

1.1 Basics

- (0) Gold [4] and Blum & Blum [1] established one perspective from which language can be studied. This perspective is developed in various ways in the texts by [5], [6], [7].

We will consider this perspective first, for two reasons:

- it provides a simple and completely clear perspective on some aspects of learning
- it allows us to focus our attention on the critical step: generalization.

(As noted earlier, by “generalization,” we mean conjecturing that the environment contains more than has been seen.)

- (1) We will think of languages as sets of expressions together with a non-expression #, where an expression is an element of Σ^* for some (finite, non-empty) vocabulary Σ .

- (2) A **(positive) text** T is an infinite sequence whose elements are expressions and #

Since we regard an infinite sequence as a function with domain $\mathbb{N} = \{0, 1, \dots\}$, we let $T(n)$ be the n 'th member of T counting up from 0.

Let $T[n]$ be the length n initial sequence of T . So $T[n] = \langle T(0), T(1), \dots, T(n-1) \rangle$.

- (3) Let the **content of text** T , $\text{content}(T)$ be the set of expressions that appear in T . We say that **text** T is **for language** L iff $\text{content}(T) = L$. Note that # is never part of the content.

Then, for every text T , and for every $x \in \text{content}(T)$, there is some finite i such that $x = T(i)$.

- (4) A **learner** is a (possibly partial) function ϕ from finite sequences of expressions and # to grammars, where each grammar G defines a language $L(G)$.

We now define a certain kind of learning, which is sometimes called **identification from positive text**, as follows.

- (5) If $x \in \text{Dom}(\phi)$ we say ϕ is **defined on** x , or $\phi(x) \downarrow$.

Learner ϕ is **defined on** T iff $\phi(T(i)) \downarrow$ for all $i \in \mathbb{N}$.

- (6) ϕ **converges** on T iff for some i , $\phi(T[j]) = \phi(T[i])$ for all $j \geq i$. In this case, we define $\phi(T)$ to be $\phi(T[i])$.

- (7) ϕ **identifies** T iff ϕ converges on T , $\phi(T) \downarrow$, and $\text{content}(T) = L(\phi(T))$.

- (8) ϕ **identifies** L iff ϕ identifies all texts for L .

- (9) ϕ **identifies** a class of languages \mathcal{L} iff for all $L \in \mathcal{L}$, ϕ identifies L .

- (10) A class of languages \mathcal{L} is **identifiable** iff some learner identifies \mathcal{L} .

Sometimes, when the context makes our intention clear, we will call such a class **learnable**.

1.2 First results

Theorem 1 The classes $\mathcal{L} = \emptyset$ and $\mathcal{L} = \{L\}$ for any r.e. language L are identifiable.

- (11) The previous theorem needs the qualification “r.e.” (recursively enumerable) since we are assuming that languages are represented by grammars (or machines) that recursively enumerate them.

Theorem 2 [4] The class \mathcal{L}_{fin} of finite languages is identifiable.

Proof: Suppose that we have an enumeration of grammars for all the r.e. languages. Then define a learner ϕ_e as follows: for any text T , $\phi_e(T[i])$ is the first grammar G in the enumeration such that $L(G) = \text{content}(T[i])$.

Consider any $L \in \mathcal{L}_{fin}$ and any text T for L . Then there is some i such that $\text{content}(T[i]) = L$. In this case, by our definition of ϕ_e , $\phi_e(T[j]) = \phi_e(T[i])$ for all $j \geq i$, so ϕ_e converges on t . And again by our definition of ϕ_e , $L(\phi_e(T[i])) = \text{content}(T)$. \square

(12) Notice that we did not define the enumeration of grammars precisely; but it can be done without too much trouble. (We return to this.)

(13) Notice also that nothing in the definition of learnability implies that learners are necessarily computable functions, and so we did not need to explain how the function ϕ_e could be computed. In this case, the decision about whether $L(\phi_e(T[i])) = \text{content}(T)$ is obviously not computable in general.

So if we are interested in a computable learner, we could, for example, define an effective enumeration of grammars for just the finite languages, and choose the first of these that matches the sample. Or we could let $\phi_e(T[i])$ be the grammar $\{S \rightarrow x \mid x \in \text{content}(T[i])\}$.

(14) Gold [4] calls the learner ϕ_e defined in the proof above an **identification-by-enumeration** learner. Notice that nothing in ϕ_e 's values signals when it has converged on the right hypothesis. It is interesting to contrast learning functions which announce their convergence (“Aha!” or, for short “ \emptyset ”), in the following way:

- a. on any text T there is a unique i such that $L(\phi(T[i])) = \emptyset$, and
- b. on any text T if $L(\phi(T[i])) = \emptyset$ then for all $j > i$, $\phi(T[i+1]) = \phi(T[j])$.

Osherson et al [9] call such learners **self-monitoring**.

Theorem 3 [3] No self-monitoring learner identifies the class $\mathcal{L}_{fin} - \emptyset$ of non-empty finite languages.

