Prof. Dr. F. Otto
Fachbereich Elektrotechnik/Informatik
Universität Kassel
34109 Kassel

Referee's Report on the Doctoral Thesis

"Machine Learning of Analysis by Reduction"

submitted by **Petr Hoffmann**

to the **Faculty of Mathematics and Physics**

of **Charles University in Prague**

This thesis addresses the problem of "learning analysis by reduction" from positive (and negative) samples. *Analysis by reduction* is a technique from linguistics that can be used to analyze sentences of a natural language with a free word-order (as, e.g., Czech). To model this technique, P. Jančar, F. Mráz, M. Plátek, and J. Vogel introduced the *restarting automaton* in the 1990s. A restarting automaton M consists of a finite-state control, a single flexible tape, and a read/write window of a fixed size $k \geq 1$. To describe a restarting automaton M in a compact way, the notion of *meta-instruction* has been coined. A *rewriting meta-instruction* is of the form $(E_\ell, u \to v, E_r)$, where E_ℓ and E_r are regular languages (given, for example, through rational expressions) and u, v are strings such that $u \to v$ is one of the possible rewrite steps of M. It applies to a tape contents of the form xuy such that $x \in E_\ell$ and $y \in E_r$, and it describes a cycle that transforms the tape contents xuy into the tape contents xvy. An *accepting meta-instruction* is of the form (E, Accept), where E is a regular language, and it describes an accepting tail computation on a tape contents of the form $x \in E$. A restarting automaton M can be described through a finite set of rewriting and accepting meta-instructions.

By now many different types and variants of restarting automata have been introduced and studied in the literature, and many important results have been obtained, among them characterizations of various well-known language classes in terms of certain types of restarting automata, closure properties, decidability and undecidability results, and results concerning the descriptional complexity of certain types of restarting automata.

Despite all these important results, restarting automata have not yet been used seriously in applications. One of the reasons for this might be the problem that it is quite a difficult task to design a restarting automaton (or an "analysis by reduction") for a given language. To overcome this difficulty, various authors have suggested ways of learning (certain types of) restarting automata automatically. Essentially two approaches have been studied. On the one hand, *genetic algorithms* have been used by S. Basovník and F. Mráz (2011) and P. Hoffmann (2002), and on the other hand, methods of *inductive inference* have been used by P. Černo and F. Mráz (2009) to learn *clearing restarting automata* and by F. Mráz, F. Otto and M. Plátek (2006) to learn *strictly locally testable restarting automata*. In these papers either positive and negative samples of sentences are provided by the teacher, or (in the latter paper) positive samples of sentences and of reductions are provided by the teacher.

As the main topic of his thesis, P. Hoffmann suggests still another way of using inductive inference to learn analysis by reduction, or to be more exact, a certain class of restarting automata. He presents a learning algorithm for *reversible restarting automata*. Here a restarting automaton is called (k-) reversible if it can be defined by a finite set of meta-instructions which satisfy the restriction that the regular languages that occur in them are all (k-) reversible. The reversible languages are a proper extension of the strictly locally testable languages, and an inductive inference method for k-reversible languages from positive samples has been provided by D. Angluin (1982).

As input for his learning algorithm P. Hoffmann uses a finite set of *fully positive analysis by reduction samples* (FPABR samples, for short), that is, a finite set of reduction sequences of the form $w_n \to w_{n-1} \to \ldots \to w_0$ such that, for all $i = 1, \ldots, n$, $w_i \to w_{i-1}$ is a reduction that can be realized by a rewriting meta-instruction of the restarting automaton M that is to be learnt, and, for

all $i = 0, 1, \ldots, n$, w_i belongs to the language $L(M)$, which is the language of all strings that are accepted by M.

1 Structure of the Thesis

After a short introduction the basic notions from formal language theory are given in Section 1. In Section 2 the (*grammatical*) *inference problem* is presented, and the notion of *learning* a class of languages *in the limit* is restated. In addition, various learning models are described in short and some results on learning regular languages are cited. Then the analysis by reduction and restarting automata are presented, and the notion of an *R-presentation* of a restarting automaton is defined. Here P. Hoffmann distinguishes between *located reductions*, which are reductions in which the rewritten factor is marked, and *sliding reductions*, in which only the effect of the reduction is provided.

