Journal of Theoretical and Applied Information Technology
31st December 2015. Vol.82. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A BACKPROPAGATION NEURAL NETWORK TO IMPROVE ARABIC STEMMING

[1] KHALDOON MEZHER, [2] NAZLIA OMAR

Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science & Technology, University Kebangsaan Malaysia, Selangor, Malaysia

E-mail: [1]khaldon1.mzher1982@gmail.com, [2] nazlia@ukm.edu.my

ABSTRACT

Stemming aims to identify the roots or stem of each word by eliminating the grammatical affixes or lexical information. Many approaches have been proposed in terms of stemming for various languages. Unlike English language, Arabic stemming is not a simple task of eliminating suffixes and prefixes. Arabic has a complex morphology where the process of stemming requires comprehensive analysis for such morphology. Several approaches have been proposed to enhance Arabic stemming. However, the state of the art of Arabic stemming algorithms such as Khoja and Light stemmer were built using rule-based approach. Hence, recent researches have attempted to examine the machine learning technique regarding to its capability of identifying word's stem based on training. Nonetheless, there is still room for improvement in terms of addressing new features that have the ability to assign accurate weight for the words. In fact, assigning an accurate weight for the words plays an essential role in terms of classifying its actual root. Therefore, this study aims to develop a set of features that would enhance the process of weighting for words. The proposed set of features consists of affixes, word length and tenses. Then, the acquired weight that would be obtained from the proposed features will be entered as an input for a backpropagation neural network classifier in order to output the actual stem. The corpus that has been used in this study is an Open Source Arabic Corpus (OSAC), which consists of 21,861 documents that are distributed among multiple topics including economy, sport, social, religion, stories, health, law, astronomy and food recipes. Two stemmers i.e. Khoja and Light stemmers have been used as baseline for comparison with the proposed method. The experimental results have shown that the proposed method has outperformed both of Khoja and Light stemmers by achieving 88% precision and 82% recall. In addition, this study proposed a new sampling method in terms of Arabic stemming evaluation.

Keywords: *Arabic Stemming, Back-Propagation Neural Network, Feature Extraction, Arabic Morphology*

1. INTRODUCTION

Information retrieval (IR) is the process of gaining information resources that is relevant to a requested information from a collection of information resources [1]. The main aim of IR is to reduce what has been called "information overload". In fact, in many text data, there are plenty of unwanted words such as punctuation, stop-words and numeric data. Moreover, words in many languages can be formed with several morphological derivations such as plural or singular, feminine or masculine and verb tenses. In order to obtain meaningful information from an unstructured text data, there is a significant task called preprocessing where noisy and unwanted data have to be removed as well as, all the morphological derivations have to be eliminated. Apparently, each word will be replaced with its roots by eliminating the prefixes and suffixes, this process called Stemming.

Stemming is a common method that has been used in information retrieval to overcome the problem of mismatching vocabularies where query words do not correspond document words [2].

Unlike English and Indo-European languages, Arabic stemming is not a simple task of eliminating suffixes and prefixes [3]. Arabic has a complex morphology where the process of stemming requires comprehensive analysis for such morphology.

Several approaches have been proposed to enhance Arabic stemming. However, the state of the art of Arabic stemming algorithms such as Khoja [4] and Light stemmer [2] were built using rule-based approach. These approaches suffer of rule restrictions where the conditions that are not

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

mentioned in the rules would be incorrectly stemmed [5].

Hence, recent researches have attempted to examine the machine learning technique regarding to its capability of identifying word's class based on training such as the approaches that have been proposed by Alserhan & Ayesh [5] and Mesleh [6]. Nonetheless, there is still room for improvement in terms of addressing new features that have the ability to assign accurate weight for the words. In fact, assigning an accurate weight for the words plays an essential role in terms of classifying its actual root. Therefore, the aims of this paper is to introduce a set of features that would enhance the process of weighting for words. The proposed weighting approach will be fed into Back-propagation Neural Network (BPNN) in order to classify the word's root. The main advantage of Back-propagation neural network method is that it can fairly approximate a large class of functions.

