Keyword Extraction Based on Word Co-Occurrence Statistical Information for Arabic Text

Mohammed Al-Kabi*, Hassan Al-Belaili**, Bilal Abul-Huda** and Abdullah H. Wahbeh***

Received on Jan. 22, 2012  Accepted for publication on June 24, 2012

Abstract

Keyword extraction has many useful applications including indexing, summarization, and categorization. In this work we present a keyword extraction system for Arabic documents using term co-occurrence statistical information which used in other systems for English and Chinese languages. This technique based on extracting top frequent terms and building the co-occurrence matrix showing the occurrence of each frequent term. In case the co-occurrence of a term is in the biasness degree, then the term is important and it is likely to be a keyword. The biasness degree of the terms and the set of frequent terms are measured using $F_2$. Therefore terms with high $F_2$ values are likely to be keywords. The adopted $F_2$ method in this study is compared with another novel method based on term frequency - inverted term frequency (TF-ITF) which tested for the first time. Two datasets were used to evaluate the system performance. Results show that the $F_2$ method is better than TF-ITF, since the precision and the recall of the $F_2$ for the first experiment was 0.58 and 0.63 respectively and for the second experiment the $F_2$ accuracy was 64%. The results of these experiments showed the ability of the $F_2$ method to be applied on the Arabic documents and it has an acceptable performance among other techniques.

Keywords: Keyword extraction, Arabic Keyword extraction, Information Retrieval, Natural language processing.

Introduction

Today, the internet contains a huge amount of electronic information such as papers, articles and news. This huge volume shows the necessity to have an effective ways to retrieve and filter the desired information. Many search engines have the ability to retrieve the most relevant document but there is a need to show a brief description of the retrieved information especially when the human’s incapable to summarize this huge amount of information. Keyword extraction techniques are important for information seekers, since it allow them to get what they want just by looking at the suggested keywords so determine what they should read. The need of keyword extraction comes from the huge growth of the Internet; the amount of information is rapidly increasing in many different languages [1] and documents in Arabic language are part of this growth. Therefore there is an increasing need for the retrieval, filtering and mining of Arabic documents through the World Wide Web.
Keyword extraction is the process of extracting a few most important words from a certain text and using these words to summarize the content. This approach has been widely studied for a long time in the natural language processing (NLP) [1]. Keyword extraction is important for many text applications, such as document retrieval, text mining, classification, clustering, extracting of semantic information, and summarization [2,3].

Many Studies such as [2-11] addressed the process of keyword extraction. The basic idea of keyword extraction is to select the most important words in the document that have the ability to describe the content of the document. By extracting appropriate keywords, we can easily choose which document to read and to learn the relationship among documents [4]. Keywords are not selected randomly, but based on some techniques for extracting features then selecting the proper terms. A popular technique for doing this is to use terms statistical information, which extracts keywords according to a corpus of documents from the same domain [12]. Other techniques can extract keywords directly from a single document without using a domain corpus [4].

Keywords enable readers and researchers to decide whether a document or a page is relevant for them or not. In addition, keywords can be used as a low cost measure of similarity between documents. Unfortunately, a large number of existing electronic documents available today do not have keywords available for them. Taking into account the difficulty and time-consuming of assigning keywords to documents, it is preferable to automate this task using different methods and techniques [13].

Current studies proposed different methods to extract keywords such as term co-occurrence, term frequency–inverse document frequency (TF-IDF), term weighted and vector space model, lexical chains, k-Nearest Neighbor (kNN) algorithm and others, but some of these techniques are not applied to Arabic texts. This study aims to build an Arabic keyword extractor based on term frequency and term co-occurrence statistical information and measure the bias degree of terms using $F_2$ measures, then selecting the candidate keywords which have high $F_2$ value [4] and applying some rules that are appropriate for Arabic language to combine meaningful terms.

The remaining sections this study is structured as follows: section 2 presents some related work in the field of keyword extraction; section 3 presents the scope of the study and the adopted methodology; section 4 presents the conducted experiments with their results and interpretations. Conclusion and possible future work are presented in section 5.

Related Work

There are many Studies in the field of building and implementing effective keyword extraction systems. In this section we focused on keyword extraction Studies based on data mining, information retrieval and other fields.
Extraction Based on Information Retrieval (IR) Techniques

Information retrieval (IR) techniques showed high accuracy and robustness for extracting the most relevant keywords from documents. Many Studies take the advantages of these techniques and applied it to different languages.

