ISSN 1996-3343

Asian Journal of Applied Sciences

http://knowledgiascientific.com

Knowledgia SCIENTIFIC
A Place to Publish Outstanding Research

Asian Journal of Applied Sciences 8 (4): 269-276, 2015
ISSN 1996-3343 / DOI: 10.3923/ajaps.2015.269.276

A Hybrid of Statistical and Machine Learning Methods for Arabic Keyphrase Extraction

Nidaa Ghalib Ali and Nazlia Omar
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi, Malaysia

Corresponding Author: Nidaa Ghalib Ali, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi, Malaysia

ABSTRACT

A keyphrase can be described as a brief phrase comprising between one to five words that correspond to significant perceptions in an article. Text summarization, automatic indexing, classification and text mining are some of the many activities that involve the function of keyphrases. A wide range of techniques have been generated over time for the purpose of keyphrase extraction and much emphasis has been placed on the automatic extraction of keyphrases involving manuscripts in English and a variety of other dialects. However, on the other side of the coin, keyphrase extraction for documents in the Arabic language has largely been neglected. Thus, for the purpose of Arabic keyphrase extraction, this study recommends a hybrid approach which involves the merger of statistical and machine learning methods. The statistical methods involve Term Frequency (TF), First Occurrence in text (FO), Sentence Count (SC), C-Value and TF-IDF, while the machine learning algorithms comprise Linear Logistic Regression (LLR), Linear Discriminant Analysis (LDA) and Support Vector Machines (SVMs). The execution of this undertaking was initiated by the utilization of Part of Speech (POS) for the extraction of noun phrases. Following this, the outcomes generated through the application of statistical methods are employed as features for the purpose of classification. The hybrid model, which is based on SVM achieves the best result with 93.9% accuracy. Through several tests, it has been substantiated that the recommended model is appropriate for extracting Arabic keyphrase.

Key words: Arabic document, machine learning, keyphrase extraction, features

INTRODUCTION

Keyphrases are commonly utilized in compilation of documents for elaborating on the subject matter of a solitary document in an uncomplicated and logical manner. This goes a long way towards the provision of a swift means to verify the informational requirements of the person reading the document. The purpose of this study was to recommend ways of utilizing the results from a number of keyphrase extraction methods as input features meant for a machine learning algorithm. This algorithm then paves the way for a decision on whether or not a term is a keyphrase. Keywords are defined as the words employed for the classification of a document, or the identification of a topic, theme or subject of a document. They also point the way towards the contents that are highly relevant to the reader (Gupta and Lehal, 2010). As mentioned earlier, a keyphrase is a brief phrase made up of between one to five words. It meets the same demands of a keyword, albeit with a wider capacity for the summarization of a concept. This author opines that a brief phrase comprising connected words is assumed to be more meaningful than a solitary word.

Compared to a multiword, also known as a compound multiword expression (MWE), keyphrases are considered more significant in documents. They offer a concise outline of the contents and define phrases that are keyphrases in one document but irrelevant in others (Witten *et al*.,1999). Compound multiword expressions are made up of two or more words, which can be syntactically and/or semantically idiosyncratic in character. Also, MWEs operate as a solitary element during several stages of linguistic investigations, where their presence is detected in a document regardless of their relevancy (Huang *et al*., 2006). Made up of one or more than one words, a keyphrase is considered an expressive and significant component in documents. Keyphrases are extensively utilized for recovering data and the processing of natural language. These procedures come with the added advantage of being applicable for all forms of text indexing, summarization and clustering (Auglul *et al*., 2012).

Keyphrases can ease the way in situations, where a quick investigation on the relevancy of a document is required. The extraction of keyphrases can be achieved through the employment of supervised or unsupervised machine learning algorithms. Bearing in mind the diversity of elements on the internet, manually assigning keyphrases can turn out to be a tiresome and time wasting affair. As a result, mechanical keyphrase production procedures are very much in demand (Abulaish and Anwar, 2012).

