Fast parsing for Boolean grammars: a generalization of Valiant's algorithm

Alexander Okhotin*

Department of Mathematics, University of Turku, Turku FI–20014, Finland
Academy of Finland
alexander.okhotin@utu.fi

Abstract. The well-known parsing algorithm for the context-free grammars due to Valiant ("General context-free recognition in less than cubic time", *Journal of Computer and System Sciences*, 10:2 (1975), 308–314) is refactored and generalized to handle the more general Boolean grammars. The algorithm reduces construction of the parsing table to computing multiple products of Boolean matrices of various size. Its time complexity on an input string of length n is $O(BMM(n) \log n)$, where $BMM(n)$ is the number of operations needed to multiply two Boolean matrices of size $n \times n$, which is $O(n^{2.376})$ as per the current knowledge.

1 Introduction

Context-free grammars are the universally accepted mathematical model of syntax, and their status is well-justified. On the one hand, their expressive means are *natural*, in the sense whatever they define is intuitively seen as the syntax of something. On the other hand, they can be implemented in a variety of efficient algorithms, including a straightforward cubic-time parser, as well as many practical parsing algorithms working much faster in special cases.

The main idea of the context-free grammars is *inductive definition* of syntactically correct strings. For example, a context-free grammar $S \to aSb \mid \varepsilon$ represents a definition of the form: a string has the property S if and only if either it is representable as awb for some string w with the property S, or if it is the empty string. Note that the vertical line in the above grammar is essentially a disjunction of two syntactical conditions. *Boolean grammars*, introduced by the author [7], are an extension of the context-free grammars, which maintains the main principle of inductive definition, but allows the use of any Boolean operations to combine syntactical conditions in the rules. At the same time, they inherit the basic parsing algorithms from the context-free grammars, including the Cocke–Kasami–Younger [7] along with its variant for unambiguous grammars [10], the Generalized LR [8], as well as the linear-time recursive descent [9].

The straightforward upper bound on the complexity of parsing for Boolean grammars is the same as in the context-free case: $O(n^3)$, where n is the length of

* Supported by the Academy of Finland under grants 134860 and 218315.

the input string [7]. However, for the context-free grammars, there also exists an asymptotically faster parsing algorithm due to Valiant [12]: this algorithm computes the same parsing table as the simple Cocke–Kasami–Younger algorithm, but does so by offloading the most intensive computations into calls to a Boolean matrix multiplication procedure. The latter can be efficiently implemented in a variety of ways. Given two $n \times n$ Boolean matrices, a straightforward calculation of their product requires n^3 conjunctions and $(n-1)n^2$ disjunctions. An improved algorithm by Arlazarov et al. [2] reduces the number of bit operations to $O\big(\frac{n^3}{\log n}\big)$, which is achieved by pre-computing products of all bit vectors of length $\log n$ with certain submatrices. An asymptotically more significant acceleration is obtained by using fast algorithms for multiplying $n \times n$ numerical matrices, such as Strassen's [11] algorithm that requires $O(n^{2.81})$ arithmetical operations, or the algorithm of Coppersmith and Winograd [3] with the theoretical running time $O(n^{2.376})$. These algorithms can be applied to multiplying $n \times n$ Boolean matrices by calculating their product in the ring of residues modulo $n+1$ [1].

Taking a closer look at Valiant's algorithm, one can see that first the entire grammar is encoded in a certain semiring, then the notion of a *transitive closure* of a Boolean matrix is extended to matrices over this semiring, so that the desired parsing table could be obtained as a closure of this kind, and finally it is demonstrated that such a closure can be efficiently computed using Boolean matrix multiplication. This approach essentially relies on having two operations in a grammar, concatenation and union, which give rise to the product and the sum in the semiring. Because of that, Valiant's algorithm as it is cannot be applied to Boolean grammars.

This paper aims at refactoring Valiant's algorithm to make it work in the more general case of Boolean grammars. It is shown that using matrices over a semiring as an intermediate abstraction is in fact unnecessary, and it is sufficient to employ matrix multiplication to compute the concatenations only, with the Boolean operations evaluated separately. Furthermore, the proposed algorithm maintains one fixed data structure, the parsing table, and whenever the matrix is to be cut as per Valiant's divide-and-conquer strategy, the new algorithm only distributes the ranges of positions in the input string among the recursive calls. This leads to an improved parsing algorithm, which, besides being applicable to a larger family of grammars, is also better understandable than Valiant's algorithm, has a succinct proof of correctness and is ready to be implemented.

