Journal of King Saud University – Computer and Information Sciences xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Parallel hardware for faster morphological analysis

Issam Damaj [a,*], Mahmoud Imdoukh [a], Rached Zantout [b]

[a] Department of Electrical and Computer Engineering, American University of Kuwait, P.O. Box 3323, Safat 13034, Kuwait
[b] Department of Electrical and Computer Engineering, Rafik Hariri University, P.O. Box 10, Mechref, Damour, Chouf, 2010, Lebanon

ARTICLE INFO

Article history:
Received 22 May 2017
Revised 5 July 2017
Accepted 11 July 2017
Available online xxxx

Keywords:
Morphological analysis
NLP
Performance
Hardware design
FPGAs

ABSTRACT

Morphological analysis of Arabic language is computationally intensive, has numerous forms and rules, and intrinsically parallel. The investigation presented in this paper confirms that the effective development of parallel algorithms and the derivation of corresponding processors in hardware enable implementations with appealing performance characteristics. The presented developments of parallel hardware comprise the application of a variety of algorithm modelling techniques, strategies for concurrent processing, and the creation of pioneering hardware implementations that target modern programmable devices. The investigation includes the creation of a linguistic-based stemmer for Arabic verb root extraction with extended infix processing to attain high-levels of accuracy. The implementations comprise three versions, namely, software, non-pipelined processor, and pipelined processor with high throughput. The targeted systems are high-performance multi-core processors for software implementations and high-end Field Programmable Gate Array systems for hardware implementations. The investigation includes a thorough evaluation of the methodology, and performance and accuracy analyses of the developed software and hardware implementations. The developed processors achieved significant speedups over the software implementation. The developed stemmer for verb root extraction with infix processing attained accuracies of 87% and 90.7% for analyzing the texts of the *Holy Quran* and its Chapter 29 – *Surat Al-Ankabut*.

1. Introduction

Natural Language Processing (NLP) is a rapidly developing field. Developments in NLP are, at the larger part, driven by the fact that the world has turned into a small village equipped with advanced transportation, media, and communication. In 2016, the total number of social network users worldwide is estimated to be around 2.2 billion with a global penetration of 31%. In the US, 78% of the population has social network profiles. Indeed, it is expected that the total number of users will grow to 2.5 billion in 2018 (Statista, 2017). In such a modern, connected, and global society, people still use different languages. The idea of having all the people use one language has been proven by practice to be impossible. Even in science fiction, as in Star Trek, a universal translator is presented to solve the challenge of automatic translation among languages. At present, NLP active areas of research are machine translation, information retrieval, text categorization, sentiment mining, to name a few (Nirenburg and Wilks, 2000; Yang et al., 2012; Agarwal and Mittal, 2016).

A Morphological Analyzer (MA) is a core subsystem in NLP applications. MAs work on identifying words being used in a specific language and study the internal structure of these words (Hamalawy, 2009). Morphology can be defined as producing a word from another by changing it to fit a new meaning. Furthermore, Morphological analysis is usually complicated, computationally intensive, and intrinsically parallel. Arabic language is well-known for having rich morphology, complex word formation and patterns (Al-Sughaiyer and Al-Kharashi, 2004).

1.1. Background

The rich morphology of Arabic enables the language to develop and grow. Arabic morphology (الإشتقاق) is categorized into small, large, larger, and the largest morphologies. The small morphology derives a word from a root but keeps similarities between the two

[*] Corresponding author.
E-mail addresses: idamaj@auk.edu.kw (I. Damaj), s00024916@alumni.auk.edu.kw (M. Imdoukh), Zantoutrn@rhu.edu.lb (R. Zantout).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2017.07.003
1319-1578/© 2017 The Authors. Production and hosting by Elsevier B.V. on behalf of King Saud University.

words in their pronunciation and meaning; such as (علم, عالم) that translates to (root: science, derived: scientist). The remaining morphologies comprise exchanging letters of the root, producing a word from another by changing one or more letters, and producing a word by combining a group of words. By far, the mostly used type of morphology in Arabic is the small morphology (Al-Sughaiyer and Al-Kharashi, 2004; Soudi et al., 2007; Dahdah, 1995; Rajhi, 1979; Hamandi et al., 2006; Al-Khalifah, 1996).

Arabic words are grouped into the three main categories of Nouns, Verbs, and Particles. The Nouns and Verbs consist of subcategories in which the main form of the word changes per its position/role in the sentence and some of the other words in the sentence. For example, verbs are categorized per time and structure. The verb times are past, present or future, while structures may be either proper or defective. The difference between words in subcategories can be as subtle as the presence/absence of a vowel or as clear as the addition/removal of letters to/from the word.

In Arabic language, small morphology can act on verbs (Al-Sughaiyer and Al-Kharashi, 2004). The roots of Arabic words have traditionally been considered to consist of three or four letters. Like other languages, in Arabic, letters can be added to the beginning (prefixes) and/or to the end (suffixes) of the root. However, in Arabic, letters can be added inside a root (infixes). Infixes complicate the morphological analysis of the Arabic language because the infix letters can also be letters in the root. In addition, a Verb in Arabic has different forms if the subject or object is masculine or feminine. Also, differences in the forms can exist if the sentence refers to one person, a group of two, or a group of more than two (Al-Khalifah, 1996). Nevertheless, Arabic verbs follow specific forms. For example, the root (درس, Study) maps to the ternary pattern (فعل), while the verb (يدرس) maps to (يفعل). Here, the addition of the prefix (ي) derives the present tense of the verb.

