A High Performance Optoelectronic Machine for Automated Arabic-English Parallel Corpus Creation and for Text Mining Processing

Samy S. A. Ghoniemy and Omar H. Karam

Abstract—In this paper a parallel optoelectronic computer architecture is proposed for large-scale parallel corpus, full text search and text mining applications, while achieving high speed and high performance and utilizing the parallel processing nature of optics. An optical technique is presented that aims to automatically construct an Arabic-English corpus through web mining. In addition, it is shown how Arabic morphology and Arabic synonymous can be used to extract the required knowledge from Arabic texts based on converting the Arabic text into a semi-structured database. The proposed system is modeled and implemented using a two-dimensional space optics symbolic correlator to perform the comparison operations of the expert system. The proposed optical engine is able to perform 3D processing for inferring conclusions to queries by representing page contents in a 2D space using mapped-template logic. A hardware realization for these systems based on commercially available components is discussed. The performance analysis of the proposed system shows its remarkably improved processing performance with respect to processing speeds and both recall and precision measures.

Index Terms—Optical computing, parallel corpus, text mining.

I. INTRODUCTION

With silicon based technologies for logic already approaching the physical limits of switching speed [1], [2], Optics may have answers for these limitations. In addition, due to continuing globalization there is also much interest in multi-language collections. The recent availability of machine translation systems is, in that context, an important development. Multi-language text mining is much more complex than it appears. In addition to differences in character sets and words, text mining makes intensive use of statistics as well as the linguistic properties (such as conjugation, grammar, senses or meanings) of a language. There are many basic assumptions about capitalization and tokenization that work for one language and that would not work for other languages. When text mining techniques are used on non-English data collections additional challenges have to be addressed. All these reasons are require the development of new computer techniques/systems to help control and harvest huge amounts of information.

Inherent features of optical components make them ideal for transferring planes comprising millions of data points simultaneously and thus achieving effective multi-dimensional input/output and parallel processing at the same time. [3]-[12]. The optical bi-stabilities recently developed with switching times in the order of Femtoseconds [13], FS, compared with milliseconds in electronics, have increased the possibility of using optics for the practical realization of new parallel optical computers for optical information processing especially in the field of Knowledge base systems (KBSs) and text mining as they have reached the Terabyte level [14]. The text mining involves processes such as information retrieval, text analysis, information extraction, clustering, categorization, visualization, and machine learning [15]. These require powerful processing methods and involve extensive parallel processing.

Cross-language information retrieval (CLIR) is another highly required application for retrieving information written in a language different from the language of the user's query. 43.4% of the web content is in English while the English web population does not exceed 26.8% [16]. Therefore, CLIR has many useful applications such as searchers with a limited active vocabulary, but good reading comprehension in a second language (e.g. English), might prefer to issue queries in their most fluent language [17]. A large-scale parallel corpus plays an important role in CLIR by providing the primary training data for the statistical translational models [18]. Most of the previous work on parallel texts has been conducted on a few manually constructed parallel corpora such as the work published by Canadian Hansard Corpus and Linguistic Data Consortium (LDC) [19].

The main problems in building parallel corpora are accuracy and processing time. Since such systems are always real time systems they still require higher processing speed. Implementation of parallel unification for high-speed matching is currently under investigation [14]. Herein lies the potential of an optical implementation of optical text mining machines that are compatible with electronics to exploit the strengths of both.

The main objectives of the system proposed in this paper are: i) constructing the optically-based parallel corpus (OPC) automatically through web mining, and ii) preparing the developed OPC to be used as training material for the translation model. The proposed system will be used for English-Arabic parallel texts to construct the corpus and can be extended to include other language pairs in a similar way.

The remainder of the paper is organized as follows: In Section II, the proposed optical parallel corpus is presented.
The architectural design of the overall text mining and parallel corpus machine and the optoelectronic design of each subsystem are discussed in Section III. In Section IV, the hardware realization of the proposed system is illustrated. Section V presents the experimental results and the performance evaluation of the proposed system. Finally, the summary and conclusions are in Section VI.

