IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 2, February 2016.
www.ijiset.com
ISSN 2348 – 7968

An Arabic Text Summarization Model Based on LSA and Arabic Word Morphology

Dr. Abdulwase M. Alezzani[1], Dr. Sharaf A. Alhomdy[2]&Dr.Ghaleb H. Al-Gaphari[3]

[1]Assistant Prof., Faculty of Computer and Information Technology, Sana'a University, Yemen.
E-mail: amalezzani71@gmail.com
[2]Association Prof. & Vice-Dean for Students Affairs, Faculty of Computer and Information Technology, Sana'a University, Yemen.
E-mail: sharafalhomdy@gmail.com
[3]Professor & Vice-Dean for Academic Affairs, Faculty of Computer and Information Technology, Sana'a University, Yemen.
E-mail: drghalebh@gmail.com

Abstract

The target of text summarization is to get the important contents of the given document and remove an information redundancy. Arabic language text summarization methods are few compared to English and European languages. In this paper, we combine two methods for text summarization based on Latent Semantic Analysis (LSA) and Arabic Word Morphological model. The suggested model is associated with sentences similarity and semantic morphology. It selects sentences for each topic and removes repeated sentences from output summarization based on three techniques, i.e. word, stem, and root, in Arabic word morphology. We also use AN4 technique to improve and enhance the obtained Arabic text summarization. Experimental results show that our model obtains higher ROUGE score which was very promising.

Keywords: *Arabic Text Summarization, Latent Semantic Analysis, Arabic Word Morphology, AN4.*

1. Introduction

In Arabic text there are many problems such as the read, character writing style and recognition. Therefore, the key problem in the text automatic summarization process (TASP) is that the target summarized text is incoherent and deviates from the context of the original text. This problem emerges when statistical techniques are used for summarization [1]. It is hard to determine the quality of a good summary since it depends on many parameters such as the user background knowledge, user's requirements and compression ratio, among others, but scientists have been able to reach a level that in which computers are able to generate human consumable summaries [2].

There are several research projects to investigate and find out the techniques in automatically summarizing English documents as well as other European languages [3]. Unfortunately, the work on Arabic automatic summarization is very limited compared to the research on English and other European languages [4]. This huge gap is due to the lack of open-source tools and Arabic resources on Arabic Text Summarization.

The suggested model is based on the combination of two methods that are LAS and Arabic word morphological model. The combined model extracts an important concept from Arabic documents and returns the result of such concept as documents summarization. It is associated with similar sentences and it removes a redundancy of sentences from a target text summarization.

2. Previous Works

It is possible to classify different summarization approaches into different categories based on specific characteristics of each approach. We provide here an overview of the major summarization approaches:

1- El-Haj proposed a model to automatically summarize Arabic text using text extraction. The model consists of four main steps: Data Acquisition, preprocessing and feature extraction, scoring, ranking and generating the summary [4].
2- Haboush & Al-Zoubi have investigated a developed automatic Arabic text summarization model. They used word root clustering as a major activity [5].
3- Ibrahim & etc. have proposed a novel hybrid model for Arabic text summarization, which combined Rhetorical Structure Theory (RST) and Vector Space Model (VSM). This model has tried attempted to take advantage of both [1].
4- Wang & Ma have LSA summarization algorithm that combines term description with sentence description for each topic. LSA uses Singular Value Decomposition (SVD) to find out the semantic meaning of sentences. This model starts with document analysis using document representation and singular value decomposition [6].
5- El-Sayed & El-Barbary, presented a technique for Arabic document summarization using a fuzzy ontology, which is a fuzzy linguistic variable ontology and Field Association words [7].
6- Hadni etl. have followed a hybrid approach for Arabic multiword term extraction. This approach

ISSN 2348 – 7968

is composed by of two main steps: the linguistic approach and the statistical one [8].

3. An Arabic Words Morphology

The morphology of the Arabic language is based on the Semitic pattern scheme of forming words. Therefore, the majority of words are generated from basic entities called roots or stems. Roots are radicals according to a predefined list of patterns called morphological balances or patterns [2] [3] [9]. Also the roots are constructed mainly from 3, 4 or 5 letters roots (derivation rules). The mechanism is performed by adding letters and/or diacritical marks to the roots. These additional letters and diacritical marks may be added at the beginning, at the middle or at the end of the root. The pattern used for generating a word determines its various attributes such as gender (masculine/feminine), number (singular/plural), tense (past, present, and imperatives), mode etc.

