Hybrid Technique for Arabic Text Compression

Arafat Awajan[1] and Enas Abu Jrai[2]

[1] Princess Sumaya Unversity for Technology

Received: 9 December 2014 Accepted: 3 January 2015 Published: 15 January 2015

Abstract

Arabic content on the Internet and other digital media is increasing exponentially, and the number of Arab users of these media has multiplied by more than 20 over the past five years. There is a real need to save allocated space for this content as well as allowing more efficient usage, searching, and retrieving information operations on this content. Using techniques borrowed from other languages or general data compression techniques, ignoring the proper features of Arabic has limited success in terms of compression ratio. In this paper, we present a hybrid technique that uses the linguistic features of Arabic language to improve the compression ratio of Arabic texts. This technique works in phases. In the first phase, the text file is split into four different files using a multilayer model-based approach. In the second phase, each one of these four files is compressed using the Burrows-Wheeler compression algorithm.

Index terms— text compression, multilayer model text compression, morphological analysis, word-based compression, burrows-wheeler algorithm.

1 Introduction

ata compression is important for data transmission and data storage. It aims at reducing the size of data in order to improve the speed of transmission and reduce the size that is needed for the storage. Data compression techniques can be classified into two general categories: Lossy and Lossless techniques. Lossless techniques themselves can be classified into two main categories: statistical compression techniques and dictionary compression techniques [1], [2].

Text compression is a subfield of data compression. It focuses on compressing natural language texts as they occur in the real world. Text compression uses mainly the different features of natural languages to improve the compression ratio and performance. Research papers concerning natural language text compression have been published during the past three decades. Their main concern were European languages such as English, French and German [3], [4] [5]. Other languages such as Japanese and Chinese were subjects of this type of research, too [6]. Few studies and published research papers focused on the compressing of Arabic text.

Each type of compression technique has advantages and disadvantages. Dictionary-based techniques are fast, but they give smaller compression ratios. On the other hand, statistically based techniques provide high compression ratios but ignore the specificities of natural language texts. Arabic and other Semitic languages are complex and rich in terms of morphological features, where tens or hundreds of words can be derived from the same root. These morphological features can be exploited to improve the compressing ratio of Arabic texts [7]. In 2008, ?tujbe [8] showed that utilizing multiple compression techniques is a superior alternative to the classic single-compressor approach. Thus hybrid approaches that combine several of these techniques in order to obtain better compression ratio have been proposed.

Studies on Arabic text compression were limited despite the fact that Arabic is one of the major international languages. This work aims at developing new compression techniques based on the exploitation of morphological and grammatical features of Arabic language to present a hybrid paradigm that will be able to improve the compression ratio and performance and to produce a new representation of text that can be more appropriate for other applications such as information retrieval.

2 II.

3 Features of Arabic Language

An Arabic word is a series of alphabet letters and diacritical marks. Thirty-six characters are used in Modern Standard Arabic (MSA): 28 basic letters and eight diacritical marks. The diacritical marks, called TASHKEEL, are optional and in general are added above or below Arabic letters. Table 1 shows the different vowelization states of the Arabic word: fully vowelized, partially vowelized and unvowelized.

4 ???????? -???????

In Arabic language, a word may be derivative or non-derivative. A derivative word is generated from a basic Arabic root according to a predefined palette or template called morphological balances. Figure 1 shows an example of some words that are derived from the root ????? k-t-b which represent the concept 'writing'. The non-derivative words are mainly functional words and nouns borrowed from foreign languages. Stop words are words that have little semantic meaning. However, they are used to explain grammatical relationships between the words within a sentence. This class of words includes pronouns, prepositions, conjunctions and interjections. The number of stop words is limited, but their frequency is very high in natural texts. They represent nearly 40% of the total number of words in a text [9]. Table 2 shows the frequency of these words in real-world text that contains one million words taken from a collection of articles from newspapers and magazines.

The morphological analysis is one of the most important techniques used in natural language processing. Its objective to analyze words in order to decompose them into their original morphemes and identify their internal structure. In the case of Arabic words, a word may be decomposed into suffix, prefixes, root or stem. In the case of derivative words, the morphological analyzers may generate the morphological pattern used for the creation of the word in addition to the other components listed before. It is a key step for many applications of natural language processing systems [10], [11], [12].

