Enhanced Clustering-Based Topic Identification of Transcribed Arabic Broadcast News

Ahmed Jafar[1], Mohamed Fakhr[1], and Mohamed Farouk[2]
[1]Department of Computer Science, Arab Academy for Science and Technology, Egypt
[2]Department of Engineering Math and Physics, Faculty of Engineering, Egypt

Abstract: *This research presents an enhanced topic identification of transcribed Arabic broadcast news using clustering techniques. The enhancement includes applying new stemming technique "rule-based light stemming" to balance the negative effects of the stemming errors associated with light stemming and root-based stemming. New possibilistic-based clustering technique is also applied to evaluate the degree of membership that every transcribed document has in regard to every predefined topic, hence detecting documents causing topic confusions that negatively affect the accuracy of the topic-clustering process. The evaluation has showed that using rule-based light stemming in combination of spectral clustering technique achieved the highest accuracy, and this accuracy is further increased after excluding confusing documents.*

Keywords*: Arabic speech transcription, topic clustering.*

Received June 17, 2014; accepted January 27, 2015

1. Introduction

The growing amount of audible news broadcasted on TV channels, radio stations and on the Internet demands reliable and fast techniques to organize and store those vast amounts of news in order to facilitate future search and retrieval.

In our previous work [12] Automatic Speech Recognition (ASR) -a technology that converts spoken words to written text- is applied to audible Arabic news documents. Then a set of pre-processing and clustering techniques are applied on the transcribed documents in order to categorize them into a set of predefined topics.

Since the transcription process is normally highly erroneous [11, 12] and as an attempt to overcome some of these errors two pre-processing steps: words formatting and stemming were considered in [12] to evaluate their impact on limiting the negative impact of such errors. Two stemming techniques were selected: root-based and light stemming and both showed a notable improvement in the clustering accuracy of the picked-out algorithms with the superiority of the root-based stemming technique.

At the clustering stage, a similarity measure based on the Chi-square method [11] is utilized. This similarity measure is designed to eliminate non informative words (usually erroneous words when applied on transcribed documents). Two clustering techniques were utilized to achieve topic identification: k-means [12, 25] and spectral clustering [12, 19]. K-means was selected as a simple and fast traditional clustering algorithm. Spectral clustering was selected as it is one of popular, effective, and simple to implement modern clustering techniques.

The topic clustering accuracy was evaluated for the two selected clustering algorithms in three situations: when the transcribed documents are clustered without applying any stemming techniques, when light stemming is applied, and when root-based stemming is used. The results showed that spectral clustering in combination with root-based stemming yield the highest accuracy of 68.9%.

Since results were not completely satisfying, the need for achieving better results emerged. As a modification to the prior work, additional stemming techniques is used "rule-based light stemming" [13] which is a hybrid technique of the prior used stemming techniques; it makes use of the strength points of each of them while limiting their drawbacks. Also a possibilistic [16] version of the fuzzy clustering technique "Gustafson-Kessel (GK)" [10] is added to measure the degree of membership of each document to every topic cluster, hence all documents that don't belong vividly to one topic can be identified and scheduled for manual topic assignment.

This research is organized as follows: in section 2, speech transcription challenges are discussed. In section 3, data set pre-processing steps are discussed. In section 4, topic identification is discussed in details. In section 5, experimental results are evaluated and then discussed in section 6. The last section concludes the research.

2. Speech Transcription Challenges

The process of transcribing audible media to textual form using ASR system confronts many challenges that are typically not present in normal textual documents [11, 12]. The main challenges include: transcription errors, grammatical errors, and Out-Of-

Vocabulary problem (OOV), or combination of the previously mentioned problems.

The occurrence of such problems can seriously restrict the transcription process efficiency and hence restricts any further analysis applied on the transcripts. This work targets overcoming the problem of transcription errors as it is the most common problem when dealing with news transcripts [12].

The transcription errors occur due to limitation in the ASR system. The correction or elimination of such errors is a challenging task and requires understanding the nature of these errors. According to authors' observation, the transcription errors regarding Arabic language can be categorized into four sets:

- *Omission errors*: happen when the ASR fails completely to recognize a word or a series of consecutive words. In this case, the words are dropped out from the transcribed text and recognition process is continued. Omission errors are irrecoverable.
- *Word insertion errors*: occur when the ASR confuses word syllables with a separate word or multiple words. In this kind of errors, the original word is irrecoverable.
- *Misidentification errors*: identifying a pronounced word as a different word similar in pronunciation. The transcribed word may or may not belong to the valid Arabic vocabulary set.
- *Minor spelling errors*: a spoken word is identified correctly, but spelled wrong in transcription. These errors usually affect the way a word should be pronounced and it may affect its meaning as well. Common minor spelling errors generated by ASR are replacing the letter 'ه' with 'ة' at the end of the word and vice versa, replacing one of the following letters with one another 'ا', 'أ', 'إ' , and 'آ', and diacritics related errors.