2 Boolean grammars

Let Σ be a finite nonempty set used as an *alphabet*, let Σ^* be the set of all finite strings over Σ. For a string $w = a_1 \ldots a_\ell \in \Sigma^*$ with $a_i \in \Sigma$, the *length* of the string is denoted by $|w| = \ell$. The unique *empty string* of length 0 is denoted by ε. For a string $w \in \Sigma^*$ and for every its partition $w = uv$, u is a *prefix* of w and v is its *suffix*; furthermore, for every partition $w = xyz$, the string y is a *substring* of w.

Any subset of Σ^* is a *language* over Σ. The basic operations on languages are the *concatenation* $K \cdot L = \{ uv \mid u \in K, \ v \in L \}$ and the Boolean set operations: union $K \cup L$, intersection $K \cap L$, and complementation $\overline{L}$. *Boolean grammars* are a family of formal grammars in which all these operations can be explicitly specified.

Definition 1. [7] *A Boolean grammar is a quadruple $G = (\Sigma, N, P, S)$, where Σ and N are disjoint finite non-empty sets of terminal and nonterminal symbols respectively; P is a finite set of rules of the form*

$$A \to \alpha_1 \& \ldots \& \alpha_m \& \neg\beta_1 \& \ldots \& \neg\beta_n, \tag{1}$$

where $m + n \geqslant 1$, $\alpha_i, \beta_i \in (\Sigma \cup N)^*$; $S \in N$ is the start symbol of the grammar.

If negation is not allowed, that is, $m \geqslant 1$ and $n = 0$ in every rule, the resulting grammars are known as *conjunctive grammars* [6]. If conjunction is also prohibited, and thus every rule must have $m = 1$ and $n = 0$, then the *context-free grammars* are obtained.

The intuitive semantics of a Boolean grammar is fairly clear: a rule (1) specifies that every string that satisfies each of the conditions α_i and none of the conditions β_i is therefore generated by A. However, formalizing this definition has proved to be rather nontrivial in the general case. In the case of conjunctive grammars (including the context-free grammars), the semantics can be equivalently defined by a least solution of language equations and by term rewriting. The definition by language equations carries on to Boolean grammars of the general form as follows.

A grammar is interpreted as a system of language equations in variables N, in which the equation for each $A \in N$ is

$$A = \bigcup_{A \to \alpha_1 \& \ldots \& \alpha_m \& \neg\beta_1 \& \ldots \& \neg\beta_n \in P} \left[\bigcap_{i=1}^{m} \alpha_i \ \cap \ \bigcap_{j=1}^{n} \overline{\beta_j} \right] \tag{2}$$

The vector $(\ldots, L_G(A), \ldots)$ of languages generated by the nonterminals of the grammar is defined by a solution of this system. In general, such a system may have no solutions (as in the equation $S = \overline{S}$ corresponding to the grammar $S \to \neg S$) or multiple solutions (with $S = S$ being the simplest example), but the below simplest definition of Boolean grammars dismisses such systems as ill-formed, and considers only systems with a unique solution; to be more precise, a subclass of such systems:

Definition 2. *Let $G = (\Sigma, N, P, S)$ be a Boolean grammar, let (2) be the associated system of language equations. Suppose that for every number $\ell \geqslant 0$ there exists a unique vector of languages $(\ldots, L_C, \ldots)_{C \in N}$ ($L_C \subseteq \Sigma^{\leqslant \ell}$), such that a substitution of L_C for C, for each $C \in N$, turns every equation (2) into an equality modulo intersection with $\Sigma^{\leqslant \ell}$.*

Then G complies to the semantics of a strongly unique solution, and, for every $A \in N$, the language $L_G(A)$ can be defined as L_A from the unique solution of this system. The language generated by the grammar is $L(G) = L_G(S)$.

This fairly rough restriction ensures that the membership of a string in the language depends only on the membership of shorter strings, which is essential for the grammars to represent inductive definitions.

Example 1. The following Boolean grammar generates the language $\{ a^m b^n c^n \mid m, n \geqslant 0, m \neq n \}$:

$$\begin{aligned} &S \to AB \& \neg DC \\ &A \to aA \mid \varepsilon \\ &B \to bBc \mid \varepsilon \\ &C \to cC \mid \varepsilon \\ &D \to aDb \mid \varepsilon \end{aligned}$$

The rules for the nonterminals A, B, C and D are context-free, and they define $L_G(AB) = \{ a^i b^n c^n \mid i, n \geqslant 0 \}$ and $L_G(DC) = \{ a^m b^m c^j \mid j, m \geqslant 0 \}$. Then the propositional connectives in the rule for S specify the following combination of the conditions given by AB and DC:

$$L(AB) \cap \overline{L(DC)} = \{ a^i b^j c^k \mid j = k \text{ and } i \neq j \} = \underbrace{\{ a^m b^n c^n \mid m, n \geqslant 0, m \neq n \}}_{L(S)}$$

