Lecture 1: Fundamentals

Learnable representations for languages

Alexander Clark

Department of Computer Science
Royal Holloway, University of London

August 2010
ESSLLI, 2010

Outline

Introduction

- Linguistics
- Linguistic Nativism
- Empirical APS

Learnability

- Gold style learning
- Probabilistic learning
- Indirect negative evidence

Indirect Negative Evidence

- Proposal

Complexity problems

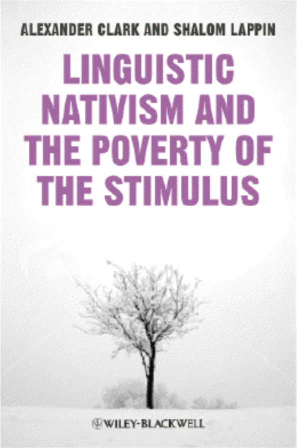
Conclusion

This lecture

Why should we care about learnability in language? What does learnability mean?

- The APS
- Linguistics and learnability
- Formal ideas of learnability:
    - Gold
    - PAC
- Negative results
- Arguments based on these negative results.

Book



ALEXANDER CLARK AND SHALOM LAPPIN

LINGUISTIC NATIVISM AND THE POVERTY OF THE STIMULUS

WILEY-BLACKWELL

What is linguistics?

The scientific study of language.

Mainstream Chomskyan view

Linguistics is a branch of psychology

Primary object of study

Adults have the ability to generate and understand sentences in their native language(s).

- We will call this “knowledge of language”
- Not interested in questions like “Is it propositionally represented?”
- Think of it as some “machinery”

Chomskyan view

Chomsky's questions (1986)

1. What constitutes knowledge of a language?
2. How is this knowledge acquired by its speakers?

The Central Debate

How does language acquisition proceed?

Empiricist hypothesis

It proceeds largely through domain-general processes of induction, generalisation and so on.

Linguistic Nativism

It proceeds largely through processes that are specific to the domain of language.

General nativism

Uncontroversial

- Lobsters can’t learn language
- Humans can
- Therefore: there is some part of our innate, genetically determined endowment that allows us to acquire and use language.

Common points

All parties (except a few philosophers) accept that there are innate mechanisms.

> *The behaviorist is knowingly and cheerfully up to his neck in innate mechanisms of learning readiness.*

Quine

Linguistic nativism

Working definition

(First) Language acquisition proceeds primarily through domain-specific mechanisms rather than through general-purpose learning mechanisms.

Linguistic nativism

Working definition

(First) Language acquisition proceeds primarily through domain-specific mechanisms rather than through general-purpose learning mechanisms.

- Distinction between domain-specific innate *knowledge* and domain-specific *processes* seems hard to sustain.
- Empirical debate not philosophical

Linguistic nativism

Working definition

(First) Language acquisition proceeds primarily through domain-specific mechanisms rather than through general-purpose learning mechanisms.

- Distinction between domain-specific innate *knowledge* and domain-specific *processes* seems hard to sustain.
- Empirical debate not philosophical
- Chomsky does not accept this definition: there are merely specific theories, rather than a general thesis.

Innateness

Philosophical analysis tends to muddy the waters. Originally a philosophical term predates discussion of genetics, but we will use it in a naive sense.

- Not learned from experience
- Genetically determined

Wexler

Chomsky's hypothesis is that many aspects of the formal structure of languages are encoded in the genome

Domain specific

- Spectrum from truly general purpose to domain specific.
- Parts of speech: clearly domain specific
- Sequence models: maths, music and other domains that might be parasitic on language.
- Hierarchical “tree” structures: occurs in nature. H Simon’s observations on *The architecture of complexity*.

Species specificity

Hauser

- Only humans know language.

Species specificity

Hauser

- Only humans know language.
- Therefore, if a cognitive ability is language specific, then it must be human specific.

Species specificity

Hauser

- Only humans know language.
- Therefore, if a cognitive ability is language specific, then it must be human specific.
- If we observe this ability in non human animals, then it is not language specific.

Universal Grammar

- Debate is about how much/what parts of language is innate.
- Important to avoid fallacy of equivocation in the term *UG*
  - UG can refer to the empirical claim that some non trivial parts of language are innate
  - UG can refer to the proportion of language is innate (and that proportion might be zero).
  - UG can refer to the initial state of the language faculty: presupposing that there is a domain specific language faculty

Opposing views

- Chomskyan family of views:
  - Principles and Parameters
  - Minimalist grammars
- Non-Chomskyan views
- Connectionist models
  - Rumelhart and McClelland
  - Elman et al. *Rethinking Innateness*
- Usage based/Emergentist/Constructionist approaches:
  - Tomasello (2000)
  - Goldberg (2003)

Minimalist Program

Principles and Parameters

All syntactic knowledge is innate except for a small number of binary parameters.

