6.863J Natural Language Processing
Lecture 21: Language Acquisition Part 1

Robert C. Berwick

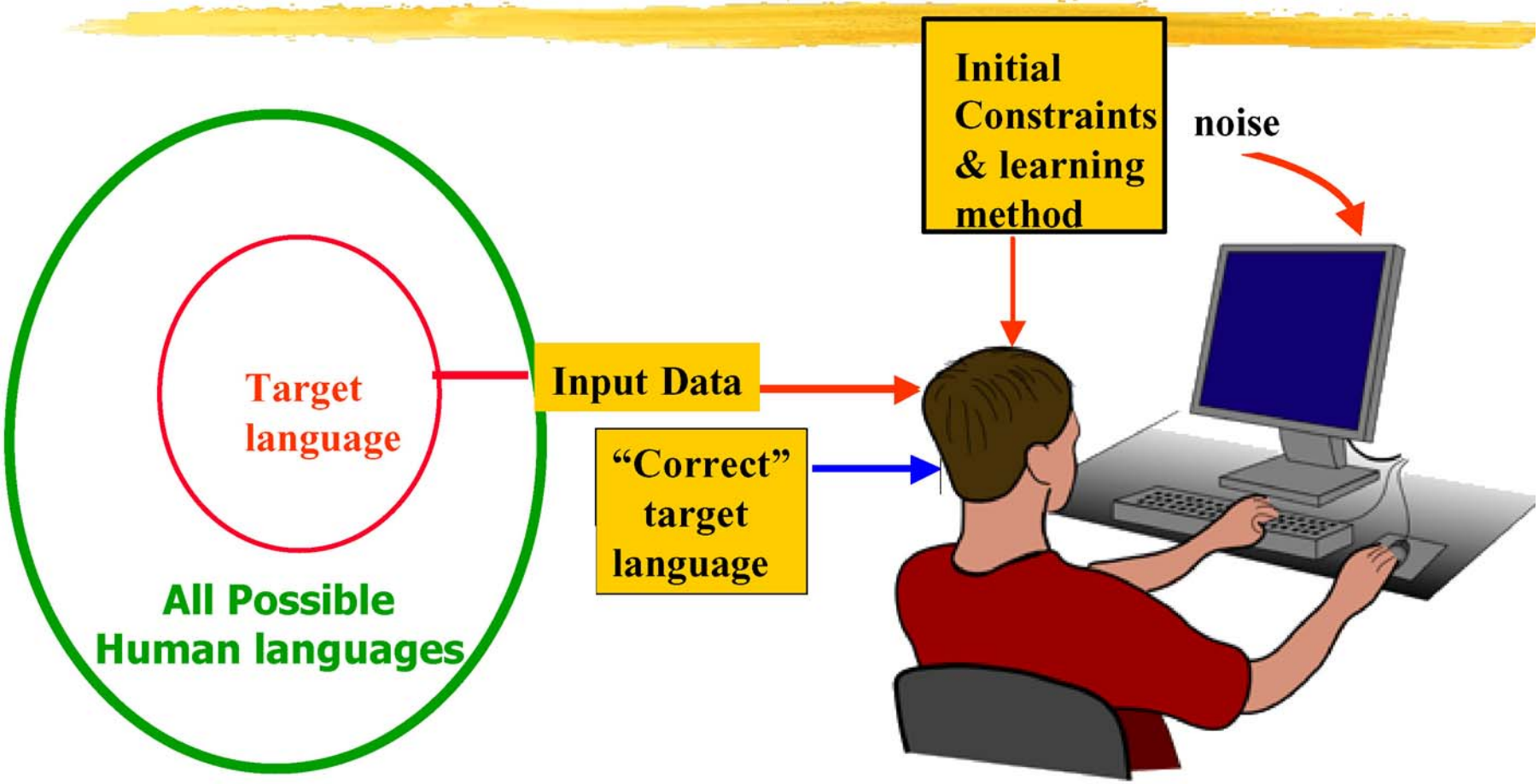
The Menu Bar

- Administrivia:
  - Project check
- The Twain test & the Gold Standard
- The Logical problem of language acquisition: the Gold theorem results
- How can (human) languages be learned?
- The logical problem of language acquisition
  - What is the problem
  - A framework for analyzing it

The Twain test

- Parents spend….

The Logical problem of language acquisition



The problem

- From finite data, induce infinite set
- How is this possible, given limited time & computation?
- Children are not told grammar rules

- Ans: put constraints on class of possible grammars (or languages)

The logical problem of language acquisition

- Statistical MT: how many parameters? How much data?
- "There's no data like more data"
- Number of parameters to estimate in Stat MT system -

The logical problem of language acquisition

- Input does not uniquely specify the grammar (however you want to represent it) = Poverty of the Stimulus (POS)
- Paradox 1: children grow up to speak language of their caretakers
- Proposed solution: target choice of candidate grammars is restricted set
- This is the theory of Universal Grammar (UG)
- (Paradox 2: why does language evolve?)