Assuming Definition 2, every Boolean grammar can be transformed to an equivalent grammar in *binary normal form* [7], in which every rule in P is of the form

$$\begin{aligned} &A \to B_1C_1 \& \ldots \& B_nC_m \& \neg D_1E_1 \& \ldots \& \neg D_nE_n \& \neg \varepsilon \\ &\qquad\qquad (m \geqslant 1,\ n \geqslant 0,\ B_i, C_i, D_j, E_j \in N) \\ &A \to a \\ &S \to \varepsilon \quad \text{(only if } S \text{ does not appear in right-hand sides of rules)} \end{aligned}$$

In the general case, the transformation requires an exponential blowup in the size of the grammar.

An alternative, more general definition of the semantics of Boolean grammars will be presented in Section 7.

3 Simple cubic-time parsing

Let $G = (\Sigma, N, P, S)$ be a Boolean grammar in binary normal form, let $w = a_1 \ldots a_n$ be an input string. The simple cubic-time parsing algorithm constructs a table $T \in (2^N)^{n \times n}$, with

$$T_{i,j} = \{ A \in N \mid a_{i+1} \ldots a_j \in L_G(A) \}$$

for all $0 \leqslant i < j \leqslant n$. The elements of this table can be computed inductively on the length $j-i$ of the substring, starting with the elements $T_{i,i+1}$ each depending only on the symbol a_{i+1}, and continuing with larger and larger substrings, until the element $T_{0,n}$ is computed. The induction step is given by the equality

$$T_{i,j} = f\Big(\bigcup_{k=i+1}^{j-1} T_{i,k} \times T_{k,j}\Big),$$

where the function $f \colon 2^{N \times N} \to 2^N$ is defined by

$$f(R) = \{A \mid \exists A \to B_1C_1\& \ldots \&B_mC_m\&\neg D_1E_1\& \ldots \&\neg D_{m'}E_{m'} \in P : \\ (B_t, C_t) \in R \text{ and } (D_t, E_t) \notin R \text{ for all } t\}.$$

In total, there are $\Theta(n^2)$ elements, and each of them takes $\Theta(n)$ operations to compute, which results in a cubic time complexity.

The full algorithm can be stated as follows:

Algorithm 1 (Extended Cocke–Kasami–Younger [6,7]) *Let $G = (\Sigma, N, P, S)$ be a Boolean grammar in the binary normal form. Let $w = a_1 \ldots a_n$, where $n \geqslant 1$ and $a_i \in \Sigma$, be an input string. For all $0 \leqslant i < j \leqslant n$, let $T_{i,j}$ be a variable ranging over subsets of N. Let R be a variable ranging over subsets of $N \times N$.*

```
1: for $i = 1$ to $n$ do
2:     $T_{i-1,i} = \{ A \mid A \to a_i \in P \}$
3: for $\ell = 2$ to $n$ do
4:     for $i = 0$ to $n - \ell$ do
5:         $R = \varnothing$
6:         for all $k = i + 1$ to $i + \ell - 1$ do
7:             $R = R \cup (T_{i,k} \times T_{k,i+\ell})$
8:         $T_{i,i+\ell} = f(R)$
9: accept if and only if $S \in T_{0,n}$
```

The most time-consuming operation in the algorithm is computing the unions $R_{i,j} = \bigcup_{k=i+1}^{j-1} T_{i,k} \times T_{k,j}$, in which $R_{i,j}$ represents all concatenations BC that generate the substring $a_{i+1} \ldots a_j$ and the index k is a cutting point of this substring, with B generating $a_{i+1} \ldots a_k$ and with C generating $a_{k+1} \ldots a_j$. If each union is computed individually, as it is done in the above algorithm, then spending linear time for each $R_{i,j}$ is unavoidable. However, if such unions are computed for several sets $T_{i,j}$ at a time, much of the work can be represented as Boolean matrix multiplication. This is illustrated in the following example:

Example 2. Let $w = a_1a_2a_3a_4a_5$ be an input string and consider the partially constructed parsing table depicted in Figure 1, with $T_{i,j}$ constructed for $1 \leqslant i < j \leqslant 3$ and for $3 \leqslant i < j \leqslant 5$, that is, for the substrings $a_1a_2a_3$ and $a_3a_4a_5$ together with their substrings. Denote by $T^A_{i,j}$ the Boolean value indicating whether A is in $T_{i,j}$ or not. Then the following product of Boolean matrices

$$\begin{pmatrix} T^B_{0,2} & T^B_{0,3} \\ T^B_{1,2} & T^B_{1,3} \end{pmatrix} \times \begin{pmatrix} T^C_{2,4} & T^C_{2,5} \\ T^C_{3,4} & T^C_{3,5} \end{pmatrix} = \begin{pmatrix} X_{0,4} & X_{0,5} \\ X_{1,4} & X_{1,5} \end{pmatrix}$$