II. PROPOSED OPTICAL PARALLEL CORPUS

The proposed optical mining system functionalities and operation sequence are illustrated in Fig. 1. The proposed English-Arabic parallel corpus system is intended to be an online real time system which starts after the user issues its query in the form of an http request (URL on a specified host). Issuing this request starts the optical host crawling algorithm which crawls the host for all pages that are written in either Arabic or English by executing a language detector algorithm which identifies a list of parallel language-pair page patterns and stores them in a structured file and filters out all other pages in the domain as shown in Fig. 2. The output from this algorithm is a set of candidate pairs. The next step is the execution of the contents optical extractor which filters out the HTML tags and other scripts and keeps only the clean page contents. Finally, content-based matching is performed to calculate the parallelism similarity between each candidate pair using an English-Arabic dictionary. The output of this system is an English-Arabic parallel corpus that is well-aligned at paragraph level with completely clean texts. This approach is straightforward, fully automated, and easy to port to any other language pair.

III. DESIGN OF TEXT MINING AND PARALLEL CORPUS MACHINE

The proposed hybrid optoelectronic architecture is shown in Fig. 3. It is capable of processing data and performing text searches optically and in real time. In this architecture, all the Web documents are considered to be stored in banks of large optical disks. The output light is then spatially modified using optical elements that form the Pattern Generation Unit (PGU). When a user issues a request (query), the participating Web pages in the specified domain are retrieved and the PGU then forms an appropriate beam distribution. The optical data are then processed by the optical data processor and the result is projected on a photodetector array.

Fig. 1. The proposed optical mining system.

Fig. 2. The proposed optical host crawling process
B. Relationship Matrices for Indexer Implementation

Since relationship matrices are processed via a matrix-vector product, a learning process can be realized simply using rules operation (matrix–matrix multiplication) and implemented optically using the above proposed optical correlator as shown in Fig. 4 [22], [23].

In this figure, an input light beams representing a set of symbols is applied to the system using a 1-D SLM. Each light beam is spread horizontally using the cylindrical lens L1 across the filter mask whose transmittances are related to the values in the relationship matrix. This mask can be implemented using a 2-D SLM. The light is then focused using the cylindrical lens L2 into a linear photodetector array that produces the output vector which represents the result of the matrix-vector inner product. Using the same methodology we can implement the matrix-matrix inner product if we used 2-D SLMs in the input and output instead of 1-D SLMs.

C. Optoelectronic Comparison Unit for Web Pages Matching and Similarity Measure

In this section we adopted the work published by Guilfoyle [23] in which optical comparisons can be performed based on the exclusive-or (EX-OR) primitive using dual-rail logic. Since both the value of the bit and its complement are needed for the comparison, each n-bit word will be represented by 2n light beams (i.e. the 4-bit word 1101 will become 10-10-01-10). Using this method, a 00 combination corresponds to a “don’t-care” character while the 11 combination always produces a “Not-Equal” result. The coding scheme for the two logical values, 1 and 0, can be light and no light respectively. The proposed optoelectronic comparison unit is shown in Fig. 5.

IV. HARDWARE REALIZATION OF TEXT MINING AND PARALLEL CORPUS MACHINE

To implement a relationship matrix form of the Web pages contents, the optical mapping templates approach is adopted. These templates are binary transmission masks (in which transparent squares represent logic 1’s and the logic 0’s are represented as opaque background squares). The first few character positions on SLM-2 will permanently contain the character strings for New-Paragraph, New-Sentence and New-Word (we used the indicator “ palavra” to refer to the New-Word character). The remaining length will hold the search argument(s). In Fig. 7, white squares represent logic “1’s”, while others represent logic “0’s”.

In this figure an input vector \( N_i \), is vertically expanded...
along the y-axis (to form an array, each row of which equals \( N_j \)), and then optically overlaid with the mapping template using imaging optics (SL), then a 2-D OSLM performs a 2-D logical AND operation. The resulting output \( N_e \) is viewed along the rows of the output 2-D photodetector array which gives the query solution.