Recent developments in the MP

- Minimise the innate basis for language
- FLN versus FLB (Hauser, Chomsky and Fitch)
- FLN may just be recursion

Many different arguments for nativism

- The Argument from the Poverty of the Stimulus
- Genetic evidence
  - the KE family and SLI
  - Williams syndrome
- Localisation in the brain
- substantive Language universals

The Argument from the Poverty of the Stimulus

APS

The general form of the APS is an argument from the premise that

1. The primary linguistic data (PLD) to which children have access is not sufficient to support the acquisition of adult linguistic competence through data driven learning procedures.

The Argument from the Poverty of the Stimulus

APS

The general form of the APS is an argument from the premise that

1. The primary linguistic data (PLD) to which children have access is not sufficient to support the acquisition of adult linguistic competence through data driven learning procedures.

to the conclusion that

2. Language acquisition requires a rich set of innate language specific learning constraints encoded in a Universal Grammar (UG).

Empirical versions of the APS

General argument

Children learn facts for which there is no evidence in the input.

Specific arguments

Auxiliary fronting
One anaphora
English auxiliaries etc.

Auxiliary Inversion: A Case Study

- Chomsky (1971, 1975), Crain and Nakayama (1987), and Crain (1991) (among others) take auxiliary inversion in polar questions to be an instance of children acquiring a structure-dependent rule without having access to the relevant evidence in the PLD.

Auxiliary Inversion: A Case Study

- Chomsky (1971, 1975), Crain and Nakayama (1987), and Crain (1991) (among others) take auxiliary inversion in polar questions to be an instance of children acquiring a structure-dependent rule without having access to the relevant evidence in the PLD.
- Crain (1991) describes this as the "parade case of an innate constraint".

Auxiliary Inversion: A Case Study

- Chomsky (1971, 1975), Crain and Nakayama (1987), and Crain (1991) (among others) take auxiliary inversion in polar questions to be an instance of children acquiring a structure-dependent rule without having access to the relevant evidence in the PLD.
- Crain (1991) describes this as the "parade case of an innate constraint".
- It is possible to specify an operation for forming polar interrogatives in English by means of two distinct rules.

Two Possible Inversion Rules

(1a) The student in the garden is hungry.

(b) Is the student in the garden hungry?

Two Possible Inversion Rules

(1a) The student in the garden is hungry.

(b) Is the student in the garden hungry?

(2a) Front the first auxiliary in the string to the beginning of the main clause.

Two Possible Inversion Rules

(1a) The student in the garden is hungry.

(b) Is the student in the garden hungry?

(2a) Front the first auxiliary in the string to the beginning of the main clause.

(b) Front the auxiliary of the main VP to the beginning of the main clause.

Linear Order vs. Constituent Structure

(2a), a linear counting rule, is the simpler rule, but (2b), a constituent structure dependent operation, is the correct one.

Linear Order vs. Constituent Structure

(2a), a linear counting rule, is the simpler rule, but (2b), a constituent structure dependent operation, is the correct one.

(3a) The student who is in the garden is hungry.

(b) Is the student who is in the garden hungry?

(c) *Is the student who in the garden is hungry?

Structure Dependence and the APS

Advocates of this instance of the APS make the following claims.

Structure Dependence and the APS

Advocates of this instance of the APS make the following claims.

- Considered as a hypothesis about the available data, (2b) is a less natural rule than (2a)

Structure Dependence and the APS

Advocates of this instance of the APS make the following claims.

- Considered as a hypothesis about the available data, (2b) is a less natural rule than (2a)
- Children do not generally make errors like (3c).

Structure Dependence and the APS

Advocates of this instance of the APS make the following claims.

- Considered as a hypothesis about the available data, (2b) is a less natural rule than (2a)
- Children do not generally make errors like (3c).
- Sentences like *Is the student who is in the garden hungry?* (3b) do not occur (or do not occur frequently) in the linguistic data available for language acquisition.

Empirical debate

Frequency

How frequent are these constructions?
such as *Is the student who is in the garden hungry?*

Pullum and Scholz

Point out that nobody really tried to count them
They are actually quite frequent

Rebuttals and retorts by Legate and Yang in special issue of Linguistic Review.

Problem

What is the relevant class of constructions?

- Impossible to decide how large the class should be without a learning theory.
- Target sentence: *Is the student who is in the garden hungry?*

Problem

What is the relevant class of constructions?

- Impossible to decide how large the class should be without a learning theory.
- Target sentence: *Is the student who is in the garden hungry?*
- Possible data:
    - *Is the student who is in the garden hungry?*

Problem

What is the relevant class of constructions?

- Impossible to decide how large the class should be without a learning theory.
- Target sentence: *Is the student who is in the garden hungry?*
- Possible data:
    - *Is the student who is in the garden hungry?*
    - *Is the professor who is in the house hungry?*

Problem

What is the relevant class of constructions?