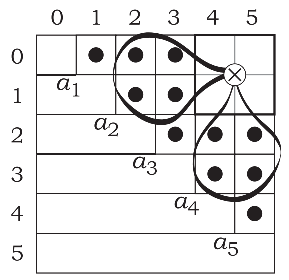

Fig. 1. Product of two Boolean matrices in Example 2.

represents partial information on whether the pair (B, C) should be in the following four elements: $\begin{pmatrix} R_{0,4} & R_{0,5} \\ R_{1,4} & R_{1,5} \end{pmatrix}$. To be precise, $X_{1,4}$ computes the membership of (B, C) in $R_{1,4}$ exactly; $X_{0,4}$ does not take into account the factorization $a_1 \cdot a_2a_3a_4$, which actually requires knowing whether C is in $T_{1,4}$; the element $X_{1,5}$ is symmetrically incomplete; finally, $X_{0,5}$ misses the factorizations $a_1 \cdot a_2a_3a_4a_5$ and $a_1a_2a_3a_4 \cdot a_5$, which can be properly obtained only using $T_{0,4}$ and $T_{1,5}$. In total, this matrix product computes 8 conjunctions out of 12 needed for these four elements of R.

Already in this small example, using one matrix product requires changing the order of computation of the elements $\{T_{i,j}\}$: the elements $T_{0,3}$ and $T_{2,5}$ need to be calculated before $T_{1,4}$. In the next section, the whole algorithm will be restated as a recursive procedure, which arranges the computation so that as much work as possible is offloaded into products of the largest possible matrices.

4 Parsing reduced to matrix multiplication

Let $w = a_1 \dots a_n$ be an input string. For the time being, assume that $n+1$ is a power of two, that is, the length of the input string is a power of two minus one; this restriction can be relaxed in an implementation, which will be discussed in the next section.

The algorithm uses the following data structures. First, there is an $(n + 1) \times (n + 1)$ table T with $T_{i,j} \subseteq N$, as in Algorithm 1, and the goal is to set each entry to $T_{i,j} = \{ A \mid a_{i+1} \dots a_j \in L(A) \}$ for all $0 \leqslant i < j \leqslant n$. The second table R has elements $R_{i,j} \subseteq N \times N$ each corresponding to the value of R computed by Algorithm 1 in the iteration $(\ell = j - i,\, i)$. The target value is $R_{i,j} = \{ (B, C) \mid a_{i+1} \dots a_j \in L(B)L(C) \}$ for all $0 \leqslant i < j \leqslant n$.

Initially, the elements of the tables are set as follows: $T_{i-1,i} = \{ A \mid A \to a_i \in P \}$ for all $1 \leqslant i \leqslant n$, and the rest of values of T are undefined; $R_{i,j} = \varnothing$. The rest of the entries are gradually constructed using the following two recursive procedures:

- The first procedure, *compute*(ℓ, m), constructs the correct values of $T_{i,j}$ for all $\ell \leqslant i < j < m$.

– The other procedure, *complete*(ℓ, m, ℓ', m'), assumes that the elements $T_{i,j}$ are already constructed for all i and j with $\ell \leqslant i < j < m$, as well as for all i, j with $\ell' \leqslant i < j < m'$; it is furthermore assumed that for all $\ell \leqslant i < m$ and $\ell' \leqslant j < m'$, the current value of $R_{i,j}$ is

$$R_{i,j} = \{ (B, C) \mid \exists k \, (m \leqslant k < \ell') : \; a_{i+1} \dots a_k \in L(B), \; a_{k+1} \dots a_j \in L(C) \},$$

which is a subset of the intended value of $R_{i,j}$.
Then *complete*(ℓ, m, ℓ', m') constructs $T_{i,j}$ for all $\ell \leqslant i < m$ and $\ell' \leqslant j < m'$.

– Matrix multiplication is performed by one more procedure, *product*(d, ℓ, ℓ', ℓ''), whose task is to add to each $R_{i,j}$, with $\ell \leqslant i < \ell + d$, and $\ell'' \leqslant j < \ell'' + d$, all such pairs (B, C), that $B \in T_{i,k}$ and $C \in T_{k,j}$ for some k with $\ell' \leqslant k < \ell' + d$. This can generally be done by computing $|N|^2$ products of $d \times d$ Boolean matrices, one for each pair (B, C).