Liu et al. [2] presented a method to assign weight to each candidate word according to a vector of four features (TF-IDF, Part-of-speech tagging (POS tagging or POST), relative position of first occurrence and $\chi^2$ statistics). Liu et al. [2] found that the feature vector weighted sum is a good choice for scoring. Test data is consisting of 2563 web pages about information technology, (1,563) of them were used for training and (1,000) for testing. A Chinese lexical analysis system for segmentation and POST were used. Results showed that the weight vector is (66.56, 22.76, 8.25, and 4.20) based on the four features (TF×IDF, the part-of-speech (PoS) tag, relative position of first occurrence and chi-square statistics) adopted by the researchers. An average of 6.8 keywords are manually assigned per text, where the top 7 words are assigned by experiment conducted, comparing results for both manual; and automatic assignments show results with 64% precision, 73% recall and 68% F-measure.

Qingguo et al. [14] preprocessed the document by Vector Space Model to represent the training documents and testing documents, and then construct a candidate keywords based on k-NN method. The weight of the candidate keyword is estimated by the sum of distances between the test document and all nearest neighbors containing the candidate keyword. The training set contains 10,221,700 documents with keywords assigned by library and information experts, the test data set is collected randomly and contains about 10,000 documents. Results show the capability of the proposed method to improve the precision and recall of keyword extraction using k-Nearest Neighbor (k-NN) algorithm relative to Pure Statistic Method, Feature Combination Method, and Naive Bayes, beside the ability to extract implicit subject efficiently using k-NN.

Lui et al. [5] study presents a novel domain independent keyphrase extraction (KE) algorithm, which uses a combination of statistical and computational linguistics techniques, a five new attributes (TF-IDF, position of term, title, proper noun, number of terms in a term phrase), and a novel neural network based learning method to distinguish keyphrases from non-keyphrases. The experiment of this work shows that it performs like other extraction tools such as GenEx introduced by Turney [15], and Kea (a keyphrase extraction algorithm based on a Naive Bayes technique) which introduced by Frank et al. [16], and MS Word 2000’s AutoSummarize. Results showed that the performance of the proposed algorithm is as the performances of GenEx and Kea, besides outperforming MS Word 2000’s AutoSummarize.

Nguyen et al. [7] described a novel keyword extracting tool depends on some characteristics like term frequency, document collection characteristics, and term distribution at the document collection. The architecture of the proposed system starts with Lexical analyzer which is responsible for scanning the input documents, break them up into meaningful words, and then collect term distribution information. Next, the Vector Model Builder which is responsible for building vector based model for
documents and building the document matrix, each vector represent an individual document. Finally; the Keyword Extractor which is responsible for extracting keywords for each document using the document matrix. This tool allows the user to determine the number of keywords as threshold from 1 to 10. The experimental results showed a satisfactory performance achieved by this tool.

Li et al. [17] proposed an iterative language pattern mining algorithm to discover meaningful patterns from online corpus where these patterns help extracting keywords at real time. The proposed algorithm consists of two steps: Language Pattern Mining (LPM) and Extract Keywords from Online Broadcasting Content. Li et al. [17] carried out experiments on some real-world data and their results showed the effectiveness of the proposed algorithm, compared with traditional statistical methods on online streams.

**Extraction Based on Hybrid and other Techniques**

A combination of neural network models and information retrieval techniques are used to develop the extraction models or methods. A research by Azcarraga et al. [18] evaluated the keyword selection methods for WEBSOM methodology that used for text archiving; this methodology uses a well-known neural network training algorithm “Self-Organizing Map” (SOM) at the core of its archiving technique. The archive contains about 100,000 nodes, which mean about 7 million documents. Azcarraga et al. [18] proposed an alternative method for extracting keywords for this archive, and compared the results with the WEBSOM extraction method. Results showed that a high percentage of the keywords extracted by the new system match the top keywords extracted for the same nodes using the WEBSOM extraction method, and that the new system is faster than the WEBSOM.

Another research by Jo et al. [19] proposed a neural network back propagation model for extracting keywords. This model based on some factors such as term frequency TF, inverse document frequency IDF, and word position WP. Applying these factors to generate vectors to select keywords. The research use a group of heterogeneous documents as simple documents to determine IDF (Inverse Document Frequency), TF (Term Frequency) and ITF (inverted term frequency) which is presented as integers greater than or equal to zero. Other factors are used and represented in binary value such as T (title), FS (first sentence) and LS (last sentence). To experiment this model a set of data collected from news articles is applied to the model. The precision was 0.7 for selected rate of 20% and 0.82 for selected rate of 40% (Selected rate is the rate of words selected as keywords of the document to the total words included in it).

