Journal of Engineering and Applied Sciences 13 (6): 1395-1399, 2018
ISSN: 1816-949X

Automatic Summarization Arabic Text Using Key Phrases Extraction

[1]Hamzah Noori Feje and [2]Mohanaed Ajmi Falih
[1]Ministery of Eduction, General Directory of Al Qadissyah Education, Al Qadissyah, Iraq
[2]Ministery of Eduction, General Directory of Babilon Education, Babilon, Iraq

Abstract: Because of the growing number of electronic documents, human being are badly in need of more rapid techniques for evaluating the link of documents. Summarization is representation of underlying written text. A full underst anding of the document is essential to form an ideal summary. However, achieving full underst anding is either difficult or impossible for computers. Therefore, selecting main sentences from the original text and introducing these sentences as a summary present the most frequent techniques in automated text summarization. This study propose using key phrase extraction module is applied to extract main important key phrases from the text that helps specify the most important sentences and find similar sentences based on similarity algorithm. It is applicable to extract one sentence from a set of similar sentences while overcoming the other similar sentences (i.e., sentences that have a greater similarity than the predefined threshold). This model is designed for single-document Arabic text summarization. The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) matrix is employed for the assessment. For the summarization dataset, Essex Arabic Summaries Corpus was used. It has many topic based articles with multiple human summaries. This model achieved accuracy more than 80%.

Key words: Text summarization, key phrase extraction, similarity, ROUGE matrix, techniques, rapid, single-document

INTRODUCTION

One simply sums up an article or a document by arriving at accurate, adequate summarization of its most essential concepts. Nevertheless, this action dem and great labor and time for both human being and computers. Two different people can arrive at different summaries of the same article built on what they believe is most important and how they recognize the article. This prompts the creation of an automated summarizer system that can summarize an article with least effort and time. Many researches have tackled the topic of automatic summarization over the last 50 years (Luhn, 1958).

Summing up a text implies formulating a summary from an article or a group of articles which tackle the same subject by picking up the most important parts of the text and then writing them in historical order. Automatic text summarization points to the creation of a shorter version of a particular document or a set of documents by machines. A text document may be reduced using several applications.

A summary should supply the key notions of the input text. Only the major sentences should be involved in the summary and the sentences ought to be interpreted on the basis of the used summarization method. Two basic methods, viz. extractive and abstractive are widely held in automatic text summarization.

Extractive summarization is confined in such a way that only the significant sentences are taken and included chronologically to create an adequate summary. The extracted parts of text may differ according to the applied summarizer. Many summarizers take sentences in contrast to paragraphs or other large text units. Extractive summarization is often applied in automatic text summarization.

Abstractive summarization employs more tools that rest on language and natural language creation nation. This technique may suitable terms that do not exist in the article that summarized. Abstractive summarization aims at copying methods used by human beings such as incorporating a concept that is present in the source article in a better and more fashioned way. Although, abstractive summarization is more influential than extractive summarization, adopting this approach is more complicated task. Therefore, an extractive concept is carried out in the proposed model.

Summarization has been investigated for over the past 50 years with most researches focusing on the

Corresponding Author: Hamzah Noori, Ministery of Eduction, General Directory of Al Qadissyah Education, Al Qadissyah, Iraq

English language (Hirao *et al*., 2007). Researches on Arabic-language documents have been implemented at a much later period and stays far behind those researches on other languages. Research on automatic summarization of Arabic-language documents has started approximately 10 years ago (Conroy *et al*., 2006). Further studies on Arabic-language resources are required. Several researches have adopted relatively advanced ways for Arabic language summarization (El-Haj *et al*., 2011a, b; Giannakopoulos *et al*., 2008).

This study propose a summarization approach to uses a key phrase extraction approach by an unsupervised machine learning algorithm to identify sentences which involve key phrases and to summarize source text documents.

Literature review: Despite all the progresses that have be made in the summarization of English-language documents, study on the syntactic and semantic summarization of Arabic-language documents remains basic. Nonetheless, some approaches and models for summarizing Arabic-language documents have been tackled.

Douzidia and Lapalme in highlight on Arabic text summarization in their generation of a model for summing up Arabic documents. They formulate an abstract model of the whole document and then created a shorter summary by selecting several important sentences from the input document. This model, labeled "Lakhas" was generated by carrying out extraction approaches to create 10 term summaries of news stories. Though this approach boasted high achievement and bearing, Lakhas could only employ 10 terms for the summarization that was regarded very limited because the summary could not tackle all the important topics.

