Restricted context-free grammars with conditions on symbol occurrence

Jiří Skácel
iskacelj@fit.vutbr.cz

October 10, 2016

Abstract

There are methods, like pumping lemma, that check whether a given language cannot be generated by a context-free grammar (CFG). To confirm that a language is generated by CFG, it is necessary to find its grammar which can be difficult for complicated languages. This paper introduces a new method of specification of context-free languages with comparatively simpler CFG defining superset of desired language and additional conditions on symbol occurrence in its derivation tree to remove undesired words.

Every sentence in the final language must meet at least one of the given conditions. Each condition consists of two sets of symbols:

- symbols required to occur somewhere in the derivation tree and
- symbols forbidden to occur anywhere.

Presented algorithm creates a new grammar which defines only those sentences that can be generated by the given input grammar and that comply to at least one additional condition. This new grammar is still context-free.

The algorithm in principle replaces symbols of each rule with extended symbols that consist of the symbol itself and two sets. First set contains forbidden symbols and the second set contains symbols required to occur in a derivation sub-tree rooted in this symbol. For every rule where a required symbol from left-hand side does not occur on the right-hand side, we introduce new rules which propagate this symbol to required set of a non-deterministically chosen symbol on the right-hand side. On the other hand if a required symbol is on the right-hand side, this propagation stops. If a forbidden symbol is spotted, instead of normal rule we add one with a blocking non-terminal, for which there is no rule and so the generation fails. At the leaves of derivation tree, there are only extended terminals with empty required set expected. Those can be simply transcribed to their original form.

This paper shows the formal algorithm and proof of combining of CFG G and additional conditions I into a new CFG G_I generating such a language $L(G_I) \subseteq L(G)$ that for every word in $L(G_I)$ there exists a derivation tree which comply to at least one condition in I.