Regular Expressions for Finite-State Syntactic Description

Lauri Karttunen
Rank Xerox Research Centre
6 Chemin de Maupertuis
F-38240 Meylan
France
lauri.karttunen@Grenoble.RXRC.Xerox.com

This paper surveys the methods for constructing finite-state analyzers from regular expressions by intersection and composition. The expressions are of two types: (1) constraining expressions allow some tag or pattern only in a specific context; (2) marking expressions introduce tags and brackets to identify a phrase as an instance of some regular pattern. The restriction operator, $\Rightarrow$, originally introduced for two-level morphology is a useful tool for syntactic constraints. Marking expressions can be constructed with the help of a special replace operator, $\theta \to \tau$. Such replace expressions yield transducers that unambiguously introduce tags or bracketing under a left-to-right, longest match regimen. We will define these concepts and illustrate their application to tokenization, filtering, and phrasal analysis.