Proof: For contradiction, assume that ϕ is a self-monitoring learner that identifies $\mathcal{L}_{fin} - \emptyset$. Consider any finite language $L \in \mathcal{L}_{fin} - \emptyset$ and any text T where $\text{content}(T) = L$. Let i be the unique point at which $L(\phi(T[i])) = \emptyset$. Then by assumption, $L(\phi(T[i+1])) = L(\phi(T[m]))$ for all $m > i$. Since $L \in \mathcal{L}_{fin} - \emptyset$, there are infinitely many strings $x \notin L$. Take one such string x , and consider the text $T' = T[i+1]xxx\dots$. That is, T' is the sequence $T[i+1]$ followed by an infinite sequence of the sentence x . Then $\text{content}(T') \in \mathcal{L}_{fin}$. By definition of the self-monitoring ϕ , $L(\phi(T'[i])) = \emptyset$, and so for all $j > i$, $L(\phi(T'[i+1])) = L(\phi(T'[i+1])) = L(\phi(T[j])) = L(\phi(T'[j])) = L$. So $L(\phi(T')) \neq \text{content}(T')$, and so ϕ does not identify T' – a contradiction! ∇ \square

Exercise 1 It is a triviality that no self-monitoring learner identifies any class \mathcal{L} that contains \emptyset , since this learner would have to have a unique “aha” point where it takes the value \emptyset , and it would also need to take that value at all later points in the text – a contradiction.

So suppose we modify the definition so that a learner ϕ is “self-monitoring” iff

1. on any text T there is a unique i such that $L(\phi(T[i]))$ is undefined, and
2. on any text T if $L(\phi(T[i]))$ is undefined, then for all $j > i$, $\phi(T[i+1]) = \phi(T[j])$.

With this definition, show that there is no self-monitoring learner identifies any class \mathcal{L} that contains \emptyset and another language. \circ

The previous exercise establishes

Corollary 1 No self-monitoring learner identifies the class \mathcal{L}_{fin} of finite languages.

(15) So although the learner ϕ_e cannot tell when it has converged on a text for $L \in \mathcal{L}_{fin}$, no learner that identifies \mathcal{L}_{fin} could do that. The learner ϕ_e is as good as possible in another respect as well: no learner can learn finite languages faster than ϕ_e does.

Definition 1 Suppose ϕ identifies L and let T be a text such that $\text{content}(T) = L$. Let the **convergence point** for ϕ in T , $cp(\phi, T)$ be the least i such that for all $j \geq i$, $\phi(T[j]) = \phi(T[i])$.

Given learners ϕ, ϕ' that identify some class \mathcal{L} , let's say that ϕ' is **uniformly faster** than ϕ on \mathcal{L} iff

1. for all texts T such that $\text{content}(T) \in \mathcal{L}$, $cp(\phi', T) \leq cp(\phi, T)$, and
2. for some text T such that $\text{content}(T) \in \mathcal{L}$, $cp(\phi', T) < cp(\phi, T)$.

Theorem 4 [4] No learner ϕ that identifies \mathcal{L}_{fin} is uniformly faster than ϕ_e on \mathcal{L}_{fin} .

Proof: Suppose for contradiction that there is a learner ϕ which identifies \mathcal{L}_{fin} uniformly faster than ϕ_e . Then there is some text T such that $content(T) \in \mathcal{L}_{fin}$ and $cp(\phi, T) < cp(\phi_e, T)$. So at the point $i = cp(\phi, T)$ where ϕ correctly identifies $content(T)$, ϕ_e has not yet converged. In fact, by the definition of ϕ_e , $content(T[i]) = L(\phi_e(T[i])) \neq L(\phi_e(T)) = L$, since ϕ_e never conjectures two different grammars for the same language. On the other hand, ϕ has converged, that is, $\phi(T[i]) = \phi(T)$ and $L(\phi(T)) = L$. Now consider another text T' such that $T'[i] = T[i]$ and $content(T') = content(T'[i])$. Then $content(T') \in \mathcal{L}_{fin}$ and $cp(\phi_e, T') = i$. Since $T'[i] = T[i]$, we know that $\phi(T'[i]) = \phi(T[i])$ and so $L(\phi(T'[i])) \neq L$. It follows that $cp(\phi, T') > i$, and so ϕ is not uniformly faster than ϕ_e . $\zeta\Box$

1.3 Confusions

- (16) “The logical problem of human language acquisition has been solved many times over.”

Suppose the ‘logical problem of language acquisition’ is the problem of finding learners that can identify human languages from the evidence readily available to children (and we might also require: with feasible computational effort). Is this problem really solved?

