

# Template Based Summarization for Arabic Documents

نماذج مرتكزة على تلخيص الوثائق العربية

By

Ala'a Abdallah Al-Zoubi

Supervisor

Prof. Ghassan Kanaan

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

College of Computer Sciences and Informatics

Amman Arab University

May/2017

## Authorization





جامعة عمان العربية  
AMMAN ARAB UNIVERSITY

Form (9)

College of Scientific Research and Graduate Studies

### Authorization

We, the undersigned, pledge to grant Amman Arab University for discretion in the publication of the academic content of the dissertation, so that the intellectual property rights of Master's thesis be back to the university in accordance with the laws, regulations and instructions relating to intellectual property and patent.

Advisor Name	Student Name
Prof. Ghassan kanaan	Ala'a Abdallah Al-Zoubi
Signature: 	Signature: 
Date: Aug. 5. 2017.	Date: Aug. 5. 2017




شارع الأردن - موبيل - عمان 11953 - هاتف: 962 7 8054 0040 - فاكس: 962 7 8054 0040 - عمان 2234 - عمان 11953 - الأردن  
Jordan Street - Mubia - Telephone +962 7 8054 0040 - P.O.Box 2234 Amman 11953 - Jordan  
Email: aaug@aaau.edu.jo / Web: www.aaau.edu.jo

## Committee Members Decision

### Committee Members' Decision

The thesis entitled: ""TEMPLATE BASED SUMMARIZATION FOR ARABIC DOCUMENTS"" was submitted by the student, Ala'a Abdallah Al-Zoubi was examined and approved on 7/6/2017.

### Committee Members

Name		Signature
Prof. Ghassan kanaan	Chair/Advisor	
Dr. Akram Almashaikhi	Member	
Dr.Tarek kanaan	External/ Member	

## DEDICATION

I dedicate this dissertation to my family, especially my father and my mother who taught me the value of persistence and challenging work and for their endless care.

To my beloved wife for her love and patience;

To Karam, Jawad, Jood and Anood my sweetheart children;

To my brothers and sisters for their support.

## Acknowledgements

I would like to acknowledge my advisor dr. Ghassan kanaan for all his support and advice and for his encourage over my years at Amman Arab university, and I would express my thanks and gratitude to those who helped me during my study.

## Table of contents

Authourization .....	II
Committee Members Decision .....	III
DEDICATION.....	IV
Acknowledgements .....	V
Table of contents .....	VI
contents .....	VII
List of Tables.....	IX
List of Figures .....	X
Abstract.....	XI
الملخص .....	XII
Chapter One Introduction.....	1
Chapter Two Literature Review.....	23
Chapter Three Methodology and approach.....	35
Chapter Four Analysis and Discussion .....	46
Chapter Five Comparative Analysis and Methodology .....	49
References:.....	57

## contents

Subject
Chapter one: Introduction
1.1 Introduction
1.2 Significance of The Study
1.3 Research Questions
1.4 Research Problem
1.5 Research Model
1.6 Procedural Definitions
1.7. Arabic Encoding
1.8. Arabic Morphology
1.9 Text summarization
1.9.1 General architecture
1.9.2 Text Collection Pre-processing types
1.9.2.1 Document Indexing
1.9.2.2 Tokenization
1.9.2.3 Stemming and Stop-word Removal
1.10 Summarization Types
1.10.1 Single-Document Summarization
1.10.2 Multi document summarization

1.11 Creating Corpus Resources
1.12 General Summary
Chapter two: Literature Review
2.1 Literature Review
2.2. Arabic language
Chapter three: Methodology and approach
3.1 Dataset Selection
Research method and Approach
3.3 Text categorization with WEKA
3.4 Outsider devices
3.5 Transformation
Stopwords
UTF-8
Chapter four: Analysis and Discussion
4.1 Discussion
Chapter five Comparative analysis and methodology
5.1 Comparing Methodologies
5.2 Conclusion
References



## List of Tables

Number	Table
1	Derivation Forms of the Word "Write" in Arabic
2	Summarization Approaches
3	Summarization Techniques
4	Method of Machine learning
5	Semantic and Syntactic
6	Diacritics for the Letter "BAA
7	Various characteristics of dataset
8	Categories of Text File
9	Number of Articles for each Score
10	The techniques used to extract the attribute values
11	Evaluation of Kanan research
12	Evaluation of this research

## List of Figures

Number	Figure
1	General Architecture for summarization
2	Selection Module for Summarization
3	Process of Index Creation
4	Tokenization algorithm
5	Stemming and Stop Word Removal Algorithm
6	Architecture of Single-document Summarization
7	General Algorithm of Single- document Summarization
8	Architecture of Multi-Document Summarization
9	General Algorithm of Multi document Summarization
10	WEKA Explorer
11	Result file from WEKA
12	Gate Tool
13	Output of Gate Tool
14	Template summarization form 1
15	Template summarization form 2
16	Political Template summarization / form 3
17	Science template summarization/ form 4
18	Art and Culture template summarization/ form 5
19	Sport template summarization/ form 6
20	The result of the summary
21	Categories Used to Show the Summary Evaluation Results
22	The Template used in Kanan research
23	The attributes in Kanan research
24	Evaluation of Kanan research
25	Evaluation of this research

## Template Based Summarization for Arabic Documents

**Prepared by**

**Ala'a Abdallah Al-Zoubi**

**Supervised by**

**Prof. Ghassan Kanaan**

### **Abstract**

Since the number of electronic documents increase quickly in the world, the need for faster techniques to assess of these documents arises. A summary is a brief representation of implicit text. To form an ideal summary, a full understanding of the document is primary. However, obtaining a full understanding is either impossible or difficult for computers. Therefore, selecting significant sentences from the original text and presenting these sentences as a summary setting the most common techniques in automated text summarization. This research focus on the NEWS article and what you want to do with the entire news article to enable the user to decide if they want to read the whole article or not, good summaries are rare, and too costly to produce manually, so the researcher writes a program to read a document collection and then produce a summary based on the designed template for this purpose

نماذج مرتكزة على تلخيص الوثائق العربية

إعداد

علاء عبدالله الزعبي

إشراف

الأستاذ الدكتور غسان كنعان

### الملخص

إن عدد الوثائق الإلكترونية يزداد بسرعة في العالم، ولهذا تنشأ الحاجة إلى تقنيات لتقييم هذه الوثائق واسترجاعها بسرعة. ومن هذه التقنيات التلخيص وهو عبارة عن عرض موجز للنص الأصلي. ولإيجاد ملخص مثالي للوثائق، فإن الفهم الكامل للوثيقة أمر أساسي. ومع ذلك، فإن الحصول على فهم كامل للوثيقة أمر مستحيل أو صعب. لذلك، اختيار الجمل المهمة والأساسية من النص الأصلي وتقديم هذه الجمل كملخص يعدّ من التقنيات الأكثر شيوعاً في تلخيص الوثائق. يركز هذا البحث على المقالات الإخبارية، فبدلاً من قراءة كامل المقال الإخباري يستطيع المستخدمون أن يقرروا ما إذا كانوا يريدون قراءة المقال كاملاً أم لا، وذلك بقراءة الملخص القائم على النموذج المبني في هذه الرسالة. واحدة من أهم القضايا التي واجهتنا هي تلخيص المقالات المكتوبة باللغة العربية، واللغة العربية هي واحدة من أصعب اللغات في العالم وأكثرها تعقيداً، فهي تتكون من 28 حرفاً، تكتب من اليمين إلى اليسار إضافة إلى قواعدها النحوية واللغوية المتشعبة، وبالتالي فإن قليلاً من الأبحاث العلمية سلطت الضوء على استخراج النماذج الملخصة باللغة العربية. وبالتالي، فإن الفكرة الرئيسية من هذه الرسالة هي بناء ملخصات باللغة العربية قائمة على النماذج، وهذه العملية تمر في عدة مراحل وصولاً إلى النموذج النهائي المصنف حسب نوعه. ومن هذه المراحل، اختيار المقالات الإخبارية وتدقيقها وادخالها على برامج استرجاع البيانات لاستخراج الخصائص الرئيسية من المقال الإخباري، مثل اسم الكاتب وتاريخ النشر ومكانه بعد ذلك تؤخذ هذه الخصائص لتوضع في نموذج مبني مسبقاً في الرسالة لبناء تلخيص مفهوم المعنى وصحيح التركيب.

# Chapter One

## Introduction

### 1.1 Introduction:

As the amount of information in the web increases, systems that can automatically summarize documents become increasingly popular. Information retrieval (IR) finds material (usually documents) of an unstructured nature (usually text) that satisfies information need from among large collections (usually stored on computers) (Christopher, 2018).

In general, summaries are used commonly in our daily life. News headlines, book reviews, store shopping guides, and even movie trailers are all examples of summaries (Mani, 2001). In general, a summary can be defined as a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that' (Radev, 2002). Despite the word 'text' in the previous definition, summaries can comprise any other documents.

One of the most important issues do for doing summaries is Arabic media such as audio, hypertext, or even videos. One of the most important issues that we do summaries of them is Arabic. Arabic Languages is one the most difficult and complex in the world, consisting of 28 letters, written from right to left with grammatical and linguistic characteristics classifications have very little in the light of technological overwhelmed by the provisions of the Arabic languages.

Text classification is a system to support browsing services for online Arabic newspapers and the process of structuring a set of documents according to the researches most classifiers and NER (Named Entity Recognition) tools works in English. However, it does not Support Arabic language; these tools are very difficult due to the nature of the Arabic language. These texts were going according to a system based on some machine learning algorithms for data mining tasks many problems during the process of doing the summary,

such problems as Headline, summarization is a difficult task because it requires maximizing text content in short summary length while maintaining grammaticality and we found that there is little research on the automatic summarization of the Arab news articles. The limited number of experts in the field summary.

Then Named Entity Recognition (NER) task is used to extract the existing information in the text and classified into specific categories such as the names of the people (the media), websites, places, organizations, the temporal and spatial conditions.... etc. It is very important for many tasks in natural language processing (NLP), such as data retrieval, where made a lot of efforts in this area to build tools (NER) support the Arabic language, which is the most difficult and being one of the most complex languages in the world's, with the use of additional information such as the use of text, Part-Of-Speech tags and Base Phrase Chunks.

## 1.2 Significance of the Study

Most systems of automatic text summarization are made to process the most known languages such as German, English etc. In other word, there are few and little systems and researches on Arabic language. Studies this area is very restricted. Therefore, there is an important need to develop systems that summarize and process electronic Arabic document

The Importance of this study stems from the importance of the problem that there is no good summarization methodology for Arabic documents, help users reducing the time search and allowing users to get a better view of search results. This will help the reader to easily know the summary about particular document.

### 1.3. Research Questions

- 1- Are our generated template summaries good?
- 2- How applying the IR and NLP methods will help generate summary?
- 3- What are the practical values of the summaries?
- 4- What are the benefits of summaries to user?

### 1.4. Research Problem

There is no good summarization framework for Arabic news articles that can satisfy a diverse user community using fully automated methods. Moreover, there are no good automatically generated summaries for Arabic news articles. For users who lack such summaries,

Determining whether an article is of interest without reading (or at least scanning) it is infeasible.

### 1.5. Research Model

This is a practical study based on summaries the set of documents in a given field so that the researcher can get the desired results in a short time.

### 1.6. Procedural Definitions

1. Automatic summarization (AS) is to create an abridged version of the text by a computer program. The result of this action still contains the most important points of the original text.
2. Natural Language Processing (NLP): A branch of science informatics that deals with natural language information.
3. Information retrieval (IR): science of searching information within documents.
4. Information extraction (IE): Is a technology used automatic summary. In this approach, the statistical inferences used to identify the most important sentences of text.

5. Extractive Summarization: using IE to create a system summary.
6. Semantic Analysis: a stage of NLP, which include extraction of context independent aspects from meaning of sentences.
7. Machine Learning: a scientific specialty interested with the development of algorithms that let computers to evolve behaviors based on experimental data.
8. Syntactic Analysis: is the operation of analyzing a text, made of a sequence of tokens to create its grammatical structure with respect to a given formal grammar.
9. Query-based Summary: a summary that presents the contents of a document that are related to a user's query.

#### 1.7. Arabic Encoding

While working with texts, there are challenges that may face us, such as how to recognize the characters programmatically (with a computer program). Encoding tends to be problematic, but by using Unicode (UTF8 for example), this difficulty will be solved effectively.

#### 1.8. Arabic Morphology

The Arabic language has a derivational and inflectional nature, to have a complex morphology (Benajiba, 2009).

Arabic root word gives verbs and nouns, and usually consists of the root word followed by a pattern to form a lemma [6].



Table 1: Derivation Forms of the Word “write” in Arabic:

Write	Writing	Writer	I write	Manuscript
كتب	كتابة	كاتب	كتبت	مكتوب

This is very common in the language that both the word “write,” and the other forms have the same base root, where the first word is the root word, the second word derived from it, but the meaning is changed by inflection, due to the suffix indicating singularity or plurality.