- Impossible to decide how large the class should be without a learning theory.
- Target sentence: *Is the student who is in the garden hungry?*
- Possible data:
    - *Is the student who is in the garden hungry?*
    - *Is the professor who is in the house hungry?*
    - *Did the man over there give me a biscuit?*
    - Any sentences that support hierarchical structure.

A general problem with this class of argument.

Outline

Introduction
- Linguistics
- Linguistic Nativism
- Empirical APS

Learnability
- Gold style learning
- Probabilistic learning
- Indirect negative evidence

Indirect Negative Evidence
- Proposal

Complexity problems

Conclusion

A priori APS

Other versions of the APS rely on formal general arguments that derive from the formal theory of learnability

> *The strongest most central arguments for innateness thus continue to be the arguments from APS and learnability theory. . . . The basic results of the field include the formal, mathematical demonstration that without serious constraints on the nature of human grammar, no possible learning mechanism can in fact learn the class of human grammars. Wexler, MIT Encyclopedia of the Cognitive Sciences.*

A lot of what people write on this subject is quite shockingly ignorant.

Some typical quotes

Names removed to protect the innocent.

Quotes from some recent papers that allude to Gold's paper:

- The input does not include reliable negative evidence, . . . logical arguments suggest that in the absence of such evidence there must be strong innate constraints on the possible forms of grammars (and then a citation to Gold)
- (Gold) provided a logical proof which concluded that, without explicit error correction, the rules of a logical system with the structural complexity of a natural language grammar could not be inductively discovered, even in theory.

Some typical quotes (II)

Names removed to protect the innocent.

- Gold (1967) . . . obtained results that implied that natural languages could not be learned only on the basis of positive evidence
- Gold showed that, for even simple classes of languages, no procedure (statistical or other) exists that could learn a language without non-trivial a priori assumptions.
- The problem is presented even more strikingly by Gold (1967) who, simulating language acquisition on a computer, argues that an unbiased learner who had to induce the rules of grammar from strings of input would require more than a human lifetime.

Some typical quotes (III)

Names removed to protect the innocent.

- Gold asked the question: under what conditions is it possible to learn the correct context free grammar of a language given a set of training instances? His most significant result was that it is impossible to learn the correct language from positive examples alone. If a blind inductive program is given an infinite sequence of positive examples the program cannot determine a grammar for the correct context free language in any finite time.
- Gold (1967) proved that a general learner who has no *a priori* knowledge of the language to be learned cannot learn any language that has an infinite number of sentences from text presentation.

APS arguments based on Gold’s theorems

- Chomsky has never used these arguments.

APS arguments based on Gold’s theorems

- Chomsky has never used these arguments.
- There have been several books and long articles on this subject.
  - “Learnability and Linguistic Theory”, Matthews and Demopoulos
  - “Formal Principles of Language Acquisition”, Wexler and Culicover
  - “Language Acquisition and Learnability.” Bertolo
- Good discussion in Johnson (2004) “Gold’s theorem in Cognitive Science”, but some technical flaws.
- All of these accept the Gold model.

The formal theory of learnability

- Using formal models to study learning problems permits one to study the formal limits of learnability under specified assumptions concerning the learning situation.

The formal theory of learnability

- Using formal models to study learning problems permits one to study the formal limits of learnability under specified assumptions concerning the learning situation.
- The assumptions define the nature of the learning process and the data on the basis of which learning is achieved.

The formal theory of learnability

- Using formal models to study learning problems permits one to study the formal limits of learnability under specified assumptions concerning the learning situation.
- The assumptions define the nature of the learning process and the data on the basis of which learning is achieved.
- These involve idealizing the learning situation to facilitate formal modeling.

The formal theory of learnability

- Using formal models to study learning problems permits one to study the formal limits of learnability under specified assumptions concerning the learning situation.
- The assumptions define the nature of the learning process and the data on the basis of which learning is achieved.
- These involve idealizing the learning situation to facilitate formal modeling.
- The results which one can prove in a model depend on the way in which one defines the object to be learned, what learning consists in, and the evidence available to the learner.

Balance

It is easy to make the model wrong

Too easy

The learner can learn anything using an unrealistic model.

Too hard

The learner cannot even learn the class of observed natural languages

In both cases, the modeling assumptions must be wrong.
We get no insight.

Formal models of language learning

Pinker (1979)

Put forward various conditions for a model of language learning.

1. Learnability condition
2. Equipotentiality condition
3. Time condition
4. Input conditions
5. Developmental condition
6. Cognitive condition

Formal models of language learning

Pinker (1979)

Put forward various conditions for a model of language learning.

1. Learnability condition
2. Equipotentiality condition
3. Time condition
4. Input conditions
5. Developmental condition
6. Cognitive condition

- Mathematical models tend to ignore the last two.
- Missing an evolutionary criterion?

The Value of Formal Learning Models

- It is reasonable to ask whether formal learning models are useful in understanding human language acquisition.