2. RELATED WORK

There are many research efforts that have been proposed for Arabic stemming for instance, Chen et al. [7] have proposed an Arabic stemmer which is called Light stemmer using TREC 2002 cross lingual corpus. First, the authors have created a list for Arabic stop-words consisting of Arabic pronouns and prepositions. This has been performed by translating English stop-words list to Arabic using Google translator. Secondly, the authors have used an English stemmer in order to stem all the words in English. In fact, all the English words are linked with their corresponding in Arabic based on the translation. Hence, the English roots that have been stemmed already will be used for clustering Arabic words. For instance, the word 'children' will be stemmed by the English stemmer into 'child' so that, each word in Arabic contain this root 'child' will be grouped together. Thirdly, the authors have built an Arabic stemmer called Light stemmer which eliminates only prefixes and suffixes. They have identified two sets one for prefixes and the other for suffixes. These two set will be eliminated based on three aspects; (i) grammatical functions of the affixes, (ii) their occurrence frequencies among the Arabic words found in the Arabic document collection and (iii) the English translations of the affixes. The proposed method has obtained an 87.94% of f-measure.

In the same manner, Larkey [2] have developed several light stemmers based on heuristic and a statistical stemmer based on co-occurrence for Arabic retrieval using TREC-2001 Arabic corpus which contains 383,872 newspaper articles in Arabic from France Press Agency. The authors firstly have transformed the data by converting it to UTF-8 encoding in order to recognize Arabic letters. Then, they have normalized the data by removing stop-words, punctuation and numeric characters. Eventually, they have developed several light stemmers for Arabic which remove a small number of prefixes and suffixes and a co-occurrence based statistical stemmer which creates large stem classes by vowel removal and then refines these classes using co-occurrence.

Rogati et al. [8] have proposed an unsupervised machine learning method for Arabic stemming using a parallel corpus. First, the authors have used an English stemmer in order to stem the half of the words in the parallel corpus. This has been performed by building a translation model that links the English stems with its corresponding in Arabic. Then the authors have built a matrix based on the translation model in order to provide the translation probabilities. Apparently, once the translation model has been built, the authors have stemmed the Arabic portion of the parallel corpus by scoring all possible stems that an Arabic word would contain, the best one will be selected.

Alserhan & Ayesh [5] is one of the few research efforts that attempt to utilize machine learning techniques for Arabic stemming. In fact, a novel neural network based approach for stemming Arabic words is proposed in this paper. This has been done by exploiting numerical relations between characters by using backpropagation neural network (BPNN). The authors have used an arbitrary 100 words for testing. The experimental results have shown an enhancement compared to the rule-based stemming algoirthms. However, the wieghting technique that has been used in this study for each word, was not consider significant features such as affixes and tenses. In addition, this study has examined the stemming for words with three and four letters. Words with five and six letters have not been considered in the experiments.

Hadni et al. [9] have proposed a hybrid method in order to enhance Arabic stemming in text classification. The proposed method consists of Khoja stemmer, Light stemmer and N-gram with some adaption for Arabic language. The proposed algorithm starts with constructing the root file containing more than 9,000 valid Arabic roots taken from a dictionary of Arabic words, and constructing of the stop word file. The next step is the normalization of documents, after the removal of punctuation, diacritics and stop-word. The result of

Journal of Theoretical and Applied Information Technology
31st December 2015. Vol.82. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

this step is used as input in the process of removing prefix / suffix, by checking if the word match on of the patterns extract the relevant word, otherwise to remove the suffix and prefix respectively, with verification of the length after each removal of affixes. Finally the valid root is found by using the bi-gram and the Dice measure similarity.

Sembok & Ata [10] have proposed a rule-based approach in order to enhance Arabic stemming. The proposed rule-based approach consists of categorized groups of rules which are prefixes, suffixes and recoding. Prefixes and suffixes rules are adopted to recognize affixes in Arabic and identify its type. Whereas, the recoding module concentrates on changing some letters to their correct form. Such changes will probably occur during the process of template formation when a word is formed from a root.

3. PROPOSED METHOD

The proposed method as shown in Fig. 1 consists of five main phases. The phases are; (i) Corpus, (ii) Pre-processing, (iii) Word weighting, (iv) Back-Propagation Neural Network (BPNN), and finally (v) Evaluation. The corpus phase concentrates on the details of the dataset that has been used in this study by identifying its source, volume and type. Pre-processing phase aims to remove all the irrelevant data such as stop-words, numbers and punctuation. On other hand, Word's weighting phase aims to analyze each word based on a utilized set of features in order to assign the word with an accurate weight. Such features are consist of affixes, tenses and word's length. The process of weighting plays an essential role on the classification method by BPNN. Hence, the word's weight will be entered as an input into the BPNN classification phase. Finally, an evaluation method will be used in order to evaluate the proposed method. The evaluation consists of three steps; words list generation, unique words and partial matching.