Algorithm 2 (Parsing through matrix multiplication)

```
Main procedure:
for $i = 1$ to $n$ do
    $T_{i-1,i} = \{ A \mid A \to a_i \in P \}$
compute$(0, n+1)$
Accept if and only if $S \in T_{0,n}$
Procedure compute$(\ell, m)$:
if $m - \ell > 4$ then
    compute$(\ell, \frac{\ell+m}{2})$
    compute$(\frac{\ell+m}{2}, m)$
complete$(\ell, \frac{\ell+m}{2}, \frac{\ell+m}{2}, m)$
Procedure complete$(\ell, m, \ell', m')$, which requires $m - \ell = m' - \ell'$:
if $m - \ell > 1$ then                                          /* see Figure 2 */
    /* compute $\mathcal{C}$ */
    complete$(\frac{\ell+m}{2}, m, \ell', \frac{\ell'+m'}{2})$
    /* compute $\mathcal{D}_1$ */
    product$(\frac{m-\ell}{2}, \ell, \frac{\ell+m}{2}, \ell')$            /* $\mathcal{D}_1 \leftarrow \mathcal{B}_1 \times \mathcal{C}$ */
    complete$(\ell, \frac{\ell+m}{2}, \ell', \frac{\ell'+m'}{2})$
    /* compute $\mathcal{D}_2$ */
    product$(\frac{m-\ell}{2}, \frac{\ell+m}{2}, \ell', \frac{\ell'+m'}{2})$       /* $\mathcal{D}_2 \leftarrow \mathcal{C} \times \mathcal{B}_2$ */
    complete$(\frac{\ell+m}{2}, m, \frac{\ell'+m'}{2}, m')$
    /* compute $\mathcal{E}$ */
    product$(\frac{m-\ell}{2}, \ell, \frac{\ell+m}{2}, \frac{\ell'+m'}{2})$       /* $\mathcal{E} \leftarrow \mathcal{B}_1 \times \mathcal{D}_2$ */
    product$(\frac{m-\ell}{2}, \ell, \ell', \frac{\ell'+m'}{2})$            /* $\mathcal{E} \leftarrow \mathcal{D}_1 \times \mathcal{B}_2$ */
    complete$(\ell, \frac{\ell+m}{2}, \frac{\ell'+m'}{2}, m')$
else if $m \neq \ell'$ then
    $T_{\ell,\ell'} = f(R_{\ell,\ell'})$
```

The partition of the matrix in *complete*() is illustrated in Figure 2.

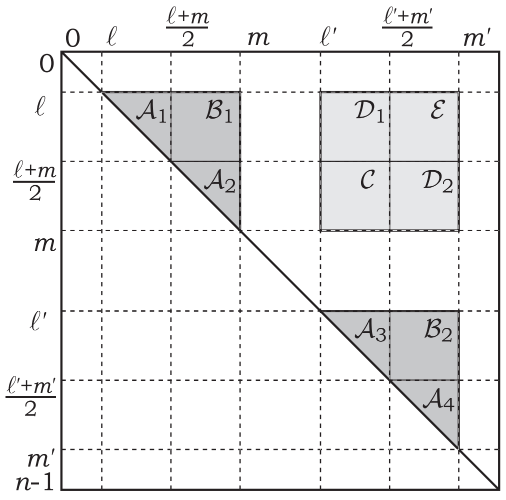

Fig. 2. Matrix partition in *complete*(ℓ, m, ℓ', m').

Lemma 1. *Let $\ell < m \leqslant \ell' < m'$ with $m - \ell = m' - \ell'$ being a power of two, and assume that $T_{i,j} = \{ A \mid a_{i+1} \dots a_j \in L(A) \}$ for all i and j with $\ell \leqslant i < j < m$, as well as for all i, j with $\ell' \leqslant i < j < m'$. Furthermore, assume that, for all $\ell \leqslant i < m$ and $\ell' \leqslant j < m'$,*

$$R_{i,j} = \{ (B, C) \mid \exists k \, (m \leqslant k < \ell') : \, a_{i+1} \dots a_k \in L(B), \, a_{k+1} \dots a_j \in L(C) \}.$$

Then complete(ℓ, m, ℓ', m') returns with $T_{i,j} = \{ A \mid a_{i+1} \dots a_j \in L(A) \}$ for all $\ell \leqslant i < m$ and $\ell' \leqslant j < m'$.