3. Dataset Pre-Processing

As any written text documents, the transcribed documents produced by the ASR system are of unstructured format. Thus it is required to transform these unstructured documents to structured format using pre-processing in order to facilitate any further analysis applied on them. The following are the steps involved in the pre-processing applied in this work.

- *Tokenization*: the process of mapping sentences from character strings into strings of words. For example, the sentence " اللغة العربية تعد من أشهر اللغات" would be tokenized into "اللغة/", "العربية/", "تعد/", "من/", "أشهر/", "اللغات/".
- *Stop words removal*: Stop words are typical frequently occurring words that have little or no discriminating power, or other domain-independent words. Stop words removal can increase the effectiveness of the information retrieval process [2, 15], especially when dealing with large volume of text [22]. Stop words identified in this work include numbers, days and months names, prepositions, pronouns, and conjunctions.
- *Words Formatting*: An extra step applied in this work to unify all different shapes of the same letter to one form and also to remove some unwanted suffixes [12].
- *Stemming*: Removes the affixes in the words and produces the root word known as the stem. Typically, the stemming process will be performed so that the words are transformed into their root form. Automatic Arabic stemming is effective technique for text processing for small collections as in [4, 5] and large collections of documents as in [17, 18]. It also can enhance clustering as in [5]. Arabic stemmers are categorized as either root-based as in [6, 14] or stem-based (light stemmers) as in [17, 18]. Also the research for hybrid techniques, like the rule-based light stemming technique [13], has evolved to minimize the drawbacks associated with standard stemming techniques.
- *Weighted matrix construction*: the process of representing the text document into a machine readable form [21].

Besides transforming the unstructured transcribed text to structured form, the pre-processing is also used as the first phase to reduce the transcription errors by either correcting or help overcoming some of these errors. This happens during the pre-processing steps: words formatting and stemming. The following discuss how applying pre-processing can correct or help overcoming some of the transcription errors.

In Figures 1 and 2 erroneous words like "بدات", "امس", "اذا", "الفتره", "التجريبيه" and "ملاحقه" are examples of minor spelling errors. According to Arabic syntactic rules [3], the correct spelling for these words should be: "بدأت", "أمس", "إذا", "الفترة", "التجريبية" and "ملاحقة".

تتويجا يوفنتوس رسميا بلقب الدوري الإيطالي لكرة القدم للمرة الثانية على التوالي وذلك بعدما اللعب على ضيفه باليرمو
وأقدم داخلة فريق المدرب انطونيو كونتي المباراة وهو بحاجة الي نقطة 1 فقط لحس اللقب للمرة ثانية على التوالي والتاسعة والعشرون في تاريخه لكوني يتصدر الترتيب بفارق ١١ نقطة عن ملاحقة نابولي قبلة أربعة مراحل على انتهى الموسم

Figure 1. Sample of transcribed text with various transcription errors.

تتويج يوفنتوس رسم بلقب دور ابطال لكر قدم مر ثان توال لعب ضيف باليرم
اقدم داخل فريق مدرب انطوني كونت مبارا بحاج نقط فقط لحس لقب مر ثاني توال والتاسع عشر في تاريخ لكون يتصدر ترتيب بفارق نقط ملاحق نابول قبل اربع مراحل انق موسم

Figure 2. Sample of transcribed text after applying pre-processing with light stemming.

Such a problem is common in Arabic ASR systems and leaving it without handling would cause problems in any further analysis as in any computer system

words like "بدات" and "بدأت" aren't the same and actually will process them as two separate words. In this kind of errors, the correct spelling is determined on syntactic rules that depend in most cases on pronunciation. The correct pronunciations depends on the meaning of the word which is determined according to its context, thus the only way to detect and correct such errors is searching massive dictionary of correctly spelled Arabic words and searching for the correct syntax if the word is misspelled. It is possible for a word to take many correct forms in different contexts, hence automatically selecting the correct form needs understanding the word context. Such process is inefficient in terms of the processing time and power required as Arabic is very complex language, moreover, many existing machine learning techniques don't require understanding the language structure to operate on text, so it is not efficient to do such thing as a pre-processing step just to correct some errors. The most suitable solution is not to correct these errors but to work around them by unifying all different formats of a letter into one form. The unification process is performed at the words formatting step. Some suffixes are also removed at the word formatting step to fine-tune the input text for the stemming step.