The illogical problem of language change

Langagis, whos reulis ben not writen as ben Englisch, Frensch and many otheres, ben channgid withynne yeeris and countrees that oon man of the oon cuntre, and of the oon tyme, myghte not, or schulde not kunne undirstonde a man of the othere kuntre, and of the othere tyme; and al for this, that the seid langagis ben not stabili and fondamentali writen

Pecock (1454) *Book of Feith*

Information needed

- Roughly: sum of given info + new info (data) has to pick out right language (grammar)
- If we all spoke just 1 language – nothing to decide – no data needed
- If just spoke 2 languages (eg, Japanese, English), differing in just 1 bit, 1 piece of data needed
- What about the general case?

Can memorization suffice?

- Can a big enough table work?
- Which is largest, a <noun>-1, a <noun>-2, a <noun>-3, a <noun>-4, a <noun>-5, a <noun>-6, or a <noun>-7?
- Assume 100 animals
- # queries = 100 x 99 x …94 = 8 x 10^{13}
- 2 queries/line, 275 lines, 1000 pages inch =
- How big?

The inductive puzzle

- Unsupervised learning
- (Very) small sample complexity
  - 1—5 examples; no Wall Street Journal subscriptions
- The Burst effect
- Order of presentation of examples doesn't matter
- External statistics don't match maturational time course

The burst effect

two-3 words, ages 1;1–1;11 ⟷ **"full" language (some residual)**

? time span: 2 weeks–2months

1;10 ride papa's neck
1;10.3 this my rock-baby
1;11.2 papa forget this

2;1.2 you watch me open sandbox
2;1.3 papa, you like this song?
2;4.0 I won't cry if mama wash my hair
2;4.3 put this right here so I see it better

What's the difference?

1. I see red one
2. P. want drink
3. P. open door
4. P. tickle S.
5. I go beach
6. P. forget this
7. P. said no
8. P. want out
9. You play chicken

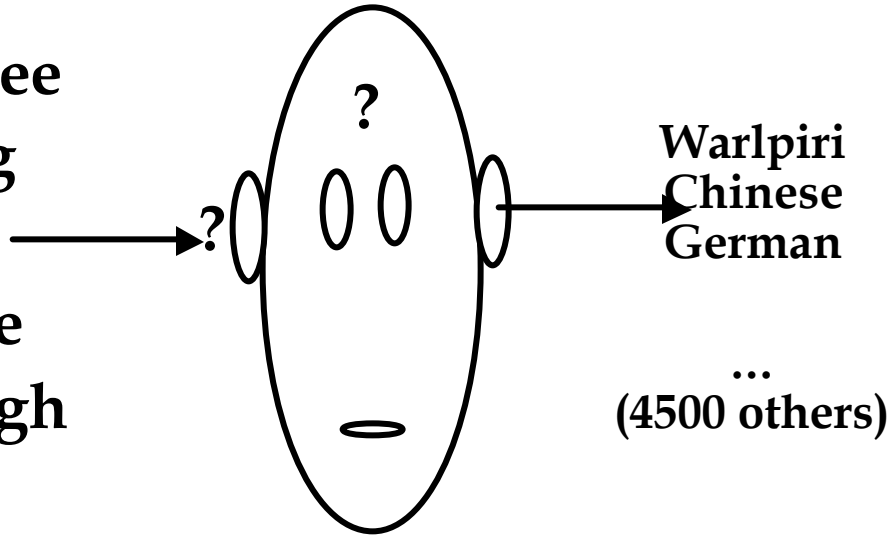
Multiple choice:
(a) Pidgin speakers; (b) apes;
(c) feral child Genie;
(d) ordinary children

Answers

1,5,9 = pidgin
2,4,8 = apes
7= Genie
3,6= children

Challenge: tension headaches

- **Essentially error-free**
- **Minimal triggering (+ examples)**
- **Robust under noise**
- **Still variable enough**



The input…

Bob just went away .