Lemma 2. *The procedure compute(ℓ, m), executed on ℓ and m with $m - \ell$ being a power of two, returns with $T_{i,j} = \{ A \mid a_{i+1} \dots a_j \in L(A) \}$ for all $\ell \leqslant i < j < m$.*

In order to determine the algorithm's complexity on an input of length $2^k - 1$, consider how many times the procedures *compute*() and *complete*() are called for subproblems of each size. For each $i \in \{1, \dots, k-1\}$, *compute*(ℓ, m) with $m - \ell = 2^{k-i}$ is called exactly 2^i times; *complete*(ℓ, m, ℓ', m') with $m - \ell = 2^{k-i}$ is called exactly $2^{2i-1} - 2^{i-1}$ times; *product*() is called for $2^{k-i} \times 2^{k-i}$ submatrices exactly $2^{2i-1} - 2^i$ times, and multiplies $O(|G|)$ pairs of Boolean matrices.

Theorem 1. *For every Boolean grammar G in binary normal form, Algorithm 2 constructs the parsing table for a string of length n in time $O(|G| \cdot BMM(n) \log n)$, where $BMM(n)$ is the time needed to multiply two $n \times n$ Boolean matrices. Assuming $BMM(n) = \Omega(n^{2+\varepsilon})$, the complexity is $\Theta(|G| \cdot BMM(n))$.*

5 Notes on implementation

The restriction on the length of the string being a power of two minus one is convenient for the algorithm's presentation, but it would be rather annoying for any implementation. This essential condition can be circumvented as follows. Let $w = a_1 \ldots a_n$ be an input string of any length $n \geqslant 1$. The algorithm shall construct a table of size $(n + 1) \times (n + 1)$, yet while doing so, it will imagine a larger table of size rounded up to the next power of two. Whenever a subroutine call is concerned entirely with the elements beyond the edge of the table, this call is skipped. The matrix products with one of the matrices split by the edge are changed to products of rectangular matrices fitting into the table. More details shall be presented in the full version of the paper.

Another question concerns the possible data structures for the algorithm. In general, not everything mentioned in the theoretical presentation of the algorithm would need to be computed for an actual grammar. First assume that the grammar is context-free. In this case, whenever a pair (B, C) is added to $R_{i,j}$, it will eventually make all nonterminals A with a rule $A \to BC$ be added to $T_{i,j}$. Accordingly, the data structure R is not needed, and all matrix multiplication procedures can output their result directly into the appropriate elements of T.

If the grammar is conjunctive or Boolean, there is a genuine need for using R, yet only for the rules involving multiple conjuncts. Simple context-free rules with a unique conjunct can be treated in the simplified way described above, with all matrix products being directly flushed into T. If there exists a rule $A \to BC \& \ldots$ with at least two conjuncts, or any rule $A \to \neg BC \& \ldots$, then all data about the pair (B, C) needs to be stored in R as described in the algorithm. This data shall be used in the calculation of f, which takes into account the complex rules.

With this optimization of the algorithm, the following data structures naturally come to mind:

- For each nonterminal $A \in N$, an $(n + 1) \times (n + 1)$ upper-triangular Boolean matrix T^A, with $T^A_{i,j}$ representing the membership of A in the set $T_{i,j}$. All matrix products computed in the algorithm shall have some submatrices of this matrix as the arguments.
- For every such pair $(B, C) \in N \times N$ that occurs in multiple-conjunct rules $A \to BC \& \ldots$ or is negated in any rule $A \to \neg BC \& \ldots$, the algorithm shall maintain an $(n + 1) \times (n + 1)$ upper-triangular Boolean matrix R^{BC}.

6 Generalized algorithm

The original Valiant's algorithm was presented in a generalized form, in which it computes a certain kind of closure of a matrix over a semiring. While the updated algorithm no longer uses any semiring, its computation can also be generalized to operations over abstract structures.

Let X and Y be two sets, let $\circ : X \times X \to Y$ be a binary operator mapping pairs of elements of X to elements of Y, let $\sqcup : Y \times Y \to Y$ be an associative and commutative binary operator on Y, and let $f : Y \to X$ be any function. Let $x = x_1 \ldots x_n$ with $x_i \in X$ be a sequence of elements of X and consider the matrix $T = T(x) \in X^{n \times n}$ defined by the following equations:

$$T_{i-1,i} = x_i$$
$$T_{i,j} = f\big(\bigsqcup_{k=i+1}^{j-1} T_{i,k} \circ T_{k,j}\big)$$

Theorem 2. *There is an algorithm, which, given a string $x = x_1 \ldots x_n$ of length n, computes the matrix $T(x)$ in time $O(BMM(n) \log n)$.*

In this generalized form, the algorithm can be applied to different families of grammars. For example, for context-free grammars in the binary normal form one can set $X = 2^N$, $Y = 2^{N \times N}$, $\circ = \times$, $\sqcup = \cup$, $x_i = \{\, A \in N \mid A \to a_i \in P \,\}$ and $f(y) = \{\, A \in N \mid \exists A \to BC \in P :\ (B, C) \in y \,\}$. For Boolean grammars, the only difference is in f, which has to take into account more complicated Boolean logic in the rules.