5 Related Work

Three approaches to research on Arabic text compression can be found in the literature. The first approach considers general-purpose compression techniques and does not take into account the features of Arabic languages. Some of these techniques proceed at the level of characters [13]. They use the frequency of characters in order to replace the most frequent characters by short codes. Therefore, they are called statistical compression methods and are developed based on the Huffman compression technique and its variants. Other techniques look at strings in the text and put pointers to strings or substrings that have already appeared [14]; these techniques are called dictionarybased techniques and are developed in general based ???????? ?????? ?????ïº?”?? ??????? ?????? ???????? ???????? ????????? ???????? ??????? ???????? ??????? ??????? ????ïº?”?? on the Lempel-Ziv technique (LZ). The third category consists of techniques that work at the frequency of the character and its neighbouring characters to decide how a character will be encoded. Examples of the last category are Burrows-Wheeler Transform (BWT) and Prediction by Partial Matching (PPM). In 2005, Khafagy [15] presented a study analyzing the results of a variety of data compression techniques applied to both English and Arabic texts. The best compression ratio had been obtained by neural compression, followed by PPM and LZW variations and Huffman-based techniques. RLE gave the worst results.

The second approach to research on Arabic text compression uses the features of Arabic language to develop new compression techniques. These techniques use either the statistical features of the languages, such as the most frequent N-grams, or the morphological features and linguistics of the language to achieve a shorter representation of the text [16], [17]. The results of these techniques are in general very limited.

The third approach to research on Arabic text compression are hybrid techniques that use the features of Arabic language in addition to general-purpose data compression techniques such as Huffman in order to achieve better results. The combinations of these techniques leads to better results as shown in [18], [19].

IV.

6 Burrows-Wheeler Compression

Several studies have proved that the compression technique based on BWT provides good results in comparison with general-purpose compressors [20]; it achieves good compression ratios combined with high speed [21].

7 a) Burrows-Wheeler Algorithm

The BWT technique was invented by Michael Burrows and David Wheeler in 1994. It converts the original blocks of data into a format that is extremely well suited for compression, through a sequence of steps [1]. Figure 2 describes the steps of the BWT technique. The first step performs the Burrows-Wheeler transform (BWT), which is done by reading blocks of text with predefined size from input and processing each block to make it easier to code the data with a simple coder. The second step implements the Move to Front transformation (MTF) to transform the characters into a list of numbers. This technique does not compress data; its aim is to decrease the redundancy of letters. The third step applies RLE on the new text that has been produced in the previous step. RLE is one of the simplest compression techniques dealing with consecutive recurrent symbols

[21], which are encoded as a pair: the length of the string and the symbol itself. After these steps, we can apply and identify the compression technique. Usually arithmetic coding or adaptive Huffman technique is used. We have suggested the adaptive Huffman technique to apply in our work.

8 b) Burrows-Wheeler Algorithm And Arabic Language

Arabic language is rich in morphology. Several surface forms may be generated from the same root according to a predefined tempaltic pattern. The order of letters may change inside the derived words. For example, the word " ?"???? -"read" may change to " ?"????? -"read," ?????"? -"reader" or " ??????? -readable." This is unlike the English language, in which the origin of the word remains unchanged and the derivations are limited to adding suffixes at the end or the beginning of the word, for example, "read," "reads," "reader," "the reader" [22].

The BWT technique is very sensitive to the structure of the word, so derivative words are not suitable for compression by this technique. Therefore, we have suggested using one of the morphological analyzers as a pre-processing step to implement (BWT) on derivative words, using the root-pattern dictionaries technique guided by the proposed method of [23], [19]. The main idea of this technique is to replace derived words with index values for their roots and their standard pattern as shown in Figure 3. Then BWT technique is applied to these components to compress the text.

9 Multilayer Model

Awajan [19] provided a multilayer model for the analysis of fully vowelized, non-vowelized and partially vowelized Arabic text. It classifies the text into three categories of words: derived, functional words and other words (i.e. non-derivative words and words that the system fails to classify into one of the categories). His approach depends on searching to determine if the word is functional or not, and using two techniques to determine the derived word; the first technique applies the pattern-based algorithm, and the second uses the dictionary for patterns and roots. This approach attaches all prefixes and suffixes to the dictionary of patterns to decrease the duration of the morphological analysis.

