

Collaborative Report DC-98-06

**Proceedings**  
**of the Prague Stringology Club Workshop '98**

*Edited by Jan Holub and Milan Šimánek*

August 1998

Department of Computer Science and Engineering  
Faculty of Electrical Engineering  
Czech Technical University  
Karlovo nám. 13  
121 35 Prague 2  
Czech Republic

## **Program Committee**

Jun-ichi Aoe, Maxime Crochemore, Jan Holub, Bořivoj Melichar, Václav Snášel,  
Bruce W. Watson

## **Organizing Committee**

Martin Bloch, Jan Holub, Martin Rýzl, Milan Šimánek, Zdeněk Troníček

## Table of contents

<b>Preface</b>	<b>v</b>
<b>A Fast Morphological Analysis Using the Extended AC Machine for Oriental Languages</b> <i>by Kazuaki Ando, Kimihiro Iwasaki, Masao Fuketa and Jun-ichi Aoe</i>	<b>1</b>
<b>The Longest Restricted Common Subsequence Problem</b> <i>by Gabriela Andrejková</i>	<b>14</b>
<b>Implementation of DAWG</b> <i>by Miroslav Balík</i>	<b>26</b>
<b>Exact String Matching Animation in Java</b> <i>by Christian Charras and Thierry Lecroq</i>	<b>36</b>
<b>Local Prediction for Lossless Image Compression</b> <i>by Ahmad Daaboul</i>	<b>44</b>
<b>On the All Occurrences of a Word in a Text</b> <i>by O.C. Dogaru</i>	<b>51</b>
<b>A Highly Parallel Finite State Automaton Processor for Biological Pattern Matching</b> <i>by Glen Herrmannsfeldt</i>	<b>58</b>
<b>Dynamic Programming for Reduced NFAs for Approximate String and Sequence Matching</b> <i>by Jan Holub</i>	<b>73</b>
<b>Validating and Decomposing Partially Occluded Two-Dimensional Images (Extended Abstract)</b> <i>by Costas S. Iliopoulos and James F. Reid</i>	<b>83</b>
<b>Application of Sequence Alignment Methods to Multiple Structural Alignment and Superposition</b> <i>by Arthur M. Lesk</i>	<b>95</b>
<b>Approximate String Matching by Fuzzy Automata</b> <i>by Václav Snášel</i>	<b>101</b>
<b>The Factor Automaton</b> <i>by Milan Šimánek</i>	<b>102</b>
<b>Directed Acyclic Subsequence Graph</b> <i>by Zdeněk Troníček and Bořivoj Melichar</i>	<b>107</b>
<b>An Early-Retirement Plan for the States</b> <i>by Bruce W. Watson and Richard E. Watson</i>	<b>119</b>



## Preface

This collaborative report contains the proceedings of the Prague Stringology Club Workshop '98 (PSCW'98), held at the Department of Computer Science and Engineering of Czech Technical University in Prague on September 3–4, 1998. The workshop was preceded by PSCW'96 which was the first action of the Prague Stringology Club and by PSCW'97. The proceedings of PSCW'96 and PSCW'97 were published as collaborative reports DC-96-10 and DC-97-03, respectively, of Department of Computer Science and Engineering and are also available in the postscript form at Web site with URL: <http://cs.felk.cvut.cz/psc>. While the papers of PSCW'96 were invited papers, the papers of PSCW'97 and PSCW'98 were selected from the papers submitted as a response to a call for papers. The papers in this proceedings are alphabetically ordered by the authors.

The PSCW aims at strengthening the connection between stringology (the computer science on strings and sequences) and finite automata theory. The automata theory has been developed and successfully used in the field of compiler construction and can be very useful in the field of stringology too. The automata theory can facilitate the understanding of existing algorithms and the developing of new algorithms.