Three stemming techniques are utilized in this work: light stemming represented in Larkey's light10 stemmer [18], root-based stemming represented in Khoja's root-based stemmer [14], and rule-based light stemmer introduced in [13]. The stemming techniques are applied in this work to unify vocabulary and also to overcome some transcription errors.

The words "تتويجا", "اللعب", "أقدم", "داخلة", "لحس", "لكونى", "قبلة", and "انتهى" in Figure 1 are examples of misidentification errors. The original words are "توج", "تغلب", "قد", "دخل", "لحسم", "لكونه", "قبل", and "انتهاء". If a word is misidentified into one of its relative forms, so there is a good chance that this mistake would be overcame when root-based stemming is applied. The light and rule-based stemmers would either transform the word to another form or leave it without any transformation in some cases [12].

Both light and rule-based stemmers actually tend to correct the error in case of occurrence of inserted letters in the start and the end of the word as long as the inserted letters exist on their prefix/suffix removal list [12]. The root-based stemmer can also correct an erroneous word if by chance the original word is the same as the root of erroneous word. The word "تتويجا" (Figure 1) is a good example, after removing the suffix 'ا' at the word formatting step the result is "تتويج" which would be transformed by the root-based stemmer to "توج" as shown in Figure 3 which is also the same as the original word in spelling.

توج يوفنتوس رسم لقب دور ايطالي كور قدم مرر ثاني ولي لعب ضيف باليرم
قدم دخل فرق درب انطوني كونتي بر ا ح وج نقط فقط لحس لقب مر ر ثاني ولي تاسع عشرون ارخ
كون صدر رتب فرق نقط لحق نابولي قبل اربعه رحل نهي وسم

Figure 3. Sample of transcribed text after applying pre-processing with root-based stemming.

All three stemming techniques fail when the erroneous word is substituted by completely another word of a different spelling and meaning like "اللعب", "أقدم", and "لحس" as in Figures 2, 3, and 4 or if by chance a letter is inserted in the middle of the word.

تتويج يوفنتوس رسم لقب دور إيطال كرة قدم مره ثان وال لعب ضيف اليرم
أقدم داخل فريق مدرب انطوني كونت مباراه حاج نقط فقط لحس لقب مره ثان توال التاسع عشر تاريخ
كون يتصدر الترتيب بفارق نقطه عن ملاحق نابول قبل اربعه مراحل على انته موسم

Figure 4. Sample of transcribed text after applying pre-processing with root-based stemming.

If such erroneous words are not repeated frequently along the whole set of documents, they would probably have poor information contribution, and hence they would be eliminated by the chi-square based similarity measure if their information contribution assessment doesn't comply with a certain threshold, otherwise they would be retained.

After applying stop words removal, words formatting and stemming, and in order to process the transcribed documents for topic identification, they must be represented in a machine readable form. Vector-Space Model (VSM) the model applied in this work because of its effectiveness in proximity estimation between text documents in addition to its conceptual simplicity [21].

In VSM all documents are represented as vectors of weights in an n-dimensional space of terms. At the recent time, there are a number of well-known methods that have been developed to evaluate term weight [24], in this work, Okapi method [20] is applied, which is a modification of a classic Term Frequency×Inverse Document Frequency (TFIDF) weighting scheme and proved to be efficient in a number of applications [1, 7].

According to the Okapi method Combined Weight (CW) of the word is calculated as in Equation 1.

$$CW(w_i|D_j) = \frac{(K+1)\times CFW(w_i)\times TF(w_i,D_j)}{K\times((1-b)\times +b\times NDL(D_j)+TF(w_i,D_j)} \quad (1)$$

The quantity $CFW(w_i) = log\frac{N}{n(w_i)}$ is the data set collection frequency weight; N is the total number of documents in the collection and $n(w_i)$ is the number of documents containing the word w_i. The quantity $TF(w_i, D_j)$ is the frequency of occurrences of word w_i in the document D_j and $NDL(D_j)$ is the length of the document D_j normalized by the mean document length. The constant b controls the influence of document length and is empirically determined to the value 0.75. The other constant K acts as a discounting parameter on the word frequency: when K is 0, the

combined weight reduces to the collection frequency weight; as *K* increases the combined weight approaches *tf*×*itf*. *K* is set to 1.25 in this work.

