

Language Learnability

Benjamin Han

March 8, 2001

Language Technologies Institute
Carnegie Mellon University

Outline

- Gold's "Identification in the Limit" (1967)
 - Learnability models
 - Results and implications
 - Proof sketches
 - Learning time
- Empirical Studies: Bohannon & Stanowicz's Experiment on Adult Feedback, and Gordon's Attack

Gold's "Identification in the Limit"
(1967)

Motivations


- Q: How to model natural languages in artificial systems?
 - Lower bound: rich enough to simulate the linguistic phenomena
 - Upper bound: a training procedure exists
- Q: What are the implications to cognitive systems brought by the artificial models?

Motivations

- Toward a formal model of natural languages
 - Step 0: Power of different classes of formal languages (naïve models)
 - Step 1: Learnability/identifiability of the formal languages
 - Step 2: Complex models for real-life linguistic phenomena
 - Step 3: Learnability of the complex models

Motivations

properly inclusive



Language	Grammar	Machine	Chomsky Hierarchy
<i>Non-computable</i>		??	
Recursively enumerable (RE)	Unrestricted	Turing machines	0
Recursive		Turing machines that always halt	
Context-sensitive	Context-sensitive grammar	Linear-bounded automata	1
Context-free	Context-free grammar	Non-deterministic pushdown automata	2
Regular	Regular expressions	Finite state automata	3

Motivations

- Q: Is *human mind* a Turing machine?
 - If yes
 - what machines can't learn, human can't learn either, and vice versa
 - Natural languages are recursive (unless humans have time-out or probabilistic reasoning capability?)
 - If no: all bets are off

Motivations

- Q: Why study the inductive inference “in the limit”
 - “in the limit” = at time
 - It’s unlikely to get answers to the question “given the information and a set of possible conclusions, at *specific time t* what are the *correct* conclusion?”
 - The ‘power’ question: the behavior of a learner in the limit

Learnability Models

- *A learnability model* consists of
 - *A definition of learnability*: what do you mean by saying that a language is learned?
 - *A method of information presentation*: how does an instructor teach the learner?
 - *A naming relation which assigns names to languages*: what is the result of the learning?

Learnability Models

- Basic concepts
 - Alphabet A is a non-empty finite set of symbols; A^* is the (infinite) set of all finite strings over A
 - Language L is a subset of A^* ; a language class \mathcal{L} is the set of languages of the same underlying machine (some language is non-computable because 2^{A^*} is uncountable but the set of all possible TMs is countable)
 - Time t is discrete ($t=1, 2, \dots$)

Learnability Models

- Basic concepts
 - A primitive recursive function is
 - a recursive function (but not vice versa, e.g., Ackermann's function)
 - composed by a finite number of applications of composition and primitive recursion over $\text{null}(0)$, successor and projection functions.
 - a total function, i.e., defined on all natural numbers
 - A string can be encoded into a single integer i.e., we can always have a $A^* \rightarrow \mathbb{N}$ function

- Learnability models

- Definitions of learnability

- A method of information presentation

- A naming relation

Learnability Models

- Learnability:

$g_t = G(i_1, i_2, \dots, i_t)$, where

- g_t is the guess of the name of unknown language at time t
- G is the guessing/learning algorithm
- i_1, i_2, \dots, i_t is the information sequence received up to time t , where i_j is an information taken from the set of all possible units I at time j

Learnability Models

- Three learnability definitions
 - Identification in the limit
 - $\forall t$ (t is finite) $g_t = g_{t+1} = \dots = g$ is correct
 - Finite identification
 - $h(i_1, i_2, \dots, i_t)$ is a decision function returning 0/1; $g = G(i_1, i_2, \dots, i_t)$ iff $h(i_1, i_2, \dots, i_t) = 1$
 - Fixed-time identification
 - $g = G(i_1, i_2, \dots, i_{\square})$ where \square is a constant

- Learnability models

- Definitions of learnability

- A method of information presentation

- A naming relation

Learnability Models

- Method of information presentation
 - For a language Σ , $I(\Sigma)$ is its set of allowable information sequences (each one has infinite length)
 - For a language Σ : i_1, i_2, \dots, i_t is a prefix of some sequence in $I(\Sigma)$

Learnability Models

- Two information presentation methods
 - Text: each i_t is a string of \square and $\square\square\square\square\square i_t = \square$.
In fact i_t is a function $\square\square\square$ and three classes of texts based on the type of functions are
 - Arbitrary: arbitrary functions
 - Recursive: recursive functions
 - Primitive recursive: primitive recursive functions
This class of texts are effectively enumerable

Learnability Models

- Informant : each i_t is a string together with a binary signal indicating if the string is in \square .
Again i_t is a function $\square \rightarrow \square$ and three classes of informants based on the type of functions are
 - Arbitrary: arbitrary functions
 - Methodical: i_t is the i -th string in A^*
 - Request: i_t is requested by the learner (or equivalently i_t is *defined* by the learner)