### 1.9. Text Summarization

We present in this research the general architecture for automatic summarization, which is important to understand this area. Then, we discuss the abstract architecture of an automatic summarization system.

Talk about single and multi-document summarization and the methodology that used in type of summarization.

#### 1.9.1 General Architecture

Figure 1 explains the overall architecture for any automatic text summarization system. The architecture is for both; single and MDS approaches. The user chooses either a single document or a set of united documents. User interface represents an interaction among the user and document supplier services, which could be a simple a search engine. The summarization operation produces a short text of the chosen document. Finally, the operation ends by generating a single summary. The index will be a repository, because the document’s information will be stored and retrieved in it.

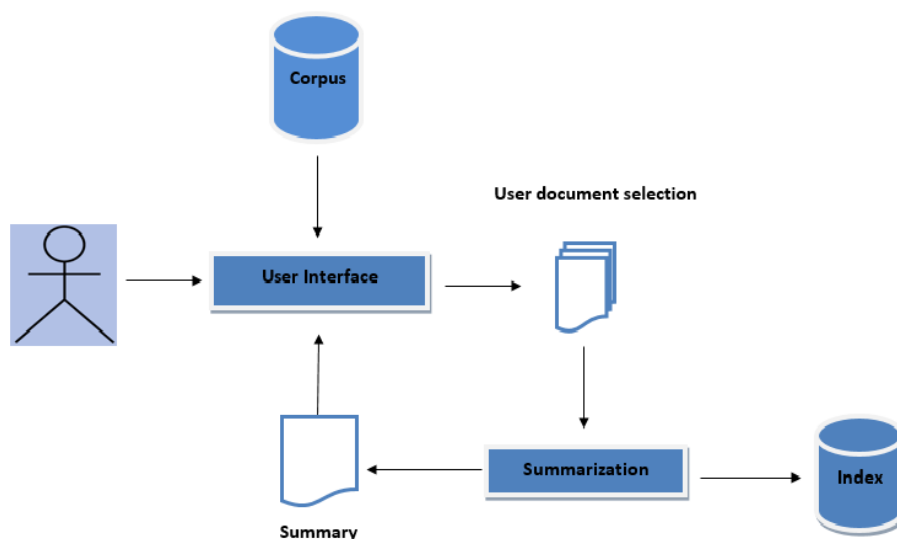


Figure 1: General Architecture for summarization

The structure differs based on the used techniques and purpose. The general diagram can be divided into two models, the first one is a Document Selection Model, which could be any approach concerns to select a relevant document from a document set to satisfy the user's need. The other one which we are most interested in, is the Document Summarization Model that could be in any form, e.g. when we use a search engine to retrieve documents where we decide which document or collection of documents to be summarized. In figure 2, first, all the chosen documents submitted a splitting operation, and then the document is being split into sentences based on a limit delimiter. Next, the sentences leave in the Sentence Matcher in order to be matched against other information, such as the first sentence in the document or a user query. Sentence selection operation is based on the sentences of ranking after the matching process. Then, the best sentences are chosen and sent to fusion operation in order to be ordered and combined to create a summary.

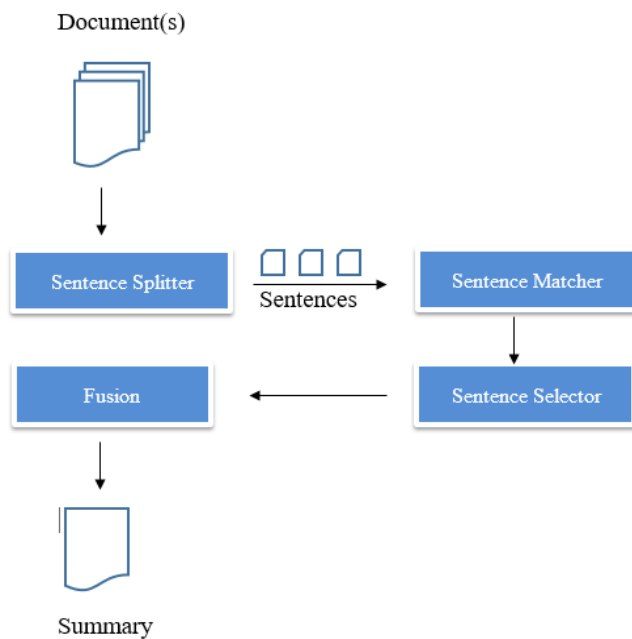


Figure 2: Selection Module for Summarization

The architecture above is for the abstract depicts of more summarizer systems in any languages, approaches, techniques used or tools. We are mainly concerned in the summarization operation, so next; we discuss the public pre-processing tools for operations text collections, which are applied for automatic summarization purpose.

### 1.9.2 Text Collection Pre-processing steps

The summarization operation needs previous several stages to create a summary. These stages include documents and language processing. The abstract stage indexes the documents, tokenization process, removes stop-words and stemming. The set of documents to be summarized is essential to afford the number of pre-processing steps.

#### 1.9.2.1 Document Indexing

Indexing is applied for a long time as an active mean for accessing the information. It can be done either automatically, or manually, by choosing by hand the index terms of a document. Automatic indexing has been greatly applied in the experiment systems such as IR systems and it has proven to make best results than manual indexing, which are a time loss and a fault susceptible process (Salton, 1989). Because the success of choosing the sentences is specified by the exact choosing of index terms, the aim of indexing for IR and automatic summarization is to locate and choose the terms in order to describe the document content as widely as possible.

Figure 3 shows the creating of an index for document summarization. As shown, the operation of indexing begins firstly with choosing automatically the documents of the collecting set. Through the process, the information in the document will be secondly sent to the index, including the document's size, title, location, and genre. Then the document will be split into sentences and the sentences will be split into tokens based on delimiters (such As. White space). Next, these tokens will be indexed and the information at all the token's location, place, weight, and frequency will be recorded. All words weight “the weight which reflects the relative significance of tokens in the sentences and the related important extraction of the token in a document based on the term frequency of this token” will be used to compute the likeness among sentences, such as the term frequency for tokens. In automatic summarization, sometimes, the weight reflects the important of this token in a sentence, so in this case, the sentence that is retrieved in this process is more important rather than the document itself.

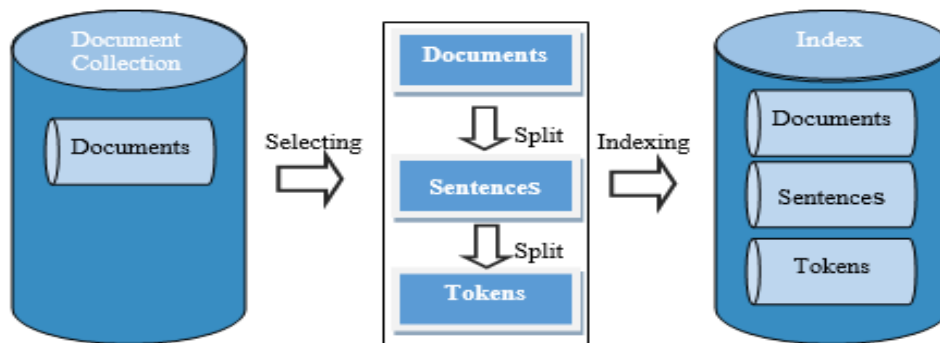


Figure 3: Process of Index Creation

The TF-IDF (Term Frequency- – Inverse Document Frequency) weight is one of the general models which is used to calculate the weight terms, TF-IDF weight uses two factors: the term frequency (TF) and the inverse document frequency (IDF)

$$W = TF \times IDF \quad \dots\dots\dots 1$$

$$(idf \ i) = \log_{10}(n / df_i), \quad \dots\dots\dots 2$$

Where n is the number of document and I is a term.

#### 1.9.2.2 Tokenization

The operation of splitting the document into tokens is called “Tokenization” which all the input documents to pre-processing starting with. The output of the tokenization according to the part whose structure is to recognize numbers, punctuation, and dates.

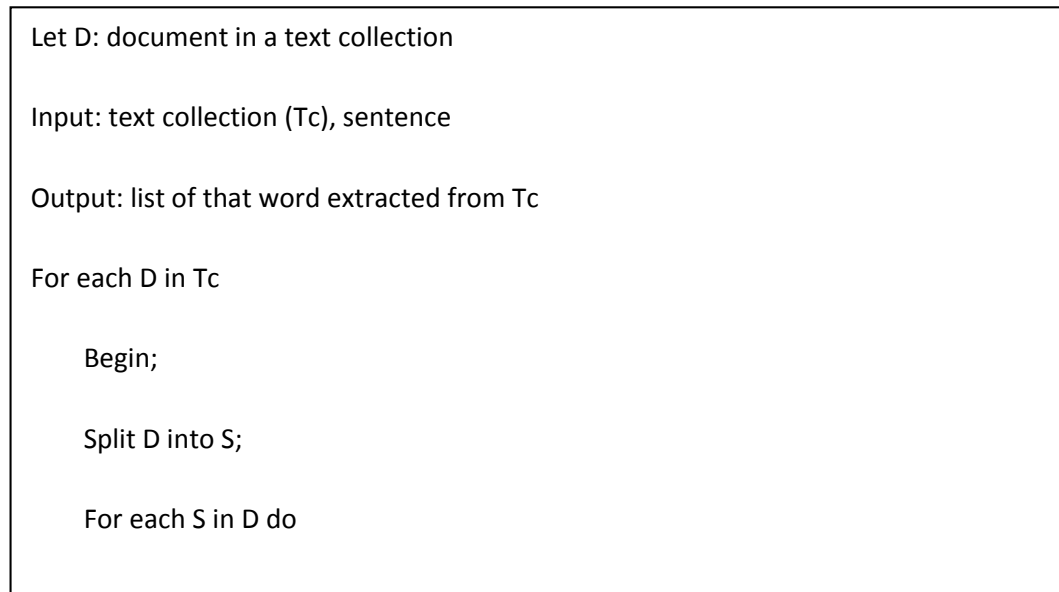


Figure 4: Tokenization Algorithm

### 1.9.2.3 Stemming and Stop-word Removal

Many words used to express a certain meaning, for example: (selects, selected, selecting, and selector), they can be collected with each other and stemmed to be selected, so all have the same conceptual significance, so to minimize the index words. Stemming is used, by stripping affixes from words automatically. This is useful for automatic summarization through locating semantically related words, which help in choosing many sentences] (Croft, 2009). There are two types of Steaming: "light stemmer" and "root extractor." Stemmer removes, suffixes, and prefixes which use pattern matching to extract a word root. In order to minimize the index volume when the index is created, we try to register significant information and ignore lower important information; also, the processing of text will be stopped when detecting any of these words, which are called "stop-words."

The advantages of disregarding stop-words are minimizing 40% of the length of the index (Gancarski, 2006), and allowing calculation of the similarity among sentences in order to be most realistic (Salton, 1986). Stop-word list is based on the language that used variously. Therefore, when working on a specific summarizer domain, stop- word list contains a word, which is considered in other domain, which is not stop-word. By processing the text collection, recording the time of all terms, and then considering all words with a frequency that greater than the certain threshold as stop-words, we can create a stop-word list [11]. A general language independent stemming and stop-word removal process are shown in figure 5.

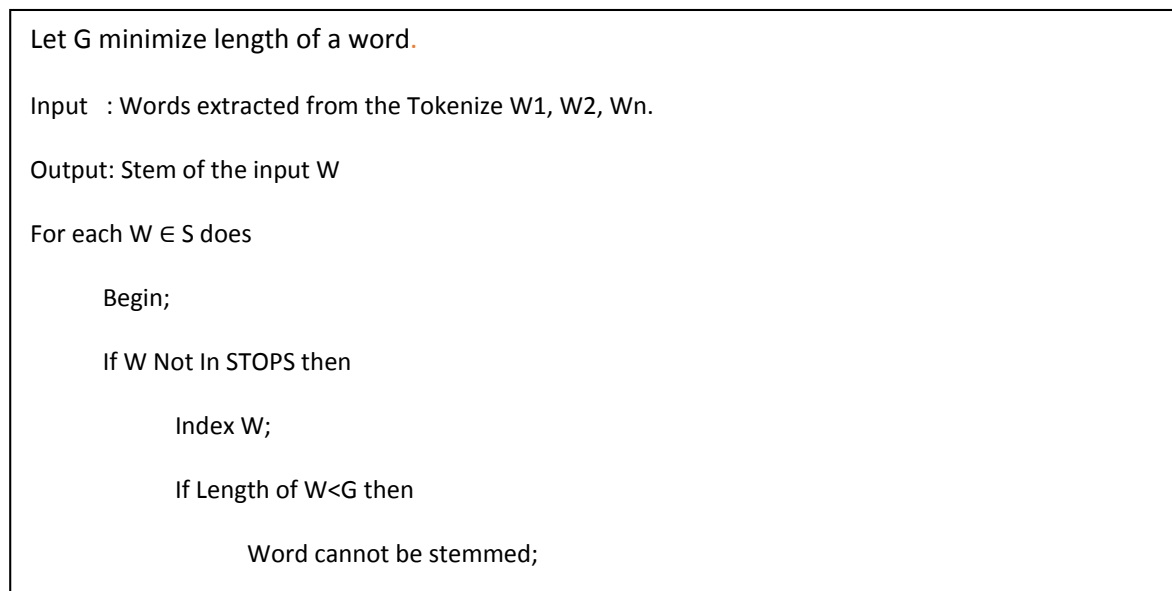


Figure 5: Stemming and Stop Word Removal Algorithm

#### 1.10 Summarization Types

There are different approaches for automatic text summarization, involving single-document and multi-document summarization, the summarization methods mentioned earlier, i.e. abstractive or extractive summarization was applied these approaches. Summarizing a text depends on Types of Summary Usage (TOSU) query based or generic.