Using this conceptual design of the optical processor that is presented in Section III, a complete optical processing machine including optical crawler, optical indexer, optical text mining unit, and mapped-template logic shown in Fig. 10 is shown in Fig. 8. The hardware design of the proposed optical processor consists of 2 E-SLMs, 2 O-SLMs, 4 spherical imaging lenses (L1, L2, L3, L4), 5 beam splitters (BS1, BS2, BS3, BS4, BS5), one optical switch/ shutter (O-SW), and a 2-D photodetector array.

![Fig. 8. The complete prototype of the mapping templates-based optical parallel corpus and text mining machine](image)

**V. EXPERIMENTAL RESULTS**

The proposed architecture is capable of performing many different kinds of text search operations. The optical part constantly performs comparisons while the electronic part is responsible for writing the necessary search arguments, manipulating and interpreting the signals generated by the optical text processor, and passing the results to the user. The following is a description of the implementation of various text operations.

**A. Count the Number of Occurrences of a Word in a Document**

As an example, the search argument ◘Egypt◘, is loaded on SLM-2. The Query Resolver is notified about the arrival of a new document at the OTP and resets a counter (N) to 0. Every new occurrence of “Egypt” will cause an increment of the counter (N) until the end of the document is reached.

This will be used for the document parallelism matching in the parallel corpus as follow:

Content-based matching between any two parallel documents must contain some token pairs that are exact translations of each other. These token pairs are known as translational token pairs. For example, in the two sentences:

- "Egypt marks the 38th anniversary of the October victory" (مصر تشهد الـ38 جدولة اكتوبر النصر" in Arabic)
- " مصر تشهد اكتوبر "

the translational token pairs are: ("Egypt" - "مصر"), ("anniversary" - "纪 niệm"), ("October" - "اكتوبر"), ("victory" - "نصر"). Therefore, a simple method to calculate parallelism for a pair of documents is to search them for translated token pairs, calculate their number as mentioned above, and then use the number of translated pairs found as a value of parallelism similarity. However, a pair in which the position of the two translational token pairs is far from each other is rarely to be a correct translation pair. In the approach proposed here, a reliable threshold \( \theta_d \) is set to conclude whether a pair of documents is a translation pair or not. Thus, for each English document, the Arabic documents are scanned until one pair with similarity exceeding \( \theta_d \) is found. The measure of similarity between a pair of documents (A, E) represented by \( \theta_d \) can be determined empirically as:

\[
\text{Sim}(A,E) = \frac{2N}{\sum \text{number of tokens}_{A,E}} \tag{1}
\]

where \( N \) is the number of translational token pairs found between A and E. It is assumed that the difference in position between a good pair of paragraphs varies from (1) to (-1). Therefore, the number of translation pairs \( N \) between two documents is based on the total number of translation pairs between paragraphs \( n_k \). Each English paragraph \( p_{e,k} \) will be compared to its 3 Arabic paragraphs neighbors \( p_{a,k-1}, p_{a,k}, p_{a,k+1} \), and the one with the maximum value of translation pairs \( n_k \) together with \( p_{e,k} \) will form a translation pair of paragraphs.

\[
N = \sum n_k \tag{2}
\]

The technique should use an English-Arabic dictionary of stemmed words. Translational token pairs are found by first stemming the English words with Porter Stemming Algorithm [26], and then looked up in the dictionary for all possible Arabic words.

**B. Search for a Word and its Derivatives in a Document, Paragraph or Sentence**

The search argument will be ◘A◘. For example ◘Terrorism◘. As soon as a match is detected the current component will be marked and kept in the buffer for further processing. The arrival of a new component will be detected by keeping track of matches. For the Arabic part of the proposed machine the preprocessor module shown in Fig. 3 is used to divide the Arabic documents into paragraphs, each paragraph is analyzed to extract its keywords by applying the search for word algorithm, and then it extracts the root of each keyword using a morphological module. The output of this module is semi-structured documents in the form of an indexed database. An example is shown in Fig. 9. The first level uses the extracted roots to point to the extracted keywords of a certain root. In the second level index, keywords in turn point to the extracted paragraphs.