The same extended algorithm can be applied to probabilistic context-free grammars, as well as to the fuzzy generalization of Boolean grammars defined by Ésik and Kuich [4]. The next section presents one more application.

7 Application to the well-founded semantics

The *well-founded semantics* of Boolean grammars was proposed by Kountouriotis, Nomikos and Rondogiannis [5]. This semantics is applicable to every syntactically valid Boolean grammar, and defines a *three-valued language* generated by each nonterminal symbol.

Three-valued languages are mappings from Σ^* to $\{0, \frac{1}{2}, 1\}$, where 1 and 0 indicate that a string definitely is or definitely is not in the language, while $\frac{1}{2}$ stands for "undefined". Equivalently, three-valued languages can be defined by pairs (L, L') with $L \subseteq L' \subseteq \Sigma^*$, where L and L' represent a lower bound and an upper bound on a language that is not known precisely. A string in both L and L' definitely is in the language, a string belonging to neither of them definitely is not, and if a string is in L' but not in L, its membership is not defined. In particular, if $L = L'$, then the language is completely defined, and a pair $(\varnothing, \Sigma^*)$ means a language about which nothing is known. The set of such pairs shall be denoted by 3^{Σ^*}.

Boolean operations and concatenation are generalized from two-valued to three-valued languages as follows:

$$(K, K') \cup (L, L') = (K \cup L, K' \cup L')$$
$$(K, K') \cap (L, L') = (K \cap L, K' \cap L')$$
$$\overline{(L, L')} = (\overline{L'}, \overline{L})$$
$$(K, K')(L, L') = (KL, K'L')$$

Two different partial orderings on three-valued languages are defined. First, they can be compared with respect to the *degree of truth*:

$$(K, K') \sqsubseteq_T (L, L') \quad \text{if} \quad K \subseteq L \text{ and } K' \subseteq L'.$$

The other ordering is with respect to the *degree of information*:

$$(K, K') \sqsubseteq_I (L, L') \quad \text{if} \quad K \subseteq L \text{ and } L' \subseteq K'.$$

It represents the fact that (K, K') and (L, L') are approximations of the same language, and that (L, L') is more precise, in the sense of having fewer uncertain strings.

Both orderings are extended to vectors of three-valued languages. The truth-ordering has a bottom element $\bot_T = \big((\varnothing, \varnothing), \ldots, (\varnothing, \varnothing)\big)$, For the information-ordering, the bottom element is $\bot_I = \big((\varnothing, \Sigma^*), \ldots, (\varnothing, \Sigma^*)\big)$.

As in the two-valued case, concatenation, union and intersection, as well as every combination thereof, are monotone and continuous with respect to the truth ordering; complementation is not monotone. With respect to the information ordering; concatenation and all Boolean operations, *including complementation*, are monotone and continuous, which extends to any combinations of these operations.

Definition 3 (Well-founded semantics [5]). *Let $G = (\Sigma, N, P, S)$ be a Boolean grammar, let $N = \{A_1, \ldots, A_n\}$. Fix any vector $K = ((K_1, K_1'), \ldots, (K_n, K_n')) \in (3^{\Sigma^*})^n$ and define a function $\Theta_K : (3^{\Sigma^*})^n \to (3^{\Sigma^*})^n$ by substituting its argument into positive conjuncts and K into negative conjuncts:*

$$[\Theta_K(L)]_A = \bigcup_{A \to \alpha_1 \& \ldots \& \alpha_m \& \neg\beta_1 \& \ldots \& \neg\beta_n \in P} \left[\bigcap_{i=1}^{m} \alpha_i(L) \ \cap \ \bigcap_{j=1}^{n} \overline{\beta_j(K)} \right],$$

for each $A \in N$. Define $\Omega(K) = \bigsqcup^{T}_{\ell \geqslant 0} \Theta_K^{\ell}(\bot_T)$ and let $M = \bigsqcup^{I}_{k \geqslant 0} \Omega^k(\bot_I)$. Then, according to the well-founded semantics of Boolean grammars, $L_G(A) = [M]_A$.