Table 2: Summarization Approaches

Summarization approach	
Single document	Multi document
Method	
Extractive	Abstractive
Type	
Generic	Query based

clustering	Statistical	Semantic and syntactic	Machine learning
------------	-------------	------------------------	------------------

The operation of summarizing a document based on the information of user's need is called Query-based (query-focused) summarization. It is working for both of single and Multi-document summarization and it is retrieving sentences that correspond a specific user query. Extraction head plays a significant role in a single document and multi-document summarization. So, some statistical scores are typically needed to evaluate the importance of the extracted units. Each unit gives a score based on features such as word frequencies (Luhn, 1958). Techniques that are more sophisticated are used by other approaches to decide which sentences to extract, involving machine learning to identify important features, and different natural language processing techniques to identify key passages and relationships among words. The most common summarization techniques and methods are clarified in Table 3, which also clarifies that because of the huge overlapping among automatic text summarization approaches.



Table 3: Summarization Techniques

Machine learning	Supervised learning
	Unsupervised learning
Clustering	Partitioned clustering
	Hierarchy clustering
	Agglomerative clustering
	Divisive clustering
Semantic and syntactic	Co-reference chain
	Graph representation
	Latent semantic
	Lexical chain
statistical	Relevance score
	Hidden markov model
	Sentence compression

Distinguishing factors among single and multi-document summarization are the input sources. One of the main differences between single and multi-document summarization is redundancy elimination. An overlapping in information could happen when selecting sentences from a set of related text, which is considered redundant and must be eliminated, so semantic, syntactic and statistical models have been used in order to eliminate this overlapping, and clustering was applied to eliminate redundancy by classifying the extracted sentences into a set of semantically related sentences. In order to detect redundant sentences, many tools and techniques have been investigated and in order to determine the quality of different similarity metrics, with regard to redundancy elimination, the locate key of sentences contains crucial information from related documents and summarization approach focuses on it.

The WorldNet distance (using a semantic lexicon for the English language), Cosine Similarity (a measure of similarity between two vectors by measuring the cosine of the angle between them) and Latent Semantic Indexing/Analysis (LSI/LSA), were examined by the three metrics. Statistical methods, many summarization systems are dependent to extract the sentences that could be relevant. For classifying text-mining summarization and supervising sentence ranking, Machine learning Summarization has been applied and the techniques and methods that are applied in machine learning are clarified in Table 4.

Table 4: method of machine learning

Approach			
Machine learning			
Method			
Supervised		unsupervised	
Technique			
Categorization	classification	generalization	Sentence ranking

Semantic correlations between sentences and semantic analysis are other approaches that have been applied extensively. Semantic analysis approach is found in Text summarization in order to locate relations between sentences. Some of the techniques, which involve textual entailment and graph representation by lexical graphs, are clarified in Table 5. Other techniques involve co-reference, semantic clustering, and lexical chains, anaphoric resolution and lexical semantic.

Table 5: semantic and syntactic technique

Semantic and syntactic		
Graph representation	Natural language processing	Lexical chains
Lexical graphs	Named entity recognition	Textual entailment
Graph matching	Latent semantic	Lexical semantic
Weighted graphs	Part of speech tagger	Co-reference chains
Un weighted graphs	Information extraction	Anaphoric resolution
	Information normalization	Wordnet

More summarization systems apply several forms of language processing, semantic and syntactic to text summaries. The application of natural language processing with syntactic and semantic models in text summarization has contributed growing the type of the generated summaries.

#### 1.10.1 Single-Document Summarization

Single document summarization is one of the early approaches towards automatic summarization. It is the operation of generating a summary for one text document without a standard length and it differs based on the implementation guidelines. For a single document summarization, a number of various techniques have been used. Both language-dependent and language-independent approaches have been decided. Language-independent approach does not depend on any language-specific knowledge resources or any manually build training data, although Language-dependent approach uses language-specific tools (such as a parser, lexical chains, knowledge-base) in order to locate semantic similarities among sentence. The summarizer (which uses a language-independent approach) must portable to modern languages or scopes (Michalcea, 2005).

Single-document summarizers include two types; generic and query-based. Generic single-document summarizers are more useful for long documents, which contain a variety of topics; they match a document's sentences to certain information extracted from the same document, which could be the document's title, or the first sentence in it (Zha, 2002). Query-based single-document summarizers in order to summarize the document around this query, they use a user query by using similarity measures and sentences that are closer to the query are being selected to be in the summary. The general architecture for an SDS is clarified in figure 6.

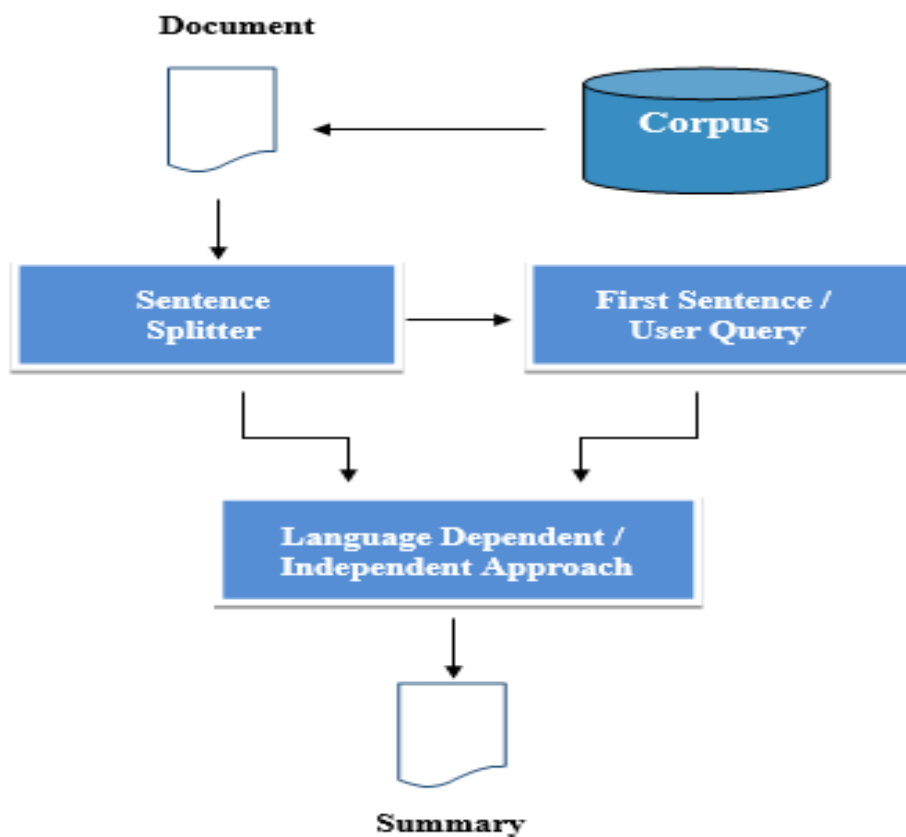


Figure 6: Architecture of Single-document Summarization

The summarization operation begins firstly by choosing one document from the text collection, which varies, and can be a simple concordance system or a sophisticated IR system. The selected document divides the document into sentences. The sentences are matched to the query or to the first sentence in a document by language dependent or language independent tools, depending on the single-document summarizer type (generic or query-based). To create the final summary, the matching operation returns the sentences that are important (the first sentence or most similar to query). The basic breakdown of the operation of single document summarization is clarified in figure 7.

$SM [i, j]$  represents a 2-dimensional array, to save similarity values; Similarity ( $S_j, I$ ) calculates the similarity value among  $S_j$  of the  $j$ th sentence in a document  $D_i$  and other information  $I$  in this document (e.g., first sentence);  $N$  represents the top maximum number of sentences or words allowed in the summary based on the similarity value;

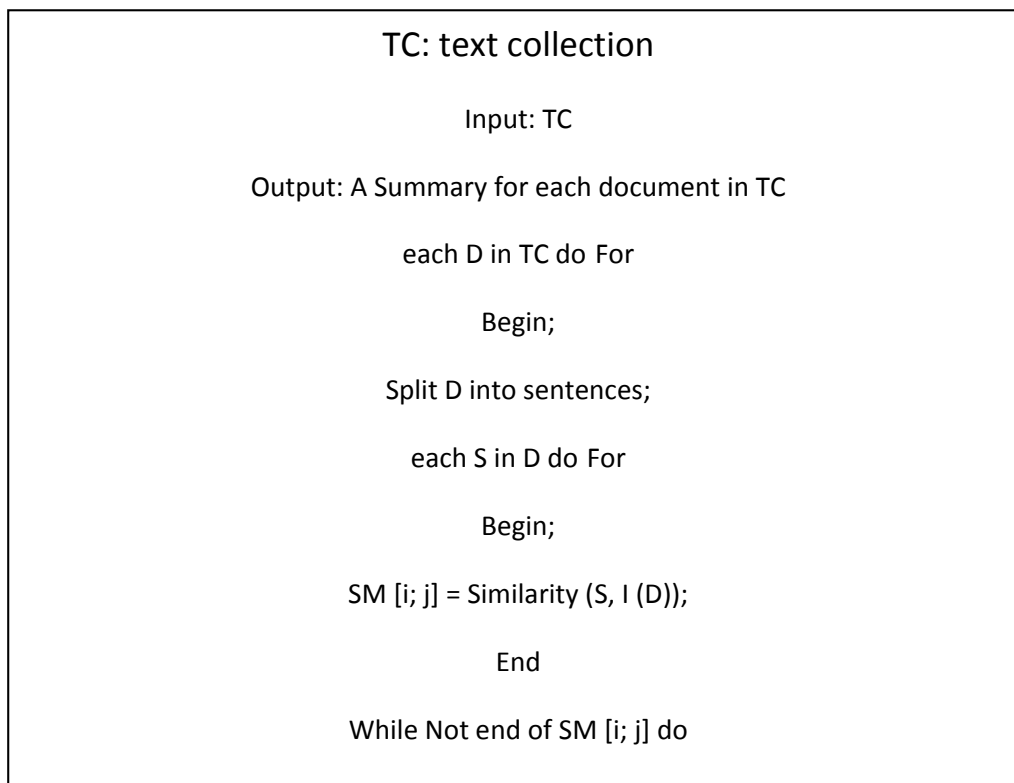


Figure 7: General Algorithm of Single-document Summarization

#### 1.10.2 Multi document summarization

It is the operation of creating a single summary for a collection of related text. Multi-document summarization needs a huge time, also an order to read the completed documents collection and then summarizing up the main events in a way to ensure coherence and maintain relevant relationships. Automatic extractive multi- document summarizer can rapidly extract the task sentences from a collection of documents and then create a single summary. The general Architecture for multi-document summarization process is clarified by figure 8.

In the diagram, after the sentence splitting operations, language dependent and language independent approaches are applied on the documents, involving tools and techniques that work for all the important extracting sentences and eliminating redundant ones. The major phases of creating a summary for a collection of related text are clarified by an algorithm shown in figure 9.

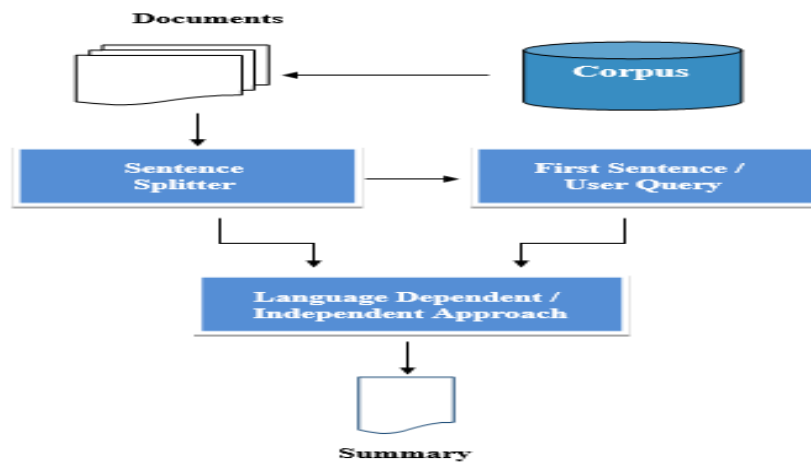


Figure 8: Architecture of Multi-Document Summarization

SM [a; b]: is a two-dimensional array, to store values of similarity;

Similarity (S<sub>j</sub>, i): calculates the similarity value between S<sub>j</sub> of the jth sentence in a document D<sub>i</sub> and other information I in this document (, first Sentence); U represents the maximum number of sentences to be selected from each document in the set of related articles; N represents the maximum number of sentences or words allowed in the summary.