Zhang et al. [20] proposed a new algorithm for keyword extraction which is called CCG (Cognition and Concept Graph), where documents are represented as maximum connected sub-graphs of the basic graph of the document and the cognition of a term is weighted with data depth. CCG is tested on a study by Zhang et al. [20] on a Chinese test documents, where 5 keywords are extracted automatically from each document, for comparison another 5 keywords with higher frequency are extracted automatically. The CCG showed that the extracted keywords can describe the document much better than keyword extracted by other algorithms.
Arabic Studies

Few Studies for Arabic language addressed the problem of document word extraction. Only one web application tool exist powered by Sakhr [21], which allows the users to have some analysis information about their documents such as keywords extraction, summarization, and categorization.

KP-Miner is a keyphrase extractor system proposed by El-Beltagy et al. [22] for both English and Arabic documents. KP-Miner built on some linguistics observations and hypotheses such as the fact that the more important a term is; the more likely it is to appear sooner in the document than later. The algorithm of proposed model consists of three steps namely; candidate keyphrase selection, candidate keyphrase weight calculation and final candidate phrase list refinement. The proposed system is evaluated by comparing it with Sakhr keyphrase extractor. Results showed that KP-Miner is better than Sakhr especially in recall (0.38 for KP-Miner and 0.32 for Sakhr) and the percentage of times title was recognized correctly as a top keyphrase was (61% for KP-Miner) and (36% for Sakhr).

Another research for the summarization of Arabic documents by Douzidia et al. [23] described a new Arabic summarization system called Lakhas. The proposed system is based on extraction methods and consists of 9 modules. The new system extracts short sentences, about 29 words in average and 16% comparison rate. Therefore authors applied sentence reduction methods. Executing these methods leads the system to reduce the length of the summarized sentence to 50% further to reach 15 words only. This work by Douzidia et al. [23] is classified by Schlesinger et al. [24] as a multilingual summarization system that used machine translated Arabic documents (MT). Beside Douzidia et al. [23] work; other Studies by [24-26] conducted for Arabic language by translating Arabic document to English and then applied sufficient techniques to summarize the translated documents.

Some Applications of Keyword Extraction

The importance of keyword extraction comes from the rapidly increasing amount of electronic information on the internet such as electronic books, papers, journals, Studies, news, reports, web pages and emails. This variety and heterogeneity of the data makes it difficult to find what the users need. If there is a keyword assigned to each material, then reaching the desired material will be easy and quickly by just looking to the corresponding keywords to determine the relevant material, and this is very important for search engines and indexing.

Another usage of keyword extraction is within documents and web page clustering, which can be performed by measuring the similarity between the keywords and the different categories, by using extracted keywords for automatic labeling of clusters [27], or using a machine learning model which is based on enumeration of all possible keywords combinations as described by [28]. Also, keywords allow us to filter web pages and extract advertising keywords as reported in [17, 29], and other electronic commercial disciplines.
Figure 1: The Methodology.
Keyword Extraction Based On Word Co-Occurrence Statistical Information for Arabic Text

The growth of the newspapers in the internet pose a big challenge to information retrieval, since every day there is a new news taking multiple forms and written in multiple languages, so there are some Studies considering this problem, and proposed methods for keyword extraction, that can help for managing and filtering these newspapers [12, 30].

Methodology

Figure 1 shows the overall steps that make up the methodology of this study. As shown in the figure, the methodology consists of three main parts namely; text preprocessing phase, keyword extraction phase, and keyword selection.

Text Preprocessing

In the text preprocessing phase the system prepare the document for the next two phases. In this phase, documents will be partitioned into sentences, and frequent words that do not have any chance to be a candidate keyword are removed.

Extraction of Phrases

The first step in the text preprocessing phase is phrase extraction from Arabic documents which include one or more paragraph, each of which contains a set of sentences that have a particular meaning or benefits to the whole document. Punctuation marks like “”, “;” “” “.” and “” are used by the built system to understand the structure of the text document. Therefore sequence of words that ends with the dot (.) is considered as a sentence, document titles, and section titles also will be considered as sentences.

Stop Words Removal

This step deals with the elimination of non-meaningful words, which does not indicate the semantic content of the document such as ("so", "with", بالإشارة), ("confirmation", بالنسبة), ("for", تاكد) or appearing frequently in the document like pronouns such as ("he", "they", "هم", "هم"). These words may have a bad effect to the Study approach which depends on statistical information and co-occurrence of the words [31]. Also, foreign words, numbers and symbols like (@, #, &, %, *) and some words that indicates a sequence of the sentences like ("firstly", "أولًا"). These words also are also recognized as stop words.