Douzidia and Lapalme carried out four main sentence-trimming approaches. Nonetheless, this approach discarded important information from the sentences. Douzidia and Lapalme translated the DUC-2004 dataset from English to Arabic by using a machine translation application. Although this application introduced a summarization model for the Arabic language, its numerous translation process created incoherent sentences that was regarded a weakness for this model.

Schlesinger *et al*. (2008) performed semantic and syntactic summarization for the Arabic language by employing CLASSY, a multiple document summarizer. Akin to their other models, the overall influence of the CLASSY-generated summaries was affected by the poor translation process.

Azmi and Al-Thanyyan (2012) postulate an extractive summarization model of Arabic text that allowed the employer to improve the entire length of the output summary. Every sentence in the first summarization was ranked and these rankings promoted the production of the output summary. Their model accomplished satisfactory results and was able to be supported.

El-Haj and Hammo (2008) explain two summarization systems in their research; the Arabic query-based text summarization system and the Arabic concept-based text summarization system. The first is a query-based single document summarizer system that takes an Arabic document and a query (in Arabic). This system gets a summary for the document in according to the organized query. While the second takes a bag-of-words st anding for a particular concept as input to the system. In both systems the summarization is conducted in accordance with the sentences that best match the query or the notion (Schlesinger *et al*., 2008).

Previous publications have been reviewed and their strengths and weaknesses have been regarded. The existing models highlight the limited a portion of research on Arabic summarization models. These models present can also be supported by means of their adequacy and influence. This study solves the problems arising in the Arabic summarization single documents by tackling the limitations of the existing summarization applications (El-Haj, 2012).

This study adduces the automatic summarization of the Arabic language, an area field that dem and too much research. We will attempt to accomplish better performance and accuracy levels than previously adopted models.

MATERIALS AND METHODS

This study describes the proposed framework for single-document Arabic text summarization. A pre-processing text, key phrase extraction is used to extract the important. Figure 1 displays the several phases in the entire architecture of our system as follows:

- Text pre-processing
- Key phrase extraction
- Important sentence extraction and filtering
- Evaluation phase

Input text: Input single-documents will be employed to extract texts from articles on various subjects such as politics, economy and sports. These documents imply texts of different sizes. Figure 1 displays the general flow of processing and copying with the document across several phases. C# language is employed to code the different stages of the adopted summarization model.

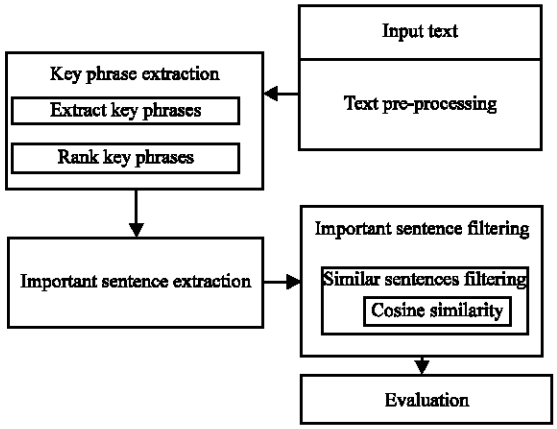

Fig. 1: Overall structure of Arabic text summarization

Text pre-processing: Pre-processing is may be the most significant step computational linguistics because the quality of the obtained summary rests on how effective the text is represented. Some experiments in this study will imply the pre-processing stage. This stage only involves four steps, namely, tokenizing, eliminating stop words, stemming and text representation and term weighting (TFIDF). In this research, used light stemmer algorithm is used to normalize words derived from the same root (Froud *et al*., 2010).

Key phrase extraction: Key phrases are significant words/phrases that express the topic of the text. The term key phrase is employed in literature to highlight that the chosen terms may be phrases. Key phrase extraction involves the following steps.

Extract key phrases: Extracts all noun phrases from the Arabic text as candidate key phrases. Our key phrase extraction algorithm only regards noun phrases as candidate key phrases. We use POS patterns during this phase to specify the Arabic noun phrases.

Rank key phrases: For each noun phrase, some set of features are taken. The following is used for ranking the candidate key phrase.

Term frequency: Frequency points to the number of occurrences of the candidate phrase. Frequency is computed employing an aggressive stemming algorithm (iterated light stemmer). Frequency is normalized by the number of noun occurrences in the document. A normalized frequency is calculated in the following equation:

$$\text{Normalized frequency} = \frac{\text{Frequency of the key term}}{\text{Total no. of words in the document}} \quad (1)$$

First occurrence in text: This feature points to the first occurrence of a word in the text. For this feature, we use sentence indexes or the index of the sentence where the term first exists. This feature is normalized by the total number of sentences in the document.

