7 Computational Grammatical Inference

Pieter W. Adriaans¹, Menno M. van Zaanen²

¹ILLC, University of Amsterdam, Amsterdam, the Netherlands, pietera@science.uva.nl
²ILK, Tilburg University, Tilburg, the Netherlands, mvzaanen@uvt.nl

Abstract

Grammatical Inference (GI) concentrates on finding compact representations, i.e. grammars, of possibly infinite sets of sentences. These grammars describe what sentences do or do not belong to a particular language. The process of learning the form of a grammar based on example sentences from the language touches several fields. Here, we give an overview of the field of GI as well as fields that are closely related. We discuss linguistic, empirical, and formal grammatical inference and discuss the work that falls in the areas where these fields overlap.

7.1 Introduction

A grammar is a finite structure that describes a possibly infinite set of sentences, called a language. In this chapter we are concerned with learning grammars from unstructured data. The unstructured data is a sample taken from the language the grammar describes. This process is called Grammatical Inference (GI) and can be researched from many different angles. Here we will relate the linguistic, empirical and formal aspects of GI with respect to the field of computational grammatical inference (CGI).

CGI is mainly concerned with finding computational models that can learn grammars. The models can have linguistic impact, but can also serve as a practical tool to experiment with the theoretical proofs taken from the field of formal GI.

In this chapter we discuss the different fields related to the research performed in the field of CGI. We do this starting from the “applied” GI in linguistics and end in the “theoretical” field of formal GI, via CGI. Of course, the boundaries between the research areas are not always so strict and clear cut. Many results are applicable to more than one field, which is really why these research topics should be considered in total. During our
discussion, you will find that, unfortunately, there has not been much interaction between researchers from the different areas. This will also be our final conclusion.

7.2 Linguistic Grammatical Inference

Within linguistics, grammatical inference has an impact in different fields. Here, we will focus mainly on first language acquisition, but much of the discussion that follows also applies to other areas.

To get an idea of the extend of the problem of grammatical inference, we mention the discussion about the complexity of human languages. The general consensus is that natural languages are at least context-free and with high probability context-sensitive. (Huybrecths, 1984; Shieber, 1985) We will mainly discuss learning context-free languages as a first approximation of human languages. So far, there has been hardly any research into the learnability of context-sensitive languages. It is unlikely that the full expressive power of the class of context-sensitive languages is necessary to describe a natural language, however, context-sensitive constructions seem to exist in some cases.

Research on language learning has led to major controversies within the field of linguistics. One very influential reason for this is the famous conjecture of Chomsky that the efficiency of human language acquisition can only be explained on the basis of an innate Universal Grammar (UG). The reason for this would be that the learner is not provided with enough linguistic information to learn the linguistic structures as fast as it does. This is called “the poverty of the stimulus”. The UG discussion is in fact a revival of the ancient philosophical debate of rationalism (Descartes) versus empiricism (Hume). This argument has attracted a lot of attention, unfortunately with surprisingly little tangible results.

Past research in this field has shown that a proper study of language acquisition will require a huge interdisciplinary enterprise. Not only is there the need for theoretical research, but also a huge empirical effort has to be made. This requires co-operation between linguists, psychologists, audiologists, neuro-physiologists, cognitive scientists, computational linguists, and computer scientists.

To get an idea of the impact of the amount of work required, imagine the creation of elementary annotated video sequences to study the language development of children. This is an investment that easily amounts to several hundreds, if not thousands of person-years. For a statistically significant experiment one needs a group of at least some twenty children