Our aim in this work is to integrate more than one technique to compress Arabic texts, by taking advantage of the morphological features of Arabic language. The most important characteristic of a multilayered model from other analyzers is that it deals with all categories of texts and all categories of Arabic words including symbols and punctuation marks.

VI.

10 Hybrid Compression Technique

The proposed compression technique consists of two phases, as shown in Figure **??**. In the first phase, the multilayer model has been selected to analyze the text. This model employs several procedures to partition the incoming text into three layers that represent three categories of Arabic words: functional, derivative and non-derivative words. The first layer is used to store the index of the stop words instead of the original word. The second layer is used to store the index of the roots and the patterns instead of derivative words. The third layer represents the words that the system failed to classify into either of the first two layers. The fourth layer, called the mask, is used during the decoding stage, to reconstruct the original text from the decoding of other layers. Suitable compression techniques were applied to the different layers in order to maximize the compression ratio.

Figure **??** : The main steps of the hybrid compression approach In the second phase, the encoding phase, the BWT technique is applied for each layer. The mask layer contains the number "zero" to indicate the position of the word in the first layer. If it contains the number "one," this means the current word in the second layer; if it contains the number "two," this means the word in the third layer. For compression, this layer we have suggested represents each number as binary code, then reads one byte to store the data. Decompression processes for both approaches are completely opposite to the compression process. It works by decoding each layer independently using the appropriate decoder, then reconstructing the original text using the mask layer.

VII.

11 Experiments and Evaluation

The main idea for the multilayer model is to split a text into smaller linguistically homogeneous layers representing the main categories of words. To evaluate the multilayer with hybrid compression techniques, several experiences were conducted. The objective was to evaluate its performance and to compare different possible implementations mainly using BWT and LZW.

A set of different categories of Arabic texts (vowelized, partially vowelized, unvowelized) was collected from multiple Internet sources. They represent stories, holy text from the Qur'an and articles from BBC Arabia news. Compression ratio, defined as the ratio of the size of the compressed text to the size of the original text, is considered to evaluate the performances of the proposed compression technique.

Three tables are used. One for storing the stop words contained 127 of the most frequently occurring stop words extracted from a corpus representing the BBC and CNN Arabic news [24]. The other two tables were constructed to represent the roots and patterns. The roots table included 4,095 of the most commonly used three-letter words, where 376,167 word types are derived from the three-letter roots [9]. The patterns table consists of the 13,600 most used patterns [25]. The later table has two entries for each pattern. One entry represents the list of consonants (LC), and the other entry represents the list of diacritics (LD) as shown in Table 3. 4 presents the compression ratio obtained at the level of the three layers using LZW and BWT compression techniques. BWT was the best technique to compress all the layers. Compression ratio for first layer was 50% when BWT was applied, 83% when LZW was applied. Compression ratio for the second layer was 54%, 75% for BWT and LZW, respectively, and for the third layer was 41%, 49% for BWT and LZW, respectively. Table 5 shows results of encoded data and size of the compressed files using LZW and BWT. These results have shown that the compression ratios are better when BWT is used with the multilayer model. On the other hand, the proposed hybrid technique for compressing Arabic texts achieved good results compared to single text data compression.

12 Conclusion

A hybrid technique for compressing Arabic texts has been developed. It integrates the multilayer model of Arabic texts with BWT. This technique relies on exploiting the morphological features of Arabic language to improve the performance of BWT, where the multilayer model was integrated with BWT. This approach gives a better compression ratio than [1]