In Section 3 the goals of the thesis are summarized which concern the *complexity* of learning analysis by reduction, an *algorithm* to learn reversible restarting automata from FPABR samples, and a *benchmark* for evaluating the quality of the obtained inference algorithm. These three goals are then addressed in the following three sections, which make up the main part of the thesis.

In Section 4, the $(\alpha, \mathsf{X}, k, \Sigma, \Gamma, R)$-*inference optimization problem* and the $(\alpha, \mathsf{X}, k, \Sigma, \Gamma, R)$-*inference decision problem* are defined for a class X of restarting automata, a size measure α for these automata, a bound k for the size of the read/write window, an input alphabet Σ, a tape alphabet Γ, for located reductions ($R = \dot{R}$) or sliding reductions ($R = \overline{R}$), and for either positive or positive and negative samples. Here two size measures are suggested: isize(M), which is simply the number of meta-instructions of M, and size(M), which is the combined size of the descriptions of all meta-instructions of M. Also the notion of *ABR inference from positive samples problem* is defined, which is the problem of finding a restarting automaton that is consistent with a given finite set of FPABR samples.

In Section 5, results on the complexity of learing restarting automata are presented. By a reduction from the graph 3-colorability problem, it is shown that the $(\alpha, \mathsf{X}, k, \Sigma, \Gamma, R)$-*inference decision problem* from positive and negative samples is NP-hard, and this result holds for both measures given above, and it holds for located as well as for sliding reductions (Theorems 5.2.2 and 5.2.3). Surprizingly, the corresponding *inference optimization problems* can be solved in polynomial time, if only positive samples are used (Theorems 5.2.4 and 5.2.8). On first glance these results are quite surprizing, since learning from positive and negative samples should be easier than learning from positive samples only. However, the clue lies in the details of what is being learnt: in the former case it is determined whether there exists a restarting automaton of type X and of size $\alpha(X) \leq n$ that is compatible with all the given positive and negative samples, while in the latter case the problem consists in finding a restarting automaton of type X that is consistent with the positive samples and that is minimal among all such automata. Because of the absence of negative samples, the latter is in fact easier to solve than the former.

Unfortunately, the degree of the polynomial time bound for the algorithms described in the proofs of Theorems 5.2.4 and 5.2.8 are not given, although it is stated that they do not appear to be applicable in practise. Therefore P. Hoffmann proposes in Section 6 to study the ABR inference from positive samples problem for a special class of restarting automata: the *single k-reversible RRWW-automata*, S-kR-RRWW-automata for short. Here a restarting automaton M is called *single*, if it satisfies the following three technical restrictions:

(S1) M has at most one accepting meta-instruction;

(S2) for each rewrite operation $u \to v$, M has at most one rewriting meta-instruction of the form $(E_\ell, u \to v, E_r)$;

(S3) for each rewriting meta-instruction $(E_\ell, u \to v, E_r)$ of M, it is the case that u and v have neither a common non-empty prefix nor a common non-empty suffix.

In addition, he restricts his attention to a particular reduction strategy, which he calls *first opportunity rewriting*. What is the consequence of this choice? Would a different reduction strategy lead to different results?

Even though the S-kR-RRWW-automata are rather restricted, it turns out that they accept all growing context-sensitive languages (Theorem 6.2.9). Based on the inference algorithm for k-reversible languages of D. Angluin (1982), P. Hoffmann presents a method, called Omega*-method, for inferring S-kR-RRWW-automata from FPABR samples (Subsection 6.4). From a given finite set S of FPABR samples, an input alphabet Σ, and a tape alphabet $\Gamma \supseteq \Sigma$, an S-kR-RRWW-automaton is determined that is consistent with S. What is the role of the chosen auxiliary symbols in this method? Observe that by definition FPABR samples do not contain any occurrences of auxiliary symbols, and in the inference method for k-reversible languages, no auxiliary symbols are introduced, either.