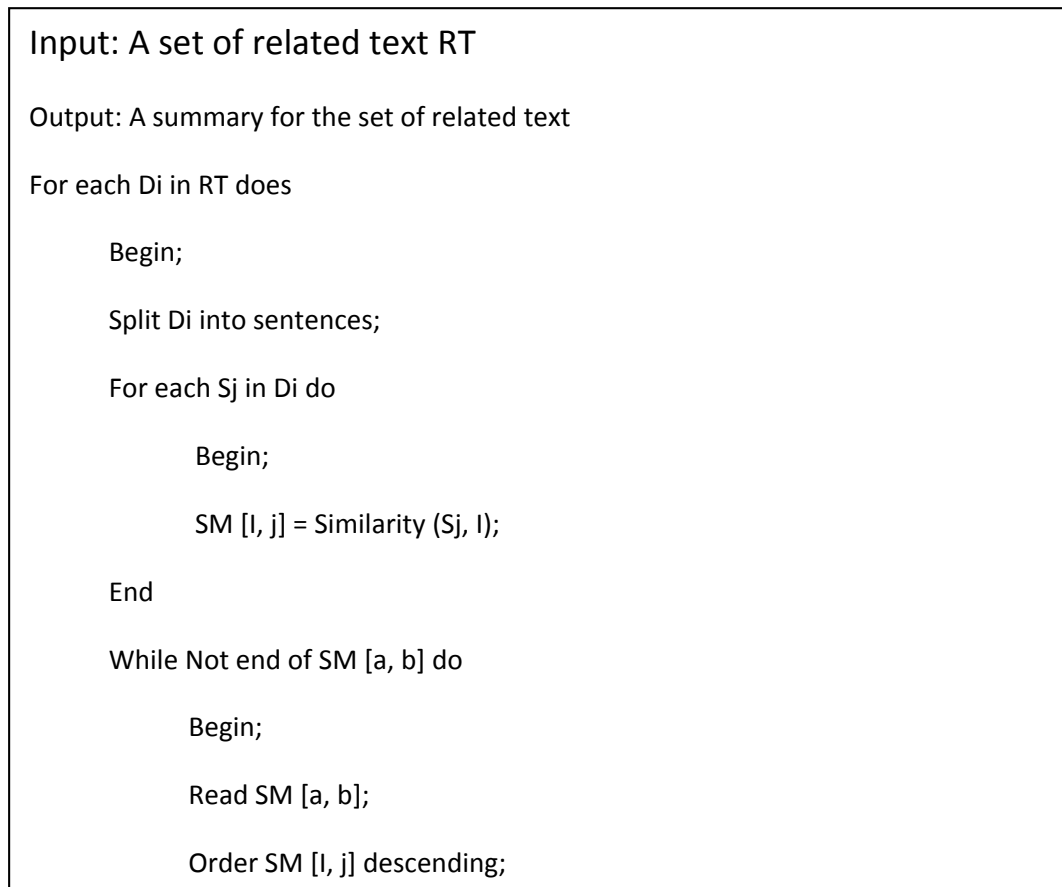


Figure 9: General Algorithm of Multi document Summarization

The MDS (Multi document Summarization) operation can use the same approaches and types of summarization that applied in single- document summarization 1.1.8. Because we deal with more than one related document at the same time, the work on multi-document summarization is much complex. When creating a multi -document summary, a collection of texts from the same subject gives an effect on the redundant sentence. We need to eliminate of those extra sentences to come up with a logical summary, so many techniques and tools to eliminate this sentence mentioned earlier.



When creating a summary in multi- document summarization, it can be a collection of text on the same topic in the redundant sentence. For getting up a logical summary, first we need to eliminate redundancy in the sentence. For this purpose, many techniques and tools like language independent and language dependent can be used. Deep semantic analysis is an example of dependent language. Similarity, measuring and statistical model are examples of independent language. If the similarity value is high; between two sentences when comparing with each other, based on the similarity threshold, one of them will be redundant and only one of the two can be selected, the redundant elimination's techniques and tools determine the decision if any sentence should stay or leave. Working on multi -document summarization opening many questions about how can we solve information redundancy without eliminating crucial sentence, and what should the extracted sentence follow. Extraction performs and sentence position is techniques used to order the sentence (Barzilay, 2001). Chronological and Closeness are the orders of the sentence according to similarity measure

#### 1.11 Creating Corpus Resources

Researchers require resources for running experiences in the field of automatic text summarization, which can be human generated such as human experts, crowd sourcing, and machine- generated using machine translation tool. Such these resources are available and published on the internet and books in English language. This allows researchers to compare their works with other works in order to evaluate the quality of summaries. Corpus resources are important for researchers to work. Any dataset can be closed in multi -document summarization, to ideally perform subgroup of related text, like a multi-source of text such as the wiki news websites. The most important part is finding the language native speaker, which the data set required.

Translation can be performed on the dataset and the participant depending on the document nature can create manual summary. In the manual summary, it is suitable for the number of summaries for any document. The resources of multi-document summarization can be created in the same approach in single-document summarization, but it is difficult in the matter of cost and time.

### 1.12 General Summary

In any automatic text summarization system, Document selection and Document summarization are the two main steps. Different tool can be used to achieve document selection like a document retrieval system. The document summarization module differs from one to another summary based on the approach used. The summarization module composed four main operations of a sentence: sentences matcher, sentences splitter, sentences selector, and sentences fusion. .this operations are officials in the generation of both single and multi-document summarization. Before summarization process, the data collection preprocessing which performs several steps like indexing, stemming, tokenization and stop word removal, are used to find the relationship between sentences and words, to reduce redundancy rank the sentences in a status of multi-document summarization.

## Chapter Two

### Literature Review

#### 2.1 Literature Review:

According to Kim J and Kim M (2004) in their Journal of intelligent information systems, text categorization defines the imperative tasks of assigning one or more thematic categories to a particular document. Essentially, it implies the application of traditional techniques to document data set. This is in contravention the latest development in the history of text categorization and classification. As opposed to the traditional technique where it was done on manual basis, the latest development over the past few decades has revolutionized the automation of the process in its entirety. It is through such automated process that the artificial intelligence has immensely developed constructive and more active techniques of the process of text categorization and classification.

Through the automated process, a document or an article can be duly analyzed using document management software such as weka and Rapid Miner software. The use of this automated processes has adequately resulted to the accomplishment of various objectives revolving the routing of the information, retrieval of the information, understanding the ontology of information domain as well as facilitating the creation and maintenance of the desired texts. This is the principal objective of summarizing texts and files in the Arabian context that this project seeks to accomplish. (Tomek, Wang, and Bowden wise, 1998) presented in their paper “The Strong Practical Text Summarization” an automated method of generating human-readable summaries from text documents such as news, technical reports, government documents and court records. They described the summarizer tool, a java-implemented prototype and its applications in various document-processing tasks. The result of their summarizer tool was very efficient and robust in summarizing (Strzalkowski, 1998).

The use of innovative text categorization model has also achieved extensive application in the classification of texts and files such as those of the Arabian contexts. The model is superior and quite imperative based on its particularity and ability to split a particular document into plausible number of different and unique passages. It is through the application of this model that many options are available for categorizing texts and files. Some of these important options include pages, paragraphs, and both overlapping as well as non-overlapping windows among others. This technique is essentially important especially under circumstances where the categorization is individually applied to each document rather than each passage. The use of innovative text categorization model is also useful in the category merging process. Here, categorized passages are merged to help in the composing of the entire document from the categorized passages. Through this technique, it becomes quite simple to categorize the various existing documents and text files based on their passage categories.

Ideally, based on research, it is quite evident that the experiments involving the use of innovative text categorization model contribute to more than five percent improvement index relative to the traditional processes of text categorization. (Mining Hu , and Bing Liu , 2004) mined and summarized in their paper “Mining and Summarizing Customer Reviews” all the customer reviews of a product by only mining the features of the product on which the customers had expressed their opinions and whether the opinions were positive or negative.

This process was performed in three steps, first: Mining product features. Second: Opinion sentences identification in each review and deciding if each opinion sentence was positive or negative. Third: the results summarization. The result indicated that summarizing the reviews was not only useful to common shoppers but also crucial to product manufacturing. They suggested a set of mining and summarizing product reviews techniques based

on the methods of data mining and natural language processing (Hu, Minqing, and Bing Liu, 2004) ideally, standard categorization of text requires additional analysis of various sentiments that are involved in the application and use of various files, texts and documents. This analysis can be achieved in various ways such as the use of computational techniques in the classification of the text. The user essentially does this through the consideration of the context of use of language. The feelings and the reactions are all important aspects that require concise consideration in the achievement of this objective. The use of words, texts or files within a particular passage, blogs or social media forms allows the analysis of the different hierarchical levels in which different texts can be categorized and classified.

One such technique used in this kind of classification is the use of explicit text aspects. This technique explores and examines the use of collation in various forms of research in a quite comprehensive manner and then presents the results in the table and outputs that comply with the recent standards as illustrated in the weka output used in the analysis section. In this regard, it also priceless to succinctly understand the primary sections involved in the analysis of sentiments implied by various texts and passages.

(Kathleen McKeown, Dragomir R. Radev, 1995) presented in their paper “Generating Summaries of Multiple News Articles” a natural language system, which summarized a series of new articles on the same event by using summarization operators that is identified through empirical analysis of a corpus of news summaries. The result was that development of a methodological framework to ease future implementation of news summarization systems (McKeown, 1995)

(Poonam Yadav, 2015) presented in his paper “Document Features-Enabled Text Summarization System For Information Retrieval” a method called: Document Features-Enabled Text Summarization System For Information Retrieval, which by it, the document database was applied to the summarization system which found score value based on the important sentences in the document by utilizing different document features such frequency-based , title-based and position-based. The result was that the proposed method achieved the highest precision value of 80% as compared to existing algorithm (Yadav, 1995).

(Octavia Marias. Ulea, Sergiu Nisioil, Liviu P.Dinul, 2016) investigated in their paper “Using World Embeddings To Translate Named Entities” the usefulness of natural word embeddings in the process of translating Named Entities (NEs) from a resource-rich language to a language low on resources relevant to the task at hand , introducing a novel, yet simple way of obtaining bilingual word vectors .Their result showed that the approach was more stable under context specific particularities (e.g. House as Assemblée or Acronyms) and it could potentially improve tools that already exist for the target language (Octavia Marias.ulea, sergiuNisioil, liviuP.Dinul, 2002).

Named Entity Recognition (NER) is a mind boggling, Information Extraction subtask, requiring a few preprocessing stages (i.e. grammatical feature tagger, tokenizer) which thusly include committed instruments. For asset rich dialects, such as English, NER is an exceptionally investigated region with the sateof-the-craftsmanship framework accomplishing close human execution: 93% F1 contrasted with the 97% F1 acquired by human annotators (Swamp and Perzanowski, 1998). For different dialects having less dialect, preparing instruments and particularly assignment physically commented on information, NER is yet a testing errand.

Word embeddings have been lately utilized as elements to enhance existing monolingual NER frameworks ((Katharina Siencnik, 2015), (Demir and Ozgur, 2014)), or to help the interpretation of NEs (Zirikly, 2015). Past to this, (Shao and Ng, 2004) revealed utilizing word embeddings as some portion of a bigger framework that concentrates named elements from similar corpora. Others have utilized arrangement models to extricate this sort of data from parallel datasets (see (Moore, 2003), (Ehrmann and Turchi, 2010)). Also to parallel or similar datasets, metadata data, whenever accessible, can likewise demonstrate helpful (Ling et al., 2011) for multilingual named entity extraction. Identified with multilingual named elements, we take note of the transliteration of NEs given outside of any relevant connection to the subject at hand, the choice on whether to transliterate or, on the other hand decipher likewise having been examined (Mahmoud Azab and Oflazer, 2013). The aftereffects of the 2015 ACL shared assignment on transliteration of named entities<sup>1</sup> uncovered that further research is important to get palatable brings about this course.

(Laila Khreisat, 2006) presented in her paper “Arabic Text Classification Using N-Gram Frequency Statistics : A comparative study “ the results of classifying Arabic text documents using the N-Gram frequency statistics technique that employed a dissimilarity measure called “ The Manhattan Distance” and Dice’s measure of similarity for comparison purposes , The Dice measure was used . The results showed that N-Gram text classification that used the Dice measure performed out classification by using the Manhattan measure. The results for the Tri-Gram method that used the Dice measure exceeded those for the Manhattan measure, reached its highest recall value (Khreisat, 2006).