- Learnability models

- Definitions of learnability

- A method of information presentation

- A naming relation

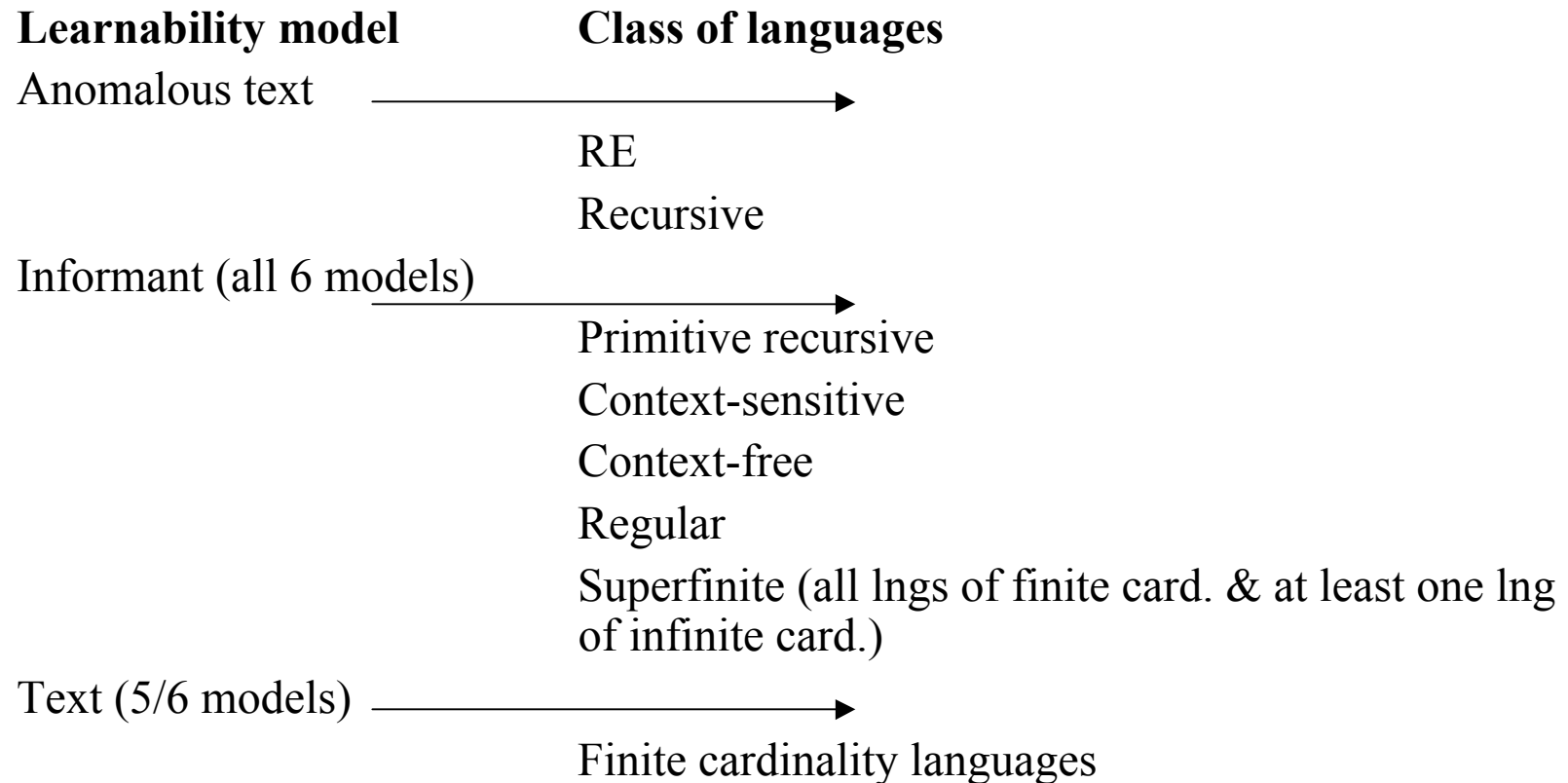
Learnability Models

- A naming relation consists of a set of names N and a function $f: N \rightarrow \Sigma^*$
- So the problem of language identification is to find a procedure by which given Σ^* and $I(\Sigma^*)$, find the name n such that $f(n) = \sigma$

Learnability Models

- Two naming relations
 - Tester (comprehensive/analysis): a binary decision procedure (TM) – 1(0) if the input string is (not) in Σ^* .
 - Generator (productive/generation): a TM generating a string in Σ^* given an input integer.
 - A tester exists to test a *recursive* language, a generator exists to generate *RE* language.
 - It is possible to translate from testers to generators, but not vice versa \square if testers are identifiable than so are the corresponding generators, but not vice versa, e.g., anomalous text.

Results and Implications



Results and Implications

- With pure positive data the model can only learn language with finite cardinality!
- If human mind is a computation device we have a problem explaining why children can pick up their mother tongues without negative evidence.
- Informant information presentation provides negative information in this framework.