In Arabic verbs, there are seven letters that can be added to the beginning of the root as prefixes; these letters are grouped in the Arabic word (فسألتني). The nine letters that can be added to the end of the root as suffixes are grouped in the Arabic word (ايتهكمون). The five letters that can be added to the inside of a root (infixes) are grouped in the Arabic word (اتوني); infixes have a more complicated set of rules with focus on the three vowel letters ا, و, and ي (Hamandi et al., 2006). In Table 1, example morphological variations from a verb root is presented with focus on the applied change on form and meaning. The same patterns shown in Table 1 can produce similar variations for the verb root (صحب, Accompany) to produce يصحبون and يصاحب in the same tense and form as in the patterns يفعلون and يفاعل. As compared to verbs, Arabic nouns are more complex due to the large variety of forms, irregularities, and number.

1.2. Related work

Al-Sughaiyer and Al-Kharashi (2004) presented a comprehensive survey of Arabic morphological analysis techniques that comprises definitions, classifications, approaches, algorithms, etc. In the literature, different MA techniques and algorithms target Arabic verbs specifically. Yaghi et al. (2003) presented a verb generation system that enables word-to-root and root-to-word lookups. The system uses coding techniques to compactly store and efficiently access a dictionary of Arabic words. Yagi and Harous (2003) details the development of a database of generated stems that supports their verb-matching system. The developed database is of multipurpose and can be used to identify stems, classify morpho-semantic and morpho-syntactic templates, and support a variety of applications. Boubas et al. presented the use of genetic algorithms to generate an MA for Arabic verb. The investigations comprised developing general verb patterns and then applying them to derive morphological rules. The reported results reflect highly accurate matching capabilities (Boubas et al., 2011).

A variety of algorithms have been developed to perform morphological analysis of verbs based on an input word, such as, sliding window algorithms (El-Affindi, 1998), word decomposition using algebraic algorithms (El-Affindi, 1991), and literals generation using permutations of the input word letters (Al-Shalabi and Evens, 1998). Other analyzers attempt to extract the root of a verb by manipulating infixes and prefixes (Hamandi et al., 2002, 2006; Khoja, 2017; Khoja and Garside, 1999; Larkey and Connell, 2006; Saad et al., 2010; Larkey et al., 2002; Asaad and Abbod, 2014; Boudlal et al., 2011; Hegaz and Elsharkawi, 1986; Hlal, 1987; Abu Shquier and Alhawiti, 2015; Sembok and Ata, 2013; Abu-Errub et al., 2014; Al-Bawab and Al-Tayyan, 1998). The extracted stems are then validated against a list of standard Arabic roots. In (Hamalawy, 2009), such a manipulation of affixes is classified under Linguistic-based (LB) stemmers. LB stemmers are usually accurate but require the preparation of lists for matching and validation. If a stemmer doesn't include analysis of infixes and root extraction, it is referred to as a light stemmer (Larkey et al., 2002).

LB stemmers attracted the attention of many researchers and enabled the development of a variety of MAs. The focus of the presented MAs is accuracy; however, almost all contributions highlight the essential need for high-performance processing. Khoja (2017) and Khoja and Garside (1999) presented the development of an LB Arabic MA algorithm; the algorithm analyzes a word by removing definite articles, prefixes, suffixes, stop words, and then matches the remaining word against the pattern of the same length to extract the root. Khoja stemmer is widely used in the literature with a reported accuracy of 96% (Khoja, 2017). Asaad and Abbod (2014) presented an extraction approach that removes prefixes, suffixes, infixes, and attempts to identify the root. The presented approach includes making a second attempt to identify an unidentified root through a procedure that handles weak, hamzated (that has the letter Hamza 'ء'), eliminated-long-vowels, and two-letter geminated words. The proposed approach produced somewhat improved accuracy over Khoja stemmer. Boudlal et al. (2011) presented an Arabic MA system that extracts roots depending on the context within a sentence. A Hidden Markov Models approach was used, where the observations are the words and the possible roots represent the hidden states. The approach achieved an accuracy of 94% in targeting the NEMLAR Arabic writing corpus with its 500,000 words. LB stemmers have a long history of reported contributions since 1985 (Al-Sughaiyer and Al-Kharashi, 2004); this includes the approaches of Hegaz and Elsharkawi (1986), Hlal (1987), Abu Shquier and Alhawiti (2015), Sembok and Ata (2013), Abu-Errub et al. (2014), El-Affindi (1998, 1991), Al-Bawab and Al-Tayyan (1998), Khoja (2017) and Khoja and Garside (1999), to name but a few. Indeed, all the reported contributions above are developed as software implementations.

Table 1
Morphological variations of the verb Study (درس) with changes on form and meaning.

Addition	Location	Morph	Pattern	Meaning	(Tense, Form)
(ي)	Prefix	يدرس	يفعل	One is studying	(Present, Singular)
(ي, ون)	(Prefix, Suffix)	يدرسون	يفعلون	Many are studying	(Present, Plural)
(ي, ا)	(Prefix, Infix)	يدارس	يفاعل	One is studying with others	(Present, Singular)

Please cite this article in press as: Damaj, I., et al. Parallel hardware for faster morphological analysis. Journal of King Saud University – Computer and Information Sciences (2017), http://dx.doi.org/10.1016/j.jksuci.2017.07.003