Mathematical Linguistics - A. Kornai

Alok Debnath

September 2019

Abstract

This is a summary of the textbook "Mathematical Linguistics" by A. Kornai. The book may be found on `http://www.helsinki.fi/esslli/courses/readers/K54.pdf`. Some parts of the book require some prerequisites which are non-crucial. I have attempted to capture those in the footnotes. For those which are crucial a relevant section has been added, which can be used. This is only a summary of the book, and I have skimmed over a lot of the parts, some of which may be of interest to the reader. If this is so, I recommend reading the textbook.

Chapter 1

Introduction

The aim of this text is to analyze the notions which are of linguistic interest, such as phonology, morphology, syntax and semantics by using ”mathematical techniques”. The author compares this to a study of mathematical physics, where some branches or subfields are studied far more than others, not because they are not interesting, more so beacuse there is no direct utility in studying them over mathematical explanations of observable phenomenon.

Definitions: Definitions in mathematical linguistics follow three basic steps, an *ostensive* definition based on examples, an *extensive* definition which dealiantes the intended scope of the notion and an *intensive* definition, which exposes the underlying mechanism.

Formalization: Much like other branches of applied mathematics, the problem of formalization of semi-formally or informally stated theories. However, the choice of a formal definitions and structures is often hard. The stochastic nature of language and linguistic rules is still a debate.

Foundations: Mathemaical linguistics has sets as the object of interest, usually finite but sometimes denumberably infinite (see: set of words). Due to emperical restrictions, however, denumerable generailzations of finite objects such as ω-words are rarely used. [1] For the purpose of this series, such constructions are not taken into account. The study of language and mathematics have evolved rather independently, the primary exception being Indian traditions of logic.

Mesoscopy: A mesoscopic system is one which is neither too small to be appropriately studied by microscopic methods and tools, and yet is not large enough for statistical approximations of macroscopic systems are necessarily accurate. Natural language may be viewed as a mesoscopic systems, which is supposedly governed by a thousands of rules, wheras the statistical methods of macroscopic levels (Markov assumption, for example) is a reasonable approximation of the behavior of Natural Language. Macroscopic techniques yeild only generalizations of mesoscopic systems. Further, statistical quantities of interest

[1] ω- words are words that can be written on an ω-automaton, a finite automaton that runs on infinite strings as inputs. Buchi automata are examples of the same.

are linked ”only very indirectly” to the objects of interest and therefore there are special techniques that must be employed in order to decide which variables should be left unmodeled.

Chapter 2

The elements

Fundamentally, enumaration of the set of objects useful to linguistics is an one of the approaches to analyzing language. This enumeration, or generation of the set of objects is an important topic of study. Along with generation, another fundamental problem being solved is one of understanding set membership and determining whether an object belongs to this set. This can be done by generating the entire set and comparing the input against each one of the generated objects, or by constructing a grammar which generates a language against which the object in question can be compared.

2.1 Generation

Given a set of primitive elements E and a set of non-procedural generation rules R, it can be stated that $\forall x, y \in E$, $r \in R$, $\exists z$ s.t. $z = r(x, y)$. Another perspective on the same is $z \rightarrow_r xy$, *i.e.* z directly generates x and y. The smallest collection of objects closed under direct generation $r \in R$ which contains all the elements of E is called the set generated from E by R.

Generative definitions have a notion of equality, which may be derivational (stronger as it depends on how the object was generated, not an abstract notion). Derivations can be mono- or multi-stratal, as in deriving via some well-defined intermediate state. In order to understand generative defintions better, a few examplesa are provided.

2.1.1 Wang tilings

Problem Definition: Let C be a finite set of colors and S be a finite set of square tiles, colored on the edges according to a generative function $e : S \rightarrow C^4$. Arrange the tiles such that the adjacent edges of the same tile are of the same color. Therefore:

$$i, j \in \mathcal{Z},\ u, v \in S, \text{ and } \pi_n \in R \forall n \in \{1, 2, 3, 4\}$$

the four production rules are:
$\pi_3(e(i,j)) = \pi_1(e(i,j+1))$
$\pi_4(e(i,j)) = \pi_2(e(i+1,j))$
$\pi_2(e(i,j)) = \pi_4(e(i+1,j))$
$\pi_1(e(i,j)) = \pi_3(e(i,j+1))$

Discussion The definition associated with Wang tiling is not entirely generative, because of the following reasons:

1. The definition relies on externally provided functions and objects,
2. The rules are well-formedness conditions, not rules of production and in order to make that production rules, external constraints are necessary,
3. The well formedness conditions are not recursively defined.

Extending the Problem: It is recursively undecidable whther a given inventory of tiles E can yeild a Wang tiling.

2.1.2 Groups as Generators

Problem Definition: Let E be a set of generators $g_1, g_2, ..., g_k$ and their inverses $g_1^{-1}, g_2^{-1}, ..., g_k{}^{-}1$. Let R contain the product and cancellation rules. Formal products with usual rules of form a free group over k, and without the inverses and cancellation rules, a free monoid is formed over k generators.

Discussion: It is in general undecidable whether a formal product of generators is included in the kernel or not. Note that the defining relations can be considered a part of production rules as well as equality relation.

2.1.3 Herbrand Universes

Problem Definition: First order languages consist of logical symbols alongwith some constraints, and function and relation symbols. In Herbrand universes, the elements E are the onject constraints on the first order logic. Rigourously,

$$\forall x_i \in E, f(\cdot) \text{ is a function constant of arity } n,\ f(x_i) \in E$$

Discussion: Herbrand universes are purely formula based and is a good example of generative definitions. However, usig first order logic abstracts away many important properties of language.

2.2 Axioms, Rules and Constraints

One of the ways of contructing a system that undertands a language's grammaticality is by propogating it through a set of rules or axioms, and defining a group of combinatorial operations that preserve this notion of grammaticality. Analogously, most transformations in grammars are seen as applications of these rules followed by "cleaning up". Similarly the characterization of the grammatical forms can be done using an axiomatic system, the difference being that the starting constructions are abstract such as words and a well defined hierarchy or representing the system (a combination rule for the set of words, for example, is established).

2.2.1 Balanced parenthesis

Problem Definition: Given a series of parenthesis, there exists a well formedness condition that determines that given a scoring function, the overall score of the sequence of parenthesis never falls below a given value, and it is a given value at the end of the series.

For example: If '(' is provided the number '+1' and ')' are provided '−1', then the value of the sequence on addition should never drop below 0 and the value at the end of the sequence should be 0.

Discussion: It is seen that the WFCs here rely on a transformation of the problem into a more computable domain, where there exists a homomorphism between both the elements and the rules that govern these elements (operations) which are in the original set.

Instead of considering a score computation function in $\mathcal{Z}$, it is better to consider a mapping to a finite structure G, with well understood rules of combination, which allows disjuntive assignments. This makes the well formedness constraints based entirely on the combiantion yeilding a desirable result. The resulting desired combiantion allows for the constitution of a certificate of membership to a grammar defined by a tranformation (in this case $c : W \to 2^G$.