Query 2.

Proposed Method:

Table (4.10) Query (2) Proposed method result

No. of Documents	Relevant in the Collection	Retrieved	Retrieved Relevant	Recall	Precision	F-measure
All	18	25	18	1	0.72	0.837
Top 15		15	12	0.67	0.8	0.729

From table above we can notice that retrieved documents are (25), relevant documents in the collection are (18), all (18) documents were retrieved after applying the proposed method. It is also noticeable that the measures have increased, Recall value equals (1) from (0.72), Precision has also increased to (0.72) from (0.65). That led F-measure value to increase as well from (0.683) to (0.837).

1.22.1.2 COMPARISON BETWEEN PREVIOUS STUDY METHOD AND PROPOSED METHOD:

In this section we will compare the results from a previous study, which we structured the idea of this research from to the results from the proposed method.

Researchers from a previous study structured a sample collection contains 30 documents, “15 documents containing the word (licenses,"تراخيص") which represent BP based on (TFaeel,” تفاعيل ”) pattern and belong to SMJ patterns, and 15 documents contain the word (license,"ترخيص"), which represent the singular of word “تراخيص”.” Results are shown in the table below:

Query 3: تراخيص عـ قود شبكات السيار

Table (4.11) Query (3) Previous study method result

No. of Documents	Relevant in the Collection	Retrieved	Retrieved Relevant	Recall	Precision	F-measure
All	8	32	5	0.625	0.1562	0.2499
Top 15		15	3	0.375	0.2	0.2608

The table above illustrates results obtained from a previous study, (32) documents were retrieved, total relevant documents in the collection are (8), (5) relevant documents were retrieved. Recall value equals (0.625), Precision equals (0.1562) and F-measure value equals (0.2499).

Table below illustrates results from the proposed method using query 1:

Query 1: تـ قارير اخبارية.

Table (4.12) Query (1) Proposed method result

No. of Documents	Relevant in the Collection	Retrieved	Retrieved Relevant	Recall	Precision	F-measure
All	18	29	18	1	0.62	0.765
Top 15		15	11	0.611	0.733	0.666

From table above, it is noticed that the retrieved documents are (29). The relevant documents in the collection are (18) and the relevant documents retrieved after the proposed method is applied are (18). Recall value equals (1), and that means that all the relevant documents in the collection were retrieved. Precision value equals (0.62), F-measure value equals (0.666).

A sample collection contains 30 documents was structured, some of the documents contain the word (قوانين) which is the BR on the pattern (فعاليل), other documents contain the word (قانون) which is the right singular form of (قوانين). Query 2 was used again.

Table(4.13) Query (2) Previous study method result

No. of Documents	Relevant in the Collection	Retrieved	Retrieved Relevant	Recall	Precision	F-measure
All	18	20	13	0.72	0.65	0.683
Top 15		15	10	0.55	0.667	0.602

The table above shows that only (20) documents were retrieved from the collection. Total relevant documents in the collection are (18), this method retrieved only (13) relevant documents, making Recall value equal (0.72) and Precision value equal to (0.65), and so the F-measure value equals (0.683). It is noticed that results from the previous study and results from the Base Line system are exactly the same, why? Because in the previous study the researchers missed to work on patterns with words like “قوانين, عناوين, طواحين” and so, they missed this problem and never noticed it to fix it. Table below shows a comparison between the method proposed in the study [4] and the proposed method for Query 2.

Table (4.14) Query (2) Proposed method result

No. of Documents	Relevant in the Collection	Retrieved	Retrieved Relevant	Recall	Precision	F-measure
All	18	25	18	1	0.72	0.837
Top 15		15	12	0.67	0.8	0.729

From table we can notice that retrieved documents are (25), relevant documents in the collection are (18), all (18) documents were retrieved after applying the proposed method. It is also noticeable that the measures increased, Recall value equals (1) which increased from (0.72), Precision also increased to (0.72) from (0.65). That led F-measure value to increase as well from (0.683) to (0.837).

1.23 SUMMARY:

This chapter reviewed results before (Base Line System) and after applying the proposed method using formulated queries, on sample collections structured from watan-2004 corpus. Results are calculated and evaluated using Recall, Precision and F-measure.