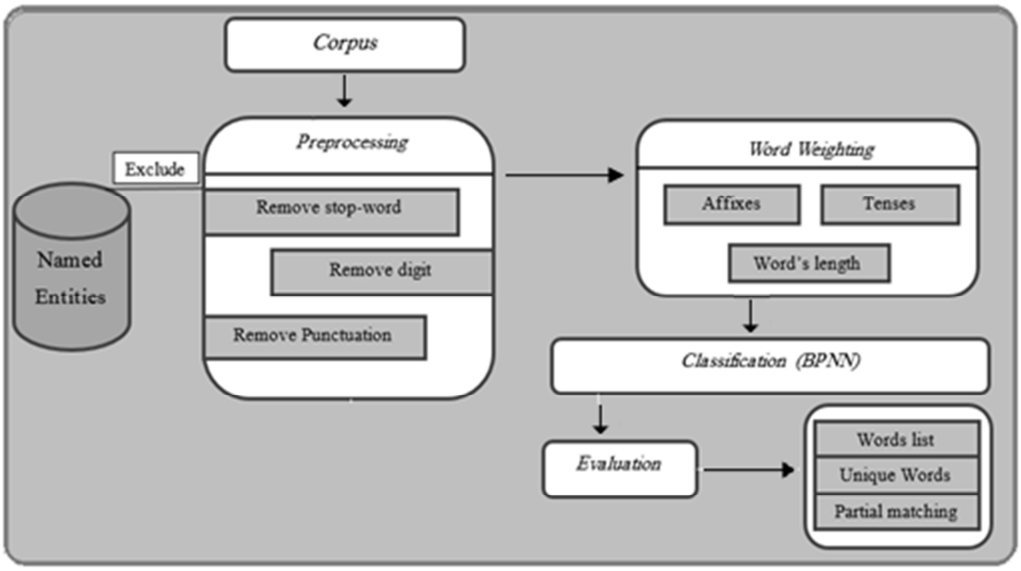

Figure.1. Architecture Of The Proposed Method

3.1. Corpus

The corpus that used in this research is an Open-Source Arabic Corpora (OSAC) which has been introduced by Saad & Ashour [11]. It contains 22,861 documents with size of 180 MB. Such corpus has been collected from several online resources such as BBC Arabic, CNN Arabic and Aljazeera newswire. The documents that have been collected are categorized into ten categories including Economy, History, Education and Family, Religion, Sport, Health, Astronomy, Law, Stories and Food recipes. Table 1 depicts the corpus details

Table 1. Corpus details

Category	No. of documents
Economy	3102
History	3233
Education and Family	3608

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Religion	3171
Sport	2419
Health	2296
Astronomy	557
Law	944
Stories	726
Food Recipes	2373
Total	22,861

3.2. Preprocessing

This phase aims to remove the irrelevant data such as numbers (*e.g.* 39402), punctuation (*e.g.* #$*%) and stop-words (*e.g.* *'من \ from', 'مع \ with', 'الى \ to', etc.*). However, there are other irrelevant data that has to be excluded from the stemming process. Such data can be represented by the named entities which are illustrated in the following sub-section.

3.2.1 Named entities

Named entities are the names of persons (e.g. عبدالله \ Abdullah), organizations (e.g. شركة زين \ Zain Corporation), locations (e.g. دبي \ Dubai) and dates (e.g. يناير \ January) [12]. Such entities should not be stemmed otherwise; the resulted word would be meaningless. Most of the stemming algorithms including Khoja and Light stemmers are incorrectly stemming named entities. Hence, this study aims to exclude the named entities before the process of learning by BPNN. The process of excluding such words has been performed by constructing a predefined list that contains the majority of Arabic named entities. For this purpose, Arabic Named Entity Recognition Corpus (ANER) [13] has been used in terms of enriching the proposed list. This corpus contains large volume of annotated named entities in Arabic language. The proposed named entity list has been used in order to contribute toward decreasing the error rate.

3.3. Word Weighting

This phase aims to assign an accurate weight for each word. Such weighting technique plays an essential role in terms of classification which will be held in the next phase of BPNN classifier. In order to obtain an accurate weight, there is a vital need to develop a set of features that could provide a comprehensive analysis for each word. These features are illustrated in the following sub-sections.

3.3.1 Affixes

This feature aims to analyze the words based on the embedded affixes including prefixes and suffixes. In Arabic, such affixes are associated with some factors such gender (e.g. feminine and masculine) and cardinality (e.g. single, dual and plural) therefore, each word will be assigned with its gender and cardinality.