The morphological module extracts the root of the each keyword in the issued query on the SLM-1. OSLM-2 then generates all the possible words that have the same root. The optical text miner then generates a group of these derivatives, which have the same semantic features as the original keyword; then the query resolver adds the selected keywords to the keywords list. This will increase the number of resultant paragraphs yielding better recall measures. An Arabic dictionary supported with semantic tags is
The proposed model is tested on an Arabic text book about the history of the Prophet Mohammad (PBUH) named "الريحق المختوم". Information retrieval systems are usually compared on the basis of the "quality" of the retrieved document sets. This “quality” is quantified using two metrics, Recall (R = r/K) and Precision (P = r/N), where r is the number relevant and retrieved paragraphs, N is the total number of retrieved paragraphs, and K is the total number of paragraphs in the answer key.

The above algorithm was applied using ten different queries and for each query both the recall and precision were measured when the queries are issued with only keywords, with keywords and derivatives with the same semantic features, and with keywords, synonyms and derivatives with the same semantic features. The overall measured performance in each case and their comparison with fully electronic based results [28] are shown in the Table I.

<table>
<thead>
<tr>
<th>The other system</th>
<th>Keywords only</th>
<th>Keywords + derivatives</th>
<th>Keywords + synonyms + derivatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>80%</td>
<td>66%</td>
<td>95%</td>
</tr>
<tr>
<td>Precision</td>
<td>95%</td>
<td>90%</td>
<td>95%</td>
</tr>
<tr>
<td>Average recall</td>
<td>95%</td>
<td>90%</td>
<td>92%</td>
</tr>
<tr>
<td>Average precision</td>
<td>92%</td>
<td>100%</td>
<td>83%</td>
</tr>
</tbody>
</table>

The extreme advantage of the proposed system is the processing time. In our proposed machine the measured average processing time is approximately $32 \times 10^{10}$ word search operation per second.

VI. CONCLUSION

This paper presented a model and an architectural design for an optoelectronic-based parallel corpus and text mining system based on mapped-template logic and template matching in the optical domain, making full use of the parallelism of optics. The proposed architecture exploits the inherent parallelism features of optics for high speed
database responses and text mining processing. A two-dimensional space optical symbolic correlator was presented based on the VanderLugt correlator to perform the matching operations. Also proposed was an optoelectronic comparison unit that consists of optoelectronic logic, a photodetector array, and the necessary lenses to perform various text operations that can perform approximately $8 \times 10^8$ 32-bit character comparisons per second.

The proposed parallel corpus creation process is executed through mining the web. It is started by classifying English and Arabic web pages, and then creating candidate parallel pairs of documents by filtering them according to their similarity using mapped-template logic and template matching in the optical domain. Finally, the technique measured the parallelism similarity between these candidate pairs. The experimental results of the text mining using an Arabic morphological analyzer and synonyms showed a remarkable performance enhancement in both Precision and Recall measures as well as the superiority of the processing time using the proposed optical system over its corresponding electronic systems. The realization of the proposed optical processor using available optical elements was also addressed. The performance evaluation results showed that the proposed systems are valid, reasonable, and promising for parallelization and speed up of search, indexing and matching.

REFERENCES


Omar H. Karam received his B.Sc. in Electrical Engineering (Communications and Electro-physics) from Alexandria University and his M.Sc. and Ph.D. degrees in Computer Engineering from North Carolina State University. He has extensive professional experience in ICT management and is the author or co-author of more than 50 publications in international scientific journals and conferences. Dr. Karam is currently an Associate Professor and the Program Director of the Faculty of Informatics and Computer Science at the British University in Egypt (BUE). He is on secondment from the Faculty of Computer and Information Sciences, Ain Shams University, Egypt. Prior to the BUE, Dr. Karam was the Director of the Egyptian Universities Network (EUN). His current research interests are Data Mining and Warehousing, Interconnection Networks, Optical Computing and IOT in Education.