A list of 745 stop words is used to complete this step. After removing all these types of stop words we got pure sentences with meaningful words and from these sentences the system extract candidate keywords.
Keywords Extraction

In this phase, the system extracts keywords from the preprocessed Arabic sentences. This phase consist of the following steps.

Selection of Frequent Terms

Frequent terms are used to show the importance of the term and its ability to play a role in the semantic of the document content. In this step the system obtain the frequent terms by counting term frequency of each unique word that forms the entire training document. Our system selects the top 10 frequent terms denoted as the set of the frequent terms $G$, and 50% of the number of the running terms $N_{total}$. As an example, table 1 shows the top ten Arabic frequent terms of training document and their normalized probability of occurrence, so the sum of all probabilities is one. After frequency calculations, the system extracts keywords using $\chi^2$ computations or a novel TF-ITF proposed by the authors.

Table 1: Frequency and Probability Distribution of the top 10 Arabic Frequent Terms.

<table>
<thead>
<tr>
<th>Frequent term</th>
<th>Oath</th>
<th>Spanish (Masculine)</th>
<th>Madrid</th>
<th>Teams</th>
<th>Union</th>
<th>Number</th>
<th>Year</th>
<th>Team</th>
<th>Ball</th>
<th>Spanish (Feminine)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>27</td>
<td>21</td>
<td>18</td>
<td>17</td>
<td>13</td>
<td>12</td>
<td>11</td>
<td>11</td>
<td>10</td>
<td>9</td>
<td>149</td>
</tr>
<tr>
<td>Probability</td>
<td>0.181</td>
<td>0.141</td>
<td>0.121</td>
<td>0.114</td>
<td>0.087</td>
<td>0.081</td>
<td>0.074</td>
<td>0.074</td>
<td>0.067</td>
<td>0.06</td>
<td>1</td>
</tr>
</tbody>
</table>

$CHI$-Square Method ($\chi^2$)

The next subsections describe the process of computing the value of $\chi^2$.

Computing the Co-occurrence

This step deals with term co-occurrence with other terms. A co-occurrence matrix is generated by counting co-occurrence frequencies of pairs of terms [4], as an example; if term $a$ co-occurred with term $b$ in 7 sentences, therefore the value of the terms co-occurrence is equal to 7.

Based on the above idea, the system generates the term co-occurrence matrix which is an $N \times 10$ matrix where:

$N$: The total number of unique terms in the document.

10: The number of terms belongs to set $G$.

For example, table 2 shows the co-occurrence of the top 10 Arabic frequent terms that belong to $G$. Co-occurrence of two terms $a$ and $b$ is denoted as $freq(a, b)$ and represent the observed probability.
### Table 2: A Co-occurrence Matrix of the top 10 Arabic Frequent Terms.

<table>
<thead>
<tr>
<th>Arabic Word</th>
<th>Oath (القسم)</th>
<th>Spanish (Masculine) (القسم)</th>
<th>Madrid (القسم)</th>
<th>Teams (الفريق)</th>
<th>Union (الفريق)</th>
<th>Number (عدد)</th>
<th>Year (سنة)</th>
<th>Team (فريق)</th>
<th>Ball (الكرة)</th>
<th>Spanish (Feminine) (القسم)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oath</td>
<td>-</td>
<td>4</td>
<td>2</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>11</td>
<td>2</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Spanish (Masculine)</td>
<td>4</td>
<td>-</td>
<td>9</td>
<td>4</td>
<td>8</td>
<td>6</td>
<td>5</td>
<td>10</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Teams (الفريق)</td>
<td>2</td>
<td>9</td>
<td>-</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Teams (الفريق)</td>
<td>6</td>
<td>4</td>
<td>0</td>
<td>-</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>9</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Union (الفريق)</td>
<td>4</td>
<td>8</td>
<td>2</td>
<td>5</td>
<td>-</td>
<td>4</td>
<td>2</td>
<td>7</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Number (عدد)</td>
<td>2</td>
<td>6</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>-</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Year (سنة)</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>-</td>
<td>7</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Team (فريق)</td>
<td>11</td>
<td>10</td>
<td>6</td>
<td>9</td>
<td>7</td>
<td>5</td>
<td>7</td>
<td>-</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Ball (الكرة)</td>
<td>2</td>
<td>8</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>Spanish (Feminine)</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>-</td>
</tr>
</tbody>
</table>