The Value of Formal Learning Models

- It is reasonable to ask whether formal learning models are useful in understanding human language acquisition.
- Isn't it the case that knowledge of this process depends entirely on the psychological and biological facts of acquisition? It's an empirical problem and will be settled by empirical data.
- The formal study of grammar induction can clarify the sort of data required, and the nature of the learning biases that are required.
- Such learnability results establish basic conditions of adequacy that theories of grammar and language acquisition must satisfy.

The Value of Formal Learning Models

- It is reasonable to ask whether formal learning models are useful in understanding human language acquisition.
- Isn't it the case that knowledge of this process depends entirely on the psychological and biological facts of acquisition? It's an empirical problem and will be settled by empirical data.
- The formal study of grammar induction can clarify the sort of data required, and the nature of the learning biases that are required.
- Such learnability results establish basic conditions of adequacy that theories of grammar and language acquisition must satisfy.

How do birds fly?

Empirical point, but aerodynamics puts some boundary conditions on it.

Computational learning theory

Two problems

- Information theoretic problems
- Computational complexity problems

Learning theory should help!

- Unfortunately, it has often been distorted and used to stifle research rather than guide it.
- If X is mathematically impossible, then don't try doing X ...
- If you have a program that seems to do X, then it is probably actually doing something else *Y*

Learning models

Basic components:

- Objects to be learned – languages/grammars
- The learner has a source of information: positive examples, queries
- The learner has to create a hypothesis using this information
- The hypothesis must converge to the (a) right answer as the learner gets more information.
- Constraints on the type of information, the amount of information, the amount of computation.

Basic assumptions

Abstract learning problem

Alphabet

A finite alphabet Σ – elements of Σ might be words, phonemes, POS tags
Not normally “letters”

Objects to be learned

Σ^* is the set of all finite strings:

- A language $L \subseteq \Sigma^*$
- A distribution over Σ^*

Grammatical sentences, phonotactically well formed sequences etc.

We can treat a lot of different linguistic learning problems as instances of the same problem.

Chomsky hierarchy

Traditional classification of languages

‘

Finite languages

Regular languages

Context-free languages

Context-sensitive languages

Recursive languages

All languages

This is the wrong way to think about language classes.

Idealisation: A single level

Natural languages have different levels. At the very least we have morphology and syntax. We conflate all of these into a single task: learning a formal language.

- Given a finite alphabet Σ, we want to learn a language $L \subset \Sigma^*$.
- Σ could be a set of phonemes, or characters, or a set of words, or a set of lexical categories (POS tags)
- The language could be the set of well-formed sentences, or the set of words that obey the phonotactics of the language .
- We have a single abstract task that represents several different real tasks.

Idealisation: A single level

Natural languages have different levels. At the very least we have morphology and syntax. We conflate all of these into a single task: learning a formal language.

- Given a finite alphabet Σ, we want to learn a language $L \subset \Sigma^*$.
- Σ could be a set of phonemes, or characters, or a set of words, or a set of lexical categories (POS tags)
- The language could be the set of well-formed sentences, or the set of words that obey the phonotactics of the language .
- We have a single abstract task that represents several different real tasks.

Use syntax to be concrete.

Example

A sequence of examples

- *ab*
- *abab*
- *ab*
- *ababab*
- λ
- ...

Information sources

Passive

- Given a sequence of unlabelled strings: $w_1, w_2 \ldots$, all in L
- Labelled strings: $(w_1, 1), (w_1, 0) \ldots$; 1 means it is in L, 0 means it is not in L
- Structural examples: positive examples and parse trees
- Syntax-semantics pairs

Active

Learner has some control

- Pick a string w and find out if $w \in L$
  Membership queries (MQ)
- Ask whether a hypothesis is right or not?
  Equivalence queries (EQ)

Convergence criterion

Gradual convergence

Define error function – distance between target and hypothesis
This must converge to zero, as the data increases.

Exact convergence

At some point we have exactly the right answer
A finite number of errors
Distance 0 if equal, 1 otherwise

One-shot learning

You produce a single hypothesis, which must be correct

Example

- Suppose the language is just $(ab)^*$ $\{\lambda, ab, abab, \ldots\}$.
- The learner receives a sequence of examples from L

Example

- Suppose the language is just $(ab)^* \ \{\lambda, ab, abab, \ldots\}$.
- The learner receives a sequence of examples from L
- $abab, abab, ab, ab, abab, \lambda, ababab \ldots$
- There are lots of possible hypotheses

Learnability

Learnable class

Given a class of languages $\mathcal{L}$
If the learner can learn every element of $\mathcal{L}$, then we say it can learn $\mathcal{L}$

Pause and note how strange this is:

- Singleton classes can trivially be “learned”
- Huge difference between the normal notion of learning and the formal notion.

