A NOTE ON THE GRAMMATICAL INFERENCE PROBLEM FOR EVEN LINEAR LANGUAGES

Erkki Mäkinen

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF TAMPERE

REPORT A-1994-9

UNIVERSITY OF TAMPERE
DEPARTMENT OF COMPUTER SCIENCE
SERIES OF PUBLICATIONS A
A-1994-9, NOVEMBER 1994

A NOTE ON THE GRAMMATICAL INFERENCE PROBLEM FOR EVEN LINEAR LANGUAGES

Erkki Mäkinen

University of Tampere
Department of Computer Science
P.O. Box 607
FIN-33101 Tampere, Finland

ISBN 951-44-3663-6
ISSN 0783-6910

A NOTE ON THE GRAMMATICAL INFERENCE PROBLEM FOR EVEN LINEAR LANGUAGES

Erkki Mäkinen
Department of Computer Science, University of Tampere,
P.O. Box 607
FIN-33101 Tampere, Finland

E-mail: em@cs.uta.fi

Abstract. This note introduces subclasses of even linear languages for which there exist inference algorithms using positive samples only.

Key Words and Phrases: formal languages, grammatical inference, Szilard language, even linear language, reversible language.

1. Introduction

Takada [11] has shown that the grammatical inference problem for even linear languages can be solved by reducing it to the grammatical inference problem for regular languages. This is possible by observing that any even linear language can be produced by a grammar schema (called universal even linear grammar in [11]) regulated by a control set which is always a regular language. Hence, instead of the even linear language we can infer the regular control set which uniquely determines the even linear language. This control set approach is used also in [12]. There are well-known inference algorithms for regular languages using both positive and negative samples but no such algorithms using positive samples only are possible [3]. (Our criterion for successful inference is "identification in the limit" [3].) The grammatical inference problem for even linear languages and an application concerning picture description languages is discussed also in [9] .

For many practical purposes it is more natural to consider inference algorithms which use positive samples only. In this note we consider special cases in which such algorithms exist for even linear languages.

The Szilard language of a grammar contains a word for every terminating derivation according to the productions of the grammar. The Szilard language of a context-free grammar is in general non-context-free, but the Szilard language of each linear grammar is regular [5,7,8]. It turns out that the form of homomorphic images of Szilard languages of linear grammars is of great interest in our problem. In fact, if it is possible to infer a certain homomorphic image of the Szilard language in question, then we can also infer the corresponding even linear language.

2. Notations and preliminaries

If not otherwise stated we follow the notations and definitions of [4]. Let $G = (V,\Sigma,P,S)$ be a context-free grammar (hereafter simply "grammar") whose productions are uniquely labelled by the symbols of an alphabet Z. If a production $A\rightarrow\alpha$ is associated with the label ρ we write $\rho{:}A\rightarrow\alpha$. If a sequence σ of labelled productions is applied in a derivation $\beta \Rightarrow^{*} \gamma$, we write $\beta \Rightarrow^{\sigma} \gamma$. The *Szilard language* Sz(G) of G is defined as

$$Sz(G) = \{ \sigma \in Z^{*} \mid S \Rightarrow^{\sigma} w, w \in \Sigma^{*} \} \text{ [5,7,8].}$$

On the other hand, we can first fix a language C over Z, and then consider the derivations defined by the words in C. This gives us the concept of control set. More formally, we can define the *language generated by G with control set C* as

$$L_C(G) = \{ w \in \Sigma^{*} \mid S \Rightarrow^{\sigma} w, \sigma \in C \}.$$

We consider reduced [4] grammars only; i.e. grammars in which each nonterminal and terminal symbol appears in some derivation from the start symbol to a terminal string. A production is said to be *terminating* if its right hand side contains no nonterminals; otherwise a production is said to be *continuing*.