Then P. Hoffmann compares S-kR-RRW-automata to S-$(k+1)$R-RRW-automata and S-kR-RR-automata to S-$(k+1)$R-RR-automata (Theorem 6.4.7), and he proves that S-kR-RR-automata are strictly less expressive than S-kR-RRW-automata, which in turn are strictly less expressive than S-kR-RRWW-automata (Theorem 6.4.9). Finally, he shows that the property of being 'single' is not a real restriction for kR-RRWW-automata (Theorem 6.4.10), and he compares the expressive power of S-kR-RR- and S-kR-RRW-automata to that of strictly locally testable R-automata (Theorems 6.4.11 and 6.4.12). All these proofs are technically quite involved and highly non-trivial. Section 6 closes with Theorems 6.4.14 and 6.4.15, which summarize the behaviour of the Omega*-method.

Finally in Section 7, the problem of evaluating the Omega*-method is discussed in detail. To compare this method to other inference methods through tests, the problems of how to choose target languages and of how to choose samples (or sample reductions) must be solved. One approach consists in simply taking example languages that other researchers have used before. In addition, one may use the same samples, or one may choose samples at random. This approach is used in Subsection 7.3.1 showing that the Omega*-method performs well for many of these simple example languages. The second approach consists in extending randomness to also generating the target languages. For the latter approach one may use grammars or automata to represent the target languages. In Subsection 7.2.1 it is shown that grammar-based methods have the drawback that it is in general undecidable whether a given rewriting meta-instruction is error-preserving with respect to a given linear language (Theorem 7.2.6). A positive result is obtained when attention is restricted to the so-called *even linear* languages (Theorem 7.2.15). However, there are even 'even linear' languages that cannot be accepted by any restarting automaton that does not use auxiliary symbols (Theorem 7.2.18).

Because of these drawbacks, P. Hoffmann suggests an evaluation method that is based on automata (Subsection 7.2.2). Specifically he uses restarting automata that are generated at random to represent target languages and to obtain training samples and testing samples. In Subsection 7.3.2 these random targets are then used to evaluate the Omega*-method experimentally. The results of 1200 experiments are reported that consisted of 100 runs for each value of reversibility $k \in \{0, 1, 2, 3\}$ and each number $r \in \{20, 40, 80\}$ of training samples. As it turns out the Omega*-method performs quite well for $k \geq 1$, and not surprizingly, the performance increases with the number of training samples used. However, the evaluation considered only measures the quality of the language learnt, but it does not measure the quality of the analysis by reduction that is being learnt. How would a benchmark look like that measures this latter quality?

2 Appraisal of the Results

The problem of learning analysis by reduction or a restarting automaton from positive (and negative) samples and sample reductions is of great theoretical and practical importance. This thesis presents a very important step towards solving this problem. The following results constitute the main contribution of the thesis:

1. The NP-hardness results on the inference decision problem from positive and negative data for restarting automata and the result that the inference optimization problem from positive data is decidable in polynomial time for restarting automata (Section 5) are new and very interesting.

2. The Omega*-method for learning S-kR-RRWW-automata from ABR samples (Section 6) extends the previous known methods to a larger class of restarting automata. In particular, the idea of using FPABR samples as training data is new and convincing.

3. The results comparing the various types of S-kR-RRWW-automata to each other are quite

nice and important, as they clearly state the relative expressive power of the various models (Section 6).

4. Finally, the benchmark based on random restarting automata proposed in Subsection 7.2.2 is new and interesting, as it presents a reasonable way for measuring the performance of methods for learning restarting automata experimentally.

3 Recommendation

In my opinion this thesis is very nice. It contains many new and interesting results, and as analysis by reduction is an important technique in linguistics, it can be expected that the results of this thesis will actually have some influence, not only on automata theory, but also on linguistics, as they may help to establish the restarting automaton as a formal model for analysis by reduction also in that field. The thesis is well written, and the proofs are given in sufficient detail, although some of them are quite hard to read. This thesis clearly shows P. Hoffmann's ability for creative scientific work. Therefore, I recommand *to accept this doctoral thesis.*

Kassel, June 3, 2013

Prof. Dr. Friedrich Otto