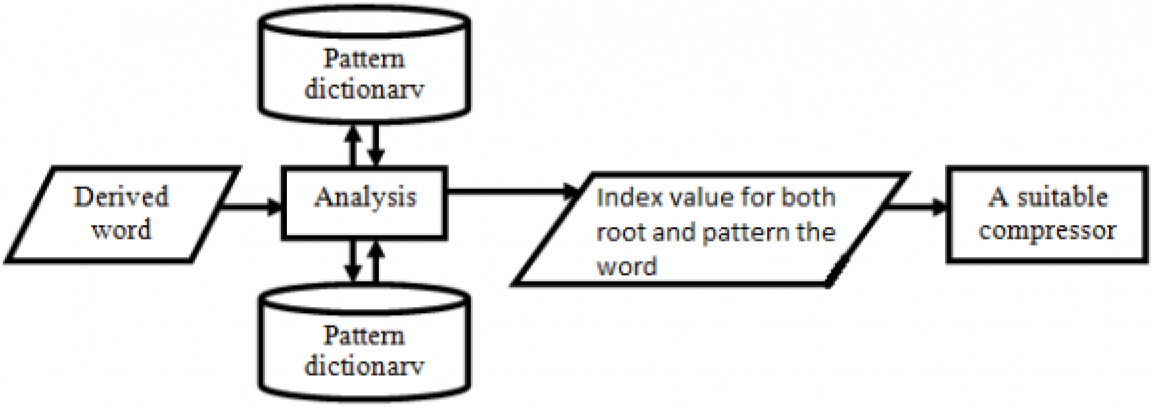
1

Figure 1: Figure 1 :

2

Figure 2: Figure 2 :

[1] Global Journal of Computer Science and Technology (C) Volume XV Issue I Version I © 2015 Global Journals Inc. (US) Year 2015



3

Figure 3: Figure 3 :

1

Vowelization States	Examples
Fully vowelized words	????? ???? ???? ??? ? ??? -??? ? ???? ???? ??? ? ???
Partially vowelized words	?????? ????? -?????? ? ???
Unvowelized words	

Figure 4: Table 1 :

2

Partially vowelized stop words		Unvowelized stop words	
Word	Frequency	Word	Frequency
?ï»?"??	292,396	????	322,239
????	269,200	?ï»?"??	301,895
???	120,060	????	132,635
?????	108,252	???	130,809
????	89,027	?????	119,639
????	83,027	?????	115,842
III.			

Figure 5: Table 2 :

3

Pattern	List of Consonants (LC)	List of Diacritical Marks (LD)
???? ? ??? ? ??ï»?"? ???? ????	????**?*?	?? ?? ? ? ??
? ??? ? ??? ? ??ï»?"? ???? ????	????***?	?? ?? ?? ?? ??
? ??? ???? ? ??ï»?"? ???? ????	????***??	? ? ? ?? ?? ??

Figure 6: Table 3 :

Figure 7: Table

4

Algorithm	First Layer	Second Layer	Third Layer
LZW	0.83	0.75	0.49
BWT	0.50	0.54	0.41

Figure 8: Table 4 :

5

Text Category	BWT	LZW	Multilayer with LZW	Multilayer with BWT
Vowelized	0.31	0.30	0.24	0.23
Unvowelized	0.35	0.32	0.23	0.26
Partially Vowelized	0.33	0.32	0.30	0.25
Average	0.33	0.31	0.26	0.25
VIII.				

Figure 9: Table 5 :

[Lourdusamy and Shanmugasundaram ()] ‘A Comparative Study Of Text Compression Algorithms’. R Lourdusamy , S Shanmugasundaram . *International Journal of Wisdom Based Computing* 2011. 1 (3) p. .

[Alasmer et al. ()] ‘A Comparison between English and Arabic Text Compression’. Z M Alasmer , B M Zahran , B A Ayyoub , M A Kanan . *Journal of Contemporary Engineering Sciences* 2013. 6 (3) p. .

[Teahan et al. ()] ‘A Compressionbased Algorithm for Chinese Word Segmentation’. J Teahan , R Mcnab , H Witten . *Computer Journal of Computational Linguistics* 2000. 26 (3) p. .

[Moronfolu and Oluwade ()] ‘An enhanced LZW text compression algorithm’. D Moronfolu , Oluwade . *Afr. J. Comp. & ICT* 2009. 2 (2) p. .

[Soudi, V. Bosch, G. Neuman (ed.) ()] *Arabic Computational Morphology*, Soudi, V. Bosch, G. Neuman (ed.) 2007. New York: Springer.