Based on these results researchers can judge for themselves that the proposed methodology could improve retrieval by enhancing recall, precision and F-measure for BR pattern.

CHAPTER 5

CONCLUSION AND FUTURE WORK

CHAPTER FIVE·

CONCLUSION AND FUTURE WORK

1.24CONCLUSION:

Identifying Arabic Broken Plurals represents one of the many challenges for IRS. Our proposed methodology depends on study [4] to identify Arabic Broken Plurals as mentioned in related work in Section 2.6.

Our study focuses only on one pattern (فعاليل) .To identify BP the researchers followed five steps. Step one is ***Pattern Reorganization ;*** the output of this step is the extraction all words from a query which come on the BP patterns, and step two is ***Word Recognition;*** the output of this step is to identify which word in the query represent plural based on some rules. The output of this step is an input to the next step. Step three is ***Singular Candidates;*** the output of this step is getting all possible singular forms of words. The output of this step is an input of the next step. Step four ***Singular Form;*** the output of this step is getting the accurate singular form of words from a query which represent BP. Step five ***Query Expansion;*** the output of this step is to expand the query with the right singular form to improve document retrieval.

After applying these steps, proposed method makes a query expansion in which the singular form is added to the existed query to retrieve all documents that contain singular form of BP word. Researchers calculated the results before and after applying the methodology using sample documents and compared with Baseline system (*Lucene*). Based on the result researchers can consider that, our proposed method has successful to identified Broken Plural words and enhanced retrieval by referring back to this research objectives , found that the study could improve information retrieval for Arabic language especially for query which contain Broken Plurals words.

1.25 FUTURE WORK:

- This study was covered only one pattern of Broken Plurals Syntax Montaha Jemoaa (SMJ), we recommend to study remaining patterns.
- The corpus that had been used in this study does not contain any queries nor a previously measured relevant judgment, we recommend using a well structured corpus with a defined relevant judgment and pre defined queries.

:REFERENCES

Christopher D. Manning, PrabhakarRaghavan, HinrichSchütze (2008). Introduction [1] to Information Retrieval. Cambridge University Press

Kareem Darwish , Walid Magdy , (2014). Arabic Information Retrieval, Qatar [2] .Computing Research Institute

Abduelbaset Goweder , Massimo Poesio, Anne De Roeck, Jeff Reynolds (2004). [3] .Identifying Broken Plural in Unvowelised Arabic Text

MohammedAlmoayed TagAlsir, (November 2015). design Broken Plurals [4] Processing Method for enhancing the performance of Arabic Information Retrival Systems. Sudan University of Science and Technology

[5] RamezElmasri, Shamkant B. Navathe.(2010). Fundamentals of Database Systems, sixth edition. The University of Texas at Arlington, Georgia Institute of Technology.

[6] Ebtihal Mustafa Elamin (October 2014) .Term translation disambiguation in CrossLanguageInformation Retrieval . Sudan University of Science and Technology

.Malek Boualem , Ramzi Abbes (2008).Information Retrieval in Arabic Language [7]

[8] Mohammed Mustafa Ali. (2013). Mixed-Language Arabic- English Information Retrieval. University of Cape Town. Cape town.

[9] Citeseer- AM Goweder, IA Almerhag, AA Ennakoa .(2008) , Arabic Broken Plural Recognition using aMachine Translation Technique.

[10] إيميل بديع يعـ قوب (2004). المعجم المفصل في الجموع .محمد علي بيضون. دار الكتب العلمية(بيروت- لبنان).الطبعة الاولى.

