Learning Distinguishable Linear Grammars from Positive Data

J.A. Laxminarayana[1], José M. Sempere[2], and G. Nagaraja[1]

[1] CSE Department, I.I.T. Bombay, India
{jalana,gn}@cse.iitb.ac.in
[2] DISC, Universidad Politécnica de Valencia, Spain
jsempere@dsic.upv.es

Abstract. We introduce a new algorithm to infer Terminal Distinguishable Linear Grammars from positive data.

1 Introduction and Preliminaries

Radhakrishnan and Nagaraja have proposed a method[4, 5] to infer a subclass of even linear languages, TDELL, from positive strings. In [6], Sempere and Nagaraja presented a characterizable method to infer a subclass of linear languages, TSDLL, from positive structural information. Recently, Laxminarayana and Nagaraja [1] proposed an incremental tabular method to infer TDEL languages from positive samples.

In this work, a method is proposed to infer a subclass of linear languages from the set of samples which are assumed to be from an even linear language. Sample set may possibly contain erroneous strings. An error model is built to capture the possible errors in the given sample set. The incremental TDELL inference algorithm [1] along with the proposed error model is used to generate a TDEL grammar which can be converted into an equivalent TSDL grammar using a transformation model. The TDEL and TSDL languages are proper subclasses of TDL (Terminal Distinguishable Languages). For further details about these concepts, the reader is referred to [4] and [2]. It is observed that TDELL $\subset$ TSDLL. Hence an equivalent TSDL grammar can be constructed for a given TDEL grammar by using the following transformation model.

Definition 1. *Let G_1 be a given TDEL grammar in normal form. An equivalent TSDL grammar G_2 can be constructed using the following finite steps: 1) For every production rule of the form $A \rightarrow aBb$ in G_1 construct productions $A \rightarrow aA'$ and $A' \rightarrow Bb$ in G_2 where A' is a new non-terminal. 2) Copy all other productions from G_1 to G_2*

2 Using an Error Model

Detection of error, formation of alternatives and selecting a best alternative are the functions involved in the proposed error model. The first function of any error model is to ascertain the existence of an error in the given string. If there is no error or if the error model fails to recognize the error then the input string will be passed to the inference algorithm without any modification. For the sake of simplicity we consider only the detection of single deletion error. Implementation and correctness issues of the error model are discussed in [3].

G. Paliouras and Y. Sakakibara (Eds.): ICGI 2004, LNAI 3264, pp. 279–280, 2004.

3 Algorithm to Infer TSDL Grammar

Algorithm 1 TSDL-INFERENCE(S^+)
Require: A nonempty set of positive sample S^+; one sample, say x, at a time.
Ensure: Inferred TSDL grammar

1: If no more strings are left in the sample set then transform the inferred TDEL grammar into an equivalent TSDL and terminate
2: If $x = \lambda$ then add the production $S \rightarrow \lambda$ to the inferred TDEL grammar; go to step 1.
3: If x is not the first string in the S^+ then using the error model obtain a preprocessed string of x
4: Using the incremental TDELL inference algorithm [1], construct a TDEL grammar and go to step 1.

Each successive string from S^+ is taken as input to the algorithm. An error model is employed to detect the occurrence of a single deletion error in the input string. If there exists a deletion error then the algorithm will continue with the *processed* string. In all the remaining cases, algorithm proceeds with the given input string. Selection of the appropriate *processed* string for the erroneous string is done using a statistical method.

The inference method proposed for the class of Terminal and Structural Distinguishable Linear Language using error correcting approach infers correctly for many languages. Further, it can be shown that, this method will not over generalize when the samples are from a higher class than linear. In addition, the method can identify any TSDL language in the limit if enough strings are provided.

References

1. J.A.Laxminarayana and G.Nagaraja, Incremental Inference of a subclass of Even Linear Languages, Eleventh International Conference on Advanced Computing and Communications, ADCOM 2003, PSG College of Technology, Coimbatore, India, 2003.
2. J.A.Laxminarayana and G.Nagaraja, Terminal Distinguishable Languages: A survey, A technical report, CSE Dept, IIT Bombay, India.
3. J.A.Laxminarayana and G.Nagaraja, Inference of a subclass of Even Linear Languages: An error correcting approach., A technical report, CSE Dept, IIT Bombay, India.
4. V. Radhakrishnan. Grammatical Inference from Positive Data: An Effective Integrated Approach. PhD thesis, Department of Computer Science and Engineering, Indian Institute of Technology, Bombay (India), 1987.
5. V. Radhakrishnan and G. Nagaraja, Inference of Even Linear grammars and its applications to Picture Description Language. Pattern Recognition, Vol 21, No.1, pp. 55-62, 1988.
6. J. M. Sempere and G. Nagaraja, Learning a subclass of linear languages from positive structural information, Proceedings of the 4th International Colloquium ICGI-98. LNAI Vol. 1433 pp 162-174, Springer-Verlag. 1998.