Based on the above fore going, an Arabic word can be represented lexically by its root, along with its morphological pattern. The letter is one element of a countable set of limited size. A pattern is defined by a set of additive letters and/or a set of diacritical marks and their positions in the generated word. Next section will cover the proposed model.

4. Suggested Model

The suggested model is based on the combination of two methods that are LAS and Arabic Word Morphological model. The combined model extracts important concepts from Arabic documents and returns the result as documents summarization. It concerns with sentences similarity and it removes redundant sentences from the resulted summarization. The major phases of this model processing are preprocessing, Arabic part of speech, LSA for Arabic, and redundancy removal.

4.1 Preprocessing phase

The preprocessing phase that includes the following:

- Removing unnecessary information and tags such as diacritics, symbols, and stop words.
- Tokenization*:*

The refined document is decomposed into individual sentences in turn each sentence is decomposed into words. For text decomposition, we use:

- A morphological decomposition based on punctuation,
- Decomposition based on the recognition of markers morphological-syntactic forms or functional words.

4.2 Arabic Part of Speech Phase

In this stage, the document is processed on the base of the ***Parts of Speech (***POS) tagger and stemming techniques over the preprocessed documents. On the one hand the POS tagger is used in identifying the parts of speech corresponding to each word in a given document. On the other hand, the document words are reduced to their stems and roots formats. The purpose of this step is to obtain some morphological semantic features. Different approaches for Arabic stemming are implemented such as light stemmer and morphological analyzers. The light stemmer removes prefixes and suffixes, whereas the morphological analyzers extract the roots of words. In this experiment, the process of stemming increases the semantic features in sentences as well as the terms frequency in documents.

4.3 LSA phase

In this phase, the input document is represented by a matrix form to perform the computation. Each cell of the matrix is filled out with values based on Eq. (1). Local, global, near and far neighbors' weights are computed.

$$AN2(t_{i,j}) = \gamma[L(t_{i,j-1}) * G(t_{i,j-1}) + L(t_{i,j+1}) * G(t_{i,j+1})] \quad (1)$$

On the one hand the near neighbors are associated with two neighbors one from the right side while another one from the left side. Eq. (2) is used to compute such frequency. Moreover, the far neighbors are associated with four neighbors; two from the right side and the other two from the left side. Eq. (3) is used to compute the frequency.

$$AN4(t_{i,j}) = \gamma[0.5 * L(t_{i,j-2}) * G(t_{i,j-2}) + L(t_{i,j-1}) * G(t_{i,j-1}) + L(t_{i,j+1}) * G(t_{i,j+1}) + 0.5 * L(t_{i,j+2}) * G(t_{i,j+2})] \quad (2)$$

Where γ is a parameter with a range of value between 0 and 1.

4.4 Summarization Refinement Phase

In this phase repeated sentences in the output text summarization are removed to avoid sentences redundancy in the output summarization. In the process of removing duplicated sentences, a threshold is considered and the similarity measure is computed between the candidate sentences. Then the candidate sentences are tested based on cosine similarity and only the sentence with highest score less than or equal to the threshold would be included in the final text summarization. The process is repeated until no more candidate sentences are left. In fact, no sentences pair is included in the final summary as long as the similarity score is less than the threshold.

ISSN 2348 – 7968

4.4.1 Sentence-to-Sentence Similarity

For each candidate sentence (S), the similarity between (S) and each summary sentence (S') is computed of the bases of the semantic similarity between the terms from the first sentence with corresponding terms from the second. The process of computing sentences similarity is implemented by TF-ISF as well as cosine similarity.

$$sim(s_i, s_j) = \frac{\sum_{k=1}^{m} w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^{m} w_{ik}^2 \cdot \sum_{k=1}^{m} w_{jk}^2}}, \quad i, j = 1,\ldots,n \qquad (3)$$

Where W_{ik} represents the cell weight of the term ***k*** in the sentence ***i*** [10]. The cell weighting is computed in the context of text summarization based on important measures that are term Frequency (TF) and Inverse Sentence Frequency (ISF) that form the multiplication of TF and ISF indicated by TF-ISF [11] that is illustrated in Eq.(4).

$$TF_i \times ISF_i = TF_i \times \log \frac{N}{ni} \qquad (4)$$

Where TF_i is the term frequency of word i in the document, N is the total number of sentences and n_i is the number of sentences in which word i occurs.