Last occurrence in text: This feature points to the last occurrence of a word in the text. For this feature, we employ sentence indices or the index of the sentence where the term last occurs. This feature is normalized by the total number of sentences in the document.

Sentence count: This feature points to the total number of sentences in which members of the key phrase can occur. We normalize this feature using the total number of sentences in the document.

Important sentences extraction: This phase extracts the most significant sentences from single-document text summarization, respectively. This phase implies the steps.

Sentences spiltter: A sentence is a sequence of letters that ends with a full stop (.), an exclamation mark (!) or a question mark (?). In this step, each document is split into several sentences using delimiters (e.g., full stop, question mark and exclamation mark).

Important sentences extraction: Not every sentence that is taken from a text is regarded important for text summarization. We presuppose that only those sentences that involve key phrases are regarded important. As result, we suppose that overlooking sentences that do not include any key phrases can provide better outcomes. We identify an important sentence using the following equation:

$$\text{Score}(\text{sentence}) = \frac{\text{Total no. of words in every key phrases in the sentence}}{\text{Sentence length in words}} \quad (2)$$

We regard a sentence important if the score (sentence) is greater than the threshold. Otherwise, we rule out the sentence from the sentence set.

Important sentences filtering: In this step, we take all important sentences by using similarity measures to look for similar sentences and then group these similar sentences into a adequate group. This study centres on

the famous measures of distance between patterns. In this consider, we employ cosine similarity as similarity or distance measures (Huang, 2008). Next, we chosen one sentence and ignore the other sentences. We measure the importance of the chosen sentence by following the score (sentence) formula in the previous step. The cosine similarity is described as follows:

Cosine similarty: Cosine similarity is one of the most famous similarity measures. The cosine similarity of two sentences, namely, $\vec{t_a}$ and $\vec{t_b}$ is calculated as follows:

$$SIM_C\left(\vec{t_a},\vec{t_b}\right)=\frac{\vec{t_a}\times\vec{t_b}}{\left(\left|\vec{t_a}\right|\times\left(\left|\vec{t_b}\right|\right)\right)} \quad (3)$$

where, $\vec{t_a}$ and $\vec{t_b}$ are the m-dimensional vectors over the term set T = {t1, tm}. Each variable st ands for the positive weight of a term in the document. In our research employ cosine similarity to compute the similarity between sentences that involve key phrases.

Evaluation: Evaluating the quality and consistency of a generated summary is hard to do since an ideal summary remains undefined as by Fiszman *et al.* (2009). Performing system evaluation may help address this problem.

Automatic evaluation methods are not employed in TAC, possibly because of the poor correlation between the outcomes gained from manual evaluation and existing automatic techniques. However, at least one of the automatic methods, namely, Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is used by many research groups in the field.

In the research, ROUGE is employed to compute the scores of a candidate summary built on the n-gram overlap between candidate and reference summaries (Lin, 2004). ROUGE consists of several metrics such as ROUGE 1-3 and so on with each corresponding to the size of the n-grams used in the evaluation. ROUGE-N scores are computed as follows:

$$\text{ROUGE-N recall} = \frac{\sum_{S\in\{\text{Reference summaries}\}}\sum_{gram_n\in S}Count_{match}(gram_n)}{\sum_{S\in\{\text{Reference summaries}\}}\sum_{gram_n\in S}Count(gram_n)} \quad (4)$$

$$\text{ROUGE-N precision} = \frac{\sum_{S\in\{\text{Reference summaries}\}}\sum_{gram_n\in S}Count_{match}(gram_n)}{\sum_{gram_n\in condidate\,summary}Count(gram_n)} \quad (5)$$

Table 1: Result of single document summarization

ROUGE-N	Recall	Precision	F-score
ROUGE 1	0.98211	0.67957	0.80329
ROUGE 2	0.93333	0.62222	0.74666
ROUGE 3	0.89830	0.61627	0.73103
ROUGE 4	0.86206	0.60975	0.71428

$$\text{ROUGE-NF-score} = \frac{2\times\text{ROUGE-N}_{recall}\times\text{ROUGE-N}_{precision}}{\text{ROUGE-N}_{recall}+\text{ROUGE-N}_{precision}} \quad (6)$$

where $count_{match}$ (n-gram) is the number of times that an n-gram from a reference summary appears in a candidate summary and count(n-gram) is the number of times that an n-gram appears in a candidate or reference summary. Thus, the recall measure calculates the proportion of n-grams from reference summaries that occur in a candidate summary whereas the precision measure calculates the proportion of n-grams from candidate summaries that exist in a reference summary. The F-score combines recall and precision into one metric. The outcomes are displayed in Table 1.