Very damaging effect on theorising

Gold (1967)

Historically the first

- In the Identification in the Limit (IIL) paradigm a language consists of a set of strings, and a learner is presented with a sequence of strings.

Gold (1967)

Historically the first

- In the Identification in the Limit (IIL) paradigm a language consists of a set of strings, and a learner is presented with a sequence of strings.
- The sequence can be specified as $s_1, s_2, \ldots,$ and it is infinite.
- Every string of a language appears at least once in a sequence.
- Every string that appears is in the language.

The Gold Paradigm

- The samples of the sequence are presented to the learner one at a time, and on the basis of this evidence, he/she must, at each step, propose a hypothesis for the identity of the language.

The Gold Paradigm

- The samples of the sequence are presented to the learner one at a time, and on the basis of this evidence, he/she must, at each step, propose a hypothesis for the identity of the language.

- Given the first string s_1, the learner produces a hypothesis G_1, in response to s_2, he/she will, on the basis of s_1 and s_2 generate G_2, and so on.

Learning in the Gold Paradigm

- For a language L and a presentation of that language $s_1, s_2, \ldots$, the learner identifies in the limit the language L, if there is some N such that for all $n > N$, $G_n = G_N$, and G_N is a correct representation of L.

Learning in the Gold Paradigm

- For a language L and a presentation of that language $s_1, s_2, \ldots,$ the learner identifies in the limit the language L, if there is some N such that for all $n > N$, $G_n = G_N$, and G_N is a correct representation of L.
- IIL requires that a learner converge on the correct representation G_L of a language L in up to the limit of an infinite period of time, on the basis of an unbounded sequence of data samples, and, after constructing G_L, he/she does not depart from it in response to subsequent data.

Learning in the Gold Paradigm

- For a language L and a presentation of that language $s_1, s_2, \ldots$, the learner identifies in the limit the language L, if there is some N such that for all $n > N$, $G_n = G_N$, and G_N is a correct representation of L.
- IIL requires that a learner converge on the correct representation G_L of a language L in up to the limit of an infinite period of time, on the basis of an unbounded sequence of data samples, and, after constructing G_L, he/she does not depart from it in response to subsequent data.
- A learner identifies in the limit the class of languages $\mathcal{L}$ iff the learner identifies in the limit every L in $\mathcal{L}$.

Convergence

- We want the hypothesis produced by the learner to converge in some sense to the right language.
- Gold uses an exact convergence criterion.

Identification in the limit

There is some number N such that for all $i > N$, $G_i = G_N$ and $L(G_N) = L$.

This must happen for every presentation.

Overall picture

Identification of a language L

We say that the language L is identified if for every presentation of L, the learner converges to the language L

Identification of a language class $\mathcal{L}$

We say that the language class $\mathcal{L}$ is identified if for every $L \in \mathcal{L}$ and for every presentation of L, the learner converges to the language L

Summary of Gold

Positive only

- Exact convergence at some point
- Positive data only
- For every possible presentation, even if chosen adversarially

The adversary/teacher can defer the presentation of positive examples indefinitely. This means it is very difficult to tell what is not in the language.
A clever adversary can often trick the learner into making an infinite number of errors.

Positive results in Gold

1

Finite languages

The class of finite languages is learnable

Rote learner

- Simply memorises all of the examples.
- $H_i = \{w_1, \ldots w_i\}$.
- Clearly converges since every presentation must eventually give all of the examples.

Positive results in Gold

2

Finite classes

Any finite class of languages is learnable
As long as we can decide whether $w \in L$

Algorithm

Order the languages so the larger languages are later.
Pick the first language that includes the observed data.

Negative results in Gold

Suprafinite language class

All finite languages and at least one infinite language

Theorem

No suprafinite class is IIL

Corollary

The class of regular languages is not IIL
The class of context-free languages is not IIL
The class of ontext-sensitive languages is not IIL

Identification in the limit

Criticism of the model

Why does Gold's framework not apply?
(Why *does* it apply?)
IIL requires asymptotic exact identification under any *presentation* of the data. Every string in the language must be presented at least once, no string not in the language can be presented.

- Too easy because it has no limit on resources: unbounded amounts of data and computation.
- Too difficult because it allows deliberately misleading presentations, and requires exact result.
- Convergence criterion is useless (too weak) since it is asymptotic: it tells you nothing for finite amounts of data

Variants: stochastic presentation, IIL with probability one, using polynomial computation, rational coefficients.

Randomness

There is a random element to the exact sentences that the child sees.

- The sentences are not picked to mislead the child
- The exact set of sentences varies from child to child
- Standard assumptions about examples being drawn independently.

PAC model

Probably approximately correct model of learning (Valiant, 1984).

Irrelevant to language acquisition

Distribution free learning of labelled examples
All information is in the labels
None in the distribution

Negative evidence

Evidence about what is **not** in the language:

- Indirect feedback: utterances generated by the learner are misunderstood.
- Corrections: utterances generated by the learner are corrected by the teacher
- Direct negative evidence: sentences generated by the teacher and marked as ungrammatical.