Bob went away .

no he went back to school .

he went to work .

are you playing with the plate ?

where is the plate ?

you 're going to put the plate on the wall ?

let's put the plate on the table .

the car is on your leg ?

you 're putting the car on your leg ?

on your other leg .

that's a car.

woom ? oh you mean voom . the car goes voom .

cars are on the road ?

thank you .

the cow goes moo ?

what color is the cow ?

what color is the cow ?

what color is the cow ?

what color

A developmental puzzle

- If pure inductive learning, then based on pattern distribution in the input
- What's the pattern distribution in the input?
  - English subjects: most English sentences overt
  - French: only 7-8% of French sentences have inflected verb followed by negation/adverb ("Jean *embrasse souvent/pas* Marie")
  - Dutch: *no* Verb first S's; Obj Verb Subject trigger appears in only 2% of the cases, yet…

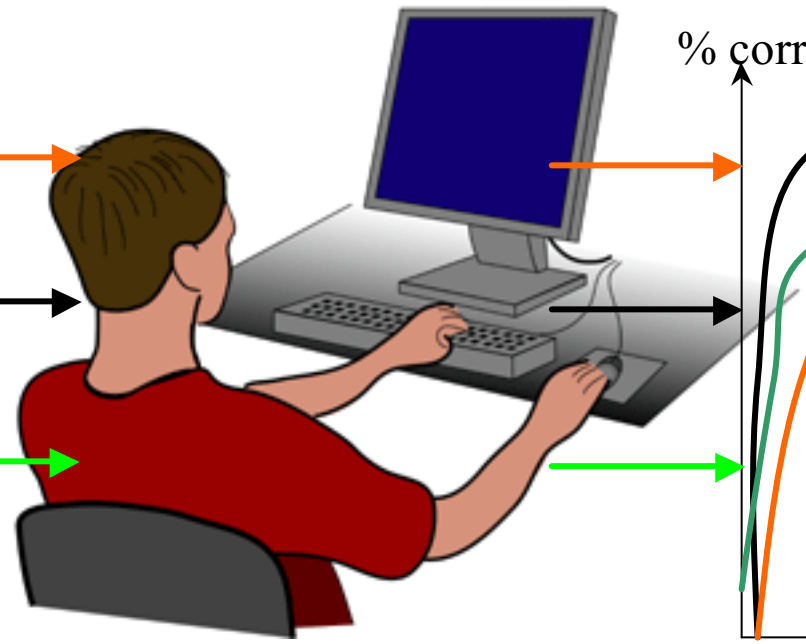
Predictions from pure induction

- English obligatory subject should be acquired *earl*y
- French verb placement should be acquired *late*
- Dutch verb first shouldn't be produced at all – because it's not very evident in the input

The empirical evidence runs completely contrary to this…

- English: Subjects acquired late (Brown, Bellugi, Bloom, Hyams…), but Subjects appear virtually 100% uniformly
- French: Verb placement acquired as early as it is possible to detect (Pierce, others), but triggers don't occur very frequently
- Dutch: 40-50% Verb first sentences produced by kids, but 0 % in input (Klahsen)
- So: what are we doing wrong?

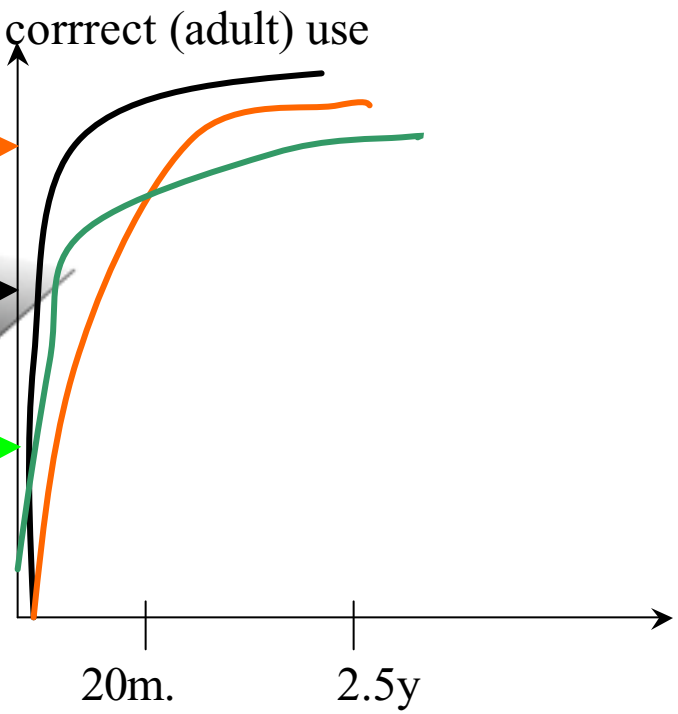
Can't *just* be statistical regularities…acquisition time course doesn't match



English Subject use
("there" sentences)
1% in Childes

French verb
raising 7-8%
(V_{fin}Adv/pas)

Dutch verb second
OVS 2%; 0 V1 pats



diffuse and *sparse*
regularities

stable inferences
and rapid time course

The language identification game

- black sheep
- baa baa black sheep
- baa black sheep
- baa baa baa baa black sheep

...

The facts

Child: Nobody don't like me.

Mother: No, say "Nobody likes me."

Child: Nobody don't like me.

Mother: No, say "Nobody likes me."

Child: Nobody don't like me.