As shown in Table 2, there are many kinds of prefixes and suffixes that could be located in Arabic words. However, this study concentrates on the type of affixes in which gender and cardinality has to be identified. For example, the word 'الذاهبات' which means 'the departed' has a suffix of 'ات'; such suffix is associated with feminine plural. Another example is the word 'يذهبون' which means 'they're going' has a suffix of 'ون'; such suffix is associated with the plural for masculine. Therefore, a predefined set of lists of affixes have been constructed in order to provide the type of each word.

3.3.2 Word Length

This feature aims to analyze the words based on the length. Basically, the length plays an essential role in terms of acquiring the actual stem [4]. Note that, this feature is mainly depending on the affixes. Without considering the affixes, many words will be incorrectly stemmed. For example, the word 'يبحثون \ yabhathoon' which means 'they're seeking' and 'مماثل \ momathel' which means 'symmetric' have the same number of letters which is five. However, the first word contains a prefix of 'ي \ ya' and a suffix of 'ون \ oon' so the process of stemming such word is represented by removing such affixes (i.e. first letter and two last letters) in order to get the root 'بحث \ bhath' which means 'seek'. Whereas, the second word does not contain affixes, but instead it has inflectional derivation so the process of stemming such word is represented by removing the first letter 'م \ mo' and third letter 'ا \ a' in order to get the stem 'مثل \ mthel' which means 'like'.

Table 2. Sample Of Affixes

Word	Transliteration	Translation	Prefix	Suffix
يذهبون	Yathhaboon	They're going	ي \ ya	ون \ oon
الذاهبات	Althahebat	The departed (plural for feminine)	ال \ al	ات \ at
الطالبان	Altaleban	The students (dual for masculine)	ال \ al	ان \ an
كالمعلمة	Kalmoalema	Like the teacher (feminine)	كال \ kal	ة \ a
بالمدرسين	Belmodareseen	Of the teachers (plural for masculine)	بال \ bel	ين \ een

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 3. Stemming Mechanism Based On Word Length

Word	Transliteration	Translation	Length	Determiner	Stem	Transliteration
اذهب	Ethhab	Go	4 letters	Prefix of (ا \ E)	ذهب	Thhab
جميل	Jameel	Beautiful	4 letters	Derivational inflection	جمل	Jaml
رائع	Raa'ea	Fabulous	4 letters	Derivational inflection	روع	Rawaa
كتبا	Kotoban	Books	4 letters	Suffix of (ا \ an)	كتب	Kotob

As shown in Table 3, the stemming process of multiple words with the same length is vary. The key characteristic behind this process lies on the determiner which check whether the word contain an affix or not. Such checking has the ability to identify the root of the word. For example, in case the word contain affix, the stemming process represented by removing such affix. In contrast, if the word does not contain affix (i.e. inflectional derived), specific process has to be took a place. Such specific process is related to the inflections that the word could be formed with it. For example, removing the middle letter such the case in the word 'جميل\ Jameel' which means 'beautiful' where the stemming represented by removing the middle letter of 'ي \ ee'. In addition, it could be replacement process of the two middle letters with one letter such in the case of the word 'رائع \ Raa'ea' which means 'fabulous' has been stemmed into the word 'روع \ rawaa'.

3.3.3 Tenses

In fact, this feature is relying on the previous two features where identifying the type of affix with the word length will facilitate the process of determining its tense. Like any languages, Arabic has multiple tenses in terms of the used verbs including past, present and organization. Identifying the tense of word has a significant role in terms of determining the actual stem of words. Table 4 shows a sample of words' tenses.

Table 4. Sample Of Word Tenses

Word	Translation	Tense
ذهب	He went	Past
ذهبت	She went	Past
يذهب	He's going	Present
تذهب	She's going	Present
اذهب	Go (order for masculine)	Organization
اذهبي	Go (order for feminine)	Organization

3.5. Back-Propagation Neural Network

Back-propagation Neural Network is the popular architectures of Artificial Neural Network. BPNN has been used widely in terms of text classification. In fact, BPNN consists of three layers; input layer, hidden layer and output layer [14]. Basically, the weighted word (from the features phase) will be entered as an input of the BPNN. Whereas, the hidden layer will construct the internal relations between input and output layers. Finally, the output layer will represent the resulted stem. The main advantage of Back-propagation neural network method is that it can fairly approximate a large class of functions [14].