Keyphrase generation is a method in order to collect major subjects of documents into a list of phrases. Automatickeyphrase generation can be classified into two categories, namely, keyphrase assignment and keyphrase extraction. In the case of keyphrase assignment, the prospective keyphrases are shown in a programmed vocabulary where, the assignment requires a categorization of documents into an array of keyphrase groupings. While keyphrase extractions are presented in the document, some that do not show are intermittently included by the authors (Turney, 1999).

Automatic keyphrase extraction is a process developed with the utilization of computers for determining the most significant phrases in a document. This process does away with the costly training data utilized for overseeing the learning of a person.

Keyphrases can be helpful in the management of procedures involving the generation and treating of a hefty load of documentary data, where they have proven their efficiency for rapid data recovery. While significant emphasis has been placed on the automatic extraction of keyphrases for documents in the English and other languages, the same cannot be said for scripts in the Arabic language which have generally been neglected (El-Beltagy and Rafea, 2009; El-Shishtawy and Al-Sammak, 2012). This situation can be attributed to the intricate nature of the Arabic language. Keyphrases can be utilized in the Arabic language for:

- Document indexing and information recovery
- Machine translation organization
- Ontology development
- Question response set-up

In comparison to most other languages, the intricacy attributed to the Arabic language proved to be a major stumbling block, when it came to the differentiation between singular and plural nouns. The overwhelming morphological diversity of words in Arabic also presented a formidable barrier as a single Arabic word can be interpreted in a variety of ways. A better understanding of this predicament can be attained through the realization that a solitary English word like "جيد"

"good" can also be interpreted as "جيدة""الجيدة""جيدان""الجيد""جيدون""الجيدون". This is dilemma is mostly felt when statistical aspects, such as frequency and TF-IDF come into play as in these situations, each arrangement involves a separate word (Farghaly and Shaalan, 2009).

The main study objective was to design a new model for automatic keyphrase extraction from single or multiple Arabic documents, by designing a hybrid method that combines several statistical methods with machine learning methods (Linear Logistic Regression, Linear Discriminant Analysis and Support Vector Machines), then evaluating the performance of the hybrid method using the selected evaluation metrics.

MATERIALS AND METHODS

Figure 1 exhibits the proposed research method which comprises these stages.

Document pre-processing: This refers to the pre-processing of the input document(s). The pre-processing phase involves the following activities.

Tokenization: Tokenization involves the exploration of the words that form a sentence. Initially, the textual data appears to be a block of characters in a sentence and data recovery from a dataset is a requirement.

Eliminate word removal: Most documents include details related to grammar topics such as pronouns, conjunctions, prepositions etc. However, while these parts of speech linguistically enhance the structure of a statement, they do little for the definition of semantic functions.

Stemming: Stemming is described as the process whereby a removed prefix is re-instated at the beginning of a word root, while a removed suffix is defined as the letter that is added to the end of a word root.

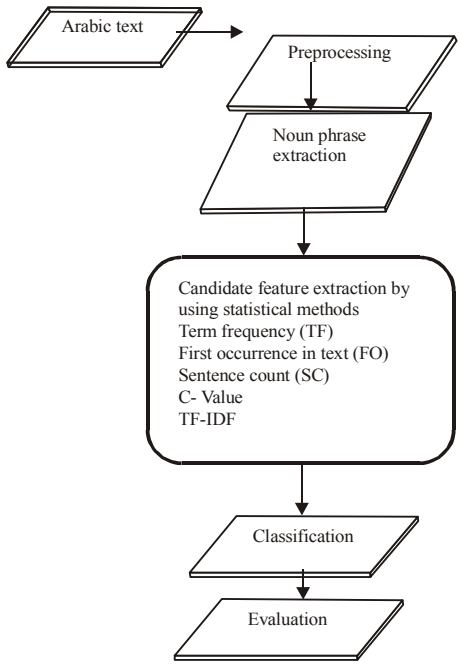

Fig. 1: Proposed method framework

Noun phrase extraction (candidate extraction): During this stage, all noun phrases extracted from the Arabic text are deemed keyphrase candidates. Our keyphrase extraction algorithm only considers noun phrases as keyphrase candidates. Some sets of features are extracted for every noun phrase. During this stage, the extraction on N-gram phrases from the Arabic phrase was executed (Kumar and Srinathan, 2008). Following this, POS patterns were utilized for the identification of candidate noun phrases.