Recall that a grammar is *linear* if each production is of the form $A\rightarrow uBv$ or $A\rightarrow u$ where A and B are nonterminals and u and v are (possibly empty) terminal strings, and *even linear* if in $A\rightarrow uBv$ we have len(u) = len(v), where len(α) stands for the length of α. A language is even linear if it can be generated by an even linear grammar. All even linear languages can be generated by even linear grammars with productions of the form $A\rightarrow aBb$, $A\rightarrow ab$, and $A\rightarrow a$, where A and B are nonterminals and a and b are terminals. Moreover, if the even linear language in question contains the empty word λ, we need the production $S\rightarrow\lambda$, where S is the start symbol [11]. Throughout this note we suppose that even linear grammars are always in this normal form.

A production with a nonterminal A in its left hand side is said to an *A-production*. Similarly, a production in an even linear grammar with terminals a and b, in this order, in its right hand side is said to be an *(a,b)-production*.

As already mentioned, the Szilard language of a linear grammar is always regular. Namely, if G is a linear grammar we can construct a regular grammar H generating Sz(G) as follows. For each production $\rho{:}A\rightarrow uBv$ (resp. $\rho{:}A\rightarrow u$) in G take the production $A\rightarrow\rho B$ (resp. $A\rightarrow\rho$) to H. It is straightforward to show that H generates Sz(G). Moreover, if G is reduced then H is reduced, too. Note that in H's productions each right-hand side begins with a unique terminal symbol. Such a grammar is said to be a *Szilard grammar*. Note also that if L is generated by a regular Szilard grammar then there are linear grammars G such that L = Sz(G). We denote the class of Szilard languages of linear grammars by LSZ.

Let $G = (V,\Sigma,P,S)$ be an even linear grammar. The corresponding *universal even linear grammar* $G^0 = (\Sigma \cup \{ S \},\Sigma,P^0,S)$ is obtained by replacing all appearances of all nonterminals in the productions by S. The language L(G) is generated by G^0 with a control set C provided that C is properly chosen. To construct the correct control set C, we define homomorphism $h: P\rightarrow P^0$ by setting $h(A\rightarrow aBb) = S\rightarrow aSb$, $h(A\rightarrow ab) = S\rightarrow ab$, $h(A\rightarrow a) = S\rightarrow a$, and $h(S\rightarrow\lambda) = S\rightarrow\lambda$. Supposing that the productions in G^0 are uniquely labelled by the symbols of an alphabet X, the homomorphism h induces another homomorphism $g{:}Z^*\rightarrow X^*$, where $g(\pi) = \rho$ if $\pi{:}A\rightarrow aBb \in P$, $\rho{:}S\rightarrow aSb \in P^0$, and $h(A\rightarrow aBb) = S\rightarrow aSb$. Homomorphism g is called the *universal homomorphism*.

Now, if $S \Rightarrow^{\sigma} w$ is a derivation in G, then $S \Rightarrow^{g(\sigma)} w$ is a derivation in G^0; we say that g *preserves derivations*. As a consequence, we have $L(G) = L_C(G^0)$. The control set constructed as above is always regular and moreover, it is unique for a given even linear language. Hence, to identify an unknown even linear language from given samples, we can infer the corresponding control set [11]. Given a positive sample from an even linear language, we can always determine the derivation in G^0 producing the sample word in question. This derivation is a homomorphic image under g of the corresponding derivation in G. These derivations (or actually, the strings of productions used in these derivations) define the Szilard language Sz(G). The purpose of this note is to consider special cases where g(Sz(G)) has a form which guarantees the existence of an inference algorithm using positive samples only.

3. Local and reversible languages

A finite automaton $A = (Q,\Sigma,d,q_0,F)$ is *reset-free* if for no two distinct states q_1 and q_2 do there exist a symbol a in Σ and a state q_3 such that $d(q_1,a) = q_3 = d(q_2,a)$. A finite automaton is *zero-reversible* if it is deterministic, has at most one final state, and is reset-free. A regular language is zero-reversible if there is a zero-reversible finite automaton accepting it [1]. Let L ($\subseteq \Sigma^*$) be a language and let w ($\in \Sigma^*$) be a word. The left-quotient $Q_L(w)$ of L and w is defined by $Q_L(w) = \{ v \mid wv \in L \}$. Let k, $k \geq 1$, be an integer. L is said to be *k-reversible* if and only if u_1vw and u_2vw in L and len(v) = k imply $Q_L(u_1v) = Q_L(u_2v)$. Angluin [1] has shown that there exists an efficient inference algorithm for k-reversible languages using positive samples only.