Long and circular debate (Marcus, 1993).

Negative evidence

- Negative evidence is not considered a problem in NLP or in unsupervised learning: either in theory or in practice.
- Direct negative evidence is completely useless in practice: almost all long strings of English words are ungrammatical.
  - Jim address tasting array umpet tag ever zoo minibeasts dodo
  - party Victoria claps wrecking weakness spanked grips apricots lunchbox bell
  - surgery gymnast taxi washable ropes cleaner measurer Scotsman rummage gracious

Indirect Negative Evidence

Chomsky (1981):

> *A not unreasonable acquisition system can be devised with the operative principle that if certain structures or rules fail to be exemplified in relatively simple expressions, where they would expect to be found, then a (possibly marked) option is selected excluding them in the grammar, so that a kind of "negative evidence" can be available even without corrections, adverse reactions etc.*

Implicitly, ML or MAP probabilistic models exploit INE.

Outline

Introduction

- Linguistics
- Linguistic Nativism
- Empirical APS

Learnability

- Gold style learning
- Probabilistic learning
- Indirect negative evidence

Indirect Negative Evidence

- Proposal

Complexity problems

Conclusion

Technical proposal

- Reconcile INE with probabilistic learning
- Provide a *formal* justification for the *informal* notion of INE.

Technical proposal

- Reconcile INE with probabilistic learning
- Provide a *formal* justification for the *informal* notion of INE.

Two steps:

1. Low observed frequency means low probability
2. Low probability means ungrammaticality (sometimes)

Probabilistic learning

Basic assumptions

IID Assumptions

- The probability distribution is the same over time
- Each sentence is independent of the previous one.

- IID assumption considered controversial in FLA

  . . .

Probabilistic learning

Basic assumptions

IID Assumptions

- The probability distribution is the same over time
- Each sentence is independent of the previous one.

- IID assumption considered controversial in FLA
  . . . because it's clearly false.
- There are dependencies between sentences: question/answer pairs etc.
- What we need is law of large numbers and large deviation bounds.
- Ergodic theory: ergodic, rapidly mixing distributions.

INE not available

In standard probabilistic models we don't get INE:

- PAC model from positive data (Shvayster, 1990)
- Probabilistic IIL (Angluin, 1988)

Distribution free assumption:

- The distribution is picked by an adversary
- Low probability does not mean ungrammaticality: EVER
- Effectively no difference between $p(w) > 0$ and $p(w) \geq 0$.

Three uncontroversial assumptions

1. Children can in fact learn at least the class of attested natural languages.

Three uncontroversial assumptions

1. Children can in fact learn at least the class of attested natural languages.
2. Data is unlabelled – not marked whether it is grammatical or ungrammatical.

Three uncontroversial assumptions

1. Children can in fact learn at least the class of attested natural languages.
2. Data is unlabelled – not marked whether it is grammatical or ungrammatical.
3. There are some ungrammatical sentences in the input.

What is the distribution?

Class of languages

A class of languages $\mathcal{L}$.

Class of distributions

For each language L in $\mathcal{L}$, we have a set of distributions $\mathcal{D}(L)$.

We must learn language L with every distribution in $\mathcal{D}(L)$.

Disjoint distribution assumption

Logical consequence of learnability from unlabelled data: the only information is the distribution.

DDA

If L_1 is different from L_2, then $\mathcal{D}(L_1)$ must be disjoint from $\mathcal{D}(L_2)$.

Alternative formulation

There must be a partial function from $\mathcal{D}$ to L.

Default assumption

Standard assumption:

$$\mathcal{D}(L) = \{D : p_D(w) > 0 \text{ iff } w \in L\}$$

Language is equal to support of the distribution.

- Too small: doesn't allow for ungrammatical sentences.
- Too big: allows arbitrarily small probabilities, that rule out INE.

Bounds on probability

Grammaticality

Larger probability correlates with grammaticality.

Express function from distribution to language as a bound:

Bound

$w \in L$ if $p_D(w) \geq g_D(w)$

Two simplistic assumptions

Zero bound

$g_D(w) = 0$
Only grammatical strings have non-zero probability.

Fixed bound

$g_D(w) = \epsilon$.
Language must be finite: at most ϵ^{-1} strings.

Neither bound depends on D or on w.

Better bounds depending on w

Exponential length bound

$g_D(w) = \alpha\beta^{|w|}$
Two fixed parameters $\alpha, \beta < 1$

- “Departure from equiprobability” Harris
- Longer sentences are rarer
- Too crude: Lexical items differ in probability

Depending on w and D

Unigram bound

$w = u_1 \ldots u_n$

$g_D(w) = p(n) \prod_{i=1}^{n} p(u_i)$

Depending on *w* and *D*

Unigram bound

$w = u_1 \dots u_n$

$g_D(w) = p(n) \prod_{i=1}^{n} p(u_i)$

Prefix suffix?