(Rania Al – Sabbagh, Roxana Girju, 2012) presented in their paper “YADAC: Yet another Dialectal Arabic Corpus” the first phase of building YADAC – a multi-genre Dialectal Arabic (DA) Corpus. They are looking in the future to extend work to other Arabic Dialects, improving performance of both POS tagging and base phrase chunking by reducing the input noise, incorporating more features and launching the information extraction web tool to YADAC.

(Samuel R. Bowman, Gabor Angeli, Christopher Potts, Christopher D. Manning, 2015) introduced in their paper “A large Annotated Corpus for Learning Natural Language Interference” the Stand ford Natural Language Interference Corpus, a new, freely available collection of labeled sentence pairs, written by humans, by doing a novel grounded- task based on image captioning. The results showed that both simple Lexicalized models and natural network models performed well, and that the representations learned by a neural network model on their corpus could be used to dramatically improve performance on a standard challenge dataset.

(Shaan K., Raza H., 2008) presented in their paper “Arabic Named Entity Recognition From Diverse Text Types” the results of their attempt at the recognition and extraction of 10 most important named entities in Arabic script; the person’s name, location, company, date, time, price, measurement, phone number, ISBN, and file name. They used a Rule –based approach to develop the system, it employed with great linguistic expertise provided a successful implementation of the NERA system by accomplishing challenges posed by Arabic language. The performance results achieved were satisfactory in terms of precision, recall and F-measure.

(Adam Kilgarriff, Gregory Grefenstette, 2003) aimed in their paper “Introduction to the Special Issue on the Web as Corpus” to survey the activities and explore recurring themes. Their results showed that the search engine would be a wonderful tool for language researchers if these constraints were removed:

The search engine results did not present enough instances (1.000 or 5.000 maximum)

They were selected according to criteria that were from a linguistic perspective, distorting (used the search term in titles and headings went to the top of the list and often occupied all the top slots).



They did not specify searches according to linguistic criteria such as the Citation Form for a word or word class.

They did not present enough contexts for each instance (Google provided a fragment of around ten words).

(Ayoub. Souleiman, Julia Freeman, 2015) investigated in their paper “Arabic News Article Summarization” Arabic PDF news articles to produce results from their new program that indexed, categorized and summarized them. The values extracted using Named Entities Recognizer (NER). They used Fusion Lucid Works (a solar based-system) to help with the indexing of their data set and to provide an interface for the user to search and brows the articles with their summaries. The results were as they expected: The interface ran seamlessly and showed the results based on the search criteria, the application was running up on the client’s machines and had been tested, and the documents that were parsed also had been imported into fusion and indexed along with the modified schema file.

(Motaz K. Saad, Wesam Ashour, 2010) presented in their paper “OSAC: Open Source Arabic Corpora “the complex nature of Arabic language, posed the problems of:

The Lack of Free Public Arabic Corpora.

The lack of High-Quality, Well-Structured Arabic Digital Contents.

The result was collection of the largest free accessible Arabic Corpus OSAC, which contained about 180 million words, and about 0.5 million-district keywords.

(Motaz K. Saad, Wesam Ashour, 2010) Furthermore, because of the lack of Arabic Morphological Analysis Tools, they presented and evaluated in their paper “Arabic Morphological Tools for Text Mining” existing common Arabic Stemming /Light Stemming Algorithms. They also implemented and integrated Arabic Morphological Analysis Tools into the leading open source machine learning and data mining tools, Weka and Rapid Miner.

(Abdulmohsen, AL-Thubaity, 2015) reviewed in his paper “A 700M+ Arabic Corpus: KACST Arabic Corpus Design and Construction” 14 Arabic Corpora categorized by their designated purpose, target language, mode of text, size, text date, location, text type, medium, text domain, representativeness and balance. The KACST Arabic Corpus was a large and diverse design criterion, it was sampled carefully and its contents were classified based on time, region, medium, domain and topic, and it could be searched and explored using these classifications. The result was that the KACST Arabic Corpus comprised more than 700 million words from the pre-Islamic Era to the present day collected from 10 diverse mediums, also it was and still freely available to explore on the Internet using a variety of tools.

(Deepali K. Gaikwad, C. Namrata Mahender, 2016) presented and summarized in their paper” A Review Paper On Text Summarization “the view of text summarization from every aspect from its beginning up to date, also they provided an abstract view of the present scenario of research work for text summarization. The results showed that text summarization had its importance in both commercial as well as research community. The abstractive summarization was a bit complex than extractive approach but provided more meaningful and appropriate summary compared to extractive. There was a lot of scope for exploring such methods for more appropriate summarization.

(Sadik Bessou, Mohamed Touahria, 2014) provided in their paper “An Accuracy –Enhanced Stemming Algorithm for Arabic Information Retrieval” a method for indexing and retrieving Arabic texts based on natural language processing. The results obtained indicated that the algorithm extracted the exact root with an accuracy rated up to 96% and hence improved information retrieval.

(Tarek Kanan, Edward, Fox, 2016) developed in their paper “Automated Arabic Text Classification With P .Stemmer, Machine Learning, And A Tailored News Article Taxonomy” a software for browsing a collection of about 237.000 Arabic news articles which should be applicable to other Arabic news collections, designed a simple taxonomy for Arabic news stories, developed tailored stemming; a new Arabic light stemmer called P. Stemmer, in connection with a Qatar National Research Fund QNRF –funded project to build a digital library community and infrastructure in Qatar. The results were classification results for Arabic textual data enhanced by using their proposed stemmer when used three classifiers: Native Bayes, SVM and Random Forest, SVM performed better than the other two, Binary classification gave better results compared to multiclass classification.

(Hmeidi Ismail, Ghassan Kanaan, Martha Evens, 1997) built in their paper “Design and Implementation of Automatic Indexing for Information Retrieval with Arabic Documents” an Arabic Information Retrieval System used to compare the results of Manual and Automatic Indexing. They carried out three separate series of experiments: one used roots as index terms, one used stems, and one used words. Their results suggested that the Arab World should consider using Automatic Indexing on a large scale because it is cheaper and faster than Manual Indexing. Also the results obtained by Al-Kharashi (1991) were confirmed by using titles only, the results obtained by Abu-Salem (1992) were confirmed by using 120 abstracts, and the roots made better index terms than stems or words, at least when phrases were not involved.

(Mohammed N. Al- Kabi, Saif, Kazakzeh, Belal, Al- Rababah, Izzat, 2015) introduced in the paper “A novel Root Based Arabic Stemmer” a new light and heavy Arabic Stemmer which compared with two well-known Arabic Stemmers. They applied three main processing phases to generate Arabic roots from words; phase one was responsible for removing prefixes

and suffixes, phase two was responsible for comparing the output to standard word sources or shapes and phase three was responsible for correcting the extracted root .The results showed that the accuracy of their Stemmer was 75.03%, which was slightly better than the accuracy yielded by each one of those two well-known Arabic Stemmers used for evaluation and comparison.

(Belal, and Asma, 2015) presented in the paper “A Rule-Based Stemmer for Arabic Gulf Dialect” a new rule based Dialect Stemmer built for the Gulf Dialect. The algorithm was built a set of predefined rules for Gulf Dialect. They showed that the result of this new Stemmer was as follows: The new Stemmer accuracy was acceptable, it gave superior results compared to other Stemming algorithms and the algorithm could handle many Dialects.

On the other hand, (Abidi Karima, Elberrichi, and Tlili, 2012) in the paper “Arabic Text Categorization: A Comparative Study Of Different Representation Modes” implemented a research on the categorization of the Arabic texts which its originality related to the use of a conceptual representation of the text , they used Arabic Word Net (AWN) as a Lexical and semantic resource .They showed the benefits and advantages of this representation compared to the most conditional methods and demonstrated that the addition of the semantic dimensions was one of the most promising approaches for the automatic categorization of Arabic.

Also (Al-Harbi, Al-Muharehb, and Al-Thuabity, 2008) in the paper “Automatic Arabic Text Classification” attempted to attain a better understanding and elaboration of Arabic text classification techniques by experimenting that on a document classification achieved on seven different Arabic Corpora using statistical methodology. The results gave better accuracy, where the tool was implemented for feature extraction, selection, and the performance of two popular classification algorithms (CVM and C5. O) Had been evaluated on classifying Arabic Corpora C5. O classifier.

## 2.2 Arabic language

Arabic language is a global language, which is written from right to left, with major differences with most popular languages such as English, Hindi, Chinese and Spanish. Arabic language includes many grammatical forms, and lots of varieties of word synonyms and different meanings for the same word, which depends about the word, and the inclusion of diacritics.

There are 28 different characters, formulation and shapes for the same letter in the Arabic language, which based on this letter in the word. In addition, these characters classify into small, which can be assigned to a letter as superscript or as a subscript, in order to add to this letter to add different grammatical formulation and meaning. These small characters are used in Modern Standard Arabic i.e. formal written Arabic version.

Arabic is the fifth most common spoken language in the world and it is used by 4.5% of the world population as their primary language. (Nationsonline, 2015). Because of the difficulties of structural, grammatical, and linguistic forms, word synonyms diversities, and different word meanings that the Arabic language has, limited work has been devoted to natural language processing involving Arabic, especially in comparison to the English language, which has been addressed by numerous studies. Arabic language necessities are not addressed by most of the computational linguistics tools, so changes and extra efforts are required for adapting to Arabic in order to make these tools work with Arabic language data. In addition, modification and extra work are required to allow handling Arabic language data because most software packages, tools, and APIs for information retrieval and natural language processing do not address Arabic language requirements.

As mentioned before, Arabic is written from right to left, unlike most languages, without capitalization, and with 28 alphabetical characters. There are multiple forms of Arabic language (Habash, 2010) such as:

Classical Arabic: This form is used in reading / reciting the holy books.

Modern Standard Arabic (MSA): This form is commonly used in writing, speech, interviewing, broadcasting, etc. Implementation is based on MSA throughout this form.

Spoken – oral dialects: These dialects significantly vary from region to region.

Vowel Marks (Tashkeel or Harakat which known as diacritics) such as those shown in Table 6 for one of the letters.

Table 6: Diacritics for the Letter “BAA”

بَ	بُ	بِ	بٌ
----	----	----	----

Such these diacritics, which can be used interchangeably, and change the meaning of the word can distinguish sounds, which are not fully specified by the Arabic letters. These diacritics hold very little value in the analysis, which carried out on texts in connection with computational linguistics because they are mostly used in the verbal exchanges or recitation context.

## Chapter Three

### Methodology and approach

#### 3.1 Dataset selection

Different datasets were collected from the Arabic newspapers and dialects and analyzed using various techniques. The processes such as parsing as well as varying filtrations were done and this process yielded many texts in the Arabic context. From these texts, various attributes were combined and analyzed independently by exploring their values as well as summarizing them on a template. This was extensively done by conducting random sampling of the collected texts. From the collection exercise, various datasets were collected as shown in table (7).

Table 7: various characteristics of dataset.

Main language	Arabic
Encoding format	UTF-8
Number of sentences	796,767
Size on the disk	900MB
Approximate number of words	7,282,897
Format of file	Text file

The extraction of the characteristics in the Arabian context was done in different ways which include, notwithstanding, the examination if various titles, topics, the date of publishing the article, the category in which the article belongs and the examination of their named identities with consideration to issues such as the person or people addressed, the organization of origin of the article and the writer of the article. This was primarily done to the article for the Arabian news. On the other hand, In the course of our dataset collection, we extensively used Arabian documentaries that we initially collected by taking news article from the newspaper and web sites, after that we filtered the collected document and butt them in a txt files as shown in Table 8

Table 8: categories of text file

Subject	Average size per doc	Number of document per folder	Overall collection size
Art & culture	4 KB	500	927 KB
Economy	5 KB	500	1270 KB
Political	7 KB	500	1800 KB
Sciences	6 KB	500	1620 KB
social Issues	3 KB	500	990 KB
Sport	3 KB	500	1340 KB

Proper and standard categorization of texts involves active placement of the various existing text documents into various categories based on their contents. This can be achieved in many ways. Mostly, the use of internet as well as automated systems has played significant roles in achieving standard and proper categorization of texts. Categorization of texts is principally done based on the application of both manual as well as automated systems. In either case, extreme care should be upheld to ensure that the method used is that which stems from the application of different fields that may be applicable and prove quite useful as far as the classification of texts is concerned.

This implies that the automated method should focus on the manifestation of various fields that provide a better tedium for performing manual processes and measure on their level of difficulty in applicability. Through this methodology, it is succinct that the classification of texts within various articles can be achieved in a more concise and precise manner. This is not only important in ensuring proper understanding of text summarization and classification but also plays a vital role in ensuring that the appreciation of various classes such as arts, economics, politics as well as sports are duly appreciated by the individuals charged with the responsibility of categorization of texts.



Other than the application of web-based categorization of texts as illustrated on the above methodology, Arabic articles can also be classified through the application of various Arabic stemmers. This can be either the light as well as root-based stemmers. Essentially, this technique is quite useful in the illustration of how effective categorization of texts is. Through this technique, it is quite possible to allow the comparison of various tools and software that are used in data mining. These tools may include, notwithstanding, Weka as well as Rapid Miner tools.

