Incremental construction of finite state automata and their utilization in natural language processing

Jan Kouřil, ikouril@fit.vutbr.cz, 19.10.2011

The main reason for incremental construction of finite automata is to have computer-assisted correction of written texts. Classical finite state automata and transducers are very important in this topic. New type of automaton, called finite state automaton with final transitions is constructed by modifying current definition. Computer-assisted correction of texts is based on deterministic acyclic finite state automata with final transitions.

Finite state automaton is a device, that can be in one of finite number of states. In certain condition it can change its state. Automaton starts at initial state. Some of finite automaton's states are final. The input is a sequence of symbols. When automaton reads the whole sequence of symbols and ends up in one of final states, it accepts the input. When automaton can change its state without reading any symbol, it is said it has ε-transitions. The symbol ε stands here as an empty symbol. An automaton, that has zero ε-transitions is called ε-free. An ε-free automaton, where there is at most one transition labeled with the same symbol is called deterministic finite state automaton. An automaton is acyclic, when it is not possible to get to the same state twice, when following transitions. For every automaton exists so called minimal automaton, that is automaton with the least number of states.

It is possible to form an alternative definition of finite state automaton, where there are no final states, but final transitions. String of symbols is accepted by finite automata with final transitions, when the last transition is in a set of final transitions. For every acyclic deterministic finite automaton, there is a transformation to this second type of automata. There is always less, or equal number of states in minimal finite state automaton with final transitions, than in a classical finite state automaton, that is minimal. Usually it is less. This occurs when there are two states in a classical finite state automaton, that has the same right language (symbols which has not been already read). These states could be merged together in a deterministic acyclic finite state automaton with final transitions.

Finite state transducers are automata having transitions labeled with two symbols. One symbol represents input and another one output. These are typical Mealy's automata. Transducers are said to translate, or transduce strings. They are used as devices computing some function. Deterministic transducers are called subsequential. For these transducers can be translating function computed by deterministically following single path in a transducer. Subsequential transducers are faster in recognition tasks, but usually take more space.

Transition graph of final automaton is a tree with initial state as root and all leaves are final. If some finite automaton is build as a tree, it is obvious, that many subtrees are isomorphic. Minimal automata is constructed so only one copy of isomorphic trees is kept. For an automata to be minimal, it is necessary and sufficient, that right languages of all states are different. Incremental construction is used for creating morphological dictionaries from regular dictionaries. Morphological dictionary is actually a tree incrementally constructed after adding every new word from a dictionary.

Main utilization of the topic lies in spelling correction, restoration of diacritics and morphology.

Acknowledgment
This work is based on paper dedicated to this topic written by Jan Daciuk.