The co-occurrence matrix helps us to identify some terms. If a term appears independently from frequent terms, then the distribution of co-occurrence of this term and the terms in $G$ is quite similar to the probability distribution of terms in $G$ that shown in table 2, so these terms is a general terms and more likely to occurred with many terms. If a term appears with high probability distribution with some frequent terms, then this term has a semantic relationship with some terms in $G$, and this kind of distribution is said to be biased, and biases usually derived from semantic or other relationships between two terms. So, terms with high biases co-occurrence may have an important meaning in a document [4]. The degree of biasness can be calculated using $\chi^2$ measures. So, terms with high $\chi^2$ values are considered to have some importance in the document.

**Computing the Expected Probability**

In this step we provide a description about how to compute the expected probability that will be used in the $\chi^2$ equation. Authors in [4] suggest two aspects to improve the calculation of $\chi^2$ values, which are variety of sentence length and robustness of the $\chi^2$ value.

Sentences lengths in different documents differ, so if the term occurs in long sentences then it will occur with many other terms. In the other extreme, if the term just occurs in short sentences then it will occur with very few terms. This diversity in
sentence lengths will be considered in the calculation of the expected probability of terms co-occurrence

The expected frequency of co-occurrence of term \( w \), and the term \( g \in G \) can be calculated using the following equation (1) [2, 4].

\[
EXP = n_w p_g
\]  

(1)

Where:

- \( p_g \) is (the sum of the total number of terms in sentences where \( g \) appears) divided by (the total number of terms in the document \( N_{total} \)).
- \( n_w \) is the total number of terms in sentences where \( w \) appears.

Calculating the \( \chi^2 \) Value

The \( \chi^2 \) is a measure by which we calculate the difference between expected counts and observed counts [7]. The degree of biasness of term co-occurrence with frequent terms can be used as an indicator of the term importance, but the problem here if the frequency of a term is small; for example if the term \( a \) appears once and co-occurred just once with only one frequent term \( g \), so the probability will be 1, also if term \( b \) appears 50 times and co-occurred with only one frequent term \( g \) 50 times, once again the probability will be 1, but here we can say term \( b \) is bias to the frequent term \( g \) because its appears many times with the same term.

To avoid such problem the \( \chi^2 \) measure used to evaluate biases between expected frequencies and observed frequencies. The value of \( \chi^2 \) of term \( w \) is defined as [4]:

\[
\chi^2(w) = \sum_{g \in G} \frac{(freq(w, g) - EXP)^2}{EXP} 
\]  

(2)

Where:

- \( freq(w, g) \): is the co-occurrence of term \( w \) with term \( g \) denoted as the observed frequency.
- \( EXP \): is the expected frequency calculated from equation (1).
- \( G \): is the set of the top frequent terms.

The second consideration by [4] is the robustness of the \( \chi^2 \) value; if a term co-occurred only with one frequent term it may has a high \( \chi^2 \) value because of the high bias degree it has, such terms is not important to the document. The importance of the term comes from co-occurrence of the term with more than one frequent term. A modification of equation (2) is made to solve such problem; as shown in equation (3):
Keyword Extraction Based On Word Co-Occurrence Statistical Information for Arabic Text

\[
\chi^2(w) = \chi^2(w) - \max_{g \in G} \left( \frac{(freq(w,g) - EXP)^2}{EXP} \right) \tag{3}
\]

Equation (3) will eliminate the term \(w\) that co-occur with only one frequent term \(g \in G\) from being a candidate keyword.

**TF-ITF Method**

The other method that the system will use to extract the candidate keywords is the novel TF-ITF which depends on the TF (Term Frequency) and ITF (Inverse Term Frequency) [6], where the extraction of keywords will be extracted according to the following equation (4):

\[
TF - ITF(w) = F_w \times \log \left( \frac{F_d}{F_w} \right) \tag{4}
\]

Where:

- \(F_w\): is the total frequency of term \(w\) in the document.
- \(F_d\): is the total number of terms in the document.

This measure is derived from TF-IDF. The only difference is that TF-IDF is used in the corpus while TF-ITF is used in a single document. Terms with highest TF-ITF values are considered candidate terms for extraction.

**Keywords Selections**

Most of the keywords are not single terms, where you can find compound terms of two or three terms. In Arabic documents the combination of words is ruled by the grammar which helps to understand the meaning of the words according to their position in different sentences. In this phase the system will take the candidate keywords and try to have a meaningful combination between them by parsing the candidate keywords in the original document and get the sequence in which they originally appear.