Mother: No, say "Nobody likes me."

Child: Nobody don't like me.

[dialogue repeated five more times]

Mother: Now listen carefully, say "Nobody likes me."

Child: Oh! Nobody don't likeS me.

(McNeill, 1966)

Brown & Hanlon, 1970

- parents correct for meaning, not form
- when present, correction was not picked up

The problem…

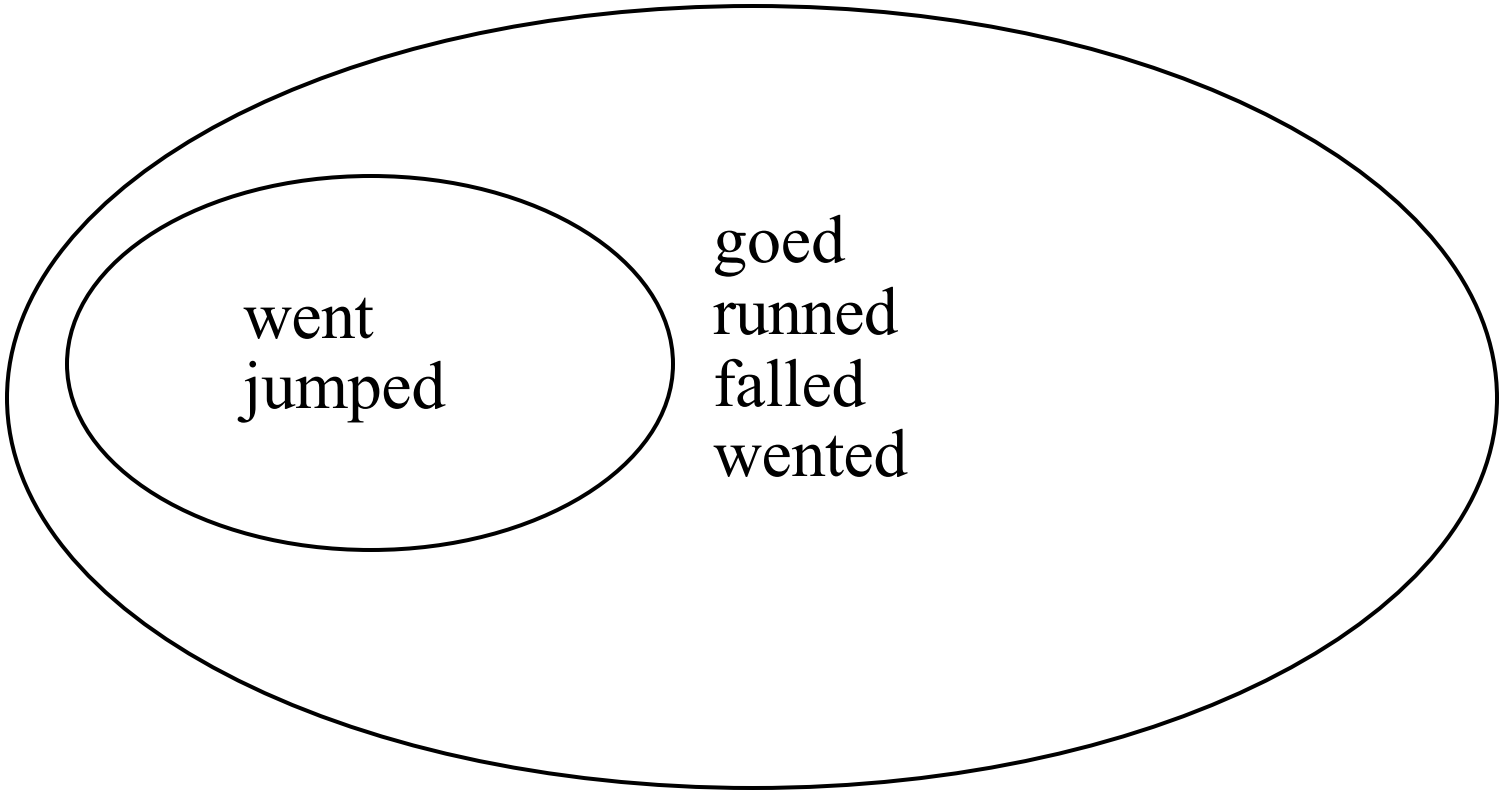
- The child makes an error.
- The adult may correct or identify the error.
- But the child ignores these corrections.
- So, how does the child learn to stop making the error?