In fact, Alserhan & Ayesh [5] have used a back-propagation neural network for Arabic stemming. The authors have encoded the Arabic letters into a binary code. Basically, the authors have encoded the letters based on the frequency of these letters where (ا ، و ، ي) letters have the most frequent occurred in Arabic, then (أ ، م ، ن ، ت) letters, and finally (ل ، ه ، س) letters respectively.

However, such distribution does not consider the probability of occurring as affixes for these letters. As mention earlier, determining that a given word contain an affix has the ability to identify its cardinality (plural, singular and dual), tense (past, present and organization) and the gender of words (feminine and masculine). As a hypothesis of this study, these features have a significant impact on clarifying the exact root of the word. Therefore, this study will manipulate the encoding of letters in order to consider the mentioned features. The letters that commonly used for prefixes and suffixes in Arabic language have been stored in a list, so that these letters will be encoded based on its occurring either as prefix or suffix. The proposed encoding can be represented in Table 5.

Table 5. Proposed Letters Encoding

Affix	Letters	Decimal code	Binary code
Prefix	ي ، يست ، ت ، است ، ن ، نست ، ال ، لل ، كال ، فال ، ك	4	100
Suffix	ي ، ية ، ك ، ه ، وا ، ت ، ات ، هم ، ها ، ان ، ون ، ين ، كما ، هما ، كن ، تن ، نا ، تا	3	011
Otherwise	-	2	010

As shown in Table 5, the distribution consider the occurring of affixes which means that it can

Journal of Theoretical and Applied Information Technology
31st December 2015. Vol.82. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

encode one or more letters in the same time such as the prefix '**يست**' which contains three letters. Note that, there are some redundant letters in both prefix and suffix such as the letter '**ك**', in this case the letter will be encoded based on its position whether in the beginning or in the ending of the word. After that, the encoding of letters will be input to the neural network, and then it will be processed in the hidden layer and finally extracted as an output. Note that, a reverse approach will be used in order to decode the output to obtain the root of the word. Figure 2 shows an example of the processing mechanism for the words using BPNN.

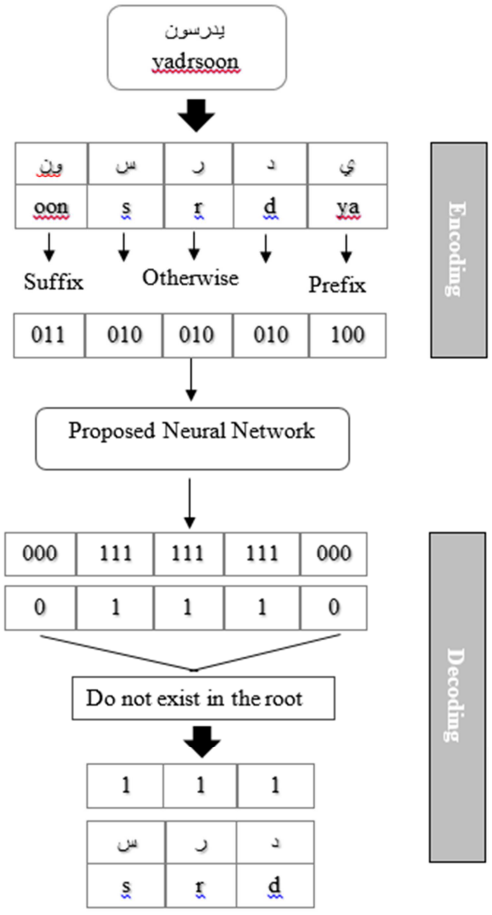

Figure 2. Example Of Encoding And Decoding

3.6 Evaluation

In order to evaluate the performance of the proposed method, a comparison has been established with the state of the art of stemming; Khoja and Light stemming algorithms. However, the evaluation of Arabic stemming is a challenging task due to the lack of available lexicon for the Arabic roots with its corresponding stems [15]. Usually, some of the literature tend to use the TF-IDF in order to evaluate the words with highest occurrences. The other tend to select an arbitrary number of words in order to evaluate. Yet, these mechanisms do not provide an overall aspect for all the words that located in the used corpus. This is due to the large number of words that could be ignored or avoided in the process of evaluation. Therefore, this study aims to construct a new sampling method that has the ability to provide an overall view for the majority of the Arabic words. Such sampling method can contribute toward enhancing the process of evaluating Arabic stemming algorithm by other researchers. Following sub-sections describe the tasks that have been performed to construct the proposed sampling method.