Results and Implications

- Possible ways out
 - Human languages is none of the classes we studied (possibly a subset of some class such as context-sensitive)
 - Babies receive negative evidence anyway – we just don't know what that is (B&S's attempt)
 - Innateness claim: hardwired constraints may infer indirect negative evidence from the data

Proof Sketches

- Definitions

- Effective vs. ineffective identification

- There exists an algorithm for the former but the latter

- The former implies the latter

- Distinguishability condition: there is no information sequence describing two different languages, i.e., $\exists \alpha_i \alpha_j \exists I (\alpha_i) \neq I (\alpha_j) \wedge \alpha_i = \alpha_j$

Proof Sketches

– Collapsing uncertainty condition

Let Σ_t be the set of languages agreeing the information received so far, i.e.,

$\Sigma_t = \{\Sigma_j : i_1, i_2, \dots, i_t \text{ is a prefix of some sequence in } I(\Sigma_j)\},$

then $\Sigma_{t_{\Sigma}} = \{\Sigma\}$ (Σ is the correct language), or

equivalently, $\Sigma \Sigma' \neq \Sigma, \Sigma' \Sigma \in \Sigma_t.$

Or intuitively, the size of Σ_t is keeps decreasing.

Proof Sketches

– Identification by enumeration

An enumeration σ is an onto function $\mathbb{N} \rightarrow \mathbb{N}$ (this assumes \mathbb{N} is countable). At time t we find the first $\sigma(n)$ in \mathbb{N} that is in \mathbb{N}_t .

In the limit we return the name of the only element left in \mathbb{N}_t , if collapsing uncertainty condition holds.

To make this *effective* we need

- an effective procedure to test $\sigma(n) \in \mathbb{N}_t$
- an effective procedure to find a name of $\sigma(n)$

Proof Sketches

- Theorem
 - Ineffective identifiability \square distinguishability
 - Collapsing uncertainty \square identification by enumeration gives ineffective identification in the limit for any enumeration
 - $\square \square I (\square)$ is countable plus distinguishability \square ineffective identification in the limit

Proof Sketches

- Informant method satisfies the collapsing uncertainty condition so all 6 models can learn almost all classes of languages
 - Q: why not recursive and RE? Are they countable? (there're countable sets which are not RE, but how about the opposite direction?)

Proof Sketches

- Text method satisfies the distinguishability condition, which alone doesn't guarantee anything
 - but if $I(\sqsupset)$ is countable then ineffective identification in the limit is guaranteed – this is part of the reasons why anomalous text (generated by primitive recursive functions, using the generator naming relation) is identifiable.

Proof Sketches

- Why text is weak?
 - For a super-finite language we can always fool the learner by provide a successive larger finite subsets of the infinite language, so the learner makes mistakes for infinite number of times.
 - How to prevent repetitions? Probabilistic assumptions?

Learning Time

- Seemingly counterintuitive result: identification-by-enumeration is the most efficient method for the identification in the limit, and none of them (each using different enumerations) performs uniformly better than the other!

Learning Time

- Let $\tau(G, \square, \mathcal{I})$ denotes the time step when the guessing algo. G correctly identifies \square , given the information sequence \mathcal{I} .
- Prove $\tau(G, \square, \mathcal{I}) < \tau(G_0, \square, \mathcal{I}) \square \tau(G_0, \square', \mathcal{I}) < \tau(G, \square', \mathcal{I})$.

Empirical Studies: Bohannon &
Stanowicz's Experiment on Adult
Feedback, and Gordon's Attack

Discrepancy

- Gold showed with pure positive data (text) only languages of finite cardinality can be identified.
- It's a common belief that parents do not give negative evidence, or, do not perform the informant role as defined by Gold.
- How to account for the discrepancy?
- Hidden assumptions: mind is a TM

Discrepancy

- Solution 1: positing innate knowledge (Chomsky, Pinker, Wexler & Culicover, etc.)
- Solution 2: there *is* negative evidence (B&S, etc.)
- The definition of “negative evidence” is somewhat relaxed in B&S, as it consists of various types of repetitions and questions)

B&S on Negative Evidence

- Adults including parents & non-parents, male and female.
- Adult responses are categorized into three types of repetitions (exact, contracted, recasts and expansions) and two types of questions (repetitious/non-repetitious) followed by three types of children's language errors (semantic, syntactic & phonological).

B&S on Negative Evidence

- Claims
 - The experiment showed adults did respond children differentially based on the linguistic errors they made.
 - The responses give more information to language learners than the negative evidence defined in the strictest sense.
 - The result undermines the belief of innate knowledge (Occam's Razor)

Gordon's Attack

- Negative evidence and innateness are two orthogonal issues: even with reliable negative evidence we might need innate knowledge to learn a “human” language (which is different from the formal languages)
- B&S's results showed substantial proportion of the ill-formed utterances were not responded, and substantial proportion of the well-formed utterances received feedback pertinent to the ill-formed ones. It is not clear without innate knowledge how children knew which to ignore.

Gordon's Attack

- There are strong evidence supporting innate knowledge: the children whose parental input was unstructured Pidgin languages still acquired structured Creole language. In this case the parents cannot provide meaningful feedback.
- (My criticism) In B&S the use of MacWhinney's claim that low frequency events do not necessarily imply they aren't important (1982) has contradicting implications: this implies the children actually have the innate ability to avoid being misled by low frequencies.