The main result justifying the correctness of the well-founded semantics, is that M is a solution of the following system of equations in three-valued languages:

$$A = \bigcup_{A \to \alpha_1 \& \ldots \& \alpha_m \& \neg\beta_1 \& \ldots \& \neg\beta_n \in P} \left[\bigcap_{i=1}^{m} \alpha_i(L) \ \cap \ \bigcap_{j=1}^{n} \overline{\beta_j(L)} \right] \quad (\text{for } A \in N).$$

The binary normal form is generalized to the well-founded semantics:

Proposition 1 (Kountouriotis et al. [5]). *Every Boolean grammar, as in Definition 3, can be effectively transformed to a grammar in the binary normal form, in which every rule is of the form*

$$A \to B_1C_1 \& \ldots \& B_nC_m \& \neg D_1E_1 \& \ldots \& \neg D_nE_n \& \neg\varepsilon$$
$$(m \geqslant 1,\ n \geqslant 0,\ B_i, C_i, D_j, E_j \in N)$$
$$A \to a \quad (a \in \Sigma)$$
$$A \to a \& U \quad (a \in \Sigma)$$
$$U \to \neg U \quad \text{(a special symbol generating uncertainty)}$$
$$S \to \varepsilon \quad \text{(only if } S \text{ does not appear in right-hand sides of rules)}$$

The transformation maintains the generated three-valued language.

Kountouriotis et al. [5] used this normal form to construct an extension of the cubic-time parsing algorithm to the well-founded semantics, which, given an input string w, computes its membership status as a value in $\{0, \frac{1}{2}, 1\}$. The data constructed in that algorithm can be computed more efficiently using matrix multiplication, which will now be demonstrated by encoding it into the abstract form of the proposed algorithm. Let $X = 3^N$, $Y = 3^{N \times N}$, $(U_1, V_1) \circ (U_2, V_2) = (U_1 \times U_2, V_1 \times V_2)$, $(Q_1, R_1) \sqcup (Q_2, R_2) = (Q_1 \cup Q_2, R_1 \cup R_2)$, $I(a) = (\{ A \mid A \to a \in P \}, \{ A \mid A \to a \in P \text{ or } A \to a\&U \in P \})$, and finally $f(Q, R) = \big(\{A \mid \exists A \to B_1C_1\& \ldots \&B_mC_m\&\neg D_1E_1\& \ldots \&\neg D_{m'}E_{m'}\&\neg\varepsilon : (B_i, C_i) \in Q \text{ and } (D_j, E_j) \notin R \text{ for all applicable } i, j\}, \{A \mid \exists A \to B_1C_1\& \ldots \&B_mC_m\&\neg D_1E_1\& \ldots \&\neg D_{m'}E_{m'}\&\neg\varepsilon : (B_i, C_i) \in R \text{ and } (D_j, E_j) \notin Q \text{ for all applicable } i, j\}\big)$. This establishes an analogue of Theorem 1 for the well-founded semantics, that is, the three-valued membership in $L(G)$ of a given string $w \in \Sigma^*$ can be computed in time $\Theta(BMM(n) \log n) = O(n^{2.376})$.

Thus, one more key algorithm for the context-free grammars has been extended to the general case of Boolean grammars, and its clarity has even been improved in the process. This provides further evidence for the author's long-time claim that Boolean grammars are the proper general case of the context-free grammars.

References

1. L. Adleman, K. S. Booth, F. P. Preparata, W. L. Ruzzo, "Improved time and space bounds for Boolean matrix multiplication", *Acta Informatica* 11:1 (1978), 61–70.
2. V. L. Arlazarov, E. A. Dinic, M. A. Kronrod, I. A. Faradzhev, "On economical construction of the transitive closure of an oriented graph", *Soviet Mathematics Doklady*, 11 (1970), 1209–1210.
3. D. Coppersmith, S. Winograd, "Matrix multiplication via arithmetic progressions", *Journal of Symbolic Computation*, 9:3 (1990), 251–280.
4. Z. Ésik, W. Kuich, "Boolean fuzzy sets", *International Journal of Foundations of Computer Science*, 18:6 (2007), 1197–1207.
5. V. Kountouriotis, Ch. Nomikos, P. Rondogiannis, "Well-founded semantics for Boolean grammars", *Information and Computation*, 207:9 (2009), 945–967;
6. A. Okhotin, "Conjunctive grammars", *Journal of Automata, Languages and Combinatorics*, 6:4 (2001), 519–535.
7. A. Okhotin, "Boolean grammars", *Information and Computation*, 194:1 (2004), 19–48.
8. A. Okhotin, "Generalized LR parsing algorithm for Boolean grammars", *International Journal of Foundations of Computer Science*, 17:3 (2006), 629–664.
9. A. Okhotin, "Recursive descent parsing for Boolean grammars", *Acta Informatica*, 44:3–4 (2007), 167–189.
10. A. Okhotin, "Unambiguous Boolean grammars", *Information and Computation*, 206 (2008), 1234–1247.
11. V. Strassen, "Gaussian elimination is not optimal", *Numerische Mathematik*, 13 (1969), 354–356.
12. L. G. Valiant, "General context-free recognition in less than cubic time", *Journal of Computer and System Sciences*, 10:2 (1975), 308–314.