3.6.1 Unigram list Generation

This task aims to separate the documents and included sentences in the corpus into series of unigram word where each word can be treated independently. This task is crucial and aims to turn the data into an internal representation.

3.6.2 Unique words

This task aims to remove all the redundant words in order to reduce the dimensionality and keep the unique words. Such reduction process facilitates the time consuming especially when carry out the stemming algorithm on a large corpora. Note that, there are redundant words but with affixes such as '**اذهب** \ *go*' and '**ذهب** \ *went*', these words cannot be removed because they have different morphology thus, the full match is cannot recognize the redundancy. Therefore, the next task 'partial matching' can handle this problem.

3.6.3 Partial (Approximate) matching

As mention earlier, there are redundant words but with some affixes. The process of removing the redundant words requires a full matching between the words. Therefore, this task aims to apply an approximate matching using Cosine similarity measure in order to identify the similarity between the words which can be an indication for redundancy. The reason behind removing the words with approximate matching because they share a same root. Cosine similarity is a distance-based approach which aims to provide a real value that implies the similarity between two words. In order to calculate the similarity between two words *s* and *r*, following equation will be used [16]

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$Cosine\ (s,r) = \frac{s.r}{|s|.|r|} \quad (1)$$

where: s and r are two words that have an approximate match (i.e. they have the same root). For example, the two words in Table 4.3 'للحقول \ lelhokool' and 'حقول \ hokool' which mean 'for fields' and 'fields' are similar to each other and share the same root. Therefore, by applying cosine similarity upon these words multiple tasks will be performed as follows:

- Combine both words as 'للحقولحقول \ lelhokoolhokool'.
- Keep the unique letters from the combination as 'لحقو \ lehko'.
- Create vectors based on the frequency using the unique letter 'لحقو \ lehko' as follows:
  a. $V_1 = \{3,1,1,1\}$ (frequency of unique letters in the first word)
  b. $V_2 = \{1,1,1,1\}$ (frequency of unique letter in the second word.
- Apply the equation of Cosine similarity as follow:

$$Cosine\ (\text{للحقول}, \text{حقول}) = \frac{V_1 \times V_2}{\sqrt{V_1^2} \times \sqrt{V_2^2}}$$

$$\frac{3*1+1*1+1*1+1*1}{(3^2+1^2+1^2+1^2)\times(1^2+1^2+1^2+1^2)} = 0.867$$

3.6.4 Word annotation

This task aims to carry out the stemming algorithm on the resulted word that obtained from the above tasks. Then, annotating each word with its class of correctness.

Moral et al. [17] have identified two kinds of stemming errors; under-stemming and over-stemming errors. Under-stemming errors are the error that occurred when the algorithm does not make enough terminations. For example, if the word '**الطلاب \ altolab**' has been stemmed into '**طلاب \ tolab**', it is considered as an under-stemming error because there is still one more procedure to turn '**طلاب \ tolab**' into '**طلب \ tlb**.

Whereas, over-stemming errors are the errors that occurred when the stemming algorithm is performing an over termination process that could affect the morphology or semantic of the word. For example, if the word '**ذهب \ thahb**' has been stemmed into '**هب \ hb**', it is considered as an over-stemming error because the word has been transformed into meaningless word.

Now the process of annotating the resulted words aims to assign 1 for the correct stem, 0 for the over-stemming stems and 2 for the under-stemming stems. Therefore, each word has been annotated with one of the mentioned values. This has been performed by expertise in Arabic language.

Now, it is possible to apply the information retrieval metrics; precision and recall in order to evaluate each stemming algorithm where the equations can be calculated as follows:

$$Precision = \frac{CS}{CS + USE} \quad (2)$$

Where CS is the number of correct stems and USE is the number of under-stemming errors.

$$Recall = \frac{CS}{CS + OSE} \quad (3)$$

Where CS is the number of correct stems and OSE is the number of over-stemming errors.

4. RESULTS

In this section, the proposed method will be evaluated based on the performance. For this purpose, both state of the art Light stemmer [2] and Khoja stemmer [4] have been applied on the resulted sample with the proposed method as well. Table 6 depicts such results.

