A Minimum Description Length Approach to Grammar Inference

Peter Grünwald*

CWI, P.O. Box 4079, 1009 AB AMSTERDAM, The Netherlands

Abstract. We describe a new abstract model for the computational learning of grammars. The model deals with a learning process in which an algorithm is given an input of a large set of training sentences that belong to some unknown grammar. The algorithm then tries to infer this grammar. Our model is based on the well-known Minimum Description Length Principle. It is quite close to, but more general than several other existing approaches. We have shown that one of these approaches (based on $n$-gram statistics) coincides exactly with a restricted version of our own model. We have used a restricted version of the algorithm implied by the model to find classes of related words in natural language texts. It turns out that for this task, which can be seen as a 'degenerate' case of grammar learning, our approach gives quite good results. As opposed to many other approaches, it also provides a clear 'stopping criterion' indicating at what point the learning process should stop.

1 Introduction

In this paper we construct an algorithm that has as its sole input a training text, that is, a very large fragment of a natural language text in for example English. It should then output grammatical rules underlying this language; so in this case, English grammar rules. Thus learning takes place using positive examples only – possibly including noise. No explicit negative example is ever given. Approaches along these lines are commonly called [8] 'Grammatical Inference'. It should immediately be noted that there are severe limits to the grammatical inference approach: we will simply use context-free grammars and the fact that they are not capable of fully capturing natural language grammars is just one of many weaknesses of the approach [8]. Nevertheless, it will turn out to be interesting to see just how far one can get with such an abstract approach.

Our algorithm is based on the Minimum Description Length (MDL) Principle, introduced in its present form by Rissanen [10]. The original ideas behind this very general principle stem from R.J. Solomonoff. In his landmark 1964 paper [11] they were already applied to grammar inference. In the next two sections we will give a very short introduction to the MDL Principle and to how it can be

* E-mail: pdg@cwi.nl. Partially supported by the European Union through NeuroCOLT ESPRIT Working Group Nr. 8556, and by NWO through NFI Project ALADDIN under Contract number NF 62-376.
used for grammar induction. Section 4 discusses our approach, first describing its algorithmic aspects and then explaining how we compute description lengths. The reader who wants to see first what this approach leads to before worrying about the details may want to have a look at Sect. 5 first, where we discuss the results of some experimental data we obtained with an implementation of our algorithm. Section 6 compares our model to some others and discusses whether our approach makes sense from a linguist's point of view. The paper ends with some concluding remarks.

2 The MDL Principle

One of the basic ideas of Solomonoff's theory of induction\(^2\) [11] is that 'learning' can, under the right circumstances, be seen as 'finding a shorter description of the observed data'. Here one views learning as finding a hypothesis that explains some observed data and makes predictions about data yet unseen. The MDL Principle, introduced in its present form by J. Rissanen [10], is a principle of statistics that is explicitly based upon Solomonoff's ideas. It goes as follows:

**Minimum Description Length Principle.** The best theory to explain a set of data is the one which minimizes the sum of

1. the length, in bits, of the description of the theory;
2. the length, in bits, of the data when encoded with the help of the theory.

The general idea here is that an 'empty' theory (the words 'theory' and 'hypothesis' will be used interchangeably here) which does not give any clue as to the sort of data we are dealing with, has no predictive value whatsoever. On the other hand, a theory that describes all of the data perfectly is very much restricted by the data and therefore does not really generalize. Therefore, it has no predictive value either. We have to find some trade-off, and in the next section we will see in a practical setting that the total description length can become much smaller using such a trade-off.

3 Grammar Learning and the MDL Principle

Consider three grammars that can account for a given natural language text [11]:

**the promiscuous grammar** This is an 'empty' grammar: given an input alphabet (vocabulary) of words, it accepts every concatenation of them as a sentence. Of course, it is a hopeless overgeneralization.

**the ad hoc grammar** This grammar accepts exactly all the sentences it has seen so far, but no single different sentence. Of course, it is a hopeless undergeneralization.

\(^2\) See [9] for a good introduction to Solomonoff's theory.