But kids do recover (well, almost)

- u-shaped curve: went - goed - went

- child must stop saying:
- “goed”
- “unsqueeze”
- “deliver the library the book”

Overgeneralization



Positive vs. negative evidence

Positive examples

	Utterance	Feedback	Result
1.	Child says “went”.	none	none
2.	Child says “goed”.	none	none
3.	Adult says “went”.	---	positive data

Positive & Negative examples

	Utterance	Feedback	Result
1.	Child says “went”.	good	positive data
2.	Child says “goed”.	bad	corrective
3.	Adult says “went”.	good	positive data
4.	Adult says “goed”.	bad	corrective

Positive & negative examples

Child:	me want more.
Father:	ungrammatical.
Child:	want more milk.
Father:	ungrammatical.
Child:	more milk !
Father:	ungrammatical.
Child:	cries
Father:	ungrammatical

Contrast…

Child:	me want more.
Father:	You want more? More what?
Child:	want more milk.
Father:	You want more milk?
Child:	more milk !
Father:	Sure, honey, I'll get you some more.
Child:	cries
Father:	Now, don't cry, daddy is getting you some.

Formalize this game…

- Family of target languages (grammars) *L*
- The example data
- The learning algorithm, *A*
- The notion of learnability (convergence to the target) in the limit

- Gold's theorem (1967): If a family of languages contains all the finite languages and at least one infinite language, then it is not learnable in the limit

Gold's result…

- So, class of finite-state automata, class of Kimmo systems, class of cfg's, class of feature-based cfgs, class of GPSGs, transformational grammars,… NOT learnable from positive-only evidence
- Doesn't matter what algorithm you use – the result is based on a mapping – not an algorithmic limitation (Use EM, whatever you want…)

Framework for learning

1. Target Language $L_t \in L$ is a target language drawn from a class of possible target languages *L*
   .
2. Example sentences $s_i \in L_t$ are drawn from the target language & presented to learner.
3. Hypothesis Languages $h \in H$ drawn from a class of possible hypothesis languages that child learners construct on the basis of exposure to the example sentences in the environment
4. Learning algorithm *A* is a computable procedure by which languages from *H* are selected given the examples

Some details

- Languages/grammars – alphabet Σ^*
- Example sentences
  - Independent of order
  - Or: Assume drawn from probability distribution μ (relative frequency of various kinds of sentences) – eg, hear shorter sentences more often
  - If $\mu \in L_t$, then the presentation consists of positive examples, o.w.,
  - examples in both L_t & $\Sigma^* - L_t$ (negative examples), I.e., all of Σ^* ("informant presentation")

Learning algorithms & texts

- A is mapping from set of all finite data streams to hypotheses in H
- Finite data stream of k examples $(s_1, s_2, \ldots, s_k)$
- Set of all data streams of length k,

$$D^k = \{(s_1, s_2, \ldots, s_k) | s_i \in \Sigma^*\} = (\Sigma^*)^k$$

- Set of all finite data sequences $D = \cup_{k>0} D^k$ (enumerable), so:

$A : D \rightarrow H$

- Can consider A to flip coins if need be

If learning by enumeration: The sequence of hypotheses after each sentence is $h1, h2, \ldots,$

Hypothesis after n sentences is h_n

Criterion of success; Learnability

- Distance measure d between target grammar g_t and any hypothesized grammar h, $d(g_t, h)$
- Learnability of L implies that this distance goes to 0 as # of sentences n goes to infinity ("convergence in the limit")
- We say that a family of languages L is learnable if each member $L \in L$ is learnable
- This framework is very general – any linguistic setting; any learning procedure (EM, gradient descent,…)

Generality of this setting

1. $L \subseteq \Sigma^*$
2. $L \subseteq \Sigma_1^* \times \Sigma_2^*$ - NO different from (1) above - (form, meaning) pairs
3. $L: \Sigma^* \rightarrow [0,1]$ real number representing grammaticality; this is generalization of (1)
4. L is probability distribution μ on Σ^* - this is the usual sense in statistical applications (MT)
5. L is probability distribution μ on $\Sigma_1^* \times \Sigma_2^*$

What can we do with this?

- Two general approaches:
  - Inductive inference (classical – Gold theorem)
  - Probabilistic – approximate learning (VC dimension & "PAC" learning)
- Both get same result that all interesting families of languages are not learnable from positive-only data!

  (even under all the variations given previously): Fsa's, Hmm's, CFGs,…,
- Conclusion: some a priori restrictions on class *H* is required.
- This is Universal Grammar

In short:

- Innate = 'before data' (data = information used to learn the language, so, examples + algorithm used, or even modify the acquisition algorithm)
- Result from Learning theory: Restricted search space must exist (even if you use semantics!)
- No other way to search for 'underlying rules' – even if unlimited time, resources
- Research question: what is *A* ? Is it domain specific, or a general method?

