T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

# 論文 / 著書情報
# Article / Book Information

| | |
|---|---|
| 題目(和文) | 先読みを持つ文法とその微分に関する研究 |
| Title(English) | Studies on Grammars with Lookahead and Their Derivatives |
| 著者(和文) | 宮嵜貴之 |
| Author(English) | Takayuki Miyazaki |
| 出典(和文) | 学位:博士(理学),<br>学位授与機関:東京工業大学,<br>報告番号:甲第12512号,<br>授与年月日:2023年9月22日,<br>学位の種別:課程博士,<br>審査員:南出 靖彦,田中 圭介,増原 英彦,鹿島 亮,脇田 建 |
| Citation(English) | Degree:Doctor (Science),<br>Conferring organization: Tokyo Institute of Technology,<br>Report number:甲第12512号,<br>Conferred date:2023/9/22,<br>Degree Type:Course doctor,<br>Examiner:,,,, |
| 学位種別(和文) | 博士論文 |
| Category(English) | Doctoral Thesis |
| 種別(和文) | 論文要旨 |
| Type(English) | Summary |

Powered by T2R2 (Science Tokyo Research Repository)

（博士課程）
Doctoral Program

# 論 文 要 旨

THESIS SUMMARY

| | | | |
|---|---|---|---|
| 系・コース：<br>Department of, Graduate major in | 数理・計算科学　系<br>数理・計算科学　コース | 申請学位（専攻分野）：<br>Academic Degree Requested | 博士<br>Doctor of　（　理学　） |
| 学生氏名：<br>Student's Name | 宮嵜 貴之 | 審査員主査：<br>Chief Examiner | 南出 靖彦 |

要旨（英文 800 語程度）
Thesis Summary（approx.800 English Words ）


Regular expressions and context-free grammars (CFGs) are central concepts in formal language theory and various important results are known. They are also important in practice and are applied in regular expression libraries and parsers. However, there are some differences between these concepts in practice and in theory. The aim of this study is to develop a theory of practical regular expressions and CFGs.

Unlike theoretical regular expressions, many regular expression libraries use regular expressions with lookahead (RELAs). Lookahead is a constraint on the following string; a positive lookahead "&e" and a negative lookahead "!e" indicate that the following string starts and does not start with "e", respectively. For a practical example, by searching for "Java!(Script)" in regular expression libraries, we can find a string "Java" that is not a prefix of "JavaScript". Lookahead do not consume letters and affect only the success or failure of the matching. It is important because it can represent intersection and complement, which are difficult to represent in regular expressions. The theoretical study of RELA was initiated by Morihata. He showed that RELA e can be converted to a DFA whose size is a double exponential of the size of e. This indicates that adding lookahead is a better choice than adding intersection and complement because the number of states of DFA obtained from regular expressions with complement is not bounded by any elementary function.

Some parsers use grammars with extensions that are not present in CFGs. Parsing expression grammars (PEGs) introduced by Ford are a prime example of such a grammar. CFG and PEG are widely used in parsing. PEG supports positive lookahead, negative lookahead, and ordered choices, and it does not support alternation. A recursive descending parsing algorithm is given for PEG, which defines the language of PEG. The algorithm runs for ordered choice "e_1 / e_2", first trying "e_1" and only trying "e_2" if it fails in "e_1". Ordered choice suppresses branching to "e_2" and removes ambiguity. PEG is unambiguous, has expressiveness beyond deterministic context-free languages and has a linear time recognition algorithm. Grammars that can handle CFG with positive lookahead and PEG with alternation without left recursion are proposed. However, there was no grammar that extended both CFG and PEG.

We develop a theory of regular expressions and context-free grammars with lookahead by using derivatives as a main tool. In Chapter 2, "Regular Expressions with Lookahead", we study regular expressions with lookahead. We introduce a language with lookahead, which is a set of pairs of strings, and we show the set of languages with lookahead forms a Kleene algebra with tests. We reformulate the semantics of RELA by a language with lookahead. We then develop derivatives of RELA and a conversion from RELA to DFA using derivatives. In Chapter 3, "Context-Free Grammars with Lookahead and Limit Semantics", we introduce context-free grammars with lookahead (CFGLAs). To accommodate negative lookahead, we give the limit semantics with the limit of iterations from the empty set. This technique is based on the technique used in Boolean grammars, which are grammars with intersection and complement. We show that the grammars are an extension of both CFGs and PEGs. We also show that the language class is closed under union, intersection, complement, marked concatenation, and marked Kleene star. In Chapter 4, "Least Interval Semantics of Context-Free Grammars with Lookahead", we introduce the least interval semantics of CFGLA. First, we point out the problems of the limit semantics: the semantics is not defined for all grammars and not preserved by substitution and replacement. In addition, it has a problem for a computational complexity. To overcome the problems, we introduce the least interval semantics of CFGLA with the least fixed point of pairs of lower and upper bounds. This is related to the study of the well-founded model of Boolean grammars. This semantics preserves many of good properties of limit semantics and also allows us to obtain a binary normal form. In Chapter 5, "Derivatives of Context-Free Grammars with Lookahead", we introduce derivatives of CFGLA. We develop a recognition algorithm for CFGLA by combining

derivatives of CFG and RELA. The recognition algorithm runs in O(n^3 |G|) time and O(n^2 |G|) space for an input string of the length n and a given CFGLA G under the least interval semantics. This result is based on the fact that recognition is reduced to the membership of the empty string, and the membership of the empty string under least interval semantics is reduced to the computation of the Fitting model of the propositional logic program.
We consider that practical aspects of the contribution include the following. Formal definitions of RELA and CFGLA allow us to distinguish between specifications and bugs. The recognition algorithm by derivatives is easy to implement and useful for prototype implementations. The introduction of CFGLA allows us to use CFG and PEG mixed together.

備考 ： 論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。
Note : Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).

注意：論文要旨は、東工大リサーチリポジトリ(T2R2)にてインターネット公表されますので、公表可能な範囲の内容で作成してください。
Attention: Thesis Summary will be published on Tokyo Tech Research Repository Website (T2R2).