Candidate feature extraction: During this stage, several conventional supervised and unsupervised algorithms, which are specifically crafted for the extraction of keyphrases are utilized.

The features utilized for the grading of keyphrase candidates are Term Frequency (TF), First occurrence in text (FO), Sentence Count (SC), C-Value and TF-IDF. Term Frequency refers to the number of times the candidate phrase appears in the text. First Occurrence in text (FO) refers to the initial appearance of the word in a text. Sentence Count (SC) refers to the sum of sentences with keyphrase associates. C-Value refers to a hybrid domain-independent procedure that merges linguistic and statistical data (with the accent on statistical) for the purpose of extracting multi-word and nested terms (terms that materialize within more extended terms and may or may not materialize spontaneously in the corpus). The TF-IDF is allocated to all keyphrases, where their relevancy is attested to in terms of the TF-IDF statistical measure.

The end result of these procedures is the employment of a feature vector for the training of machine learning classifiers.

Classification: The fundamental plan involves the supplying of statistical and linguistic data (feature vector) or the output from a number of extraction methods to a machine learning classification structure. The keyphrase extractions are utilized as input to pave the way for these features to evaluate and determine whether or not a term is considered a keyphrase. An appraisal was conducted on a number of classification algorithms for the merging of keyphrase extraction methods. The following are elaborations on these classification algorithms.

Linear Logistic Regression (LLR): Logistic regression envisages the likelihood of this outcome that can only reciprocate in a dual manner. Logistic regression can also cope with numerous predictors (numerical and categorical). The following are the logistic regression models:

$$\text{Logit (y is kp)} = \beta^0+\beta_1 x_1+...+\beta_k X_k \quad (1)$$

The form identifies the predicted probability as:

$$f(x) = p(x \text{ is kp}) = \frac{\exp^{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k}}{1+\exp^{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k}} \quad (2)$$

where, the coefficient β_i manages the influence of the predictor. The farther β_i falls from 0, the greater the influence of the predictor x_i.

Linear Discriminant Analysis (LDA): Linear Discriminant Analysis (LDA) is a widely utilized instrument for multiclass discriminative dimensionality reduction. The key purpose of LDA is to locate a well predictor for the category of any sample of a similar distribution. This involves locating a one-dimensional projection through a vector that amplifies the category division.

This process enhances the ratio between class variance to the within-class variance in any specific data set, thereby assuring maximum separation:

$$\max_{v} \frac{v^t S_B v}{v^t S_W v} \tag{3}$$

Support Vector Machines (SVM): Support vector machines (SVM) are recommended as a solution to the two-class predicament by locating the optimally separating hyper-plane that is flanked by two groupings of data. Assume that X is an array of labelled training points (features vector) (x1, y1),..., (xn, yn), where every training point $x_i \in$ RN is provided with a label $y_i \in \{-1, +1\}$, where i = 1, . . ., n.

SVMs seek to arrive at an approximation of a function f (x) = $w.x_i$+b and locate a classifier which can be resolved via the following convex optimization:

$$\min_{w,b} \sum_{i=1}^{n} \left[1 - y_i(w.x_i + b)\right] + \frac{\lambda}{2} \|w\| \tag{4}$$

where, λ is a regularization parameter.