$g_D(uv) = \alpha p_D(u\Sigma^*) p_D(\Sigma^* v)$

Context substring!?

$g_D(lur) = \alpha p_D(l\Sigma^* r) p_D(\Sigma^* u \Sigma^*)$

Consequences

- If there is such a bound then low probability sometimes leads to ungrammaticality.
- Combined with IID assumptions, or similar, we can then infer ungrammaticality from low frequency.
- This means we can simulate membership queries for some strings.

- Estimate $\hat{p}(w) = n(w)/N$
- Calculate $g_D(w)$ if it doesn't depend on D
- With polynomial N, can show $|p_D(w) - \hat{p}(w)| \leq \epsilon$ with prob $> 1 - \delta$.
- If $\hat{p}_D(w) \leq g_D(w) - \epsilon$ then w is ungrammatical.

Probabilities

Conclusion

	Inefficient	Efficient
Positive data and MQs	Gold (1967)	?
Stochastic data	Horning (1969) Angluin (1988) Chater and Vitanyi (2007)	?

These results suggest the presence of probabilistic data largely compensate for the absence of negative data. (Angluin, 1988)

Gold and probabilistic learning

Going forward:

- Positive data only chosen adversarially is unrealistically hard. (but we will still look at some algorithms of this type)
- Positive data plus membership Queries (which is slightly too easy but convenient)
- Realistic probabilistic models (which are technically quite hard, but the most convincing)

real situation is of course more complex.

Outline

Introduction
- Linguistics
- Linguistic Nativism
- Empirical APS

Learnability
- Gold style learning
- Probabilistic learning
- Indirect negative evidence

Indirect Negative Evidence
- Proposal

Complexity problems

Conclusion

Two problems of grammar induction

First problem

Information theoretic problems

A general problem of learning:

- Absence of negative data (Gold, 1967)
- VC-dimension (Vapnik, 1998), covering numbers
- Sparsity, Noise etc.

We know how to attack these problems: MDL, NPB, MaxEnt

Not specific to grammatical inference

Two problems of grammar induction

Second problem

Computational problems

Complexity of finding a good hypothesis given this information

- Gold (1978), Kearns and Valiant (1989) ...
- Often based on embedding cryptographic problems in learning problems
- Specific to certain classes of representation

Tractable Cognition Thesis (van Rooij, 2008)

Human cognitive capacities are constrained by the fact that humans are finite systems with limited resources for computation.

This is the crucial problem: given good information about the language, can we efficiently construct a representation?

Negative results

Negative results in Machine Learning (PAC-learning) come in two types

- Information-theoretic bounds on sample complexity: VC-dimension

Negative results

Negative results in Machine Learning (PAC-learning) come in two types

- Information-theoretic bounds on sample complexity: VC-dimension
- Cryptographic limitations on computational complexity. Standard assumptions that there are no efficient algorithms for common cryptographic problems:
  - factoring Blum integers
  - inverting RSA function
  - recognizing quadratic residues
  - noisy parity assumption

Negative results

Negative results in Machine Learning (PAC-learning) come in two types

- Information-theoretic bounds on sample complexity: VC-dimension
- Cryptographic limitations on computational complexity. Standard assumptions that there are no efficient algorithms for common cryptographic problems:
  - factoring Blum integers
  - inverting RSA function
  - recognizing quadratic residues
  - noisy parity assumption
- Even if you have enough information, learners could not work out what the correct grammar is.

Cryptographic limitations on learning Boolean formulae and finite automata

Kearns and Valiant (1989)

- PAC-learning deterministic finite state automata is not possible. (cryptographic assumptions)
- Even with positive *and* negative evidence
- Distribution free setting

Abe and Warmuth (1992)

On the complexity of approximating . . .

- ML training of HMMs on an arbitrary training set is NP-hard.
- Similarly for PCFGs.
- As the alphabet size grows, finding the model with maximum likelihood become difficult.
- There are still approximate training algorithms (EM algorithm).

Kearns et al. (1994)

- “On the learnability of discrete distributions”
- It is not possible to approximate distributions generated by a simple acyclic FA.

Summary

Most of these results are again a little too pessimistic: they rely on learning under unhelpful distributions.

- If the distribution is benign then learning can be easier
- Dana Angluin and Michael Kharitonov (1995)
    - Draws an interesting dviding line: some models (DFAs) are learnable, and some (NFAs, CFGs) are not.

These results actually give us some insight: they give a definite direction.