Once *CW* is calculated for all words, it is easy to calculate the Weight of a Document (DW) the same way. *DW* can be calculated for the whole document or any of its parts via applying in Equation 2.

$$DW(D_i) = \sum_{w_k \in X_i} CW(w_k) \quad (2)$$

4. Topic Identification

Topic identification is the process of assigning one or more labels to text documents chosen from a pre-defined list of topics assuming that each documents is topically homogeneous (e.g., a single news story). There exist two general styles of topic identification: topic classification/clustering and topic detection. In topic classification, it is assumed that a predetermined set of topics has been defined. In topic clustering the topics set may be predefined or open for discovery. In both scenarios each document will be categorized as belonging to one and only one topic from the topics set. This style is sometimes referred to as single-label categorization. In topic detection, it is assumed that a document can relate to any number of topics and an independent decision is made to detect the presence or absence of each topic of interest. This style is sometimes referred to as multi-label categorization.

In this work topic identification is achieved by means of clustering of data using a similarity measure, and the labels are chosen from a pre-defined list of topics. A Chi-square based similarity measure is used along with k-means and spectral clustering algorithms.

The Chi-square similarity measure determine the word co-occurrences between matching transcripts by sorting all words in transcripts by their weights and retain only those whose weights are greater than some empirically preset threshold. Thus non-informative words including low frequently repeated erroneous words should appear at the bottom of the sorted list and hence eliminated according to the empirically determined threshold. The Chi-square similarity is calculated as in Equation 3.

$$sim(Inter(D_i, D_j) = \sigma \times Inter(D_i, D_j), \quad (3)$$

Where σ is given by evaluating the Chi-square test in Equation 4 and $Inter(D_i, D_j)$ is given by Equation 5. The calculated similarity will range between 0 and 1 and it will be equal to 1 if and only if $D_i = D_j$.

$$X^2 = \sum_{w_k \in Di \cap Dj} \frac{(CW(w_k \mid w_k \in D_i) - CW(w_k \mid w_k \in D_j))^2}{CW(w_k \mid w_k \in D_j)} \quad (4)$$

$$Inter(D_i, D_j) = \frac{DW(D_i \cap D_j)}{DW(D_i)}, \quad (5)$$

Such that in case of $D_i \neq D_j$ is true, then the inequality in Equation 6 is also true.

$$Inter(D_i, D_j) \neq Inter(D_j, D_i) \quad (6)$$

The widely used cosine similarity in Equation 7 measure is also used. The clustering accuracy is, then, compared to the accuracy achieved via using the Chi-square measure.

$$S_c(D_i, D_j) = \frac{\sum_{k=1}^{N} (w_{ki} \times w_{kj})}{\sqrt{\sum_{k=1}^{N} w_{ki}^2 \times \sum_{k=1}^{N} w_{kj}^2}} \quad (7)$$

For Topic identification process, two clustering methods are used: hard clustering and possibilistic clustering.

4.1. Hard Clustering

Hard clustering means partitioning the data into a number of subsets (clusters) such that an object either belong or doesn't belong to a cluster. Two hard clustering techniques are utilized in this work: *k*-Means, spectral clustering.

K-means is based on the idea that a center point (centroid) can represent a cluster. It is one of the most popular traditional data clustering algorithms because of its simplicity and computational efficiency. The main problem with this clustering method is its tendency to converge at a local minimum and the final results highly depends on the initial choices of centroids.

Spectral clustering reformulation of the clustering process takes place using a similarity graph G=(V, E) where the goal is to find a partition of the graph such that the edges between different groups have very low weights, and the edges within a group have high weights. The similarity graph used in this work is the fully connected graph because the Chi-square similarity measure itself models local neighborhoods, so it best suite this kind of graphs as described in [12, 23].

4.2. Possibilistic Clustering

Besides the fact that the transcribed data are being erroneous, there is also a possible chance of the existence of topic overlaps and/or noisy documents that don't belong to any predefined topic, which limits the effort to correctly cluster such data into topics using hard clustering techniques. Thus the need emerged for a clustering method that allows a document to belong to more than one cluster simultaneously with different membership degrees. Fuzzy clustering method [26] allows object memberships satisfying the following constraints:

$$\mu_{ij} \in [0,1], \forall_i, \forall_j \quad (8)$$

$$0 < \sum_{j=1}^{N} \mu_{ij} < N, \forall_i, and \quad (9)$$

$$\sum_{i=1}^{C} \mu_{ij} = 1, \forall j. \quad (10)$$

The parameter μ_{ij} is the degree of membership of the feature point x in cluster β_i. C denotes the number of classes, and N denotes the total number of feature points.