4.4.2 Sentence selection:

The process of sentence selection depends on an enhanced algorithm that is based on Wang and Ma algorithm [6] that as illustrated in Fig. 4.1.

5 Experiments and Evaluation

In this section the data set, experiment results analyses and evaluation are considered where the output of the suggested model is compared by the human ideal summarization (references) that are ROOT_BREFAN4 with POS tagging, STEM_BREFAN4 with POS tagging and WORD_BREFAN4.

5.1 Datasets

The data set used in this experiment is created and distributed by the Linguistic Data Consortium (LDC) at the University City of Penn USA. The LDC distributes two Arabic collections, the Arabic GIGAWORD and the Arabic NEWSWIRE-a corpus. The contents documents are represented as UTF-8 files; those documents include meta-data and tags. The dataset includes documents which are used as input for the proposed model.

5.2 Results Analysis

This section describes two approaches to evaluation a text summary. These approaches are a human evaluation and automatic evaluation.

5.2.1 Human Evaluation

In this approach Arabic specialists from different backgrounds took part in evaluating different text summarization for the same document. Participants manually evaluate summaries based on the TAC responsiveness metric consisting of content and readability/fluency measures. Each participant was given a document with three summaries, one created using root, the second using stem and the third using word. Table (1) shows the human evaluation results for the proposed Model. The ranks and scores for the used Stem were comparatively better than the other two that used Word and Root. This was expected since the summarizer is extractive and no modifications were made to the sentences. Information redundancy, diversity and coherence are the main factors affecting linguistic quality and overall responsiveness.

Input: *DocumentSentences , Matrix U, Matrix V, MaxSummarySentences M, redundancy_limit*
Output: *Set of Summary Sentences (SummarySentences).*
Set SummarySentences =∅, k=1, N_k=0
While *NumberOfSentence in SummarySentences< M*
Find the index of Max value (L) in V_k row.
For *each (Sent in SummarySentences)*
{
If (similarity(Sent, $Sent_L$) <redundancy_limit)
{
Add $Sent_l$ to SummarySentences
Delete V_l form V^T
Set N_k= N_k +1
Get p,q,r in U_k, T={$term_p$,$term_q$,$term_r$}
T_0=T ∩ $Sent_L$, T=T-T_0
Remove $Sent_L$ from DocumentSentences list
}
}
While *T ≠ ∅*
If N_k<3 and NumberOfSentence in SummarySentences< M
Find the index of the Next Max value (L) in V_k row.
For each (Sent in SummarySentences)
{
If (similarity(Sent, $Sent_L$) <redundancy_limit)
{
Add $Sent_l$ to SummarySentences
Delete $V_l$form V^T
Set N_k= N_k +1
T_0=T ∩ $Sent_L$, T=T-T_0
Remove $Sent_L$ from DocumentSentences list
}
}
Else
Set T ≠ ∅
Endwhile
Set k=k+1
Endwhile
Return *SummarySentences*

Fig. 4.1. Algorithm for Suggested model

Table 1: Scores of Human Evaluation of the Proposed Approach

Method	Scores					Mean
	0	1	2	3	4	
	V. Poor	Poor	Fair	Good	V. Good	
Root	0.00%	14.29%	85.71%	0.00%	0.00%	1.8571
Stem	0.00%	28.57%	0.00%	71.43%	0.00%	2.4286
Word	14.29%	28.57%	0.00%	42.86%	14.29%	2.1431

ISSN 2348 – 7968

Table 2: ROUGE results with LDC datasets

Method	ROUGE-1	ROUGE-2	ROUGE-S4	ROUGE-SU4
RPOS	0.483191631	0.37292383	0.372542987	0.342770777
WPOS	0.368213499	0.280878777	0.260479323	0.300205989
SPOS	0.27464243	0.250382939	0.131031473	0.144389557

5.2.2 Automatic Evaluation

In this section, we use the Arabic specialist evaluation for a text summarization as ideal summaries. Also we have used F scores for evaluation purposes, as they represent both precision and recall measures, for different matches: ROUGE-1, ROUGE-2, ROUGE-S4, and ROUGE-SU4. Standard summarization evaluation using ROUGE route with POS (RPOS), stem with POS (SPOS) and word with POS (WPOS) is show in table (2).