[Al-Sughaiyer and Al-Kharashi ()] ‘Arabic Morphological Analysis Techniques: A Comprehensive Survey’. A Al-Sughaiyer , I A Al-Kharashi . *Journal of the American Society for Information Science and Technology* 2004. 55 (3) p. .

[Omer and Khatatneh ()] ‘Arabic Short Text Compression’. E Omer , K Khatatneh . *Journal of Computer Science* 2010. 6 (1) p. .

[Khafagy ()] *Arabic Text Data Compression*, M A M Khafagy . 2005. Zagazig University (PhD thesis)

[Awajan ()] ‘Arabic Text Preprocessing for the Natural Language Processing Applications’. Awajan . *Arab Gulf Journal of Scientific Research* 2007. 25 (4) p. .

[Saad ()] *Arabic-Corpora*, M Saad . `http://sourceforge.net/projects/ar-text-mining/files` 2011. 2013.

[Available ()] `http://www.cs.cmu.edu/~guyb/realworld/compression.pdf` *Available*, 2013.

[Wiseman and Gefner ()] ‘Conjugation-based Compression for Hebrew Texts’. Y Wiseman , I Gefner . *Computer Journal of ACM Transactions on Asian Language Information Processing* 2007. 6 (1) p. .

[Altarawneh and Altarawneh ()] ‘Data Compression Techniques on Text Files: A Comparison Study’. H Altarawneh , M Altarawneh . *International Journal of Computer Applications* 2011. 26 (5) p. .

[Hasan ()] ‘Data Compression using Huffman based LZW Encoding Technique’. R Hasan . *International Journal of Scientific & Engineering Research* 2011. 2 (1) p. .

[Ghwanmeh et al. ()] ‘Efficient data compression scheme using dynamic Huffman code applied on Arabic language’. S Ghwanmeh , R Al-Shalabi , G Kanaan . *Journal of Computer Science* 2006. 2 p. .

[Abel (2003)] *Improvements to the Burrows-Wheeler Compression Algorithm: After BWT Stages*, Abel . 2003. March 2013. (Available:www.juergenabel.info/Preprints/Pr eprint_After_BWT_Stages)

[Pauw and Schryver (2008)] ‘Improving the Computational Morphological Analysis of a Swahili Corpus for Lexicographic Purposes’. G D Pauw , G.-M D Schryver . *The 13th International Conference of the African Association for Lexicography*, (Republic of South Africa) July 2008. p. .

[Blelloch ()] *Introduction to Data Compression*, G E Blelloch . 2010. p. . Computer Science Department Carnegie Mellon University

[Akman et al. ()] ‘Lossless Text Compression Technique Using Syllable Based Morphology’. H Akman , S Bayindir , Z Ozleme , Akin , Sanjay Misra . *The International Arab Journal of Information Technology* 2011. 8 (1) p. .

[Daoud ()] ‘Morphological Analysis and Diacritical Arabic Text Compression’. M Daoud . *The International Journal of ACM Jordan* 2078-7952. 2011. 1 (1) p. .

[Awajan ()] ‘Multilayer Model for Arabic Text Compression’. Awajan . *The International Arab Journal of Information Technology* 2011. 8 (2) p. .

[Sawalha ()] *Open-source Resources and Standards for Arabic Word Structure Analysis: Fine Grained Morphological Analysis of Arabic Text Corpora*, M S Sawalha . 2011. The University of Leeds

[?tujbe ()] *Practical data compression, Master’s thesis*, V ?tujbe . 2008. Bratislava. Commenius University

[Published by the Arab League Educational, Cultural and Scientific Organization ()] *Published by the Arab League Educational, Cultural and Scientific Organization*, `http://www.reefnet.gov.sy/ed4-2.htm` 2013. (Arabic Language Derivation and Morphological System)

[Jurafsky and Martin ()] *Speech and Language Processing, 2nd*, D Jurafsky , J H Martin . `http://www.cs.-colorado.edu/~martin/SLP/Updates/1.pdf` 2008. 2013. Prentice Hall.

[Radescu ()] ‘Transform methods used in lossless compression of text files’. R Radescu . *Romanian Journal of Information Science and Technology* 2009. 12 (1) p. .