Let Σ be an alphabet, let I and F be subsets of Σ, and let T be a subset of Σ^2. Languages of the form

$$I\Sigma^* \cap \Sigma^*F \setminus \Sigma^*T\Sigma^*$$

are referred to as *local languages*. The class of *k-testable languages in the strict sense* (k-TLSS, for short), $k \geq 3$, is obtained when I and F contain strings of length at most k - 1, and T contains strings of length k. k-TLSS's can be inferred from positive samples only [2, 13].

LSZ is clearly a proper subclass of zero-reversible languages. Indeed, each such Szilard language can be recognised by finite automata in which the transitions are uniquely labelled. Hence, we can write

Theorem 1 [6]. Let G be a linear grammar. Then Sz(G) is zero-reversible.

Similarly, LSZ is a subclass of local languages. Namely, let I be the set of labels corresponding to productions with the start symbol on the left hand side, let F be the set of labels corresponding to terminating productions, and let T be the set

$$T = \{ \pi\rho \mid \pi{:}A \to uBv \text{ and } \rho{:}C \to wDy,\ B + C, \text{ or } \pi{:}A \to u \text{ is terminating} \}.$$

Then each L in LSZ has the form $I\Sigma^* \cap \Sigma^*F \setminus \Sigma^*T\Sigma^*$. We have

Theorem 2. Let G be a linear grammar. Then Sz(G) is a local language.

Before considering homomorphic images of the languages in LSZ in greater detail, we mention a little result concerning morphic representations of regular languages. A classical result of the field states that every regular language R (not containing the empty word) can be represented in the form R = h(L), where h is a letter-to-letter homomorphism and L is a local language [10, p. 97]. Similar representation is possible when L is in LSZ. Namely, if $\rho{:}A \to aB$ (resp. $\rho{:}A \to a$) is a production in the regular grammar, then we can define $h(\rho) = a$. We have also seen that LSZ is a proper subclass of local languages. Hence, our representation strengthens the classical result. Salomaa's simple proof [10] uses finite automata; proofs using regular expressions are evidently much more complicated. We have seen that grammars allow a simple proof as well.

4. Terminal-fixed and almost terminal-fixed languages

Let L be a language in LSZ and h be a homomorphism. If h(L) were always k-reversible or k-TLSS for some k, then we could infer all even linear languages by using positive samples only. This is naturally impossible, since the class of even linear languages contains all regular languages. We first consider a simple example which shows that there indeed are cases where the homomorphic image is neither k-reversible nor k-TLSS.

Let G be an even linear grammar with productions π:S→aAb, θ:A→cBd, ρ:B→cBd, and τ:B→ab. We have Sz(G) = $\{ \pi\theta\rho^n\tau \mid n \geq 0 \}$. Since θ:A→cBd and ρ:B→cBd have common terminals in their right hand sides, they are both mapped to S→cSd when constructing the control set C of G^0. If the productions in G^0 are labelled as π:S→aSb, ψ:S→cSd, and τ:S→ab, we have g(Sz(G)) = $\{ \pi\psi^n\tau \mid n \geq 1 \}$. This language is not zero-reversible. Similarly, for each k, k = 1, 2,.., there are homomorphic images $\{ \pi\psi^n\tau \mid n \geq k + 1 \}$ which are neither k-reversible nor (k+1)-TLSS (or local languages when k = 1).

When k = 1, we have a grammar with productions π:S→aAb, θ:A→cBd, ρ:B→cCd, τ:C→cCd, and υ:C→ab. When k = 2, we again lengthen the string of (c,d)-productions by one, and so on.