Essentially, these tools have been duly used in this research and their output results have been duly illustrated and analyzed as shown in the diagrams below. One major strength in the use of these imperative data mining tools is enhanced accuracy in the classification of texts. This can provide valid and sound data that can be dependable and used in the future years to come. Ideally, it is through the application of these tools that the makes the comparison of various forms of text classification or categorization easy to explore especially with regards to the exploration of classifications involving Arabic stemmers.

### 3.2 Research Method and Approach

In this research, three approaches were used. These include:

Text categorization with WEKA

Text categorization with NER

Text categorization with GATE

Text categorization with WEKA

We started working on program summary template depending on previous results set by corpus, classification and NER in order to improve after classification and NER summary template by a high quality and being clear for the user. It depends a ton of the domain you are working in. You need to characterize the components in light of the domain. Say in a web index you are chipping away at figuring out how to rank issue, creating a dynamic rank, the NE's will not give you any advantage here. It largely relies upon the domain that you are working and furthermore the yield arrangement marks (managed learning) characterized.

Presently say you are chipping away at classifying documents relating to Soccer or Movie or Politics et cetera. For this situation Named Entities can work. I will give you a case here; say you are utilizing a Neural Network, which orders documents into Soccer, Movie, and Politics and so on. Presently say an archive comes in "Lionel Messi was welcome to go to the head of "The Social Network", likewise display were the thrown and group including Jesse Eisenberg, Andrew Garfield and Justin Timberlake" Here the association between named entities (input elements) and film (yield characterized) will be more grounded and thus it will be delegated a report on Movie.

Another illustration, say our record is "Tom Cruise is depicting the character of Lionel Messi in the motion picture "The last soccer match". Here comes the advantage say your neural system has learnt that when a performing artist and footballer meets up in one report there is high likelihood of it being a film. Again, it relies upon the information and preparing it might be other route round as well (however, that is what is realizing about; seeing the past information)

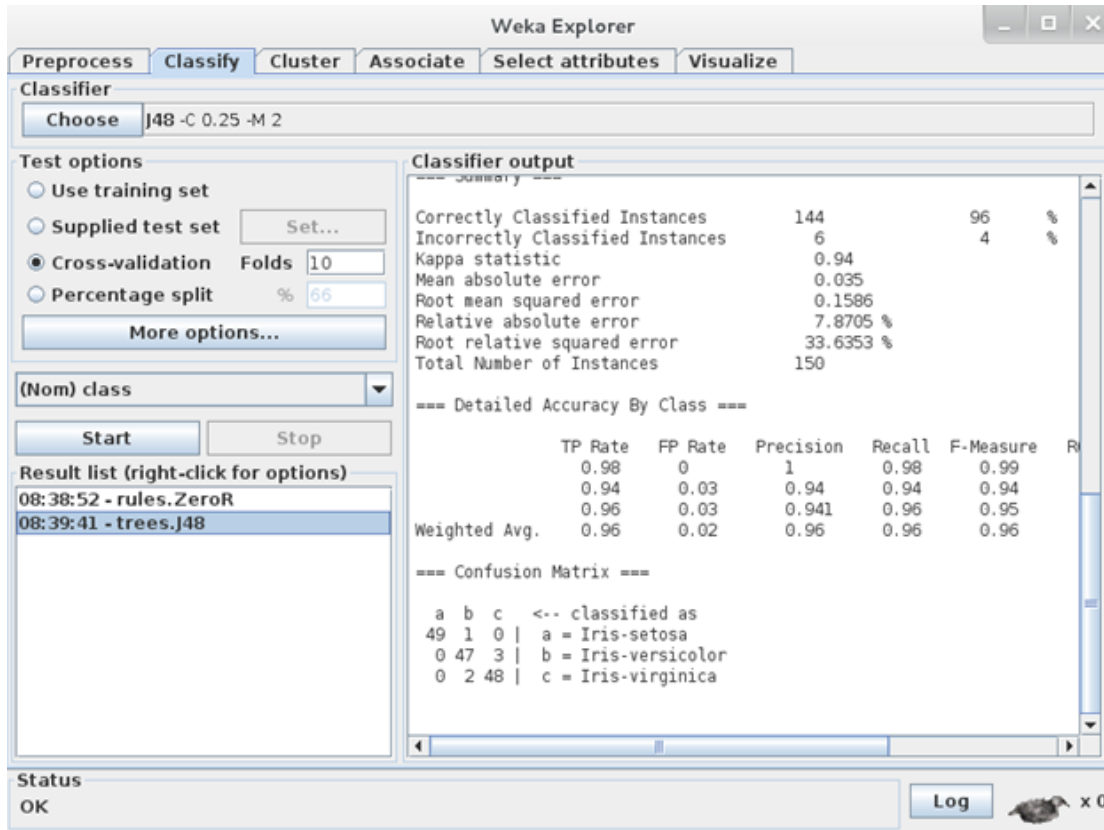


Figure 10. WEKA Explorer

Summary template depends on accurate information to reach the correct and content summary based on a clear and comprehensive model for the most important parts in query status. As a matter of course, non-numerical traits get foreign made as NOMINAL characteristics, which is not really fancied for printed information, particularly on the off chance that one needs to utilize the StringToWordVector channel. To change the text to STRING, one can run the NominalToString channel (bundle weka. filters.unsupervised.attribute) on the information, determining the trait file or scope of records that ought to be changed over (NB: this channel does not bar the class property from transformation!). With a specific end goal to hold the quality sorts, one needs to spare the document in ARFF or XRFF organize (or in the packed variant of these configurations).

### 3.4 Outsider devices

Tag Helper Tools, which enables one to change writings into vectors of stemmed or unstemmed unigrams, bigrams, grammatical form bigrams, and some client-characterized elements, and after that spares this portrayal to ARFF. As of now forms English, German, and Chinese. Spanish and Portuguese are in advance.

Working with literary information

### 3.5 Transformation

Most classifiers in Weka cannot deal with String qualities. For these learning plans, one needs to handle the information with fitting channels, e.g., the String to Word Vector channel, which can perform TF/IDF change.

The String to Word Vector channel puts the class characteristic of the produced yield information toward the start. In the event that you had to jump at the chance to have it as last quality once more, you can utilize the Reorder channel with the accompanying setup:

weka. Filters. Unsupervised. Attribute. Reorder - R 2-last, first. What's more, with the MultiFilter you can likewise apply the two channels in one go, rather than in this manner. Makes it less demanding in the Explorer for example.

### 3.6 Stop words

The String to Word Vector channel can likewise work with an alternate stop word list than the inherent one (in view of the Rainbow framework). One can utilize the – stop words alternative to stack the outer stop words document. The arrangement for such a stop word record is one stop word per line; lines beginning with "#" are deciphered as remarks and overlooked.

Note: There was a bug in Weka 3.5.6 (which presented the help of outside stopwords records), which disregarded the outer stopwords list. Later forms or depictions from 21/07/2007 on will work accurately.

### 3.7 UTF-8

In the event that you are working with content records containing non-ASCII characters, e.g., Arabic, you may experience some show issues under Windows. Java was intended to show UTF-8, which ought to incorporate Arabic characters. As a matter of course, Java utilizes code page 1252 under Windows, which distorts the show of different characters. Keeping in mind the ultimate goal to settle this, you should change the java charge line with which you start up Weka (taken from this post):

```
java - Dfile.encoding=utf-8 - classpath ...
```

The - Dfile.encoding=utf-8 advises Java to unequivocally utilize UTF-8 encoding rather than the default CP1252.

In the event that you are beginning Weka by means of begin menu and you utilize a current form (no less than 3.5.8 or 3.4.13), at that point you will simply need to adjust the fileEncoding placeholder in the RunWeka.ini as needs be.

After that, we worked on the WEKA for training the data and then get text categories to be matched as it shown in figure 11 to summarize text, WEKA will add categories on the text file.

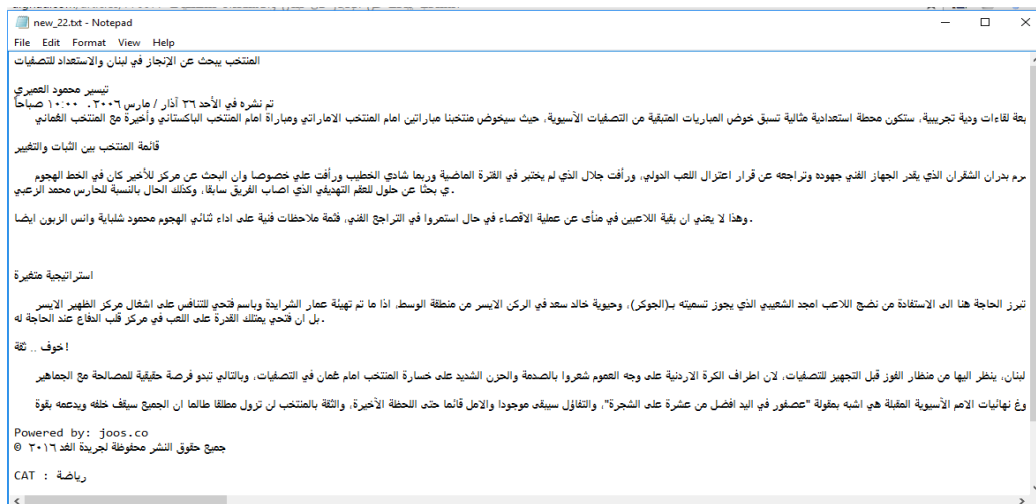


Figure 11: Result file from WEKA

After we summarized the text and labeled it by the categories, we give the results to NER tools (GATE), as it shown in figure 12.

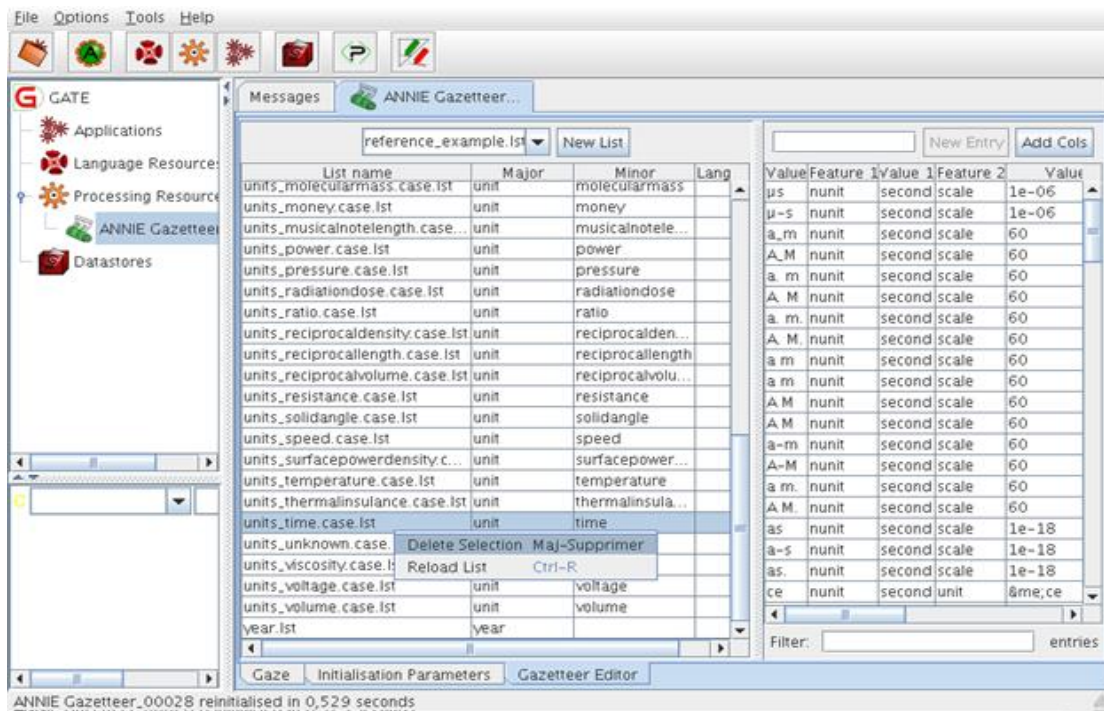


Figure 12: Gate Tool

In this tool, the result was concluded as for arguments, which are PRN, ORG, LOC and DATE as it shown in figure 13



Figure 13: Output of Gate Tool

After that, several Templates were built for each categories to be combined with the original summarized document. English template summary was built for explanation purposes .