The inductive inference approach (Gold's theorem)

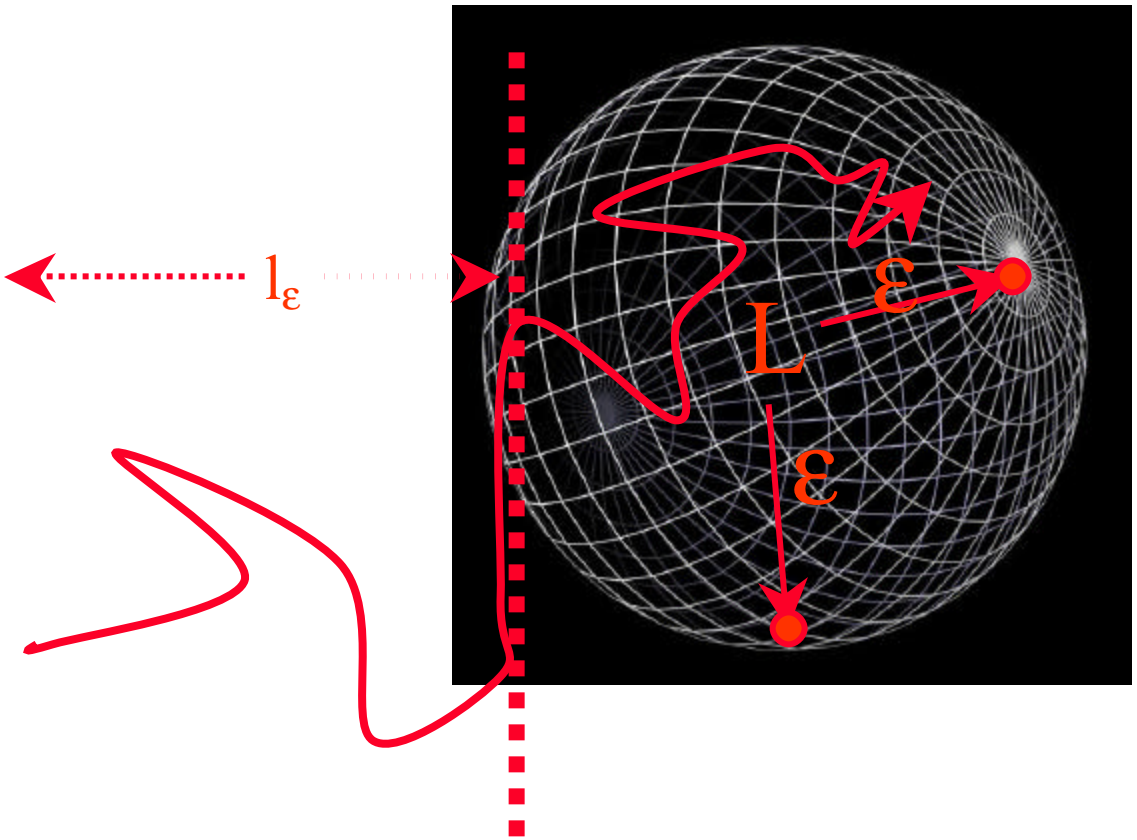
- Identification in the limit
- The Gold standard
- Extensions & implications & for natural languages
- We must restrict the class of grammars/languages the learner chooses from, severely!

ID in the limit - dfns

- Text t of language L is an infinite sequence of sentences of L with each sentence of L occurring at least once ("fair presentation")
- Text t_n is the first n sentences of t
- Learnability: Language L is learnable by algorithm A if for each t of L if there exists a number m s.t. for all $n>m$, $A\ (t_n)= L$
- More formally, fix distance metric d, a target grammar g_t and a text t for the target language. Learning algorithm A identifies (learns) g_t in the limit if

$$d(A\ (t_k), g_t\ ) \rightarrow 0\ _{k \rightarrow \infty}$$

ε–learnability & "locking sequence/data set"



Ball of radius ε

Locking sequence:
If (finite) sequence l_ε gets within ε of target & then it stays there

Relation between this & learnability in limit

- Thm 1 (Blum & Blum, 1975, ε-version) If A identifies g in the limit, then for every $\varepsilon > 0$, there exists a locking data set that comes within ε of the target

Gold's thm follows...

- Theorem (Gold, 1967). If the family *L* consists of all the finite languages and at least 1 infinite language, then it is not learnable in the limit
- Corollary: The class of fsa's, cfg's, csg's,... are not learnable in the limit