An extreme case appears when the terminal symbol combinations on the right hand sides of continuing productions of an even linear grammar are unique. In this case the inference problem for even linear languages reduces to that for LSZ which can be easily solved in linear time [6].

More formally, we say that an even linear grammar is *terminal-fixed* if A→aBb and C→aDb implies A = C and B = D. An even linear language is terminal-fixed if there is a terminal-fixed even linear grammar generating it.

We have made no assumptions concerning terminating productions. Since the last continuing production in every derivation is unique, we can combine the last continuing production and the (possibly non-unique) terminating production as a new unique production. For example, if a derivation ends with A→aBb and B→c, we can consider these as a (unique) terminating production A→acb. This gives us a Szilard language, and we can write the following.

Theorem 3. Terminal-fixed even linear languages can be inferred from positive samples in linear time.

It is clear that terminal-fixed languages form a proper subclass of even linear languages. Similarly, it is also clear that terminal-fixed even linear languages and regular languages are incomparable.

We can prove a little more than in Theorem 3. Let G be an even linear grammar. If A→aBb and C→aDb implies B = D, we say that G is *almost terminal-fixed*. An even linear language is almost terminal-fixed if there is an almost terminal-fixed even linear grammar generating it.

The situation shown in Figure 1 obviously holds for the families of languages considered in this note.

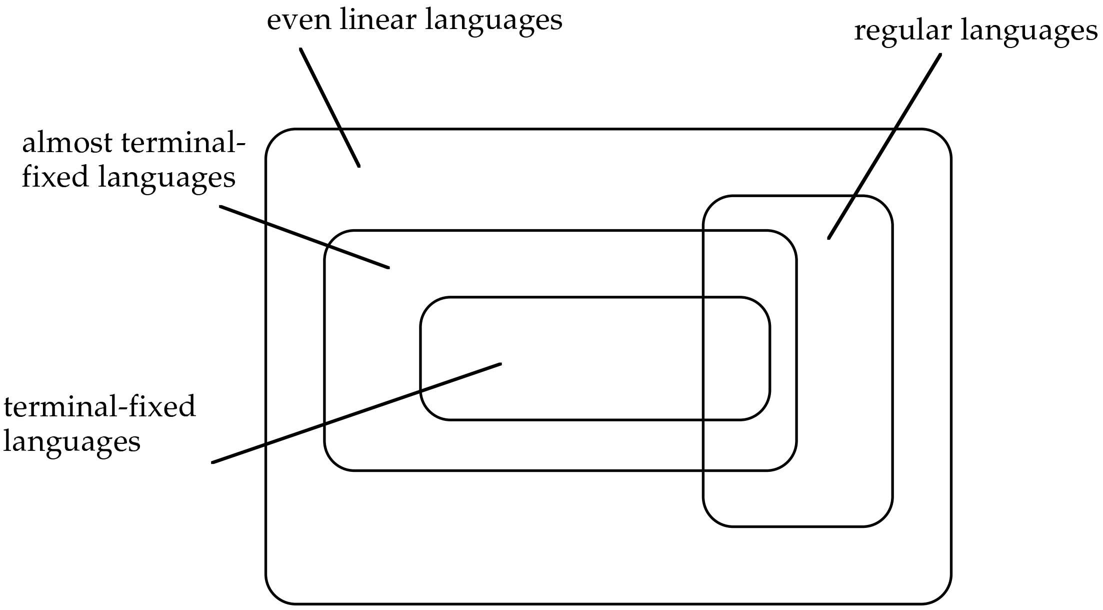

Figure 1. The relationship between some classes of languages.

Reconsider now the sample grammars discussed in the beginning of this section. We noticed that if G is almost terminal-fixed, then it is possible that g(Sz(G)) is not zero-reversible. However, when we constructed a grammar G with g(Sz(G)) not even 1-reversible, we need productions (A→cBd, B→cCd, C→cCd) not allowed in almost terminal-fixed grammars. This observation is generalized in the theorem below.