Journal of Theoretical and Applied Information Technology
31st December 2015. Vol.82. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 6. Results Of Light, Khoja And The Proposed Stemming Algorithms

	Light stemmer		Khoja stemmer		Proposed BPNN	
Document	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
Economy	0.653	0.747	0.823	0.745	0.886	0.842
History	0.621	0.701	0.779	0.712	0.879	0.801
Family	0.596	0.637	0.714	0.687	0.814	0.762
Religion	0.659	0.776	0.852	0.751	0.903	0.812
Sport	0.642	0.719	0.796	0.734	0.837	0.791
Health	0.694	0.742	0.823	0.786	0.877	0.856
Astrology	0.712	0.801	0.878	0.801	0.930	0.874
Laws	0.711	0.802	0.876	0.803	0.920	0.875
Stories	0.722	0.793	0.869	0.813	0.903	0.863
Food recipe	0.621	0.727	0.804	0.712	0.880	0.824
Average	0.663	0.745	0.821	0.754	0.884	0.829

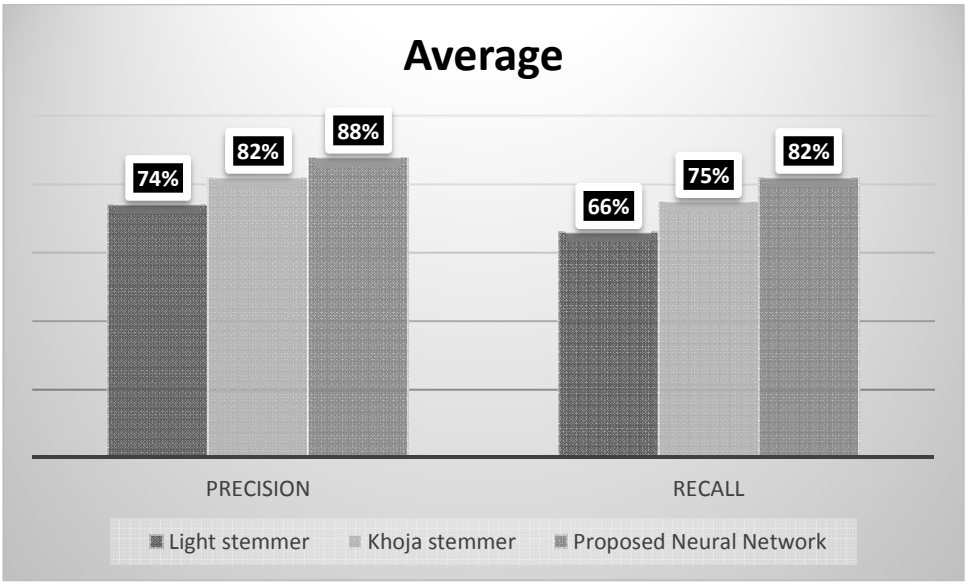

Figure 3. Comparison Among The Three Stemming Algorithms Based On The Average Results

As shown in Table. 2, the greatest value of precision has been obtained by bi-gram candidates for both of using NE pattern and without using which reported 81%, 82% and 83% for NC-value, NTC-value and NLC-value respectively. Moreover, NLC-value has outperformed the other association measures in terms of bi-gram, tri-gram, 4-gram, 5-gram, with and without using NE pattern.

As shown in Table 6, the results of all documents using the three stemming algorithms based on precision and recall have been obtained. First, it can be noticed that the Law and Astrology documents have the highest results of precision and recall using the three stemmers. This is due to the vast amount of named entities and foreign words that could be used in such domains. Hence, removing such entities has a significant impact on reducing the error rate.

In terms of Light stemming algorithm, it can be shown that the precision values are smaller than the recall values. This is due to the under-stemming errors that are frequently occurred by Light stemming algorithm. In contrary, in terms of Khoja stemming algorithm, it can be shown that the recall values are smaller than precision values. This is due to the over-stemming errors that are frequently occurred by Khoja stemming algorithm. In the same manner, the proposed BPNN stemming algorithm have also recall values that are smaller than

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

precision values. This is because the proposed BPNN has been trained using Khoja stemming.

As shown in Figure 4.2, the Khoja stemmer has outperformed the light stemmer by achieving 82% precision and 75% recall. This has been expected from the study of Goweder et al. [18] whose made a comparison between Khoja and Light stemmers and reported that Khoja has better stemming results that the Light stemmer. On other hand, the proposed method has outperformed the Light and Khoja stemmers by achieving an 88% precision and 82% recall. This is because the BPNN has the ability to correct the errors through the training. This can demonstrate the use of machine learning in terms of enhancing Arabic stemming. In addition, the proposed feature set has provide more proper encoding for the words which leads to reduce the over- stemming that usually produced by Khoja algorithm.