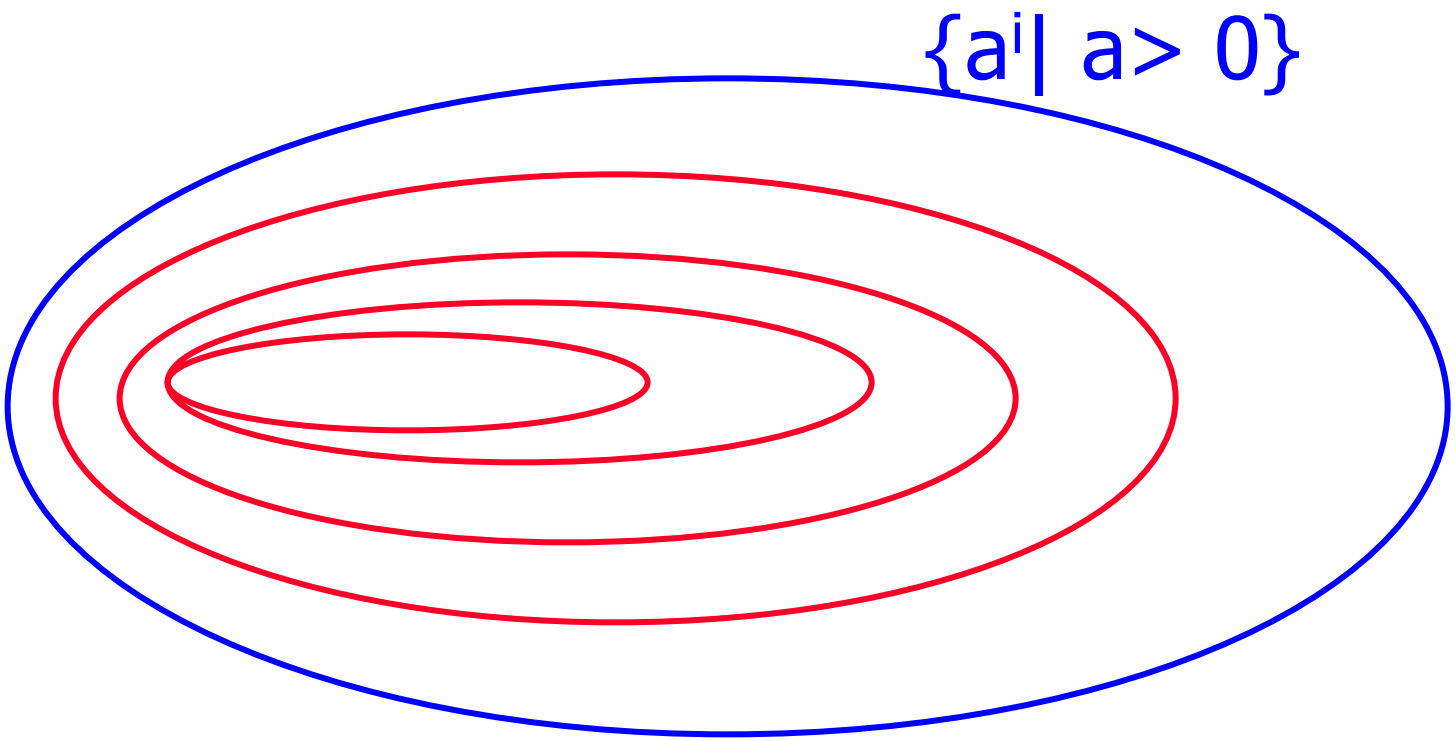
- Proof by contradiction...

Gold's thm

- Suppose *A* is able to identify the family *L*. Then it must identify the infinite language, L_{inf} .
- By Thm, a locking sequence exists, σ_{inf}
- Construct a finite language $L_{\sigma_{\text{inf}}}$ from this locking sequence to get locking sequence for $L_{\sigma_{\text{inf}}}$ - a different language from L_{inf}
- *A* can't identify $L_{\sigma_{\text{inf}}}$, a contradiction

Picture

One Superfinite L, all finite L's



But what about…

- We shouldn't require exact identification!
- Response: OK, we can use ε notion, or, statistical learning theory to show that if we require convergence with high probability, then the same results hold (see a bit later)
- Suppose languages are finite?
- Response: naw, the Gold result is really about information density, not infinite languages

But what about… (more old whine in new bottles)

- Why should you be able to learn on every sequence?
- Response: OK, use the "Probably approximately correct" (PAC) approach – learn target with high probability, to within epsilon, on 1-δ sequences
- Converge now not on every data sequence, but still with probability 1
- Now $d(g_t, h_n)$ is a random variable, and you want weak convergence of random variables
- So this is also convergence in the limit

Stochastic extensions/Gold complaints & positive results

- To handle statistical case – rules are stochastic – so the 'text' the learner gets is stochastic (some distribution spits it out…)
- If you know how language is generated then it helps you learn what language is generated
- Absence of sentence from guessed L is like negative evidence: although approximate, can be used to reject guess ("indirect negative evidence")

Results for stochastic case

- Results:
  - Negative evidence: really needs all the text (enough sampling over negative examples s.t. child can really know it)
  - If you don't know the distribution – you lose – estimating a density function is even harder than approximating functions…
  - If you have very strong constraints on distribution functions to be drawn from the language family, then you can learn fsa's, cfg's…
  - This constraint is that the learner knows a function d, s.t. after seeing at least d(n) examples, learner knows what membership of each example sentence in every sequence