Theorem 4. Let G = (V,Σ,P,S) be an almost terminal-fixed even linear grammar. Then g(Sz(G)), where g is the universal homomorphism, is 1-reversible.
Proof. Language g(Sz(G)) can be generated by a regular grammar which has a production $A \rightarrow g(\pi)B$ (resp. $A \rightarrow g(\pi)$) for each production $A \rightarrow \pi B$ (resp. $A \rightarrow \pi$) in the Szilard grammar generating Sz(G). Since G is almost terminal-fixed and g is the universal homomorphism, the terminal appearing in the right hand side uniquely determines the nonterminal appearing in the right hand side. Hence, for each terminal symbol θ, $u_1\theta w$ and $u_2\theta w$ in g(Sz(G)) imply $Q_L(u_1\theta) = Q_L(u_2\theta)$. This means that g(Sz(G)) is 1-reversible. $

Corollary 1. Almost terminal-fixed even linear languages are inferable from positive samples.

By Theorem 4 we can infer the control set of the universal even linear grammar related to an almost terminal-fixed language by using the inference algorithm of [1]. The merging operations typical for the inference algorithm appear in situations of the following type: Words acedb and ccfdc in the sample imply derivations $S \Rightarrow aA_1b \Rightarrow acA_2db \Rightarrow acedb$ and $S \Rightarrow cA_3c \Rightarrow ccA_4dc \Rightarrow ccfdc$. We can merge A_2 and A_4.

It is interesting to notice that the same result as above can be also obtained when considering the subclass of even linear grammars defined by the following condition: productions $A \rightarrow aBb$ and $C \rightarrow aDb$ always implies A = C. On the other hand, we leave it open whether or not g(Sz(G)) is k-TLSS for some k.

5. Concluding remarks

We have seen that in some special cases Takada's idea of "indirect inference" works also with positive samples only. However, a price must be paid for not using negative samples: the class of almost terminal-fixed languages is considerably more restricted that the class of even linear languages.

Acknowledgement. The author wishes to thank Dr. Yuji Takada for his helpful comments.

References

[1] Angluin, D., Inference of reversible languages, *J. ACM* **29** (1982), 741-765.

[2] Garcia, P., Vidal, E., and Oncina, J., Learning locally testable languages in the strict sense, In: *Proc. of the First International Workshop on Algorithmic Learning Theory*, 1990, 325-338.

[3] Gold, E.M., Language identification in the limit, *Inform. Contr.* **10** (1967), 447-474.

[4] Harrison, M.A., *Introduction to Formal Language Theory,* Addison-Wesley, 1978.

[5] Mäkinen, E., On context-free derivations, *Acta Univ. Tamper. Ser. A* **198**, 1985.

[6] Mäkinen, E., The grammatical inference problem for the Szilard languages of linear gramamrs, *Inf. Process. Lett.* **36** (1990), 203-206.

[7] Moriya, E., Associate languages and derivational complexity of formal grammars and languages, *Inform. Contr.* **22** (1973), 139-162.

[8] Penttonen, M., On derivation languages corresponding to context-free grammars, *Acta Inform.* **3** (1973), 285-291.

[9] Radhakrishnan, V., and Nagaraja, G., Inference of even linear grammars and its application to picture description languages, *Pattern Recogn.* 21 (1988), 55-62.

[10] Salomaa, A., *Jewels of Formal Language Theory*, Computer Science Press, 1981.

[11] Takada, Y., Grammatical inference for even linear languages based on control sets, *Inform. Process. Lett.* **28** (1988), 193-199.

[12] Takada, Y., Inferring paranthesis linear grammrs based on control sets, *J. Inf. Process.* **12** (1988), 27-33.

[13] Yokomori, T., Learning local languages from positive data, In: *Proc. of the FUJITSU IIAS-SIS Workshop on Computational Learning Theory '89.*