Evaluation measures: The implementation of a gold standard evaluation requires a manual detection of keyphrases in Arabic documents. A manual annotation of the keyphrase collocations is executed through a native speaker. An assessment of the extraction algorithm against the reference array of keyphrases was conducted. Here, every document is manually ascertained by an individual for the purpose of allocating the keyphrases according to precise guidelines called the gold standard.

In order to gauge the efficiency level through the employment of these classification methods, test results are separated into the following categories: True Positive (TP) is the array of phrases rightly deemed keyphrases, False Positive (FP) is the array of phrases that were wrongly deemed keyphrases, False Negative (FN) is the array of keyphrases that were wrongly not deemed keyphrases and True Negative (TN) is the array of phrases rightly not deemed keyphrases. The effectiveness of the procedure is judged by staking the quantity of phrases extracted through the automatic keyphrase extraction method against gold standard (precision) and the quantity of identified phrases against all likely correct phrases (recall). The F-measure (F-score) is the harmonic mean of precision and recall and it is also utilized for automatic keyphrase extraction assessment. These metrics are calculated as follows:

$$\text{Precision (Pr)} = \frac{TP}{TP + FP}$$

$$\text{Recall (Re)} = \frac{TP}{TP + FN}$$

$$\text{F} - \text{measure} = \frac{2\,\text{Pr} \times \text{Re}}{\text{Pr} + \text{Re}}$$

RESULTS AND DISCUSSION

Experiments were conducted to assess the effectiveness of keyphrase extraction algorithms utilizing some conventional supervised and unsupervised algorithms which have been constructed solely for the purpose of keyphrase extraction. Only nouns and noun phrases were considered for classification as candidate phrases.

To evaluate the proposed models, this study used a gold standard dataset. This data was manually developed and all keyphrases were manually annotated by a native speaker. The reference dataset contains 174 text files which were selected from multiple domains (health, news, sport , education, political and history).

The initial experiment was conducted for the purpose of assessing the effectiveness of the statistical systems that were utilized. The features employed for gauging the performance of keyphrase candidates are: Term Frequency (TF) refers to the number of times the candidate phrase occurred, First Occurrence in text (FO) refers to the initial occurrence of the word in the text, Sentence Count (SC) refers to the number of sentences with keyphrase members, C-Value refers to the hybrid domain-independent process which merges linguistic and statistical data (with focus on the statistical area) in order to extract multi-word and nested terms (terms that are located within other more extended terms and may or may not be naturally present in the corpus), TF-IDF which were allocated to the keyphrases, where it is of consequence in relationship to the TF-IDF statistical measure.

The performance metrics (Precision, recall and F-score) for all methods are portrayed in Table 1.

The results from these experiments were utilized as feature vectors in the training of machine learning classifiers.

The second experiment involved an assessment of the machine learning techniques employed for automatic keyphrase extraction. Six statistical and linguistic features were forwarded to signify the significance of every candidate phrase.

All machine learning classification procedures proved to be efficient. Linear logistic regression envisages the likelihood of an outcome with only a dual response for the purpose of calculating a variety of predictors, LDA locates the ideal predictor for the category of any sample of a similar distribution, SVMs recommended the establishment of a superior separating hyper plan flanked by two classes of data as a solution to the two-class dilemma. These experiments revealed that the SVM algorithm was of the highest order with a F1-measure of 93.90%. These values are considerably lofty and are comparable to preceding Arabic language keyphrase extraction procedures. The outcomes (Precision, recall and F1-measure) generated by all classification processes are displayed in Table 2.

By comparing Table 2 with other experimental results, machine learning methods performed better than the statistical methods. Linear discriminant analysis in Arabic automatic keyphrase

Table 1: Experimental result of the statistical features

Statistical features	Values
Precision	67.02
Recall	96.7
F1	79.17

Table 2: Performance of the used machine learning for keyphrase extraction

Methods	Precision	Recall	F1
SVM	91.36	97.20	93.90
LDA	72.35	97.20	82.65
LLR	90.92	97.20	93.36

SVM: Support vector machine, LDA: Linear discriminant analysis and LLR: Linear logistic regression

extraction performed worse than both LLR and SVM, because the performance of LDA, in its normal conditions is poor with limited categories variables, computes the addition of multivariate distribution and suffers multicollinearity.