**Results and Evaluation**

In this section experiments have been conducted, and an overall evaluation of the proposed system is presented. The system's overall performance is judged in terms of precision and recall.

**Evaluation Methods**

Many different measures for evaluating the performance of information retrieval systems have been proposed, the performance of the extractor system is measured by comparing the generated keywords by the system for each document with the extracted suggested keywords by an expert. The performance measure is based on the number of matches between the system generated keywords and the human generated keywords [32].

85
Information retrieval Researchers use the precision and recall to measure the system performance. Figure 2 shows how the precision and recall used to evaluate the system. Let A be the set of keywords extracted by the system, B the set of keywords extracted by the human, C the intersection set between these two sets (A & B) that shows the number of correct keywords extracted by the system.

![Keywords Intersection Diagram](image)

**Figure 2: Keywords Intersection.**

Recall is the fraction of keywords made by the human that are extracted by the system, and given by the following equation (5) [32, 33]:

$$R = \frac{|C|}{|B|}$$

Precision is the fraction of keywords extracted by the system that are made by the human, and given by the following equation (6) [32, 33]:

$$P = \frac{|C|}{|A|}$$

The above two measures clearly trade off against each other; therefore F-measure can be used to measure the trade off of precision versus recall as follows [32, 33]:

$$F = \frac{2PR}{P + R}$$

The accuracy measure will be used to measure the accuracy of the system to extract keywords, which are concepts in specific domain. The accuracy is defined as the percentage of correctly extracted keywords, and given by the following equation (8) [33]:

$$Accuracy = \frac{CT}{N}$$
Keyword Extraction Based On Word Co-Occurrence Statistical Information for Arabic Text

Where:

\( CT \): is the total number of corrected terms extracted by the system that match concepts in the thesaurus.

\( N \): is the total number of terms extracted by the system.

**Evaluation Dataset**

Two datasets are used to evaluate the proposed system. The first set consists of two types of documents; large documents and small documents; where each group consists of five documents collected from known Arab newspapers websites. The collected documents may be political, economical or religious, where a number of keywords are assigned manually to each of the collected documents. These documents are used to test the effectiveness of the proposed system to extract proper and significant keywords. Table 3 exhibits the results of the test on the collected documents.

<table>
<thead>
<tr>
<th>Document Type</th>
<th>No. of Sentences</th>
<th>No. of Words</th>
<th>Average of Words per Doc.</th>
<th>No. of Words After Preprocessing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large Docs.</td>
<td>165</td>
<td>8,969</td>
<td>2,242</td>
<td>5,354</td>
</tr>
<tr>
<td>Small Docs.</td>
<td>98</td>
<td>2,762</td>
<td>690</td>
<td>1,932</td>
</tr>
<tr>
<td>Total</td>
<td>263</td>
<td>11,731</td>
<td>----</td>
<td>7,286</td>
</tr>
</tbody>
</table>

The second dataset consists of five agricultural documents collected from the Arab agriculture ministry’s websites. The aim of using such collection is to see whether the extracted keywords map the concepts found in the agriculture vocabulary AGROVOC dataset offered by the food agriculture organization of the United Nation (FAO) [34]. Large numbers of keywords mapping the concepts in AGROVOC give an indication about the extractor system capability of extracting terms in specific domain. Table 4 shows results related to AGROVOC dataset.

<table>
<thead>
<tr>
<th>Document Type</th>
<th>No. of Sentences</th>
<th>No. of Words</th>
<th>No. of Words After Preprocessing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agriculture Docs</td>
<td>104</td>
<td>5462</td>
<td>4416</td>
</tr>
</tbody>
</table>

**Implementation and Experiment Environment**

The extractor system was implemented using the C# language which is part of the Microsoft visual studio 2005. The extractor system was experimented using a computer with the following specification: an AMD Turion-X2 2.1 MHz CPU, 4 GB RAM, 250 GB HDD 5400 rpm, and windows Vista Home premium Edition.
Evaluation Results

The first experiment was conducted on the first dataset; where the results for the $\chi^2$ and TF-ITF methods are shown in table 5 and table 6 respectively.