5. CONCLUSION

This study has proposed a set of feature with Back-Propagation Neural Network in order to enhance the Arabic stemming performance. OSAC dataset has been used for the experiments. In addition, two stemmers including Light stemmer and Khoja stemmer have been applied also on this dataset in order to compare the results with the proposed method. On other hand, this study has proposed a new approach of sampling for the test purposes. The experimental results have revealed that the proposed hybrid feature set with BPNN has outperformed the Light stemmer and Khoja stemmer. This has demonstrated that the machine learning techniques yield promising outcomes in terms of enhancing the Arabic stemming performance. In addition, the proposed features have demonstrated an improvement in terms of providing an appropriate representation for the words.

REFERENCES

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern information retrieval* vol. 463: ACM press New York, 1999.

[2] Leah S Larkey, Lisa Ballesteros, and Margaret E Connell, "Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis," in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 2002, pp. 275-282.doi:Retrieved from.

[3] Peter Willett, "The Porter stemming algorithm: then and now," *Program,* vol. 40, pp. 219-223, 2006Retrieved from.

[4] Shereen Khoja and Roger Garside, "Stemming arabic text," *Lancaster, UK, Computing Department, Lancaster University,* 1999Retrieved from.

[5] Hasan M Alserhan and Aladdin S Ayesh, "An application of neural network for extracting Arabic wordroots," *WSEAS Transactions on Computers,* vol. 5, pp. 2623-2627, 2006Retrieved from.

[6] Abdelwadood Moh'd A Mesleh, "Chi square feature extraction based SVMs Arabic language text categorization system," *Journal of Computer Science,* vol. 3, pp. 430-435, 2007Retrieved from.

[7] Aitao Chen and Fredric C Gey, "Building an Arabic Stemmer for Information Retrieval," in *TREC*, 2002, pp. 631-639.doi:Retrieved from.

[8] Monica Rogati, Scott McCarley, and Yiming Yang, "Unsupervised learning of arabic stemming using a parallel corpus," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, 2003, pp. 391-398.doi:Retrieved from.

[9] Meryeme Hadni, Abdelmonaime Lachkar, and S Alaoui Ouatik, "A new and efficient stemming technique for Arabic Text Categorization," in *Multimedia Computing and Systems (ICMCS), 2012 International Conference on*, 2012, pp. 791-796.doi:Retrieved from.

[10] Tengku Mohd T Sembok and Belal Abu Ata, "Arabic Word Stemming Algorithms and Retrieval Effectiveness," in *Proceedings of the World Congress on Engineering*, 2013Retrieved from.

[11] Motaz K Saad and Wesam Ashour, "OSAC: Open Source Arabic Corpora," in *6th International Symposium on Electrical and Electronics Engineering and Computer Science, Cyprus*, 2010, pp. 118-123.doi:Retrieved from.

[12] Y. Benajiba, M. Diab, and P. Rosso, "Arabic Named Entity Recognition: A Feature-Driven Study," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 17, pp. 926-934, 2009.doi:10.1109/TASL.2009.2019927Retrieved from.

[13] Yassine Benajiba, Paolo Rosso, and José Miguel Benedíruiz, "Anersys: An arabic named entity recognition system based on maximum entropy," in *Computational Linguistics and Intelligent Text Processing*, ed: Springer, 2007, pp. 143-153.

[14] Naji F Mohammed and Nazlia Omar, "Arabic named entity recognition using artificial

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

neural network," *Journal of Computer Science,* vol. 8, p. 1285, 2012Retrieved from.

[15] Jaffar Atwan, Masnizah Mohd, and Ghassan Kanaan, "Enhanced arabic information retrieval: Light stemming and stop words," in *Soft Computing Applications and Intelligent Systems*, ed: Springer, 2013, pp. 219-228.

[16] Vikas Thada and Vivek Jaglan, "Comparison of Jaccard, Dice, Cosine Similarity Coefficient To Find Best Fitness Value for Web Retrieved Documents Using Genetic Algorithm," *International Journal of Innovations in Engineering and Technology,* 2013Retrieved from.

[17] Cristian Moral, Angélica de Antonio, Ricardo Imbert, and Jaime Ramírez, "A Survey of Stemming Algorithms in Information Retrieval," *Information Research: An International Electronic Journal,* vol. 19, p. n1, 2014Retrieved from.

[18] A Goweder, H Alhami, Tarik Rashed, and A Al-Musrati, "A hybrid method for stemming Arabic text," *Journal of Computer Science, http://eref. uqu. edu. sa/files/eref2/folder6/f181. pdf,* 2008Retrieved from.