Comparing the results obtained by the state of Arabic art keyphrase extraction system developed by El-shishtawy and Al-Sammak (2012) (precision 56%, recall 40% and measure 49.52%), the research extends the comparison of keyphrase extraction method results to KP-Miner (precision 21.4% and recall 7.7%) (El-Beltagy and Rafea, 2009), shows that our model outperforms previous models.

CONCLUSION

We proposed a keyphrase extraction method to be utilized for the Arabic language. The candidate keyphrases were categorized through the employment of classification algorithms (linear logistic regression, linear discriminant analysis and support vector machines). Each term was classified as a keyphrase term, or a non-keyphrase term. An assessment was carried out through an array of tests and the outcomes revealed that the SVM classifier is clearly superior to other keyphrase extraction methods for Arabic language extraction. Generally, the outcomes demonstrate that a hybrid of statistical features and machine learning methods proved to be significantly efficient for keyphrase extraction.

This study focused on the development of automatic keyphrases and includes a concise synopsis on the contents of the document. In future, the Arabic word stemming can be further developed to facilitate the extraction of core words. Arabic stemming differs from that of English stemming and this invariably paves the way towards the achievement of more precise data. Furthermore, the hybrid automated extraction procedure can be confirmed on a domain specific corpus and this leads to increased accuracy in the assessment of the system.

ACKNOWLEDGMENTS

The study wish to thank University Kebangsaan Malaysia (UKM) and Ministry of Educa.

REFERENCES

Abulaish, M. and T. Anwar, 2012. A supervised learning approach for automatic keyphrase extraction. Int. J. Innov. Comp. Inform. Control, 8: 7579-7601.

Auglul, I., N. Gicekli and I.Gicekli, 2012. Searching documents with semantically related keyphrases. Proceedings of the 6th International Conference on Advances in Semantic Processing, September 23-28, 2012, Calabria, Italy.

El-Beltagy, S.R. and A. Rafea, 2009. KP-Miner: A keyphrase extraction system for English and Arabic documents. J. Inf. Sys., 34: 132-144.

El-Shishtawy, T. and A. Al-Sammak, 2012. Arabic keyphrase extraction using linguistic knowledge and machine learning techniques. Proceedings of the 2nd International Conference on Arabic Language Resources and Tools, December 8-10, 2012, Cairo, Egypt, pp: 144.

Farghaly, A. and K. Shaalan, 2009. Arabic natural language processing: challenges and solutions. ACM Trans. Asian Language Inform. Process. Assoc. Comput. Mach., 8: 1-22.

Gupta, V. and G.S. Lehal, 2010. A survey of text summarization extractive techniques. J. Emerging Technol. Web Intell., 2: 258-268.

Huang, C., Y. Tian, Z. Zhou, C.X. Ling and T. Huang, 2006. Keyphrase extraction using semantic networks structure analysis. Proceedings of the 6th International Conference on Data Mining, December 18-22, 2006, Hong Kong, pp: 275-284.

Kumar, N. and K. Srinathan, 2008. Automatic keyphrase extraction from scientific documents using N-gram filtration technique. Proceedings of the 8th ACM Symposium on Document Engineering, September 16-19, 2008, Sao Paulo, Brazil, pp: 199-208.

Turney, P.D., 1999. Learning to extract keyphrases from text. National Research Council, Institute for Information Technology, pp: 25-29.

Witten, I.H., G.W. Paynter, E. Frank, C. Gutwin and C.G. Nevill-Manning, 1999. KEA: Practical automatic keyphrase extraction. Proceedings of the 4th ACM Conference on Digital Libraries, August 11-14, 1999, California, USA., pp: 254-256.