Jan Holub and Milan Šimánek, editors



# A Fast Morphological Analysis Using the Extended AC Machine for Oriental Languages<sup>1</sup>

Kazuaki Ando, Kimihiro Iwasaki, Masao Fuketa and Jun-ichi Aoe

Department of Information Science & Intelligent Systems  
University of Tokushima  
2-1 Minami-Josanjima-Cho  
Tokushima-Shi 770-8506  
Japan

e-mail: {ando, aoe}@is.tokushima-u.ac.jp

**Abstract.** This paper presents a fast morphological analysis for oriental languages by extending an Aho and Corasick's pattern matching machine. Our method is a simple and efficient algorithm to find all possible morphemes in an input sentence and in a single pass, and it stores the relations of grammatical connectivity of adjacent morphemes into the output functions. Therefore, the costs of checking connections between the adjacent morphemes can be reduced by using the connectivity relations. Furthermore, the construction method of the relations of grammatical connectivity is described. Finally, the proposed method is verified by a theoretical analysis, and an experimental estimation is supported by the computer simulation with a 100,267 words dictionary. From the simulation results, it turns out that the proposed method was 49.9% faster (CPU time) than the traditional trie approach. As for the number of candidates for checking connections, it was 25.5% less than that of the original morphological analysis.

**Key words:** morphological analysis, oriental language, dictionary lookup, trie structure, AC machine, grammatical connectivity

## 1 Introduction

An intelligent natural language interfaces enable users to communicate with the computer in English, Japanese or other human languages. Morphological analysis [ABE86, AKI94, KUR94, LEE97, MAR94, MOR96, SAN94] is the first step of natural language processing in the applications of natural language interfaces such as Information Retrieval [AOE91], Database Queries [KAP84], Expert Systems and so on. In general, the morphological analysis means segmentations of the input sentence into words (morphemes) and attachments of part-of-speech to them. Therefore, although morphological analysis for European languages, especially for English, plays only a minor role in a natural language processing system, in the analysis of oriental languages

---

<sup>1</sup>This work was supported by the Grant-in-Aid of the Ministry of Education, Science and Culture, Japan.

such as Japanese, Chinese and Korean it plays an important role because oriental languages are agglutinative languages, that is the language do not have explicit word boundaries between the words [ABE86, AKI94, KUR94, MAR94, MOR96, SAN94].

The procedure of morphological analysis of oriental languages consists of two steps. The first is to detect all possible morphemes, which are the smallest meaningful units, in a given input sentence. The second is to find the possible connections between adjacent morphemes by using a connection cost or probability based on the grammaticality [ABE86, AKI94, SAN94]. In the first step, the morphological analysis involves a large number of dictionary lookup. In general, a well-known technique for dictionary lookup is to use a trie structure [AOE91, AOE96, KUR94]. The trie is a tree structure in which each transition corresponds to a key character in the given keys set and common prefixes of keys can be shared. Therefore, the trie can search all keys made up from prefixes in an input string without the need of scanning the structure more than once. However, it is not so effective to use the trie for the morphological analysis [KUR94, MAR94, MOR96]. In order to detect all possible substrings in a given input sentence, the dictionary access must be tried repeatedly at each character position in the input sentence. Therefore, some characters may be scanned more than once for different starting positions and the number of dictionary accesses is increased. In the second step, the morphological analysis checks grammatical connectivity between adjacent words in order to find all possible connections [ABE86, SAN94]. This grammatical connectivity can be easily checked by using a grammatical table [ABE86]. However, this process requires considerable cost to check the grammatical connectivity, because it includes some checks of unnecessary connections, for example, checking connection between NOUN and CONJUGATION, since many words as part of speech have different grammatical interpretations. In order to achieve a fast morphological analysis, the mentioned problems should be solved.

This paper proposes a high speed morphological analysis of oriental languages by extending a pattern matching machine based on Aho and Corasick machine (called AC machine) [AHO75]. The proposed method is a simple and fast algorithm to find all possible substrings in an input sentence, and during only a single scan. Moreover, since the proposed method stores relations of grammatical connectivity of adjacent words into the output functions, the cost of checking connections between the adjacent words can be reduced by using the connectivity relations.

In the following sections, our ideas are described in detail. In Section 2, we describe the dictionary lookup method using a trie structure for the morphological analysis. Section 3 presents the high speed morphological analysis by extending the AC machine. Section 4 shows the theoretical analysis, and the experimental evaluations verified by the computer simulations with a 100,267 words dictionary. Finally, the results are summarized and the future research is discussed.

## **2 Dictionary Lookup Method using Trie in the Morphological Analysis**

Morphological analysis of oriental languages is very different from that of English [ABE86, AKI94, KUR94, MAR94, MOR96, SAN94], because the languages do not have explicit word boundaries between the words as shown Fig. 1. Therefore, in order