[11] نجاة عبد الرحمن اليازجي (2007) . صيغ الجموع في اللغة العربية وفي اللغة الانجليزية). (دراسة تـ قابلية المجلد الثامن .جامعة الملك فيصل (العلوم الانسانية والادارية).

https://sourceforge.net/projects/arabiccorpus/ ,retrieved ,(Monday , [[12] 10/10/2016), 3:09 PM

https://rapidminer.com/, retrieved Friday, 12/08/2016, 6:02 PM [13]

[14] http://www.lucenetutorial.com/basic-concepts/ Sunday 2/10/2016, 12:09PM

APPENDICES

APPENDIX A

List of Arabic Stop words

و قالت	لدى	بين	كما	ما	تم	في	حين	ب
وكانت	نحو	به	لها	مع	ضد	فى	الى	ا
فيه	هذه	تم	منذ	هذا	بعد	كل	انه	أ
لكن	وان	اف	و قد	واحد	بعض	لم	اول	،
وفي	واكد	ان	و لا	واضاف	حتى	لن	انها	عن
ولم	كانت	او	ل قاء	واضافت	اذا	له	ف	عند
ومن	واوضح	اي	م قابل	فان	احد	من	و	عندما
وهو	يوم	بها	هناك	قبل	بان	هو	و 6	على
وهي	فيها	منها	و قال	قال	اجل	هي	قد	عليها
	يكون	يمكن	و كان	كان	غير	قوة	لا	عليه

APPENDIX B

Example of watan-2004 corpus document

الرياض رويترز: - قال وكيل وزارة العمل السعودية لشؤون العمل احمد الزامل انه تم اغلاق ما يزيد عن 1000 شركة تستقدم العمال للمملكة منذ بداية العام الحالي لمخالفتها قوانين العمل . و قال الزامل في تصريح نقلته صحيفة الجزيرة السعودية الصادرة امس ان مفتشي الوزارة يعملون بشكل يومي على مراقبة أوضاع الشركات بشكل عام. وأوضح وكيل الوزارة أن احصائيات وزارة العمل التي يتم تحديثها اسبوعيا تشير الى أن عدد الشباب السعوديين الذين تم توظيفهم في القطاع الخاص ارتفع حاليا الى نحو 550 ألف شاب من نحو 200 الف قبل خمس سنوات. وأشار المسؤول الى أن أي سعودي يبحث عن عمل تلتزم الوزارة بان توفره له شريطة عدم اشتراط الراتب او طبيعة العمل من قبل طالب الوظيفة. واوضح أن الوزارة أصدرت كتيبا مطبوعا يشمل صور 100 شاب سعودي يعملون في مهن مختلفة منهم من يعمل في مجال النظافة ومزارع الابقار والسباكة وغيرها وهذا دليل على أن الشاب السعودي يقبل على العمل في أي وظيفة. واشار اخر احصاء سكاني في المملكة نشر في شهر نوفمبر الماضي الى أن عدد سكان السعودية بلغ نحو 22.67 مليون نسمة بينهم 6.14 مليون أجنبي. الا أن وزير العمل السعودي غازي القصيبي كان قد ذكر في وقت سابق من هذا العام أنه يوجد 8.8 مليون أجنبي في المملكة يعمل معظمهم في مختلف القطاعات في المملكة وأن نسبة اعداد الباحثين عن العمل بين السعوديين تبلغ نحو 9.6 بالمائة من اجمالي قوى العمل السعودية لمن تبلغ أعمارهم 15 سنة فأكثر.

APPENDIX C

Code we added to *Lucene* search code to apply the methodology

```
String out[ ];
String q = line;
int row = 0;
int col = 0;
String word;
Cell cell2;
String sing;
try{
        out = line.split(" ");
        Workbook workbook =
        Workbook.getWorkbook(new File("D:/project codes/DictionaryOld.xls"));
        JOptionPane.showMessageDialog(null, "Got excel");
        Sheet sheet = workbook.getSheet(0);
        row = sheet.getRows();
        col = sheet.getColumns();
        for (int i = 0; i< out.length; i++){
             for (int j = 0; j < col; j++){
                  for (int g = 0; g < row; g++){
                          Cell cell = sheet.getCell(j, g);
                          word = cell.getContents();
                          if((out[i].contains(word)) == true){
                            cell2 = sheet.getCell(1, g);
                            sing = cell2.getContents();
                           JOptionPane.showMessageDialog(null, sing);
```

```
                    q = q + " " + sing;
                }
            }//row for
        }//col for
    }//main for
} catch (Exception e){
    e.printStackTrace();
 }
```