Finite language case

- Result doesn't really depend on some subtle property of infinite languages
- Suppose finite languages. Then Gold framework – learner identifies language by memorization - only after hearing all the examples of the language
- No possibility of generalization; no extrapolation – not the case for natural languages

A simple example…

Simple finite case

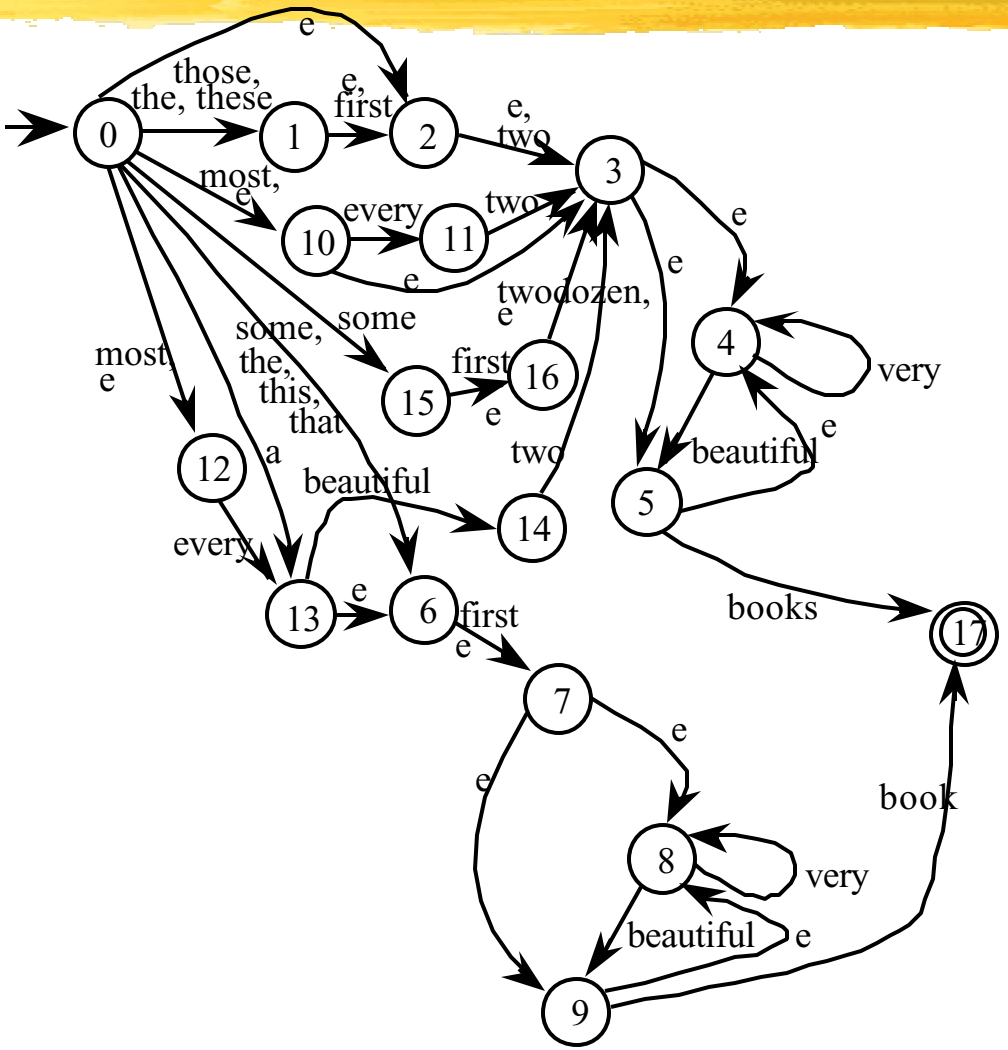
- Finite set of finite languages
- 3 sentences, s_1, s_2 , s_3, so 8 possible languages
- Suppose learner A considers all 8 languages
- Learner B considers only 2 languages: $L_1 = \{s_1 , s_2 \}$, $L_2 = \{s_3 \}$
- If A receives sentence s_1 then A has no information whether s_2 or s_3 will be part of target or not – only can tell this after hearing all the sentences
- If B receives s_1 then B knows that s_2 will be part of the target – extrapolation beyond experience
- Restricted space is requirement for generalization

How many examples needed?

- Gold (again): even if you know the # of states in an fsa, this is NP-hard

- Restrictions on class of fsa's make this poly-time (Angluin; Pilato & Berwick)
- If fsa is backwards deterministic

Example inferred fsa (NP specifiers)



OK smarty…

- What can you do?
- Make class of a priori languages finite, and small…
- Parameterize it
- How?