Outline

Introduction
- Linguistics
- Linguistic Nativism
- Empirical APS

Learnability
- Gold style learning
- Probabilistic learning
- Indirect negative evidence

Indirect Negative Evidence
- Proposal

Complexity problems

Conclusion

Tension

Chomsky, 1986

> *To achieve descriptive adequacy it often seems necessary to enrich the system of available devices, whereas to solve our case of Plato's problem we must restrict the system of available devices so that only a few languages or just one are determined by the given data. It is the tension between these two tasks that makes the field an interesting one, in my view.*

Tension

Chomsky, 1986

> *To achieve descriptive adequacy it often seems necessary to enrich the system of available devices, whereas to solve our case of Plato's problem we must restrict the system of available devices so that only a few languages or just one are determined by the given data. It is the tension between these two tasks that makes the field an interesting one, in my view.*

Principles and Parameters

No good learning model; No agreement on parameters; No tension.

Unsupervised Learning

Fundamental problem of linguistics

Chomsky's questions

1. What constitutes knowledge of a language?
2. How is this knowledge acquired by its speakers?

Pinker (1990)

To understand how X is learned, you first have to understand what X is.

Crain and Pietroski (2001)

First, one tries to find principles that characterize human grammars; *then* one tries to determine which aspects of these grammars could plausibly be learned from experience, and which are more likely to be innately specified.

Standard methodology

- Step 1: Construct a descriptively adequate representation
- Step 2: Try to design learning algorithms for those representations

Step 1

Construct a descriptively adequate grammar

This failed

- No-one ever managed to make a descriptively adequate grammar for any natural language, not even English.
- In order to account for new facts (e.g. Swiss German) representations were made more powerful and expressive.
- Statistical parsers do not separate grammatical from ungrammatical sentences (Okanohara and Tsujii, 2007; Berwick and Fong, 2008)
- Generative grammarians have largely abandoned the task of constructing large scale grammars.

Step 2

Come up with a learning algorithm

This also failed.

- Learning even regular grammars is computationally hard: Angluin and Kharitonov (1995)
- We have some heuristic algorithms that can induce crude constituent structure (Klein and Manning, 2004)
- The classes of representations we need have even richer, deeper and more abstract hidden structure: (LTAG, $ACG_{2,4}$, ...)
- It is out of the question to construct learning algorithms for these classes.

PSGs were meant to be learnable

Chomsky (1968/2006)

“The concept of "phrase structure grammar" was explicitly designed to express the richest system that could reasonable be expected to result from the application of Harris-type procedures to a corpus.”

Linguists don’t know what the representations are

A Cambridge quote

“At the most fundamental level, it is not clear that there is any meaningful empirical motivation for the representational assumptions of any current formal model of syntax.”
(Blevins, J., 2009)

Linguists cannot agree whether the head of “the cat” is “the” or “cat”. Nor can they produce any empirical evidence to decide between the two.
(Matthews, P.; 2007)

Linguists don’t know what the representations are

A Cambridge quote

“At the most fundamental level, it is not clear that there is any meaningful empirical motivation for the representational assumptions of any current formal model of syntax.”
(Blevins, J., 2009)

Linguists cannot agree whether the head of “the cat” is “the” or “cat”. Nor can they produce any empirical evidence to decide between the two.
(Matthews, P.; 2007)

- We don’t know what the representations are but we do know that they are learnable!

Reasonable Research Strategy

Slogan

Put learnability first!

- If you construct a super-powerful class of languages with no thought of learnability, you won't be able to learn them.
- Rather, design representations from the ground up to be learnable.

Reasonable Research Strategy

Slogan

Put learnability first!

- If you construct a super-powerful class of languages with no thought of learnability, you won't be able to learn them.
- Rather, design representations from the ground up to be learnable.

Strategy

- Step 1: build simple learnable representations
- Step 2: gradually try to increase their expressive power, while maintaining learnability

Reverse direction

Normal direction

Function from representation to language

Context free grammar $G \rightarrow$ context free language $L(G)$

Non-terminal → set of strings derived from non-terminal

Reverse direction

Normal direction

Function from representation to language
Context free grammar $G \rightarrow$ context free language $L(G)$
Non-terminal → set of strings derived from non-terminal

Opposite Direction

Function from language to representation
$L \rightarrow R(L)$
From set of strings → representational primitive of formalism

Ideally $L(R(L)) = L$.

Empiricist models

Slogan

The structure of the representation should be based on the structure of the language, not something arbitrarily imposed on it from outside.

- Identify some structure in the language
- Show how that structure can be observed
- Construct a representation based on that structure
- Richer structures will give you more powerful representations

Objective representations

Program

Given a language *L*

1. Define a collection of sets of strings as primitives
2. Define a derivation relation, based on algebraic properties of these sets.
3. Define a representation based on this derivation relation

Learnability

- Grammatical inference is crucial and representation classes need to be designed to be learnable
- The structure of the representation should be based on the structure of the data
- Applying this approach gives efficient algorithms
  - Congruence based approaches using CFGs
  - Syntactic concept lattice
  - Distributional Lattice Grammars
- These are the only efficient algorithms for large classes of context free languages.