Because of the restriction in Equation 10, the generated memberships degree don't always correspond well to the actual degree of belonging of the data, thus it is difficult to detect outliers in a noisy environment using fuzzy clustering. As a solution to this problem, possiblistic clustering is used, in which the restriction in Equation 10 is relaxed [6] as in Equation 11. A possiblistic version of the fuzzy GK algorithm is utilized in this work.

$$max_i \, \mu_{ij} > 0, \forall_j \quad (11)$$

GK extended the standard fuzzy c-means algorithm [8] by employing an adaptive distance norm, in which each cluster has its own norm-inducing matrix A_i, which yields the inner-product norm in Equation 12. The choice of the norm-inducing matrix A determines the cluster shape; hence the employing of adaptive distance norm adds the capability of detecting clusters of different geometrical shapes [10].

$$d_{ij}^2 = || x_j - c_i ||^2 \, A_i = (x_j - c_i)^T A_i (x_j - c_i), \quad (12)$$

Where d_{ij}^2 is the distance feature point x_j to cluster center c_i.

The possibilistic version of the GK algorithm is derived from the general form of possiblistic algorithms introduced in Equation 6. The possibilistic GK algorithm is explained in [16].

5. Experimental Results

The dataset used in this research consists of audio news stories collected and recorded manually from various Arabic news broadcast networks: Al-Jazeera, Al-Arabiya, and BBC Arabic. The dataset size is about 30 hours of recorded Arabic news stories. The average length of the news story is two minutes. The news stories are transcribed generating 1000 text files divided into five topics: culture and arts, economics, politics, science, and sports.

The reason behind the manual selection of the news stories is to minimize speaker related problems, like unclear pronunciation and grammatical errors. The collected news are then transcribed into text documents using ASR system "Dragon Dictation" [9], and then pre-processed for topic-clustering.

After applying pre-processing steps on the documents and performing clustering techniques, the accuracy of clustering is evaluated using F-Measure in Equation 13, a measure that combines the recall and precision ideas from information retrieval [20].

$$F = \sum_i \frac{n_i}{n} Max\{F(i,j)\} \quad (13)$$

The *max* is taken over all clusters at all levels, and n is the number of documents and $F(i,j)$ is defined by:

$$F(i,j) = \frac{2 \times Recall(i,j) \times Precesion(i,j)}{Recall(i,j) + Precesion(i,j)} \quad (14)$$

$$Recall(i,j) = n_{ij} / n_i \quad (15)$$

$$Precision(i,j) = n_{ij} / n_j \quad (16)$$

The quantities *Recall* and *Precision* are calculated as in Equations 15 and 16, where n_{ij} is the number of members of class i in cluster j, n_i is the number of members of class i, and n_j is the number of members of cluster j.

The dataset is divided into subsets of sizes ranged from 50 to 200 documents per category. Experiments are carried out on each of these subsets four times for each clustering algorithm: when no stemming is applied, when light-stemming is applied, when root-based stemming is applied, and finally when rule-based light stemming is applied. Each clustering algorithm is run twice: one time with the use of the Chi-square similarity measure, and the other time with the use of the popular cosine measure. The accuracy of the clustering is evaluated for each subset, and then the average accuracy is calculated among all the subsets shown in Table 1.

Table 1. Accuracy evaluation of the topic clustering of the transcribed documents using hard clustering methods.

Clustering Approach/Similarity Measure	Average Accuracy			
	Non-Stemmed	Light-Stemmed	Root-Stemmed	Rule-Stemmed
k-Means /Cosine	39.42%	44.61%	54.41%	60.04%
k-Means/Chi-square	44.3%	47.6%	56.5%	63.35%
Spectral Clustering/Cosine	45.62%	50.96%	65.57%	71.33%
Spectral Clustering/Chi-square	46.5%	53.8%	68.9%	76.11%

The dataset is divided into subsets to consider the effect of the change in dataset size and hence the change in the amount of information contained in each subset on the average clustering accuracy. The reason why clustering is first applied on non stemmed data is to measure the impact of applying stemming techniques on improving the accuracy of the clustering algorithms operating on such erroneous data. The use of two similarity measures is to ensure that the Chi-square measure makes a positive effect on clustering the erroneous data into topics.