RESULTS AND DISCUSSION

The corpus used to generate single-document summaries was extracted from the Arabic language version of Wikipedia and two Arabic newspapers, namely, Alrai from Jordan and Alwatan from Saudi Arabia. The 10 subject fields were politics, sports, art and music, the environment, health and medicine, science and technology, finance and insurance, religion, education and tourism and travel. The entire number of documents used was 153, moreover, 765 human-generated extractive summaries of those articles were used. These summaries were created by using mechanical Turk. The total number of words is 18,264 and each document contains an average of 380 words with a minimum word count of 116 words and a maximum of 971 words. In the experiment, we applied our system in single documents for each of the aforementioned topics, the recall was (0.98), precision (0.68) and F-score (0.80).

CONCLUSION

In this study, a new automatic Arabic text model of summarization is introduced, discussing the structure of the proposed frameworks for single-document Arabic text summarization. This research relies on extract key phrases in the text. Later on specifies the sentences content key phrases. We also describe these sentences as important sentences. Similarity algorithm, namely, cosine similarity is employed to choose one sentence from each set of similar sentences while ignoring the other

sentences. These sentences will be used to represent the summarized text. This research achieved best result as compared with other systems.

REFERENCES

Azmi, A.M. and S. Al-Thanyyan, 2012. A text summarizer for Arabic. Comput. Speech Lang., 26: 260-273.

Conroy, J.M., J.D. Schlesinger, D.P. O'leary and J. Goldstein, 2006. Back to basics: Classy 2006. Proceedings of the 6th Conferences on Document Understanding (DUC'06) Vol. 6, June 8-9, 2006, National Institute of Standards and Technology, Gaithersburg, Maryland, USA., pp: 150-158.

El-Haj, M., 2012. Multi-document Arabic text summarization. Ph.D Thesis, University of Essex, Colchester, England.

El-Haj, M., U. Kruschwitz and C. Fox, 2011a. University of Essex at the TAC 2011 multilingual summarisation pilot. Text Anal. Conf., 1: 1-7.

El-Haj, M., U. Kruschwitz and C. Fox, 2011b. Multi-document Arabic text summarisation. Proceedings of the 3rd Conference on Computer Science and Electronic Engineering (CEEC'11), July 13-14, 2011, IEEE, Colchester, England, UK., ISBN:978-1-4577-1300-2, pp: 40-44.

El-Haj, M.O. and B.H. Hammo, 2008. Evaluation of query-based Arabic text summarization system. Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering (NLPKE'08), October 19-22, 2008, IEEE, Beijing, China, ISBN:978-1-4244-2779-6, pp: 1-7.

Fiszman, M., D. Demner-Fushman, H. Kilicoglu and T.C. Rindflesch, 2009. Automatic summarization of MEDLINE citations for evidence-based medical treatment: A topic-oriented evaluation. J. Biomed. Inform., 42: 801-813.

Froud, H., R. Benslimane, A. Lachkar and S.A. Ouatik, 2010. Stemming and similarity measures for Arabic documents clustering. Proceedings of the 5th International Symposium on I/V Communications and Mobile Network (ISVC'10), September 30-October 2, 2010, IEEE, Rabat, Morocco, ISBN:978-1-4244-5996-4, pp: 1-4.

Giannakopoulos, G., V. Karkaletsis, G. Vouros and P. Stamatopoulos, 2008. Summarization system evaluation revisited: N-gram graphs. ACM. Trans. Speech Lang. Process., 5: 1-39.

Hirao, T., M. Okumura, N. Yasuda and H. Isozaki, 2007. Supervised automatic evaluation for summarization with voted regression model. Inform. Process. Manage., 43: 1521-1535.

Huang, A., 2008. Similarity measures for text document clustering. Proceedings of the Sixth New Zealand Conference on Computer Science Research Student (NZCSRSC2008), April 14-18, 2008, University of Canterbury, Christchurch, New Zealand, pp: 49-56.

Lin, C.Y., 2004. Rouge: A package for automatic evaluation of summaries. Proceedings of the 2004 Workshop on Text Summarization Branches Out, July 25-26, 2004, Information Sciences Institute (ISI), California, USA., pp: 1-24.

Luhn, H.P., 1958. The automatic creation of literature abstracts. IBM J. Res. Dev., 2: 159-165.

Schlesinger, J.D., D.P. O'leary and J.M. Conroy, 2008. Arabic-English Multi-Document Summarization with Classy-the Past and the Future. In: Computational Linguistics and Intelligent Text Processing, Gelbukh A. (Eds.). Springer, Berlin, Germany, ISBN: 978-3-540-78134-9, pp: 568-581.