**Table 5: Performance of the $\chi^2$ Method.**

<table>
<thead>
<tr>
<th>Docs Type</th>
<th>$P$</th>
<th>$R$</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large</td>
<td>0.66</td>
<td>0.661</td>
<td>0.65</td>
</tr>
<tr>
<td>Small</td>
<td>0.5</td>
<td>0.601</td>
<td>0.56</td>
</tr>
<tr>
<td>AVG</td>
<td>0.58</td>
<td>0.63</td>
<td>0.61</td>
</tr>
</tbody>
</table>

Table 5 shows the performance of the $\chi^2$ method. Results showed that the large document set achieved the highest F-measure with a value of 0.65, followed by the small document set with an F-measure value of 0.56.

**Table 6: Performance of the TF-ITF Method.**

<table>
<thead>
<tr>
<th>Docs Type</th>
<th>$P$</th>
<th>$R$</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large</td>
<td>0.48</td>
<td>0.484</td>
<td>0.473</td>
</tr>
<tr>
<td>Small</td>
<td>0.46</td>
<td>0.548</td>
<td>0.498</td>
</tr>
<tr>
<td>AVG</td>
<td>0.47</td>
<td>0.52</td>
<td>0.49</td>
</tr>
</tbody>
</table>

Table 6 shows the performance of the TF-ITF method. Results showed that the small document set achieved the highest F-measure with a value of 0.498, followed by the large document set with an F-measure value of 0.473.

Sometimes the length of the document plays an important role in the performance of the extractor method, so figure 3 and figure 4 show a comparison between the two methods in the large documents and in the small documents respectively to give us indication about how these methods deals with variant document lengths.

![Figure 3: Performance of $\chi^2$ and TF-ITF among Large Documents.](image)
Figure 4: Performance of $\chi^2$ and TF-ITF among Small Documents.

The above figures show that the $\chi^2$ performed better than TF-ITF when dealing with large documents, the highest value was for Recall with a value of about 0.661. Also the value of $\chi^2$ was better than the value of TF-ITF within small documents yields 0.6 Recall.

The second experiment was conducted on the second dataset. This dataset was experimented on the system and the correct matches of keywords extracted by $\chi^2$ and TF-ITF are shown in table 7 and 8 respectively. The results obtained from this experiment show the ability of the system using the $\chi^2$ method to extract keywords that are concepts in AGROVOC.

Table 7: Extracted Arabic Keywords from Agriculture Documents using $\chi^2$ Method.

<table>
<thead>
<tr>
<th>Quantities</th>
<th>Production</th>
<th>Palm</th>
<th>Cultivation</th>
<th>Pests</th>
</tr>
</thead>
<tbody>
<tr>
<td>River نهر</td>
<td>Potato البطاطا</td>
<td>Membership العضوية</td>
<td>International العالمية</td>
<td>Insects الحشرات</td>
</tr>
<tr>
<td>Storage خزن</td>
<td>Crop المحصول</td>
<td>Food الغذاء</td>
<td>Grain الحبوب</td>
<td>Fruit الفاكهة</td>
</tr>
<tr>
<td>Water مياه</td>
<td>tubers الدرنات</td>
<td>nuclei النوى</td>
<td>Production إنتاج</td>
<td>manna المن</td>
</tr>
<tr>
<td>Arab العربي</td>
<td>Heat الحرارة</td>
<td>Palm النخلة</td>
<td>Areas المناطق</td>
<td>Worms ديدان</td>
</tr>
<tr>
<td>Cubic مكعب</td>
<td>Agriculture الزراعة</td>
<td>Region منطقة</td>
<td>Food الغذائي</td>
<td>Studies دراسات</td>
</tr>
<tr>
<td>Motherland الوطن</td>
<td>Fibers الياف</td>
<td>Kingdom المملكة</td>
<td>Wheat القمح</td>
<td>Cotton القطن</td>
</tr>
<tr>
<td>Hectare هكتار</td>
<td>Lands الأراضي</td>
<td>Marketing تسويق</td>
<td>Increase زيادة</td>
<td>Pesticide مبيد</td>
</tr>
</tbody>
</table>
Table 8: Extracted Keywords from Agriculture Documents using TF-ITF method.