The experimental scenarios applied on the transcribed news documents are also applied on the error-free version of those transcribed documents, and then the average accuracy is calculated as shown in Table 2. This step is carried on to gain knowledge about the sensitivity of the original data to clustering,

thus an assessment to what extent the techniques proposed in this work have managed to overcome transcription errors can be performed.

Table 2. Accuracy evaluation of the topic clustering of the error-free version of the transcribed documents using hard clustering methods.

Clustering Approach/Similarity Measure	Average Accuracy			
	Non-Stemmed	Light-Stemmed	Root-Stemmed	Rule-Stemmed
k-Means /Cosine	62.2%	64.63%	68.06%	76.84%
k-Means/Chi-square	65.9%	67.97%	72.84%	79.05%
Spectral Clustering/Cosine	72.2%	74.97%	80.77%	85.15%
Spectral Clustering/Chi-square	74.87%	76.85%	82.74%	87.21%

By comparing the accuracy results in Tables 1 and 2, and by observing the clustering confusion matrix for each clustering scenario for both original and transcribed date, it is concluded that in both sets of data, there are documents causing clustering confusion. The existence of topic overlaps in the original data is the main cause of such confusion. The information loss due to the transcription errors is increasing the confusion even more in the transcribed data.

In the next phase of experiments possibilistic GK algorithm is applied on both the transcribed and the original data, and the membership matrix is analyzed to evaluate the amount of confusing documents in each topic as in Figures 5 and 6.

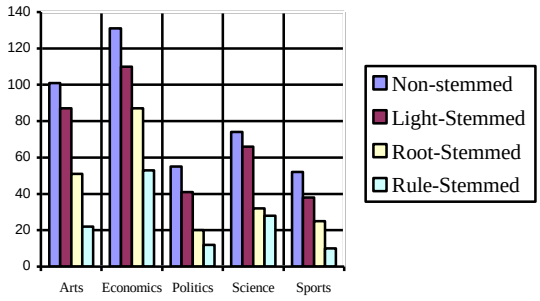

Figure 5. Confusing documents detected after applying the possibilistic GK algorithm on the transcribed data.

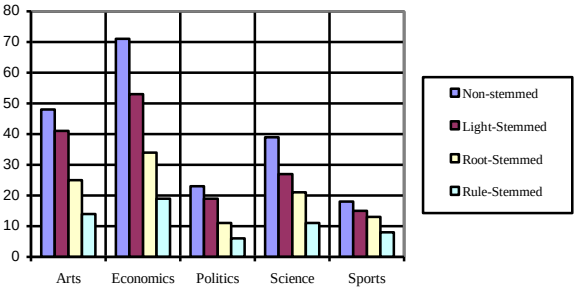

Figure 6. Confusing documents detected after applying the possibilistic GK algorithm on the original data.

A document is considered confusing to the clustering process if its membership degrees to all clusters are under a certain predefined threshold, or if its membership degrees to more than one cluster are convergent. By determining which documents are affecting the clustering accuracy, they can be excluded and the rest of the documents are maintained. Doing such exclusion, would improve the clustering accuracy for the rest of the documents. After re-applying the experiential scenarios on the remaining data on both transcribed and original data, the average clustering accuracy improved to a maximum of 85.62% and 92.26% respectively when spectral clustering is used on rule-based stemmed data. Manual categorization can be considered a solution to categorize the excluded documents.

6. Discussion

The results have showed that stemming techniques have improved the accuracy of all clustering algorithms. Rule-based light stemming has improved the efficiency of the clustering process more than the other stemming techniques. The reason behind that is the nature of the Arabic language and its related transcription errors. Light-stemming techniques only removes certain prefixes and suffixes which will not truly and effectively transform all similar words to one root, hence limit its ability to overcome misidentification errors. Light stemming is also known for causing high mis-stemming and under-stemming errors. Mis-stemming occurs when an original part of the word is confused with an affix, and hence is removed. Under-stemming occurs when stemming two word with the same root, but instead, the stemmer produce two different roots. In contrast to light stemming, root-based stemming transforms all similar words to one root, except for some limitations that may exist in the algorithm performing the transformation, which makes it more efficient than light-stemming in overcoming errors. Root-based stemming causes high over-stemming errors where two words with different roots are transformed to one root. Rule-based light stemming has more ability to overcome transcription errors than light stemming, but less ability than root-based stemming. It also has the benefit of balancing between the stemming errors caused by the light and root-based stemming techniques; hence it leads to the best performance in the clustering phase. The ability to balance between stemming errors can be explained because of the fact that rule-based light stemming technique is basically a light stemming technique guided by rules that distinguish whether an identified affix is originally part of the word or not. It also has the ability to transform plural words into its singular form before transforming it to its stem. For those reasons, mis-stemming and under- stemming errors are reduced. Additionally, it doesn't cause much over-stemming

errors because it is basically a light stemmer and as mentioned earlier, light stemmers are not known for causing much of this kind of errors.