- People sometimes seem to say things roughly like this, unless you read carefully! E.g. [8, p.883]: “Using demonstrably available positive data, simple learning procedures can be formulated for each of the syntactic structures that have traditionally motivated invocation of the logical problem.” At first this sounds dramatic, but considered more carefully it sounds dodgy, and maybe irrelevant. Cf. [11, 10]

- Another idea is that the logical problem of language acquisition has been solved by ‘discovering’ that the number of human languages is finite. In a sense of course the set of human languages is finite – the whole universe is finite, and the whole set of utterances ever made by any individual is finite... But the linguist is interested in the fact that humans are mortal, but rather in the kinds of patterns we find in human languages. These patterns do not cut off at a finite length, and as far as we know the range of grammatical constructions does not have a principled bound either. That’s why, for linguistic purposes, we regard languages as infinite. We could also guarantee a finite range of possible grammars by stipulating that only a certain range of variation counts as ‘important’, part of the ‘core’ grammar. But the infinite perspective plays an important role in learnability theory by forcing us to look at generalization. To learn an infinite language, the learner must generalize. That’s what we are interested in. To say that the problem is solved because we are only interested in the “core” leaves all the interesting questions unilluminated.

Some will respond that some exciting research has been done in the tradition that assumes a finite number of finitely valued parameters are set in human language learning. And that is true. But if you look carefully, the interest of, e.g., the psychological research on language acquisition comes from discovering that certain specified parameters are set in certain ways, and it is really not relevant whether or not the authors think the whole set of parameter settings is finite in some sense. What this research really shows is the usefulness of looking carefully at how people generalize from the data.

- (17) “The Gold paradigm is too narrow in the sense that it excludes sentence fragments, phrases, single words, intrusions from different dialects and languages, and so on, from the set of texts.”

This is a confusion. While it is true that we could let a text consist of exactly the grammatical sentences, we could also let it consist of sentence-fragments, of all phrases, of all expressions, of phrase-meaning pairs, etc. Many basic facts about learning can be established before we pay any attention about what, exactly, those things in the texts actually are.

The point about intrusions of other languages is partly right: while the distribution of items in the text is not usually required to be similar to the distribution of items in the child’s environment, we could consider texts in which the set of items presented contains only a small number of expressions from other languages. For learners that seek generalizations across elements, having some elements that are dissimilar from others may not be a problem. We will return to this question in the Gold paradigm, and then we will consider it from the perspective of stochastic learning schemes.

- (18) “The Gold paradigm allows texts that psycholinguists would want to exclude... (psycho)linguists might consider many such texts of limited interest due to their complexity.”

While this is true, it has no bearing on the applicability of the Gold paradigm.

The Gold criterion of learning says a language is learnable only if it can be learned from every text for the language. Furthermore, we can (and will) consider what languages can be learned by learners that ignore

all sentences that exceed a given complexity bound. In that case, the presence of very complex expressions in the text is irrelevant.

- (19) Chomsky [2] is right to point out that we do not know, in advance, that the class of languages (as linguists would want to characterize them) is a learnable class. It makes perfect sense, and might well be the case, that the class of languages that are similar to existing languages in linguistically relevant respects is unlearnable, with humans capable of learning only some subset of this class which differs from the rest of the class in respects that are practically but not linguistically relevant (e.g. the larger class might only be learnable from really long sentences, or really deep recursion,...) We will return to this important question later.

Exercise 2 See if you can prove that every finite collection of languages is learnable (even if some or all of these languages are infinite). This is not so easy, but you might see already some ways to show this. ○

References

- [1] BLUM, L., AND BLUM, M. Toward a mathematical theory of inductive inference. *Information and Control* 28 (1975), 125–155.
- [2] CHOMSKY, N. *Lectures on Government and Binding*. Foris, Dordrecht, 1981.
- [3] FREIVALD, R., AND WIEHAGEN, R. Inductive inference with additional information. *Elektronische Informationsverarbeitung und Kybernetik* 15 (1979), 179–185.
- [4] GOLD, E. M. Language identification in the limit. *Information and Control* 10 (1967), 447–474.
- [5] JAIN, S., OSHERSON, D., ROYER, J. S., AND SHARMA, A. *Systems that Learn: An Introduction to Learning Theory (second edition)*. MIT Press, Cambridge, Massachusetts, 1999.
- [6] KANAZAWA, M. *Learnable Classes of Categorical Grammmars*. CSLI Publications, Stanford, California, 1998.
- [7] LAIRD, P. D. *Learning from Good and Bad Data*. Kluwer, Boston, 1988.
- [8] MACWHINNEY, B. A multiple process solution to the logical problem of language acquisition. *Journal of Child language* 31 (1989).
- [9] OSHERSON, D., WEINSTEIN, S., AND STOB, M. *Systems that Learn*. MIT Press, Cambridge, Massachusetts, 1986.
- [10] PERFOR, A., TENENBAUM, J. B., AND REGIER, T. The learnability of abstract syntactic principles. MIT and University of Chicago, 2007.
- [11] PULLUM, G. K., AND SCHOLZ, B. C. Empirical assessment of stimulus poverty arguments. *Linguistic Review* 19 (2002), 9–50.