In addition, the AN4 and POS tagging are used to improve the LSA results and it they haves achieved better results than the approaches that hadn't used POS tagging. Table (3) shows a comparison between ROUGE results using route with POS (RPOS) and without POS (R) tagging.

Table 3: comaprision between ROUGE results using POS and without POS

Module Type	ROUGE-1	ROUGE-2	ROUGE-S4	ROUGE-SU4
RPOS	0.483192	0.372924	0.372543	0.342771
R	0.368213	0.280879	0.260479	0.300206

It has been observed that the method which used the Root pattern has the highest score among the others two which were based on Word and Stem patterns as indicated by in Table (2) and Table (3). Therefore, our model has achieved better results than the approach of Wang & Ma [6].

6 Conclusion

In this paper, the suggested model combines two techniques to improve the performance of Arabic text summarization. Such techniques are LSA and Arabic word morphology. It selects sentences for each topic and removes the duplicated sentences from the output text summarization. The experimental result shows that the suggested model improving the text summarization performance. It obtained higher ROUGE scores as indicated in section 5, table (2) and table (3). The future work will deal with ontology based document summarization.

References

[1] Ibrahim, A., Elghazaly, T., & Gheith, M. (2013). A Novel Arabic Text Summarization Model Based on Rhetorical Structure Theory and Vector Space Model. International Journal of Computational Linguistics and Natural Language Processing Vol 2 Issue 8 August, 480-485.

[2] Welgama, W. V. (2012). Automatic Text Summarization for Sinhala. Master Thesis, University of Colombo School of Computing, December.

[3] Al_Gaphari, G., Ba-Alwi, F. M., & Moharram, A. (2013). Text Summarization using Centrality Concept. International Journal of Computer Applications (0975 – 8887) Vol. 79 – No.1, October.

[4] El-Haj, M. (2012). Multi-document Arabic Text Summarization. PhD Thesis, University of Essex.

[5] Haboush, A., & Al-Zoubi, M. (2012). Arabic Text Summerization Model Using Clustering Techniques. World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741, Vol. 2, No. 3, 62 – 67.

[6] Wang, Y., & Ma, J. (2013). A Comprehensive Method for Text Summarization Based on Latent Semantic Analysis. Springer-Verlag Berlin Heidelberg, 395-400.

[7] El-Sayed, A., & El-Barbary, O. (2014). ARABIC DOCUMENT SUMMARIZATION USING FA FUZZY ONTOLOGY. International Journal of Innovative Computing, Information and Control, Vol. 10(No. 4), 1351-1367.

[8] Hadni, M., Lachkar, A., & Ala, S. (2014). MULTI-WORD TERM EXTRACTION BASED ON NEW HYBRID APPROACH FOR ARABIC LANGUAGE. Computer Science & Information Technology (CS & IT), pp. 109–120. doi:10.5121/csit.2014.4410.

[9] Deshpande, A. R., & Lobo, L. (2013, August). Text Summarization using Clustering Technique. International Journal of Engineering Trends and Technology (IJETT), 4(8), 3348-3351. doi:ISSN: 2231-5381.

[10] Dahale, 2014;

[11] ÖZSOY, M. G. (2011). TEXT SUMMARIZATION USING LATENT SEMANTIC ANALYSIS. Master Thesis, MIDDLE EAST TECHNICAL UNIVERSITY, FEBRUARY.

[12] Froud, H., Lachkar, A., & Ouatik, S. A. (2013). ARABIC TEXT SUMMARIZATION BASED ON LATENT SEMANTIC ANALYSIS TO ENHANCE ARABIC DOCUMENTS CLUSTERING. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.1, January 2013, 79-95.

[13] Kwaik, K. A. (2011). Automatic Arabic Text Summarization System (AATSS) Based on Semantic Feature Extraction. Master Thesis, Islamic University of Gaza, August 2011.

[14] HAMMO, B. (2011). A Hybrid Arabic Text Summarization Technique Based on Text Structure and Topic Identification. International Journal of Computer Processing Of Languages Vol. 23, No. 1, 39–65.

[15] Bawakid, A. (2011). Automatic Documents Summarization Using Ontology based Methodologies. PhD Thesis, UNIVERSITY OF BIRMINGHAM.