The spectral clustering algorithm achieved more accuracy than the *k*-means algorithm in all cases which may be explained due to the nature of the data set. In contrast to spectral clustering, *k*-means tends to perform best in linearly separable data. Since the topics chosen are general and limited in number, thus the chance of existence of cross topic documents is increased and therefore the process of linearly separating data becomes more difficult.

The results also have showed that Chi-square similarity method has showed superiority over the popular and traditional cosine similarity and it is best utilized by the spectral clustering algorithm.

Applying the possibilistic GK algorithm on both the transcribed and original data has revealed some of the characteristics of the data. By analyzing the membership matrix and manual observation of the detected confusing documents, it is notable that the Economics topic has the biggest number of confusing documents; this may be explained due to the numerical-based nature of the content involved in this topic. Thus considering that the news stories are relatively short, it is hard to extract unique features that can qualify such document to be assigned vividly to a topic. Arts and Science have the second and third places in the number of occurrences of confusing documents. This may be explained to the possibility of existence of sub-topics within them, in addition to the relatively small size of the dataset that doesn't cover all of those sub-topics well. Although Politics topic has the least confusing documents, it is the most topic that received wrong-clustered documents from all other categories. This may be explained due to the interference of politics in many aspects of life, so it is possible for an economical decision to be based on some political background and both are melded into one news story.

The number of confusing documents is increased in the transcribed documents due to transcription errors, especially when such errors occur frequently in named entities (names of persons, organizations, places, etc) which usually represent important features for guiding the clustering process. It is also notable that stemming played an important role in reducing the amount of confusing documents in all topics of the transcribed and original data.

7. Conclusions

In this research a set of transcribed textual documents obtained from a set of spoken documents are clustered into topics, and the impact of applying three stemming techniques along with the Chi-square and cosine similarity measures on the accuracy of the topic-clustering process is measured. The confusion nature of the transcribed data is investigated and compared to its original correct form by the use of a possibilistic version of the GK fuzzy algorithm which showed that possibilistic clustering is suitable for this kind of erroneous confusing data as it can detect those confusing members and hence give the option for excluding them.

The clustering accuracy evaluation showed that the best hard clustering results are achieved by applying the spectral clustering algorithm in combination with rule-based stemming scoring average accuracy of 76.11% which is higher than the results obtained in the prior work. This accuracy is further improved to 85.62% by excluding the confusing document.

References

[1] Abberley D., Renals S., and Cook G., "Retrieval of Broadcast News Documents with the THISL System," *in Proceeding of IEEE International Conference on Acoustic, Speech, and Signal Processing*, Washington, pp. 3781-3784, 1998.

[2] Abu El-Khair I., "Effects of Stop Words Elimination for Arabic Information Retrieval: a Comparative Study," *International Journal of Computing and Information Sciences*, vol. 4, no. 3, pp. 119-133, 2006.

[3] Al-Fares W., *Arabic Root-Based Clustering: an Algorithm for Identifying Roots Based on N-Grams and Morphological Similarity*, Thesis PHD, University of Essex, 2002.

[4] Al-Kharashi I. and Evens M., "Comparing Words, Stems, and Roots as Index Terms in an Arabic Information Retrieval System," *Journal of the American Society for Information Science*, vol. 45, no. 8, pp. 548-560, 1994.

[5] Al-Shammari E. and Lin J., "A Novel Arabic Lemmatization Algorithm," *in Proceeding of 2nd workshop on Analytics for Noisy Unstructured Text Data*, New York, pp. 113-118, 2008.

[6] Awde N. and Samano P., *The Arabic Alphabet: How to Read and Write It*, Lyle Stuart, 2000.

[7] Coden A. and Brown E., "Speech Transcript Analysis for Automatic Search," *in Proceeding of 34th Annual Hawaii International Conference*, Washington, pp. 9-12, 2001.

[8] Dave R., "Boundary Detection through Fuzzy Clustering," *in Proceeding of IEEE International Conference on Fuzzy Systems*, California, pp. 127-134, 1992.