<table>
<thead>
<tr>
<th>Hectare (هكتار)</th>
<th>Lands (الأراضي)</th>
<th>General (العام)</th>
<th>Cultivation (زراعة)</th>
<th>Insects (الحشرات)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nile النيل</td>
<td>Domain مملكة</td>
<td>City مدينة</td>
<td>Food الغذاني</td>
<td>manna المanna</td>
</tr>
<tr>
<td>Storage خزن</td>
<td>Fibers الياف</td>
<td>Food الغذاء</td>
<td>Ton طن</td>
<td>Research بحوث</td>
</tr>
<tr>
<td>Water مياه</td>
<td>Alascbrbin الإسكربين</td>
<td>Kingdom المملكة</td>
<td>Production إنتاج</td>
<td>Department قسم</td>
</tr>
<tr>
<td>Hands أيدي</td>
<td>Tubers الدميات</td>
<td>Palm النخلة</td>
<td>States الدول</td>
<td>Bank صفة</td>
</tr>
<tr>
<td>Hasbani الحصبياني</td>
<td>Heat الحرارة</td>
<td>Region منطقة</td>
<td>Prices أسعار</td>
<td>Pests الأناتج</td>
</tr>
<tr>
<td>Tigris ذلقة</td>
<td>Production الإنتاج</td>
<td>Marketing تسويق</td>
<td>Rise ارتفاع</td>
<td>Studies دراسات</td>
</tr>
<tr>
<td>And Euphrates والفرات</td>
<td>Crop المحصول</td>
<td>Varieties أصناف</td>
<td>Increase زيادة</td>
<td>Egypt مصر</td>
</tr>
<tr>
<td>Needs الاحتياجات</td>
<td>Heavy البذول</td>
<td>Palm النخلة</td>
<td>International العالمية</td>
<td>Various مختلط</td>
</tr>
<tr>
<td>Season فصل</td>
<td>Cultivation الزراعة</td>
<td>Dates التمور</td>
<td>Wheat الفم</td>
<td>Division تقسيم</td>
</tr>
</tbody>
</table>

Figure 5 shows the accuracy of the two methods to capture important concepts within specific domain. The $\chi^2$ method yields better results than TF-ITF. The $\chi^2$ method yields 64% accuracy while the TF-ITF method yields 52% accuracy. This percentage gives us a good indication about the capability of the system to capture important concepts from a set of Arabic documents that belongs to specific domain.
The experiments results give a positive indication about the ability of the system to extract significant keywords from Arabic documents, as well as the capability of the system to extract important terms that reflect the domain of each document.

**Conclusion and Future work**

This Study proposed an automated Arabic keyword extractor using the knowledge of the $\chi^2$ and a novel TF-ITF measures. The system can extract Arabic keywords simply from a single document without the need for a corpus. Also, the proposed system is very useful for extracting keywords that is related to a specific domain as demonstrated in table 7 as well as the system is also useful for extracting keywords that are domain independent. The $\chi^2$ measure is used to show the importance of the term by computing the degree of biasness between the term and the top frequent terms, and the novel TF-ITF method is effective in showing how the term is more important than other terms in the overall document.

Precision, recall and F-measure are used to judge both $\chi^2$ and TF-ITF methods, and the results show that the $\chi^2$ method is better than TF-ITF when dealing with small and large documents. The Precision, Recall and F-measure for the overall performance were 0.58, 0.63 and 0.61 respectively. Also, the two methods ($\chi^2$ and TF-ITF) showed their ability to extract keywords that are concepts in the domain that the documents belong to. The $\chi^2$ method was better than TF-ITF method with 64% accuracy.

The experimental results show the ability of the system to extract significant keywords from the Arabic documents and the ability of the $\chi^2$ to extract important terms that have low frequency. Both the $\chi^2$ and TF-ITF methods yield acceptable results for Arabic text keyword extraction, but the $\chi^2$ method is better than TF-ITF method.

As a future work, some techniques can implemented to enhance the ability of the system to extract significant words, a good stemming method for this purpose can be used to show the exact frequency of the word which appears in different patterns. Also, an improvement to this technique can be done by taking the position of the words into account.

Figure 5: Capability of $\chi^2$ and TF-ITF methods to Capture Important Concepts.
accounts; as known the words that appear in the introduction and conclusion or appear in the titles and subtitles may have a higher value than other words.

Finally, the drawbacks of the proposed system are summarized by two main points: how to extract the final keywords from the set of the candidate keywords, and the lack of Arabic documents that have its corresponding keywords, which force us to have human judgment for the system.

<table>
<thead>
<tr>
<th>Keyword extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>(term co-occurrence statistical information)</td>
</tr>
<tr>
<td>(co-occurrence matrix)</td>
</tr>
<tr>
<td>(biasness degree)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>The drawback is that our proposed system has not been tested with Arabic documents, and we need to have human judgment for the system.</td>
</tr>
</tbody>
</table>
Keywords Extraction Based On Word Co-occurrence Statistical Information for Arabic Text

References