Document as it shown in figure 14 - 19

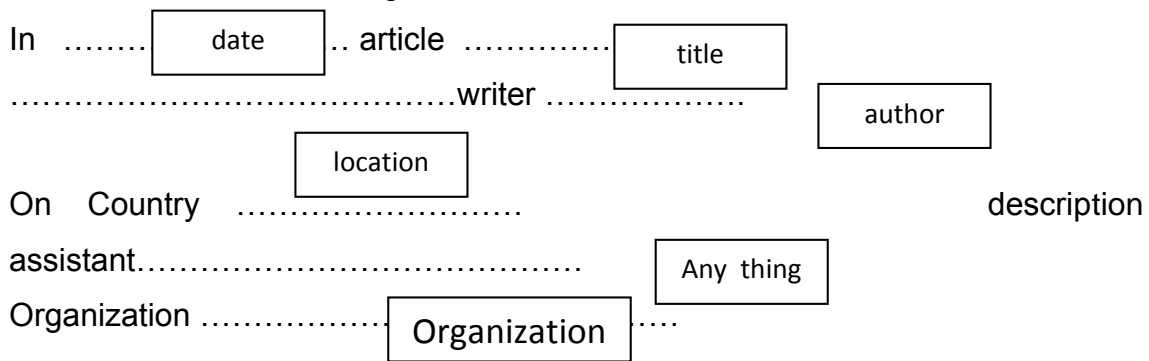
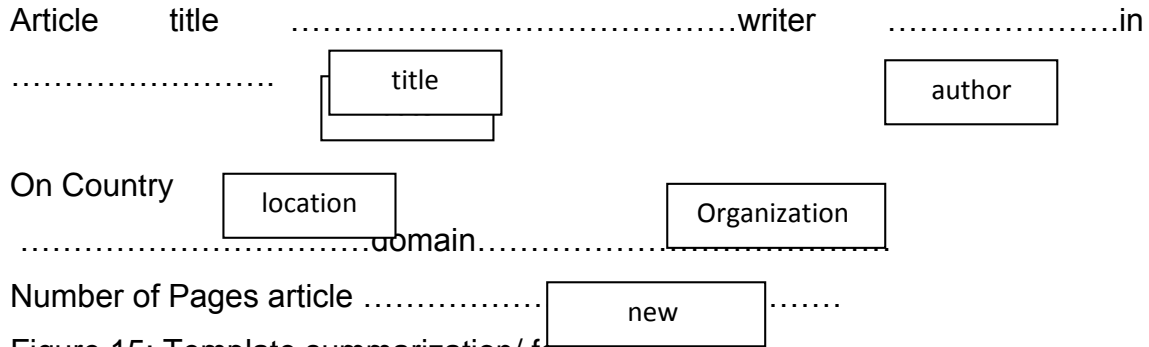


Figure 14: Template summarization/ form 1



..... في عام ..... قام الكاتب / المؤلف ..... بطرح موضوع في السياسة .....

..... وذلك في مقال نشره بعنوان..... ناقش العديد من القضايا المتعلقة ب.....

..... تم ذكر العديد من الأشخاص / الدول المعينون بذلك .....

Figure 16: Political Template summarization / form 3

..... الكاتب / المؤلف ..... مقال بعنوان ..... تاريخ النشر .....

..... تطرق الكاتب إلى ..... المجلة العلمية المشار إليها .....

..... في مكان النشر .....

Figure 17: Science template summarization/ form 4

..... التاريخ ..... مقال بعنوان ..... للكاتب ..... والذي نشر

..... بخصوص ..... مقالات / أشخاص / مؤسسات ذات صلة .....

..... والمتعلقة ب.....

Figure 18: Art and Culture template summarization/ form 5

..... كتب الكاتب..... في مكان النشر ..... بتاريخ النشر ..... في مجال .....

..... عن ان المنتخبات الرياضية ..... وكتب المقال في .....

Figure 19: Sport template summarization/ form six





## Chapter Four

### Analysis and Discussion

#### 4.1 Discussion

After collecting the news article and retaining the corpus to check the accuracy of rule-based program that was built, the document was and inputted to WEKA to categories the text file, after that GATE tool extracted the person, location, organization and date.

Our program extracts the person, location, organization and date and combine with the template we built to produce a summary for each news article.

For our summary evaluation, we generate 150 summaries and distributed them randomly to MA students, we provided each student with a number of articles, along with the corresponding summaries (that we generated automatically). We asked them to read the article first, then read the summary, and estimate the quality of the summary using a (rating) scale. We asked each student to assign a rating for each summary based on its quality, between 1-10, with 1 for substandard quality or even irrelevant summary, and 10 for excellent quality. Then we averaged the results. After collecting the data from students, we evaluate the quality of our summaries and thus of our template method.

Figure 21 and Table 9 show the evaluation results of 150 articles.

Table 9: Number of Articles for each Score

Rate Value	1	2	3	4	5	6	7	8	9	10
Number of Articles	3	2	2	9	14	16	19	22	36	30

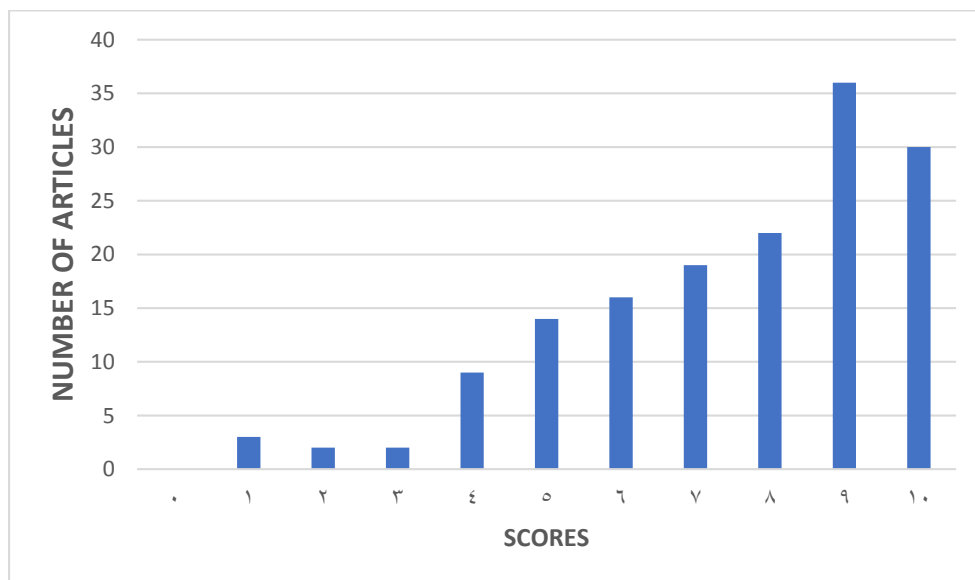


Figure 21: Categories Used to Show the Summary Evaluation Results

The student satisfaction standard was evaluated to a numerical range

The result was (7.50).

The result provides a proof that the proposed work can be used to summarize Arabic text summarization template based.

To compare the proposed work with other related one it is worth to mention the following:

The resulted summary is well structured and well written when it compares with others work.

The proposed work measurement user satisfaction by distributed 150 generated summaries and the average of the user satisfaction was (7.5).

The summary that we generate include feature that does not exist in other summaries, for example the Date attribute

## Chapter Five

### Comparative Analysis and Methodology

#### 5.1 Comparing Methodologies

The ability to create bias-free and accurate data is based on the need to compare the proposed methodology with previous successful ones. In this research, a Template Based Summarization for Arabic Documents is developed. For any developed system, there is a need to compare it with previous ones. For example, the developed systems in this research is compared against the work of Kanan (2015) [add reference here]; the work of this research has certain similarities with Kanan's work. First, both researches use almost similar methods of data collection and extraction. The data that is collected for both researches are passed through a filtration and parsing processes to create combined attribute values that would serve as a template for purpose of producing an efficient summary.

Both studies have also depicted the use of random sampling for ensuring that the required sample set does not become biased. For example, Kanan (2015) has employed an extraction process that is involved a seven-attribute-value system which are (title, topic, publishing date, category, name entity, organization and location). Only six attribute values were generated in this research (i.e. title, author, date, location, Organization, category) All of the attributes from both studies were used to create the template for creating a meaningful summary for a selected article. In this research, several templates were prepared for the categories used.

For example, there is a template for the sport category and another one for the social issues, and so on so fourth. Whereas, there is only one general template that is used in Kanan research. Figure (22) represents the template that is used in Kanan research.

{العنوان}: تحقيق الامن والاستقرار واعاده بناء مصر من جديد على اساس  
المساواه والعداله الاجتماعيه  
{تاريخ النشر}: ٢٥/أغسطس/٢٠١٢  
{الكاتب}: أنور الخطيب  
{الأشخاص المشار اليهم}: محمد مرسي, عبد الحميد الانصاري, موزه المالكي,  
ابراهيم ال ابراهيم, عيسى ال اسحاق, ربيع الكواري  
{المؤسسات المشار إليها}: رئيس مصر, جامعه قطر, حزب الحره والعداله,  
جماعه الاخوان المسلمين, مؤسس حمد الطيبه  
{التصنيف العام}: السياسيه  
{الكلمات في الموضوع الرئيسي}: حكم, محمد, مرسي, الاخوان, المسلمين,  
مصر, الرئيس, والاستقرار, السياسيه, رئيس

Figure 22: The Template used in Kanan research

To ensure that all attributes were extracted accurately, tools for extraction were employed. Both studies employed Name Entity recognizer (NER) that is used to extract values for of the Writer, Person, and Organization attributes (Kanan, 2015). The figure below contains the attributes that are generated in Kanan research. These attributes were used thereafter for creating a meaningful summary for Arabic documents.

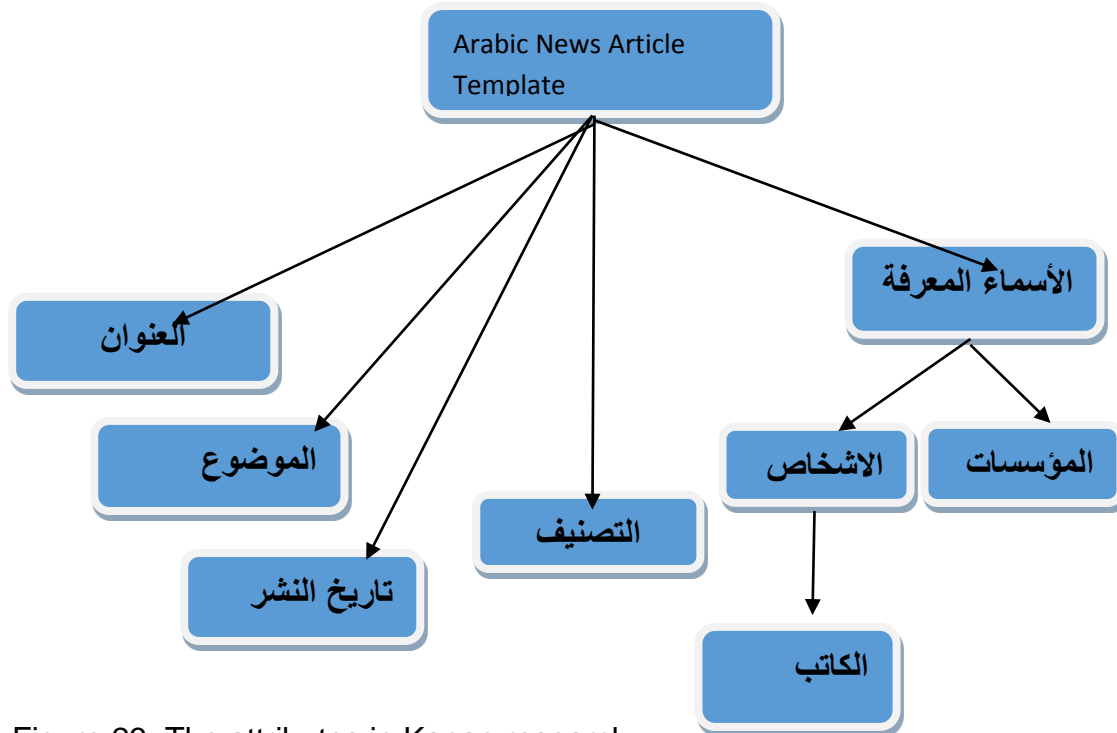


Figure 23: The attributes in Kanan research.

Still on the extraction of the values, web-based extraction applications were deemed appropriate for my study. Similarly, Kanan has employed several web-based applications to extract some attributes. Some of these include the punctuation and machine-learning methods for categorization of the articles for extracting the category attribute. These data extraction and categorization tools are deemed to produce sound and accurate data free from bias and errors. In addition, Kanan used regular expression for extracting the date attribute whereas in this research a new developed tool is used to extract the date attribute. This tool is developed using C# programming language. Seven attributes were collected from each article, the purpose of these attributes is to serve in creating a template summary for the news article Kanan (2015). In both cases, several methods have been employed to conduct the extraction of the predefined attributes. WEKA and GATE is used in this research. In addition, NER is used in Kanans methodology.

Table (10) summarizes the extraction techniques that were used by both researches in order to extract the attribute values.

Table (10): The techniques used to extract the attribute values.