[9] Dragon Dictation App home page on iTunes store, https://itunes.apple.com/us/app/dragon-dictation/id341446764?mt=8, Last visited 2014.

[10] Gustafson D. and Kessel W., "Fuzzy Clustering with a Fuzzy Covariance Matrix," *in Proceeding of IEEE CDC*, California, pp. 761-766, 1979.
[11] Ibrahimov O., Sethi I., and Dimitrova N, "A novel Similarity based Clustering algorithm for Grouping Broadcast News," *in Proceeding of SPIE Conference Data Mining and Knowledge Discovery: Theory, Tools, and Technology IV*, Orlando, pp. 294-304 2002.
[12] Jafar A., Fakhr M., and Hesham M., "Clustering-Based Topic Identification of Transcribed Arabic Broadcast News," *in Proceeding of 9th International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering*, New York, pp. 253-260, 2015.
[13] Kanaan G., Al-Shalabi R., Ababneh M., Al-Nobani A., "Building an Effective Rule-Based Light Stemmer for Arabic Language to Improve Search Effectiveness," *The International Arab Journal of Information Technology*, vol. 9, no. 4, pp. 368-372, 2012.
[14] Khoja S. and Garside R., *Stemming Arabic text*, Lancaster, UK, Computing Department, Lancaster University, 1999.
[15] Korfhage R., *Information Storage and Retrieval*, John Wiley, 1997.
[16] Krishnapuram R. and Keller J., "A Possibilistic Approach to Clustering," *IEEE Transactions on Fuzzy Systems*. vol. 1, no. 2, pp. 98-110, 1993.
[17] Larkey L., Ballesteros L., and Connell M., "Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-Occurrence Analysis," *in Proceeding of 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Finland, pp. 275-282, 2002.
[18] Larkey L. and Connell M., "Arabic information retrieval at UMass in TREC-10," *in Proceeding of Tenth Text REtrieval Conference (TREC-10)*, pp. 562-570, 2001.
[19] Luxburg U., "A Tutorial on Spectral Clustering," *Springer Statistics and Computing*, vol. 17, no. 4, pp. 395-416, 2007.
[20] Robertson S., Walker S., Jones S., Hancock-Beaulieu M., and Gatford M. "Okapi at TREC-3," *in Proceeding of 3rd Text REtrieval Conference*, New York, pp. 109-126, 1994.
[21] Salton G., *Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, 1989.
[22] Schauble P., *Multimedia Information Retrieval: Content-Based Information Retrieval from Large Text and Audio Databases*, Kluwer Academic Publishers, 1997.
[23] Shi J. and Malik J., "Normalized Cuts and Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888-905, 2000.
[24] Singler M., Jin R., and Hauptmann A., "CMU Spoken Document Retrieval in Trec-8: Analysis of the Role of Term Frequency TF," *in Proceeding of 8th Text REtrieval Conference*, Gaithersburg, pp. 1-10, 1999.
[25] Steinbach M., Karypis G., and Kumar V., "A Comparison of Document Clustering Techniques," *KDD Workshop on Text Mining*, University of Minnesota, 2000.
[26] Yang M., "A Survey of Fuzzy Clustering," *Mathematical and Computer Modelling journal*, vol. 18, no. 11, pp. 1-16, 1993.

Ahmed Jafar obtained his B.Sc in Computer Science from Faculty of Information systems and Computer Science, October 6 University, Egypt in July 2006. He received his MS.c. Degree in Computer Science from College of Computing and Information Technology, Arab Academy for Science and Technology and Maritime Transport, Cairo, Egypt in September 2014. He is currently a teaching assistant at the Faculty of Information systems and Computer Science, October 6 University.

Mohamed Fakhr received his Ph.D. in Electrical Engineering from Electrical and Computer Engineering department, University of Waterloo, Waterloo, Canada in May 1994. He is currently a full professor at College of Computing and Information Technology, Arab Academy for Science and Technology and Maritime Transport, Cairo, Egypt starting from September 2013 - Present. His fields of interest are Image Processing, Audio Processing, Pattern recognition, Sparse Coding, Sparse Recovery, and Machine Learning.

Mohamed Farouk was graduated from Electronics Engineering Dept., at Cairo University, Egypt on 1982. He had his M. Sc. And Ph.D. in Engineering physics on 1988 and 1994, respectively. Presently, he is a full professor of Engineering Physics at Cairo University. His research interests are in the areas of Acoustic Scattering and Speech Processing. He is also cross-appointed professor at faculty of Information Systems and Computer science, October 6 University.