	The used technique to extract an attribute						
	Date	Title	Name/ location	Organizati on	Write r	Category	Subjec t
Kanan Resear ch (2015)	Regular Expressi on	Simple text extraction method using punctuatio n	Named Entity Recogniz er	Named Entity Recognize r	NER	classificat ion algorithm s	Arabic LDA tool
This researc h	In house develop ed tool	Using Weka	GATE	GATE	NER	Weka	N/A

Eventually, the extracted articles were analyzed in order to get their characteristics. For example, name, size, and their subject. This process is performed to get certain information about the contents of the sample articles. The content will be used then during the summarization process; both studies stored the filtered data in a text file. The text files that were produced with be used during the process of generating the template summarization article.

Moreover, it is worth mentioning that the source of the sample articles that are used in this research is collected from the internet, specifically, from online Arabic Newspapers After the attributes were combined in Kanan (2015), they were made to serve as a template summary for the news article, each of them was attached to its corresponding article



and all of them were saved in the same file. In the proposed methodology, the various attributes were only combined and analyzed independently by exploring their values and they were summarized on a template where each category has its own template. Even though, both researches have considered the random selection of the sample articles, Kanan selected a larger sample size than the sample that is used in this research.

In addition to the above-mentioned differences, the researcher has also noticed that Kanan (2015) categorized the attributes into two categories. The named entities, which include the writer, organization, and person and the unnamed entities such as the title, topic, category, and date. Another notable difference that Kanan has with this research is that the used methodology with the articles that do not have topics. For example, Kanan used the Arabic topic generation tool that generates topics for the articles that do not have one.

Lastly, Kanan used students who have a wide understanding of the Arabic language to help in evaluating and proofing the methods and the summaries that are used in his study. The study of this research lacks the human input that Kanan does. The use of human factor with other related methodologies for evaluating the proposed methodology will give a higher chance of producing relevant template summaries (Kanan, 2015). It is worth mentioning to compare between the results of both researches. Table (11) and Figure (24) depicts the results of Kanan research and figure (25) depicts the result of this research.

Table (11): evaluation of Kanan research

Rate	0	1	2	3	4	5	6	7	8	9	10
Number of	0	0	0	4	24	99	269	300	237	65	2

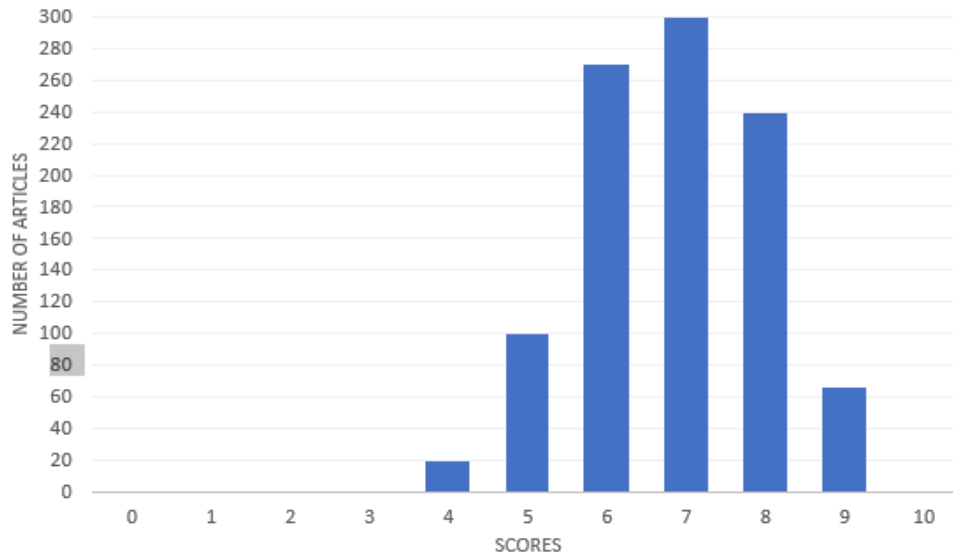


Figure 24: evaluation of Kanan research

Table (12): evaluation of this research

Rate	1	2	3	4	5	6	7	8	9	10
Value										
Number of Articles	3	2	2	9	14	16	19	22	36	30

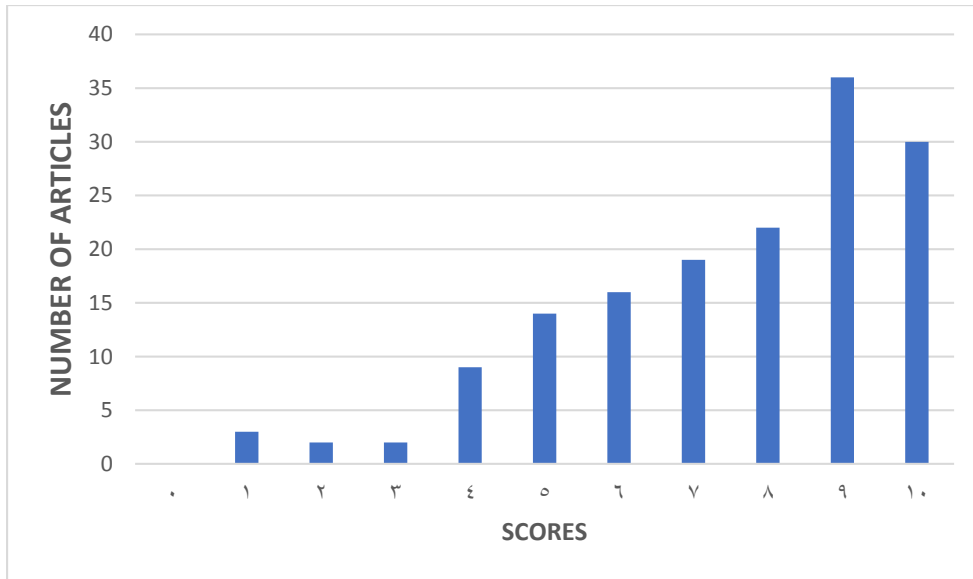


Figure (25) evaluation of this research

Kanan use student to evaluate his proposed system, the collected articles were divided into 11 categories. The ten participants evaluate the system where each one give his/her opinion on two hundreds articles. Each category is evaluated by two participants. The evaluation scores of each article and the frequency of each score are appeared in the above figures.

The more the value of the score the more relevant summary. From the figures above, it can be concluded that applying the template summarization methods on Arabic news can produce a good template-based summary.

In this research, the approach of evaluating the proposed system is like that one used by Kanan where. In this research, 150 summaries were generated and distributed randomly to a set of master's students where each student is provided by number of articles, along with their corresponding summaries the students were asked to rank the relevancy of the summary with the original articles. The evaluation is collected and analyzed as appears in the above figures. Notice that Kanan used a larger set to documents in the evaluation process.

## 5.2 Conclusion

The use of multiple facets to categorically sort out data in the two studies is critical to ensuring that the results are well filtered to provide accurate conclusions. It is also obvious that both methodologies have utilized a random sampling design to ensure that all the words and the seven-attribute system are well informed. This process defines a well-researcher format of data collecting. The use of data extraction tools is also a plus for both studies. Apart from saving time, labor and ease of access, these tools also ensure error free extraction for the purposes of maintaining viability in the entire research. However, there are somethings that the methodology should have considered, such as defining solid criteria of obtaining online newspapers with a random sampling method as advocated for by Kanan (2015). Moreover, there is no proper explanation of the number of sample articles that are used in both studies.

This study should try to tabulate the number of newspaper extracted sources., Also, there should be a defined manner in which all the proposed approaches to be integrated to produce a more reliable summarization system for Arabic-based documents where the main objective for all of us is to build up a reliable system, Eventually, the study should incorporate a system of validating the results of the study. In other words, there should be a control to ensure that the analyzed data is well tested before the presentation. This brings to us a new field of research that others can pursue.

## References:

- Saad, K., & Ashour, W. (2010, November). Osac: Open source arabic corpora. In 6th ArchEng Int. Symposiums, EEECS (Vol. 10).
- Abuata, B., & Al-Omari, A. (2015). A rule-based stemmer for Arabic Gulf dialect. *Journal of King Saud University-Computer and Information Sciences*, 27(2), 104-112.
- Al-Harbi, S., Almuhareb, A., Al-Thubaity, A., Khorsheed, S., and AlRajeh, A., (2008). Automatic Arabic Text Classification, 9th International journal of statistical analysis of textual data, pp. 77-83,
- Al-Kabi, N., Kazakzeh, A., Ata, A., Al-Rababah, A., & Alsmadi, M. (2015). A novel root based Arabic stemmer. *Journal of King Saud University-Computer and Information Sciences*, 27(2), 94-103.
- Al-Sabbagh, R., & Girju, R. (2012, May). YADAC: Yet another Dialectal Arabic Corpus. In LREC (pp. 2882-2889).
- Al-Thubaity, O. (2015). A 700M+ Arabic corpus: KACST Arabic corpus design and construction. *Language Resources and Evaluation*, 49(3), 721-751.
- Ayoub, S., & Freeman, J. (2015). Arabic News Article Summarization.
- Barzilay, R., & McKeown, R. (2001, July). Extracting paraphrases from a parallel corpus. In Proceedings of the 39th annual meeting on Association for Computational Linguistics (pp. 50-57). Association for Computational Linguistics.
- Benajiba, Y., (2009). Arabic named entity recognition, Ph.D. dissertation. Universidad Politécnica de Valencia. Valencia, Spain.
- Bessou, S., & Touahria, M. (2014). An Accuracy-Enhanced Stemming Algorithm for Arabic Information Retrieval. *Neural Network World*, 24(2), 117.
- Bowman, R., Angeli, G., Potts, C., & Manning, D. (2015). A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326.
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze: Introduction to information retrieval –(2008) Cambridge, England, Cambridge University Press,

- Croft, W. (2015). Force dynamics and directed change in event lexicalization and argument realization. In *Cognitive science perspectives on verb representation and processing* (pp. 103-129). Springer International Publishing.
- Fox, C. (1989, September). A stop list for general text. In *Acm sigir forum*(Vol. 24, No. 1-2, pp. 19-21). ACM.
- Gançarski, L., Doucet, A., & Henriques, R. (2006). Attribute grammar-based interactive system to retrieve information from XML documents. *IEE Proceedings-Software*, 153(2), 51-60.
- Habash, Y. (2010). Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1), 1-187.
- Hmeidi, I., Kanaan, G., & Evens, M. (1997). Design and implementation of automatic indexing for information retrieval with Arabic documents. *JASIS*, 48(10), 867-881.
- Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177). ACM.
- Kan'an, G. (2015). Arabic News Text Classification and Summarization: A Case of the Electronic Library Institute Seer Q (ELISQ).
- Kanan, T., & Fox, E. (2016). Automated Arabic Text. Classification with P-Stemmer. Machine Learning, and a Tailored News Article Taxonomy. *J. Assoc. Inf. Sci. Technol.*
- Karima, A., Zakaria, E., Yamina, G., Mohammed, S., Selvam, P., & VENKATAKRISHNAN, V. (2012). Arabic text categorization: a comparative study of different representation modes. *Journal of Theoretical and Applied Information Technology*, 38(1), 1-5.
- Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational linguistics*, 29(3), 333-347.
- Khreisat, L. (2006). Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study. *DMIN*, 2006, 78-82..

Luhn, P. (1958). The automatic creation of literature abstracts. IBM Journal of research and development, 2(2), 159-165.

Mani, I. (2001). Automatic summarization (Vol. 3). John Benjamins Publishing.

McKeown, K., & Radev, R. (1995, July). Generating summaries of multiple news articles. In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 74-82). ACM.

Mihalcea, R. (2005, June). Language independent extractive summarization. In Proceedings of the ACL 2005 on Interactive poster and demonstration sessions (pp. 49-52). Association for Computational Linguistics.

Nationsonline, (2015). Most widely spoken Languages in the World. [Cited03/4/2017]. Available:

[http://www.nationsonline.org/oneworld/most\\_spoken\\_languages.htm](http://www.nationsonline.org/oneworld/most_spoken_languages.htm)

Octavia Marias.ulea, sergiuNisoiu, liviuP.Dinul, (2015). the Using Word Embeddings to Translate Named Entities.

Radev, R., Hovy, E., & McKeown, K. (2002). Introduction to the special issue on summarization. Computational linguistics, 28(4), 399-408.

Saad, K., & Ashour, W. (2010). Arabic morphological tools for text mining. Corpora, 18, 19.

Salton, G. (1986). Another look at automatic text-retrieval systems. Communications of the ACM, 29(7), 648-656.

Salton, G. (1989). Automatic text processing: The transformation, analysis, and retrieval of. Reading: Addison-Wesley.

Shaalán, K., & Raza, H. (2008). Arabic named entity recognition from diverse text types. Advances in Natural Language Processing, 440-451.

Strzalkowski, T., Wang, J., & Wise, B. (1998). A robust practical text summarization. In Proceedings of the AAAI Symposium on Intelligent Text Summarization (pp. 26-33).

Yadav, P. (2015), Document Features-Enabled Text Summarization System for Information Retrieval, International Journal of Computer Systems, vol. 02, no.

04

Zha, H. (2002). Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 113-120). ACM.